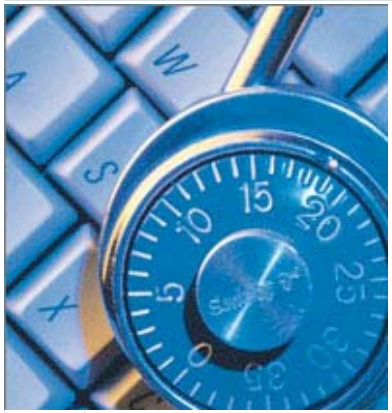
Improving Cyber-Security through Predictive Analysis

Peter Frometa
SPSS

21 November 2005

Applications of DM & TM in Security



Identify Cyber Threats

Identify Insider Threats

Detect Money Laundering Behavior

Detect Fraud

Detect Smuggling and Drug Trafficking Associations

Predict Non-Compliance of Foreign Visitors

Identify True Identity of Individuals

Better Allocate Security Resources

Ensure Maritime Domain Awareness

Discover Potentially Suspicious Inbound Cargo

Discover Terrorist Organization Affiliation

Predict Risk Potential for Threats to Safety

Data Mining Defined

What is Data Mining?

A hot buzzword for a class of techniques that find patterns or relationships that have not previously been detected

A user-centric, interactive process which leverages analysis technologies and computing power

Not reliant on an existing database

A relatively easy task that requires knowledge of the organization's problem and subject matter expertise

Text Mining Defined

What is Text Mining?

A user-centric process which leverages analysis technologies and computing power to access valuable information within unstructured text data sources

Driven by the use of Natural Language Processing and Linguistic based algorithms…not a search engine

No real value unless used in conjunction with Data Mining

Ultimately, eliminates need to manually read unstructured data sources

Beyond Historical Reporting

REACTIVE

Historical Reporting

- What are the top source/destination countries?
- What are the top attacked products?
- What are the top offending ISP's
- What IP's have led to the most events?
- What ports are scanned the most?
- What are the IDS event counts for the past day, week, month?

PROACTIVE

Predictive Analytics

- What activity within the unidentified data of the top source/destination countries is likely to be malicious?
- Of those identified to be malicious, which is likely to be the worst?
- What are the characteristics/ vulnerabilities of products likely to be attacked?
- Which activity did not get flagged that is likely to be malicious?
- Which flagged items are most likely to be false positives?
- What can we expect the IDS event counts to be next week, month, etc.?
- What activity is likely to be associated with worm emergence?
- Which isolated events are likely to be associated with originating from the same source?

NetFlow Data

	srcip	dstip	srcport	dstport	protocol	packets	bytes	flags	strttime	duration	endtime	sc
1	66.94.234.13	192.35.251.135	53	113	17	1	135	A	28-SEP-04	0	28-SEP-04	PM
2	216.239.57.99	206.33.100.4	53	2050	17	1	135	A	28-SEP-04	1	28-SEP-04	PM
3	17.254.3.183	192.35.251.74	53	30555	17	1	121	A	28-SEP-04	3	28-SEP-04	PM
4	209.202.248.202	192.35.251.74	53	30555	17	1	170	A	28-SEP-04	0	28-SEP-04	PM
5	208.45.133.23	200.150.13.120	53	1004	17	1	194	A	28-SEP-04	1	28-SEP-04	PM

sIP – source IP address (IP address of host that sent the IP packet)

dIP – destination IP address

sPort – port used by the source IP address

dPort – port used by the destination IP address

pro – IP protocol (primarily TCP/UDP in this data)

packets – packet count

bytes – byte count

flags – Transfer Control Protocol header flags

sTime – start time of flow (GMT)

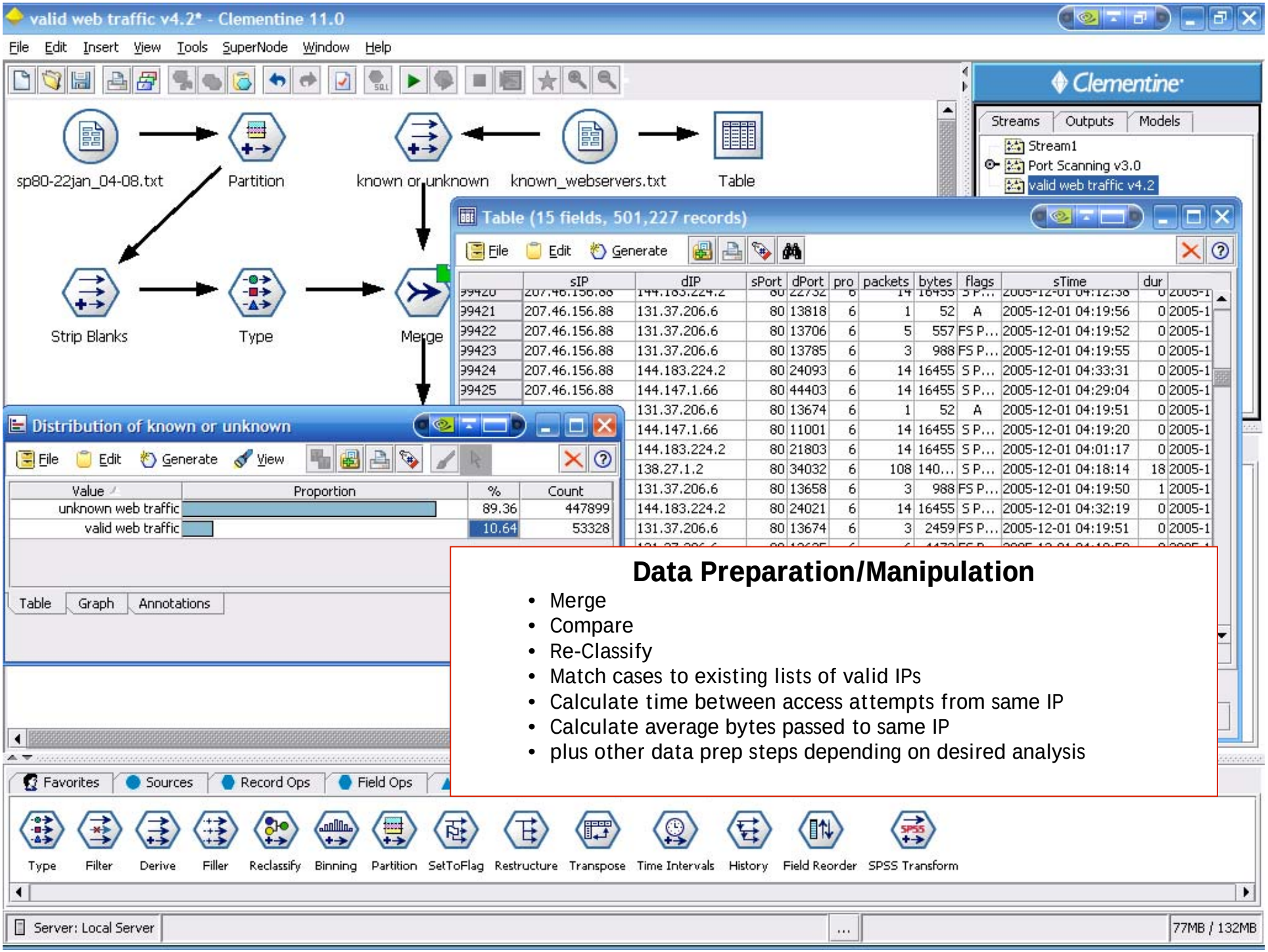
dur – duration of flow (eTime – sTime)

