2. DATA

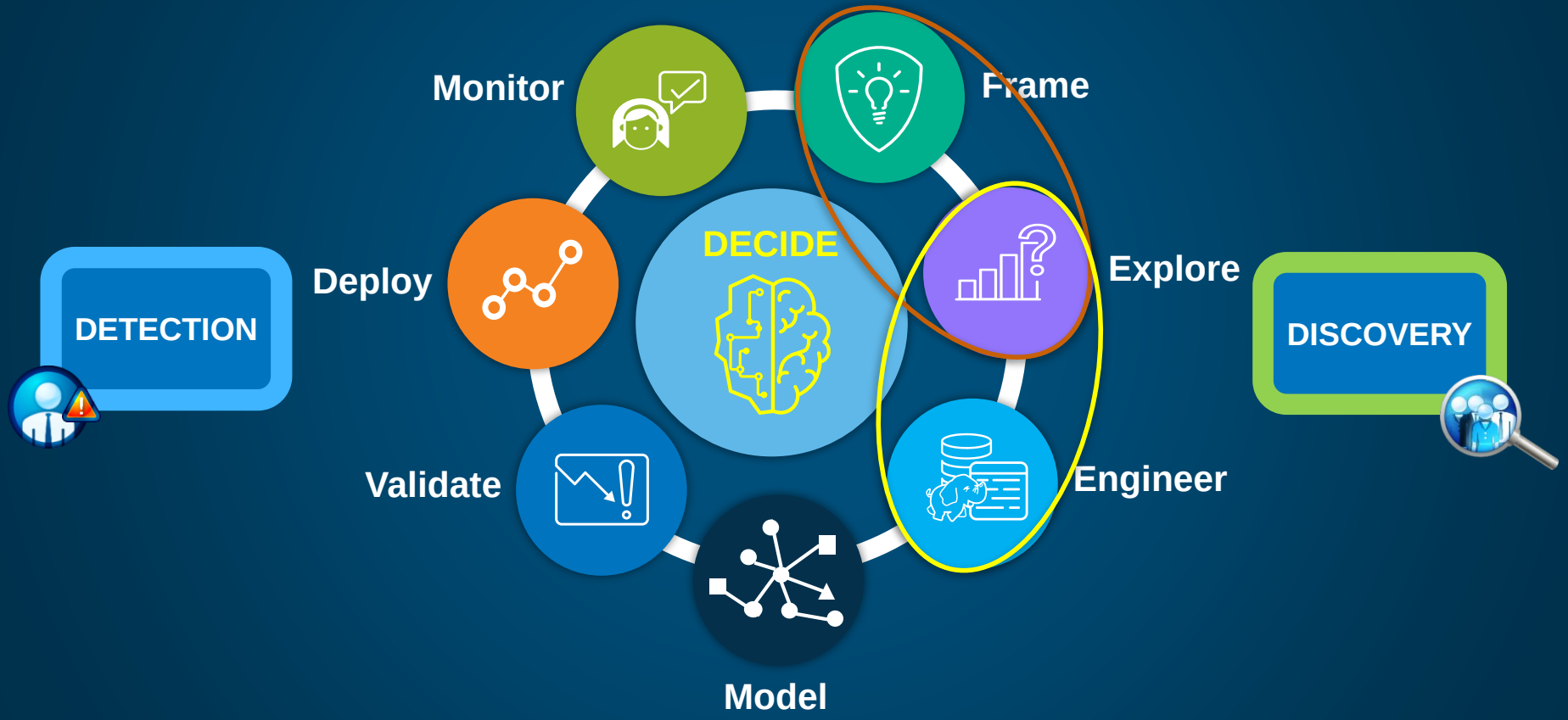
Challenges, Sources, Features, Methods

Cybersecurity Data Science (CSDS)

TOPIC
1. FRAME
2. DATA
3. DISCOVER
4. DETECT
5. DEPLOY

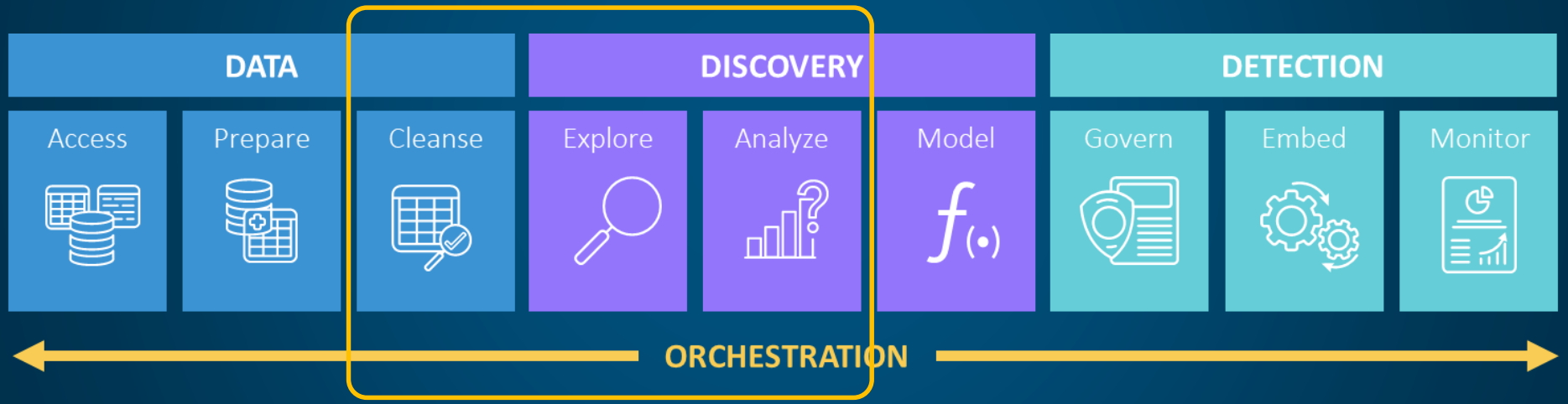
Learning Objectives

Cybersecurity Data Science (CSDS) Lifecycle

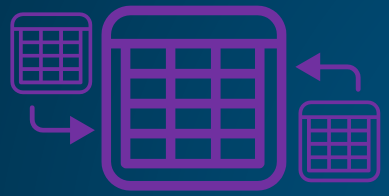


CSDS Process

Unified Orchestration



Data Management



Preparation

Quality

Structure

Integration

Importance of Data Preparation

Which data set would you rather for analytics?

Before Data Preparation

ID	NAME	COUNTRY	EMAIL_ADDRESS
101	CARLOS VIERA	US	CARLOS.VIERA@XXX.COM
102	Richard Schmidt	US	richard.schmidt@xxx.com
103	Michael Jameson	United States	michael.jameson@xxx.com
104	Albert Moore	US	AL.Moore@xxx.com
105	Harvey L Monk	United States	harvey.monk@xxx.com
106	Shelley Holmes	USA	sholmes@xxx.com
107	Al Moore	US	al.moore@xxx.com
108	MIKE JAMESON	USA	Michael.Jameson@xxx.com
109	Shelly Holms	USA	sholmes@xxx.com
110	Anne Horton	US	Anne.Horton@XXX.com