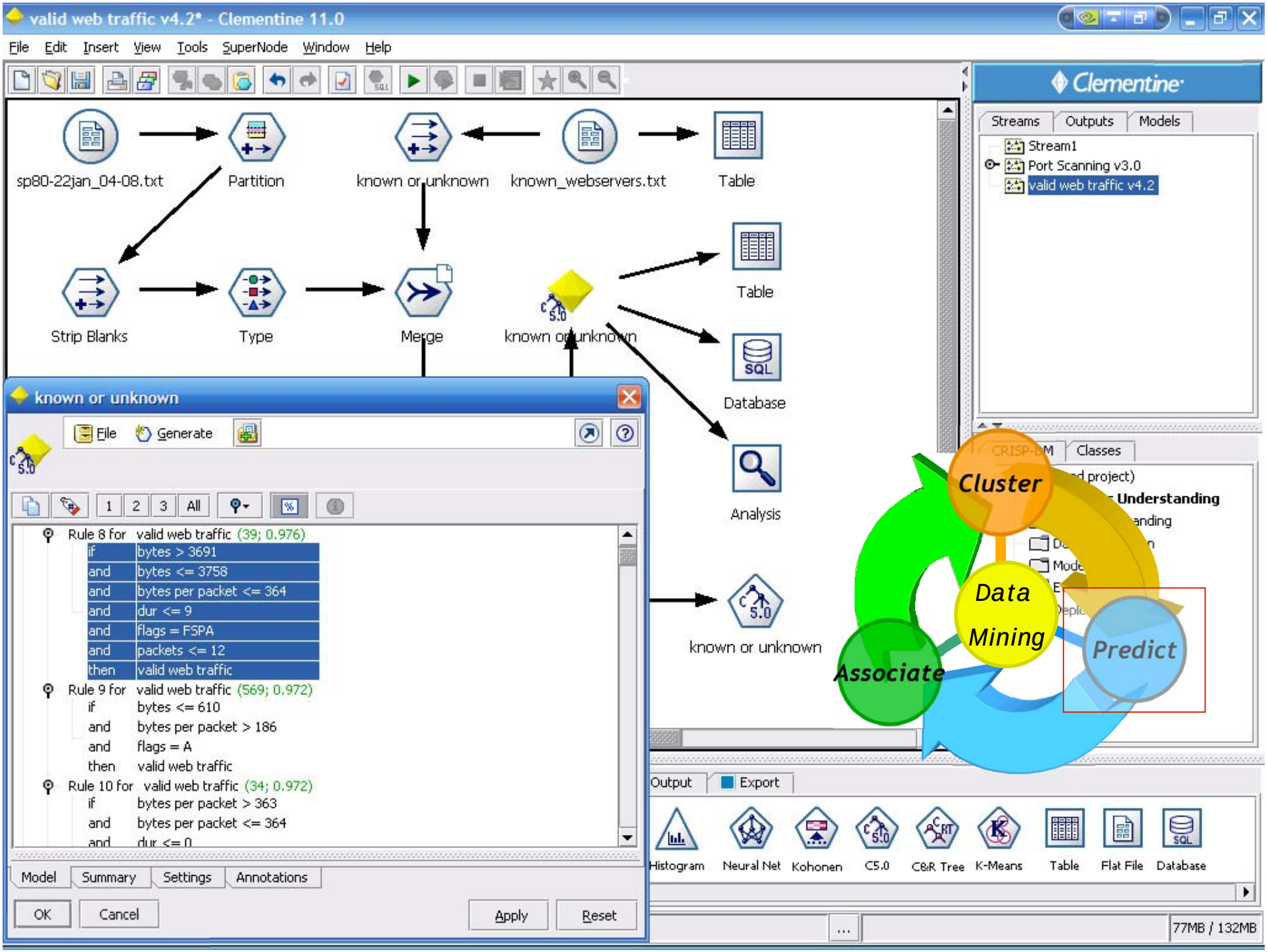
eTime – end time of flow (GMT)

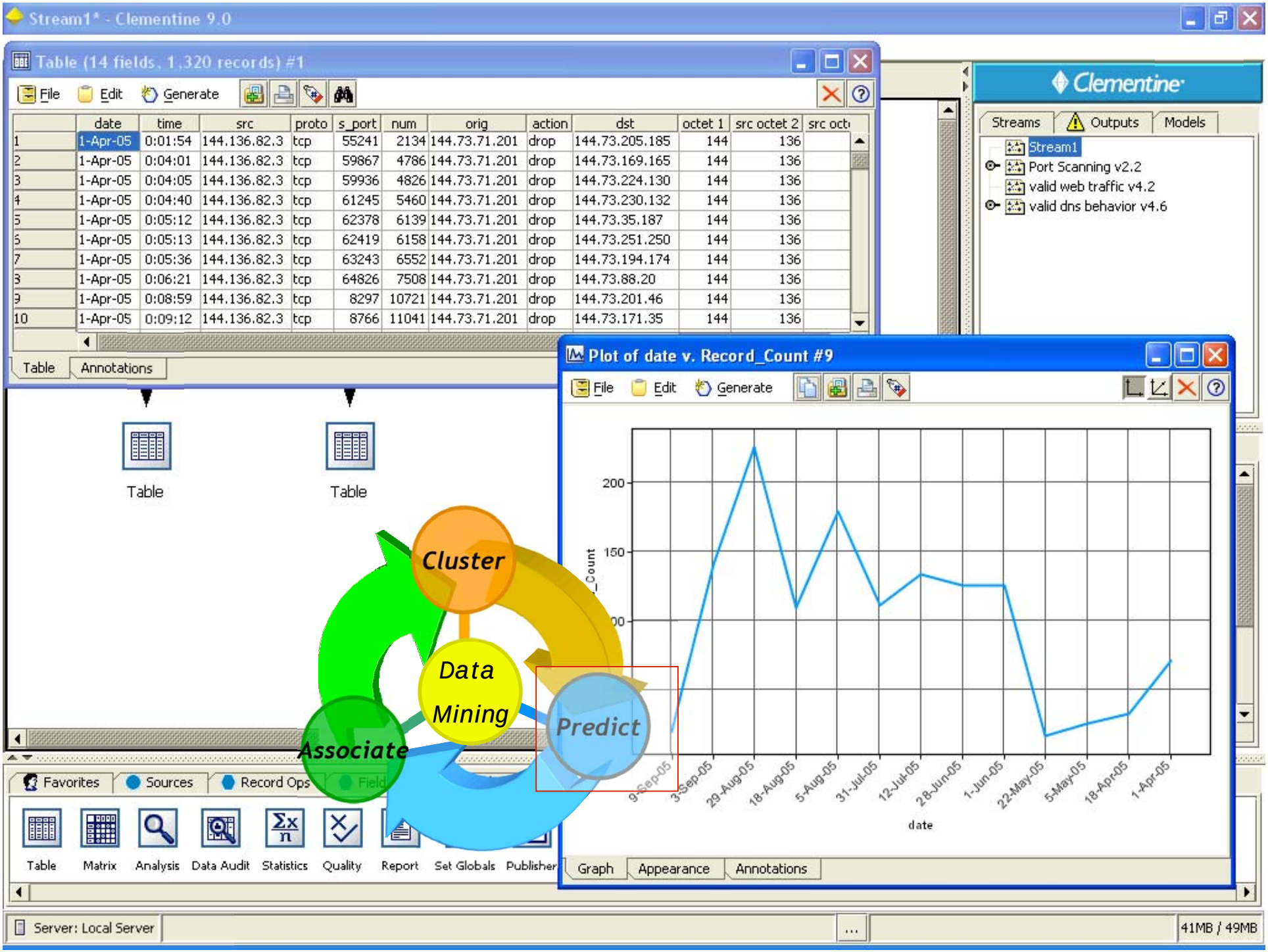
sr – source name or ID of sensor that picked up the data