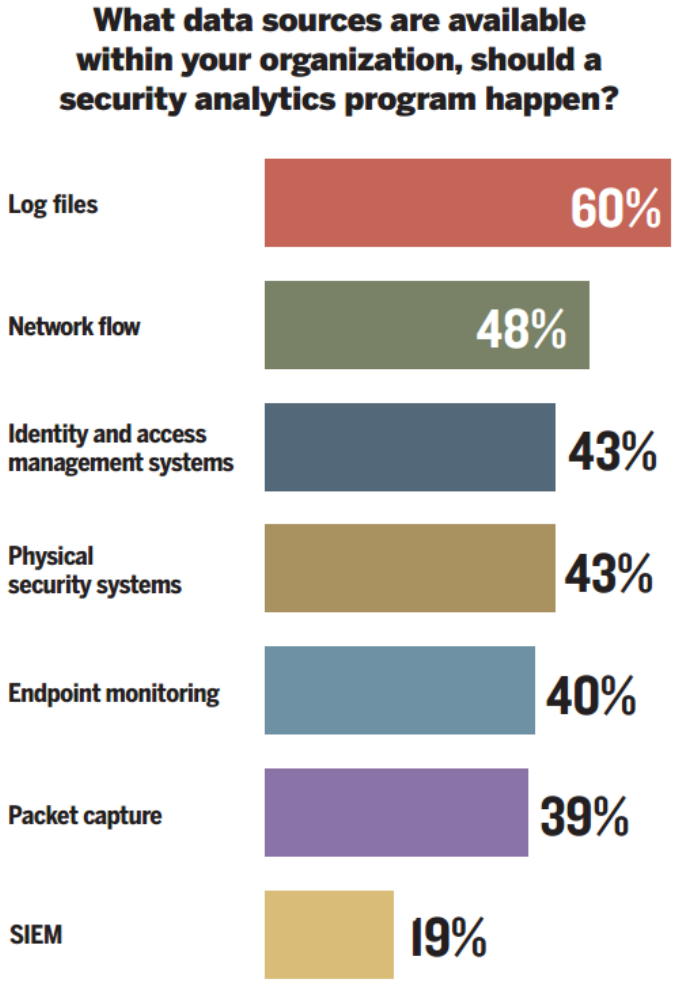
After Data Preparation

ID	NAME	COUNTRY	EMAIL_ADDRESS	GENDER
101	Carlos Viera	USA	carlos.viera@xxx.com	M
102	Richard Schmidt	USA	richard.schmidt@xxx.com	M
103	Michael Jameson	USA	michael.jameson@xxx.com	M
104	Albert Moore	USA	al.moore@xxx.com	M
105	Harvey L Monk	USA	harvey.monk@xxx.com	M
106	Shelley Holmes	USA	sholmes@xxx.com	F
110	Anne Horton	USA	anne.horton@xxx.com	F

Objectives of Data Discovery

- Framing data in the cybersecurity analytics lifecycle
- Refining data: ‘feature extraction / selection’
  - Assessing and exploring to gain a basic insights
  - Collecting, consolidating, and cleaning
  - Transforming and extracting new measures
  - Establishing a foundation for pattern analysis
  - Reducing and refining variables
- Hands-on data exploration / extraction

Cybersecurity Data Challenges



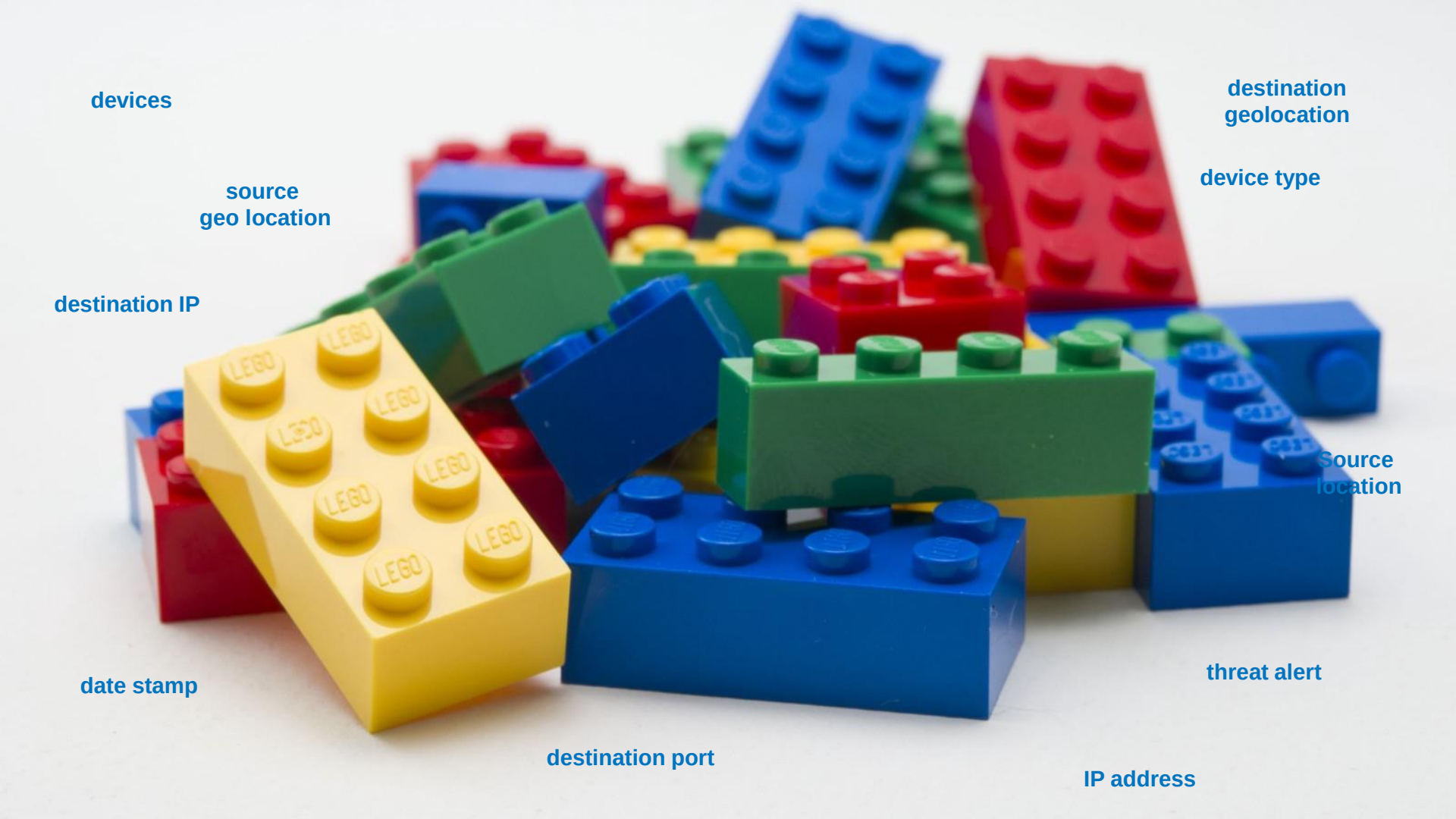
SOURCE
Security Brief Magazine. (2016). "Analyze This! Who's Implementing Security Analytics Now?" Available at https://www.sas.com/en_th/whitepapers/analyze-this-108217.html