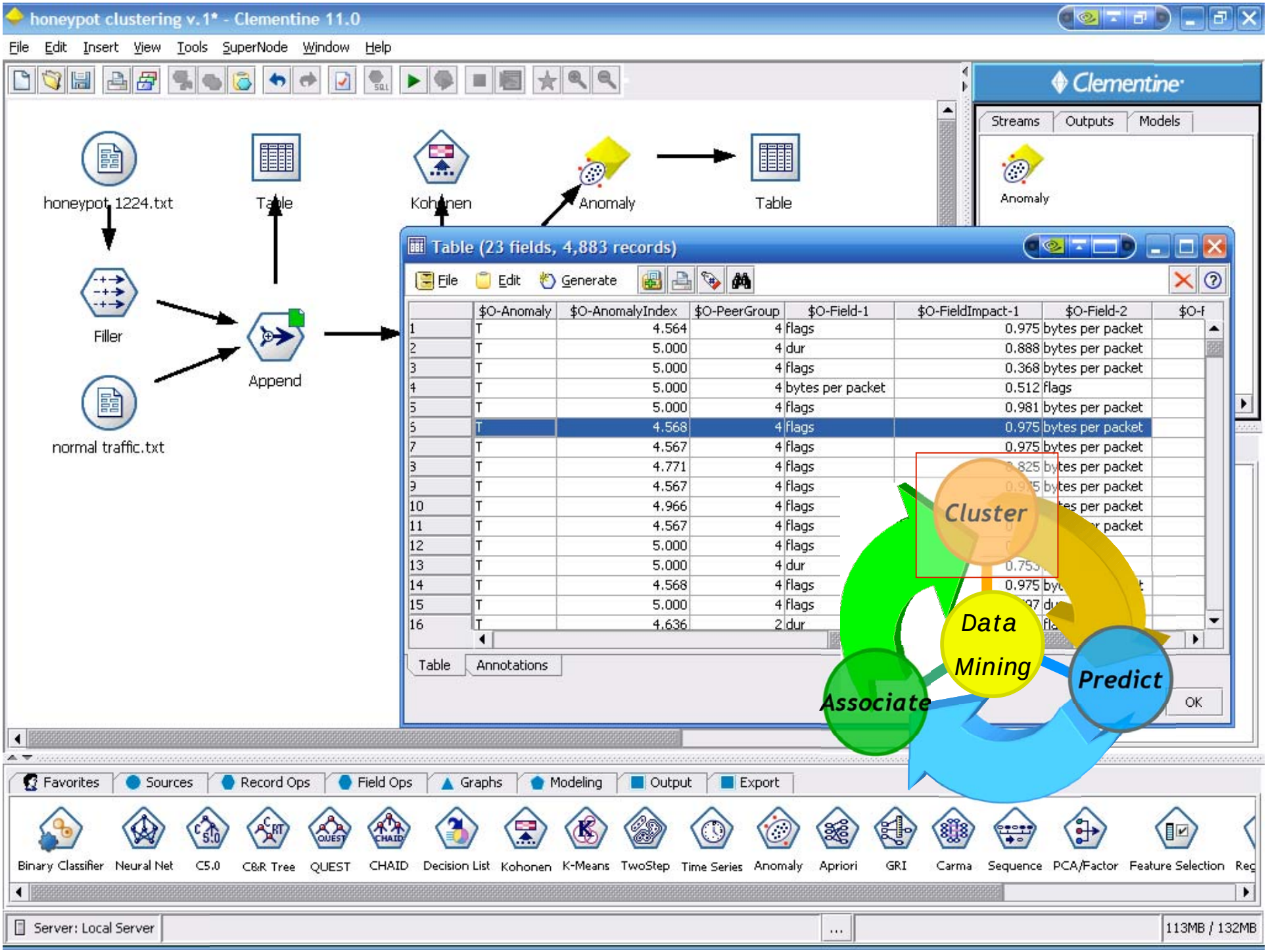
Where Does Predictive Analytics Fit?

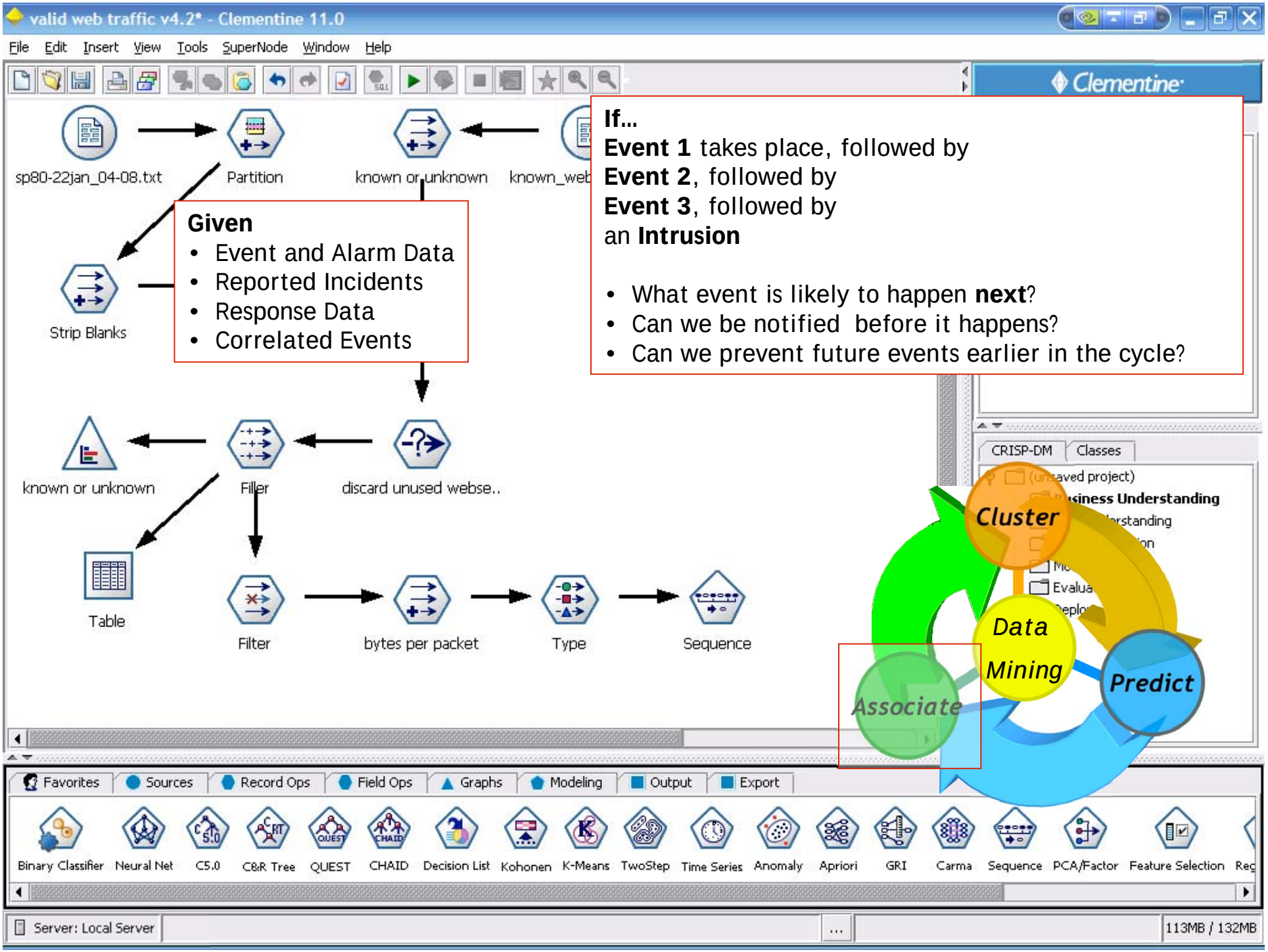
Level 1 Analysis	Level 2 Analysis
Stop an intrusion event	Associate multiple similar intrusions
Log Data	Classify organized intrusions
Report	Identify state sponsored cyber terror
Based on:	
Net Flow data (source and destination IP address, protocol, packets, bytes, flags…)	Net Flow data Response data Correlated Incidents, historical data Keystroke logging other…

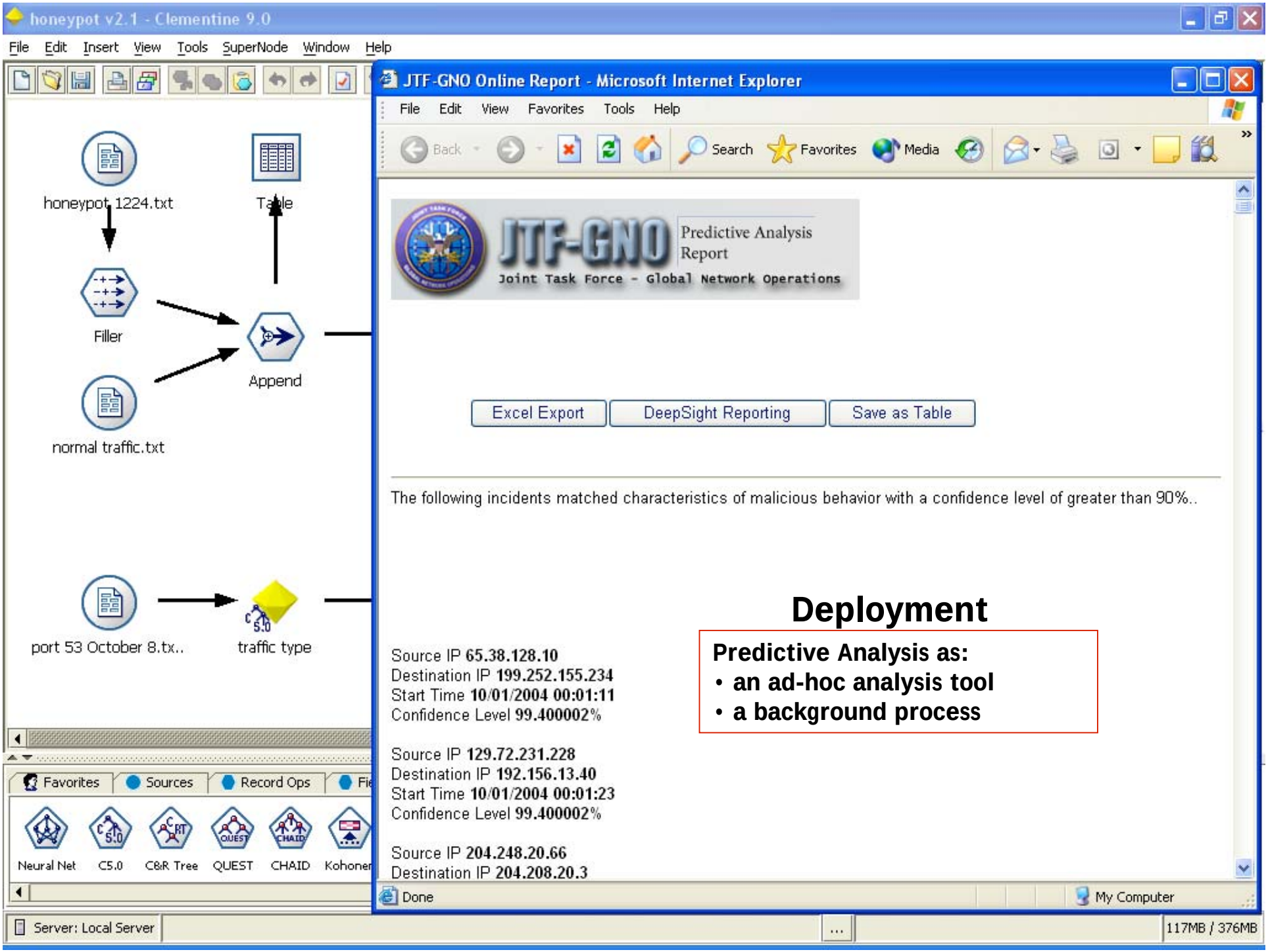

valid web traffic v4.2* - Clementine 11.0
File Edit Insert View Tools SuperNode Window Help
sp80-22jan_04-08.txt
Partition
known or unknown
known_webservers.txt
Table
Strip Blanks
Type
Merge
Clementine
Streams Outputs Models
Stream1
Port Scanning v3.0
valid web traffic v4.2
Table (15 fields, 501,227 records)
File Edit Generate
sIP dIP sPort dPort pro packets bytes flags sTime dur
99421 207.46.156.88 131.37.206.6 80 13818 6 1 52 A 2005-12-01 04:19:56 0 2005-1
99422 207.46.156.88 131.37.206.6 80 13706 6 5 557 FS P... 2005-12-01 04:19:52 0 2005-1
99423 207.46.156.88 131.37.206.6 80 13785 6 3 988 FS P... 2005-12-01 04:19:55 0 2005-1
99424 207.46.156.88 144.183.224.2 80 24093 6 14 16455 S P... 2005-12-01 04:33:31 0 2005-1
99425 207.46.156.88 144.147.1.66 80 44403 6 14 16455 S P... 2005-12-01 04:29:04 0 2005-1
131.37.206.6 80 13674 6 1 52 A 2005-12-01 04:19:51 0 2005-1
144.147.1.66 80 11001 6 14 16455 S P... 2005-12-01 04:19:20 0 2005-1
144.183.224.2 80 21803 6 14 16455 S P... 2005-12-01 04:01:17 0 2005-1
138.27.1.2 80 34032 6 108 140... S P... 2005-12-01 04:18:14 18 2005-1
131.37.206.6 80 13658 6 3 988 FS P... 2005-12-01 04:19:50 1 2005-1
144.183.224.2 80 24021 6 14 16455 S P... 2005-12-01 04:32:19 0 2005-1
131.37.206.6 80 13674 6 3 2459 FS P... 2005-12-01 04:19:51 0 2005-1
Distribution of known or unknown
File Edit Generate View
Value Proportion % Count
unknown web traffic 89.36 447899
valid web traffic 10.64 53328
Table Graph Annotations
Data Preparation/Manipulation
Merge
Compare
Re-Classify
Match cases to existing lists of valid IPs
Calculate time between access attempts from same IP
Calculate average bytes passed to same IP
plus other data prep steps depending on desired analysis
Favorites Sources Record Ops Field Ops
Type Filter Derive Filler Reclassify Binning Partition SetToFlag Restructure Transpose Time Intervals History Field Reorder SPSS Transform
Server: Local Server
77MB / 132MB


valid web traffic v4.2* - Clementine 11.0
File Edit Insert View Tools SuperNode Window Help
sp80-22jan_04-08.txt
Partition
known or unknown
known_webservers.txt
Table
Strip Blanks
Type
Merge
known or unknown
Table
Database
Analysis
known or unknown
Clementine
Streams Outputs Models
Stream1
Port Scanning v3.0
valid web traffic v4.2
Classes
Understanding
Cluster
Data Mining
Predict
Associate
known or unknown
File Generate
1 2 3 All
Rule 8 for valid web traffic (39; 0.976)
if bytes > 3691
and bytes <= 3758
and bytes per packet <= 364
and dur <= 9
and flags = FSPA
and packets <= 12
then valid web traffic
Rule 9 for valid web traffic (569; 0.972)
if bytes <= 610
and bytes per packet > 186
and flags = A
then valid web traffic
Rule 10 for valid web traffic (34; 0.972)
if bytes per packet > 363
and bytes per packet <= 364
and dur <= 0
Model Summary Settings Annotations
OK Cancel Apply Reset
Output Export
Histogram Neural Net Kohonen C5.0 C&R Tree K-Means Table Flat File Database
77MB / 132MB


Stream1* - Clementine 9.0
Table (14 fields, 1,320 records) #1
File
Edit
Generate
date time src proto s_port num orig action dst octet 1 src octet 2 src oct
1 1-Apr-05 0:01:54 144.136.82.3 tcp 55241 2134 144.73.71.201 drop 144.73.205.185 144 136
2 1-Apr-05 0:04:01 144.136.82.3 tcp 59867 4786 144.73.71.201 drop 144.73.169.165 144 136
3 1-Apr-05 0:04:05 144.136.82.3 tcp 59936 4826 144.73.71.201 drop 144.73.224.130 144 136
4 1-Apr-05 0:04:40 144.136.82.3 tcp 61245 5460 144.73.71.201 drop 144.73.230.132 144 136
5 1-Apr-05 0:05:12 144.136.82.3 tcp 62378 6139 144.73.71.201 drop 144.73.35.187 144 136
6 1-Apr-05 0:05:13 144.136.82.3 tcp 62419 6158 144.73.71.201 drop 144.73.251.250 144 136
7 1-Apr-05 0:05:36 144.136.82.3 tcp 63243 6552 144.73.71.201 drop 144.73.194.174 144 136
8 1-Apr-05 0:06:21 144.136.82.3 tcp 64826 7508 144.73.71.201 drop 144.73.88.20 144 136
9 1-Apr-05 0:08:59 144.136.82.3 tcp 8297 10721 144.73.71.201 drop 144.73.201.46 144 136
10 1-Apr-05 0:09:12 144.136.82.3 tcp 8766 11041 144.73.71.201 drop 144.73.171.35 144 136
Table
Annotations
Clementine
Streams
Outputs
Models
Stream1
Port Scanning v2.2
valid web traffic v4.2
valid dns behavior v4.6
Table
Table
Cluster
Data Mining
Predict
Associate
Plot of date v. Record_Count #9
File
Edit
Generate
200
150
Count
9-Sep-05
3-Sep-05
29-Aug-05
18-Aug-05
5-Aug-05
31-Jul-05
12-Jul-05
28-Jun-05
1-Jun-05
22-May-05
5-May-05
18-Apr-05
1-Apr-05
date
Graph
Appearance
Annotations
Favorites
Sources
Record Ops
Field
Table
Matrix
Analysis
Data Audit
Statistics
Quality
Report
Set Globals
Publisher
Server: Local Server
41MB / 49MB


honeypot clustering v.1* - Clementine 11.0
File Edit Insert View Tools SuperNode Window Help
Clementine
Streams Outputs Models
Anomaly
honeypot_1224.txt
Table
Kohonen
Anomaly
Table
Filler
Append
normal traffic.txt
Table (23 fields, 4,883 records)
File Edit Generate
$O-Anomaly $O-AnomalyIndex $O-PeerGroup $O-Field-1 $O-FieldImpact-1 $O-Field-2 $O-f
1 T 4.564 4 flags 0.975 bytes per packet
2 T 5.000 4 dur 0.888 bytes per packet
3 T 5.000 4 flags 0.368 bytes per packet
4 T 5.000 4 bytes per packet 0.512 flags
5 T 5.000 4 flags 0.981 bytes per packet
6 T 4.568 4 flags 0.975 bytes per packet
7 T 4.567 4 flags 0.975 bytes per packet
8 T 4.771 4 flags
9 T 4.567 4 flags
10 T 4.966 4 flags
11 T 4.567 4 flags
12 T 5.000 4 flags
13 T 5.000 4 dur
14 T 4.568 4 flags 0.975
15 T 5.000 4 flags
16 T 4.636 2 dur
Table Annotations
OK
Cluster
Data Mining
Predict
Associate
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export
Binary Classifier Neural Net C5.0 C&R Tree QUEST CHAID Decision List Kohonen K-Means TwoStep Time Series Anomaly Apriori GRI Carma Sequence PCA/Factor Feature Selection
Server: Local Server
113MB / 132MB


valid web traffic v4.2* - Clementine 11.0
File Edit Insert View Tools SuperNode Window Help
Clementine
sp80-22jan_04-08.txt
Partition
known or unknown
known_web
Strip Blanks
Given
• Event and Alarm Data
• Reported Incidents
• Response Data
• Correlated Events
If...
Event 1 takes place, followed by
Event 2, followed by
Event 3, followed by
an Intrusion
• What event is likely to happen next?
• Can we be notified before it happens?
• Can we prevent future events earlier in the cycle?
known or unknown
Filter
discard unused webse..
Table
Filter
bytes per packet
Type
Sequence
CRISP-DM
Classes
Cluster
Data Mining
Predict
Associate
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
Binary Classifier
Neural Net
C5.0
C&R Tree
QUEST
CHAID
Decision List
Kohonen
K-Means
TwoStep
Time Series
Anomaly
Apriori
GRI
Carma
Sequence
PCA/Factor
Feature Selection
Server: Local Server
113MB / 132MB

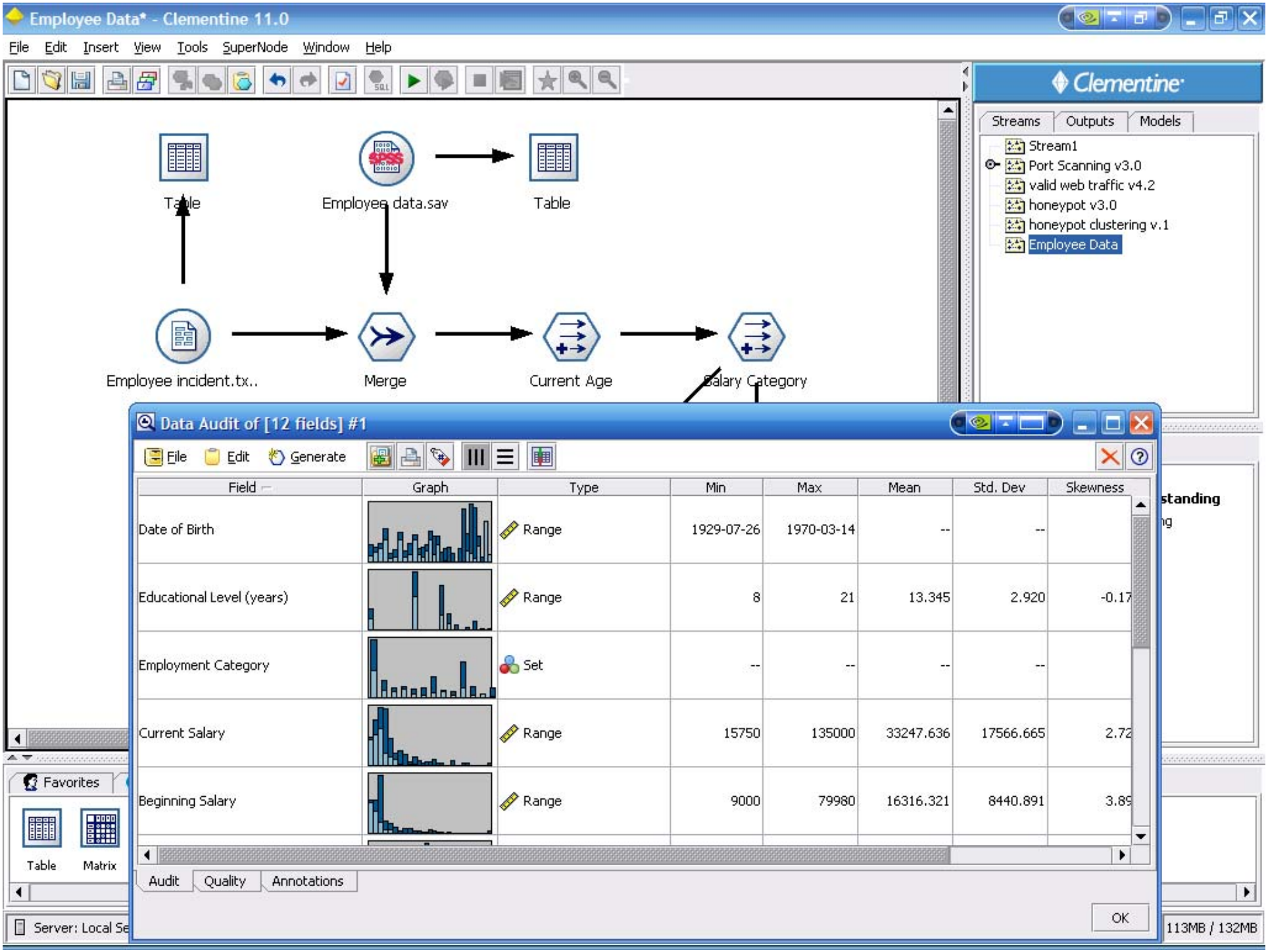


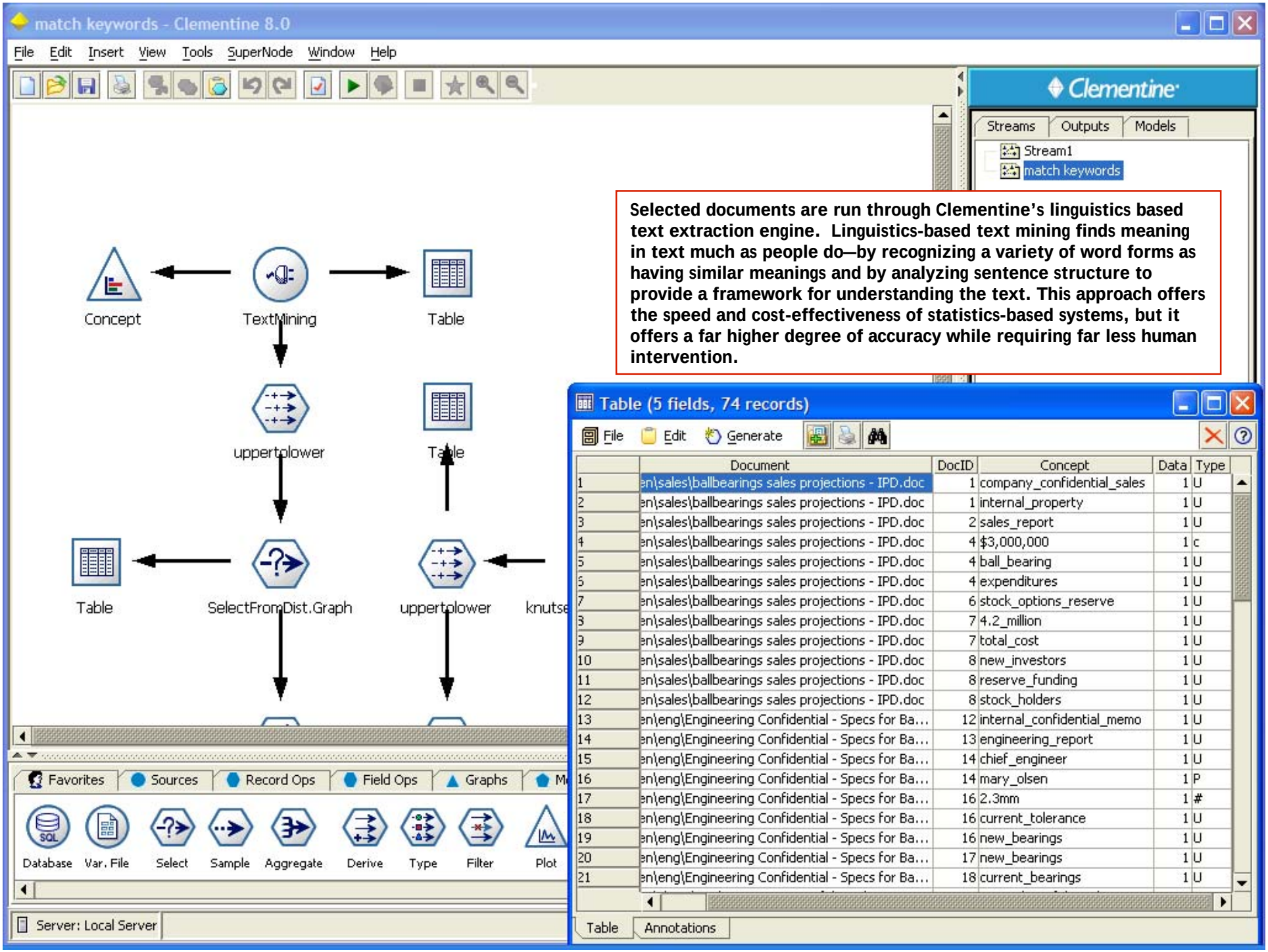
Deployment

Predictive Analysis as:

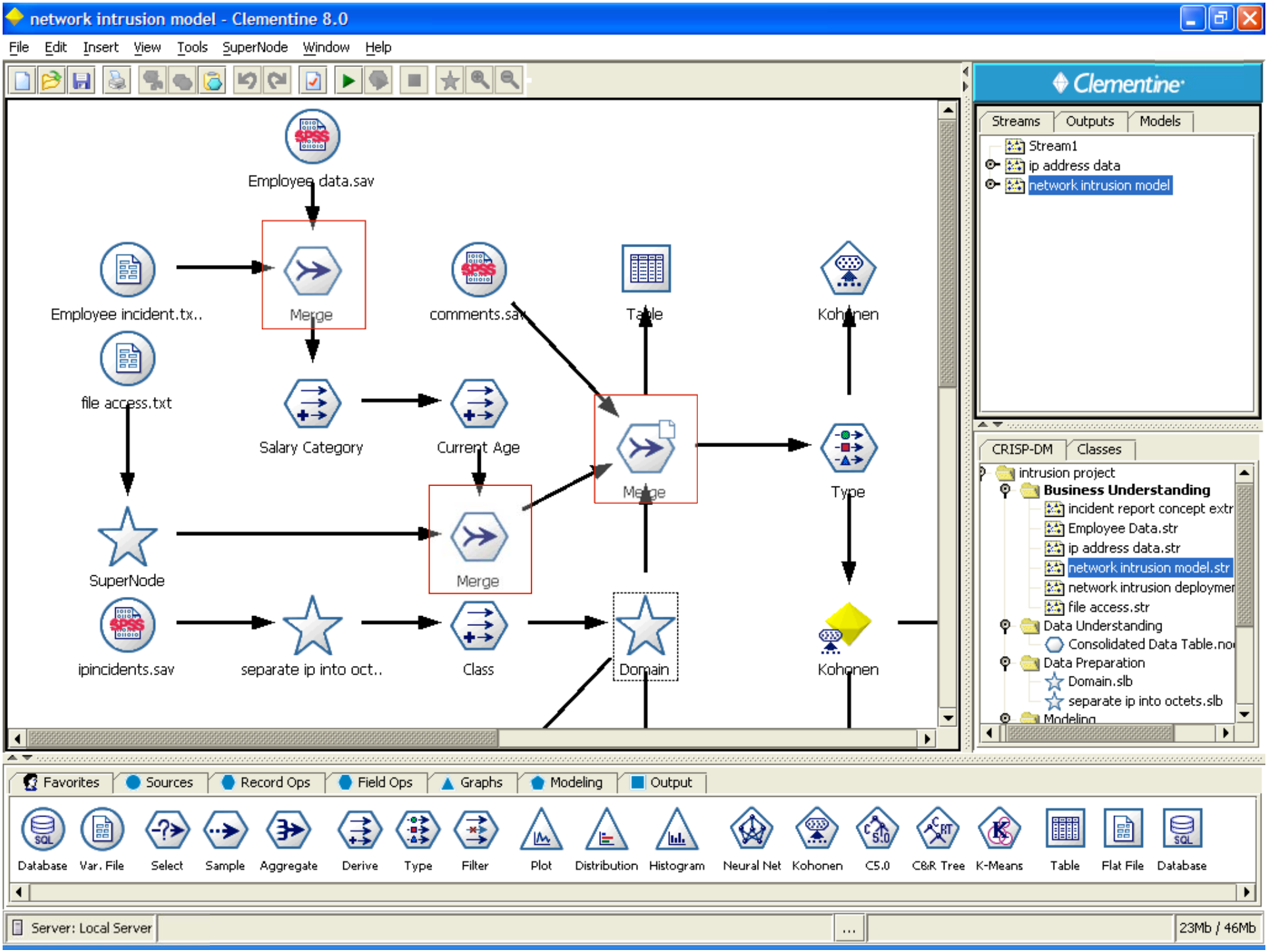
- **an ad-hoc analysis tool**
- **a background process**

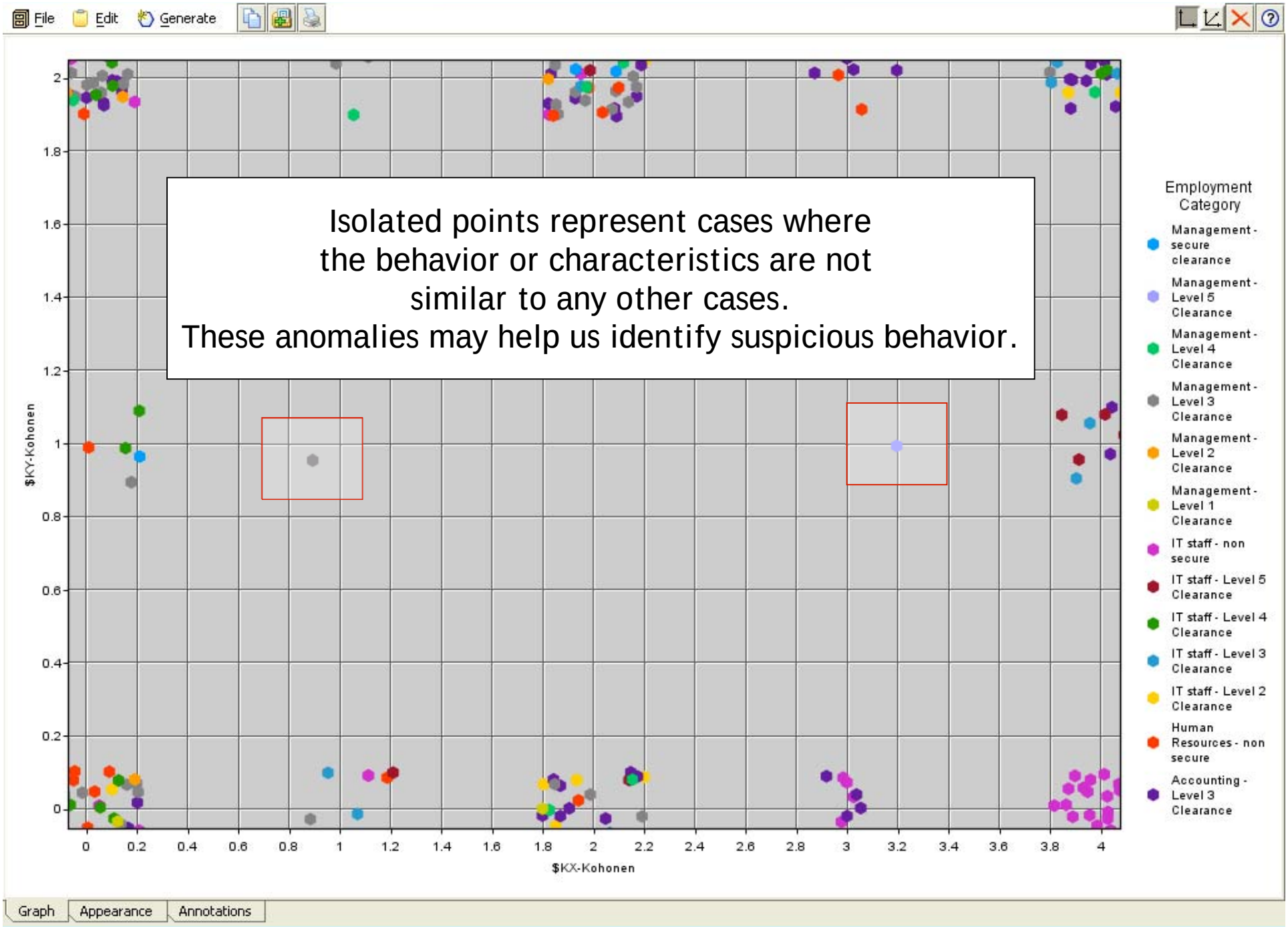
Improving Insider Threat Detection through Predictive Analysis

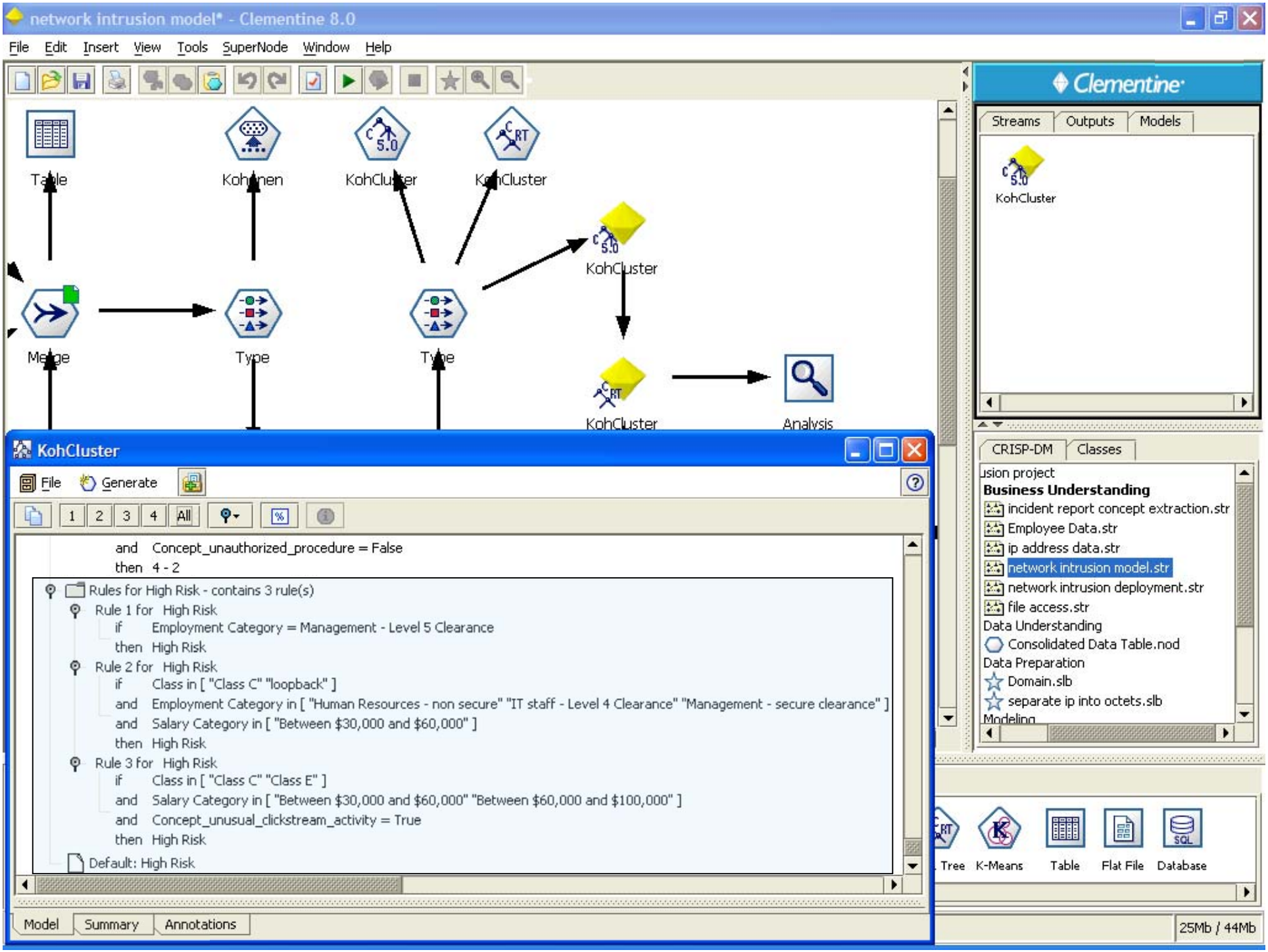

Employee Data* - Clementine 11.0
File Edit Insert View Tools SuperNode Window Help
Clementine
Streams Outputs Models
Stream1
Port Scanning v3.0
valid web traffic v4.2
honeypot v3.0
honeypot clustering v.1
Employee Data
Table
Employee data.sav
Table
Employee incident.tx..
Merge
Current Age
Salary Category
Data Audit of [12 fields] #1
File Edit Generate
Field Graph Type Min Max Mean Std. Dev Skewness
Date of Birth Range 1929-07-26 1970-03-14 -- --
Educational Level (years) Range 8 21 13.345 2.920 -0.17
Employment Category Set -- -- -- --
Current Salary Range 15750 135000 33247.636 17566.665 2.72
Beginning Salary Range 9000 79980 16316.321 8440.891 3.89
Audit Quality Annotations
OK
Favorites
Table
Matrix
Server: Local Se
113MB / 132MB

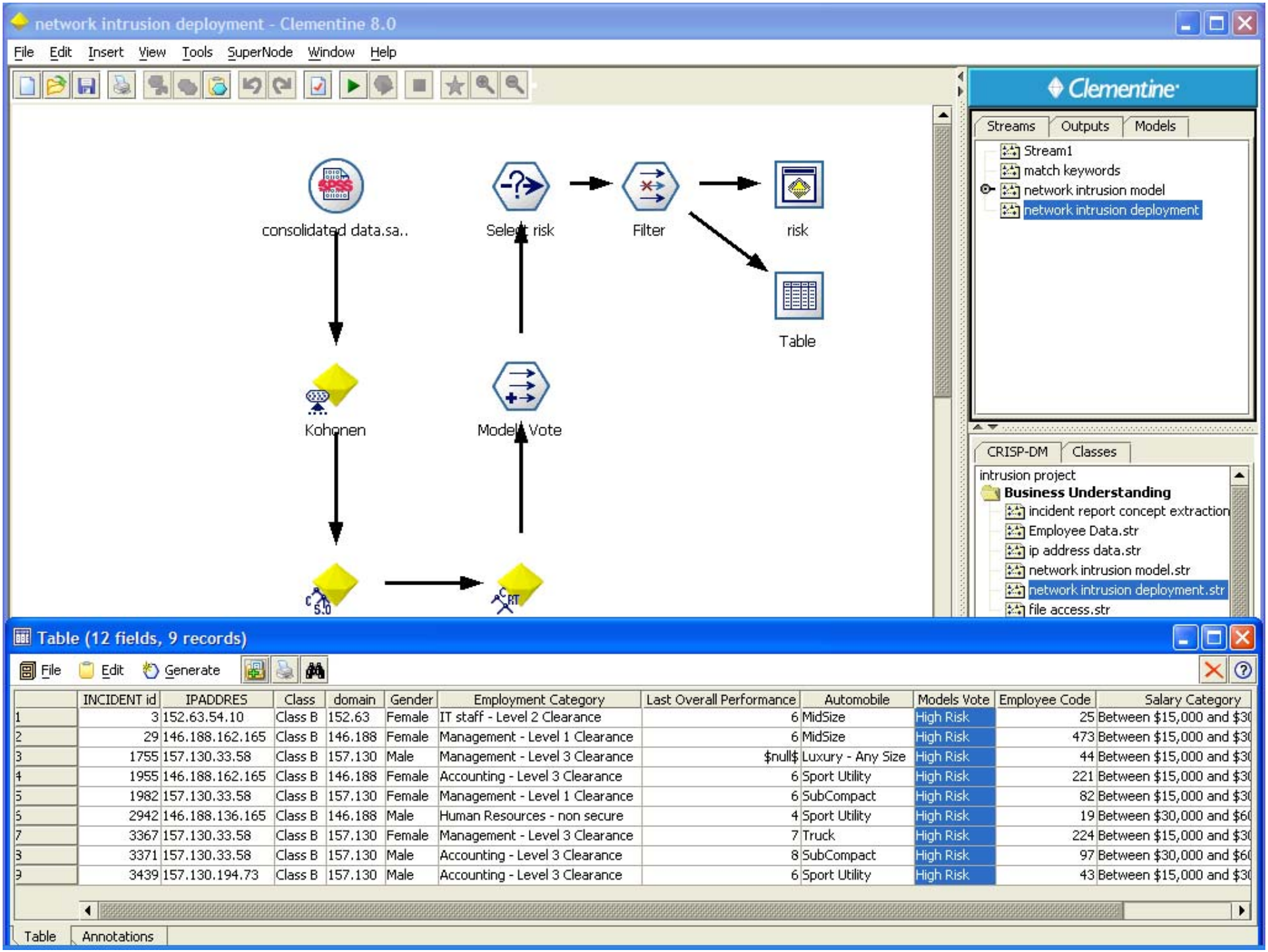


	Document	DocID	Concept	Data	Type
1	en\sales\ballbearings sales projections - IPD.doc	1	company_confidential_sales	1	U
2	en\sales\ballbearings sales projections - IPD.doc	1	internal_property	1	U
3	en\sales\ballbearings sales projections - IPD.doc	2	sales_report	1	U
4	en\sales\ballbearings sales projections - IPD.doc	4	$3,000,000	1	c
5	en\sales\ballbearings sales projections - IPD.doc	4	ball_bearing	1	U
6	en\sales\ballbearings sales projections - IPD.doc	4	expenditures	1	U
7	en\sales\ballbearings sales projections - IPD.doc	6	stock_options_reserve	1	U
8	en\sales\ballbearings sales projections - IPD.doc	7	4.2_million	1	U
9	en\sales\ballbearings sales projections - IPD.doc	7	total_cost	1	U
10	en\sales\ballbearings sales projections - IPD.doc	8	new_investors	1	U
11	en\sales\ballbearings sales projections - IPD.doc	8	reserve_funding	1	U
12	en\sales\ballbearings sales projections - IPD.doc	8	stock_holders	1	U
13	en\eng\Engineering Confidential - Specs for Ba...	12	internal_confidential_memo	1	U
14	en\eng\Engineering Confidential - Specs for Ba...	13	engineering_report	1	U
15	en\eng\Engineering Confidential - Specs for Ba...	14	chief_engineer	1	U
16	en\eng\Engineering Confidential - Specs for Ba...	14	mary_olsen	1	P
17	en\eng\Engineering Confidential - Specs for Ba...	16	2.3mm	1	#
18	en\eng\Engineering Confidential - Specs for Ba...	16	current_tolerance	1	U
19	en\eng\Engineering Confidential - Specs for Ba...	16	new_bearings	1	U
20	en\eng\Engineering Confidential - Specs for Ba...	17	new_bearings	1	U
21	en\eng\Engineering Confidential - Specs for Ba...	18	current_bearings	1	U


network intrusion model - Clementine 8.0
File Edit Insert View Tools SuperNode Window Help
Employee data.sav
Employee incident.tx..
Merge
comments.sav
Table
Kohonen
file access.txt
Salary Category
Current Age
Merge
Type
SuperNode
Merge
ipincidents.sav
separate ip into oct..
Class
Domain
Kohonen
Clementine
Streams Outputs Models
Stream1
ip address data
network intrusion model
CRISP-DM Classes
intrusion project
Business Understanding
incident report concept extr
Employee Data.str
ip address data.str
network intrusion model.str
network intrusion deploymer
file access.str
Data Understanding
Consolidated Data Table.no
Data Preparation
Domain.slb
separate ip into octets.slb
Modeling
Favorites Sources Record Ops Field Ops Graphs Modeling Output
Database Var. File Select Sample Aggregate Derive Type Filter Plot Distribution Histogram Neural Net Kohonen C5.0 C&R Tree K-Means Table Flat File Database
Server: Local Server
23Mb / 46Mb




network intrusion model* - Clementine 8.0
File Edit Insert View Tools SuperNode Window Help
Table
Kohonen
KohCluster
KohCluster
KohCluster
Merge
Type
Type
KohCluster
Analysis
Clementine
Streams Outputs Models
KohCluster
CRISP-DM Classes
Business Understanding
incident report concept extraction.str
Employee Data.str
ip address data.str
network intrusion model.str
network intrusion deployment.str
file access.str
Data Understanding
Consolidated Data Table.nod
Data Preparation
Domain.slb
separate ip into octets.slb
Modeling
KohCluster
File Generate
1 2 3 4 All
and Concept_unauthorized_procedure = False
then 4 - 2
Rules for High Risk - contains 3 rule(s)
Rule 1 for High Risk
if Employment Category = Management - Level 5 Clearance
then High Risk
Rule 2 for High Risk
if Class in ["Class C" "loopback"]
and Employment Category in ["Human Resources - non secure" "IT staff - Level 4 Clearance" "Management - secure clearance"]
and Salary Category in ["Between $30,000 and $60,000"]
then High Risk
Rule 3 for High Risk
if Class in ["Class C" "Class E"]
and Salary Category in ["Between $30,000 and $60,000" "Between $60,000 and $100,000"]
and Concept_unusual_clickstream_activity = True
then High Risk
Default: High Risk
Model Summary Annotations
Tree K-Means Table Flat File Database
25Mb / 44Mb



	INCIDENT id	IPADDRES	Class	domain	Gender	Employment Category	Last Overall Performance	Automobile	Models Vote	Employee Code	Salary Category
1	3	152.63.54.10	Class B	152.63	Female	IT staff - Level 2 Clearance	6	MidSize	High Risk	25	Between $15,000 and $30
2	29	146.188.162.165	Class B	146.188	Female	Management - Level 1 Clearance	6	MidSize	High Risk	473	Between $15,000 and $30
3	1755	157.130.33.58	Class B	157.130	Male	Management - Level 3 Clearance	$null$	Luxury - Any Size	High Risk	44	Between $15,000 and $30
4	1955	146.188.162.165	Class B	146.188	Female	Accounting - Level 3 Clearance	6	Sport Utility	High Risk	221	Between $15,000 and $30
5	1982	157.130.33.58	Class B	157.130	Female	Management - Level 1 Clearance	6	SubCompact	High Risk	82	Between $15,000 and $30
6	2942	146.188.136.165	Class B	146.188	Male	Human Resources - non secure	4	Sport Utility	High Risk	19	Between $30,000 and $60
7	3367	157.130.33.58	Class B	157.130	Female	Management - Level 3 Clearance	7	Truck	High Risk	224	Between $15,000 and $30
8	3371	157.130.33.58	Class B	157.130	Male	Accounting - Level 3 Clearance	8	SubCompact	High Risk	97	Between $30,000 and $60
9	3439	157.130.194.73	Class B	157.130	Male	Accounting - Level 3 Clearance	6	Sport Utility	High Risk	43	Between $15,000 and $30

Final Thoughts and Questions

- Visit www.spss.com
- Future Questions:
  - Peter Frometa – pfrometa@spss.com