IP address

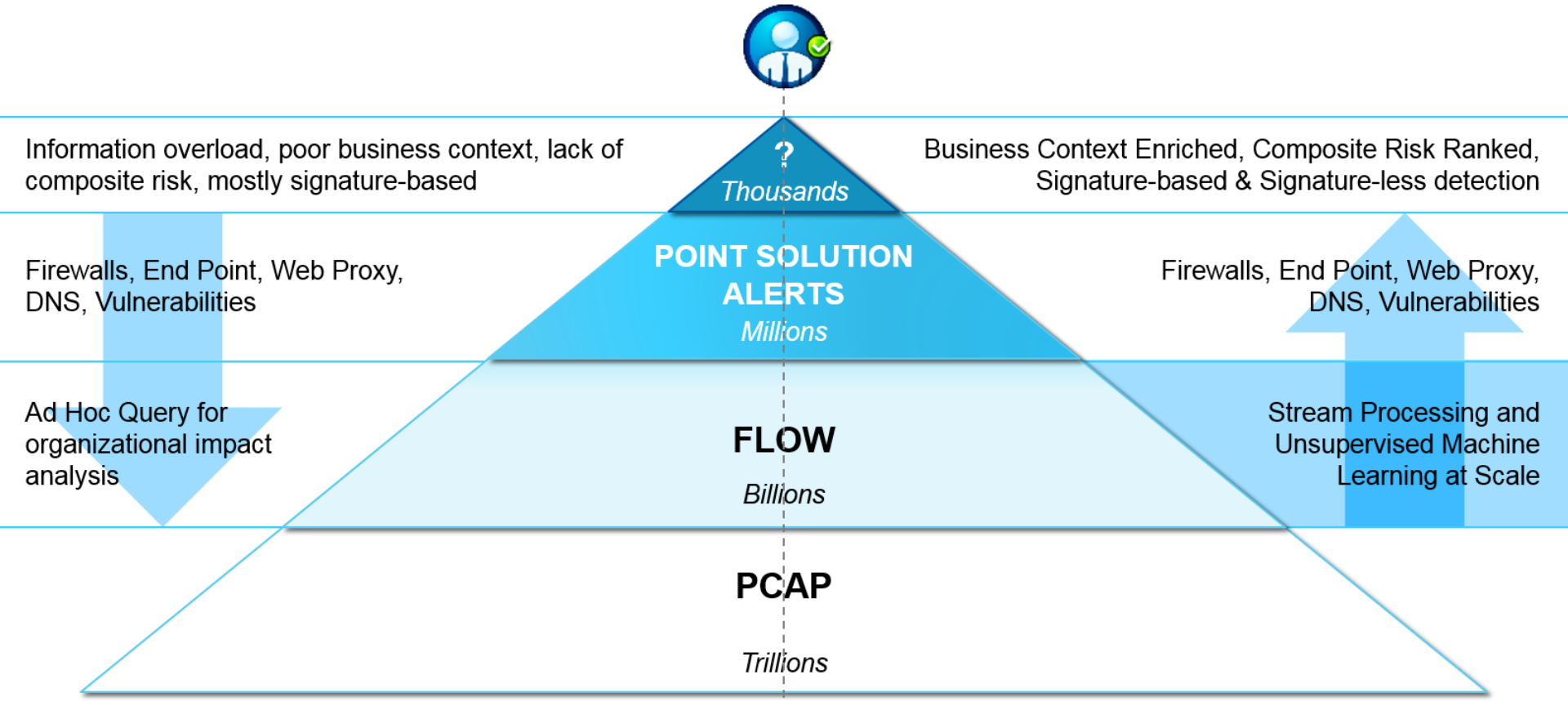
time stamp

userid

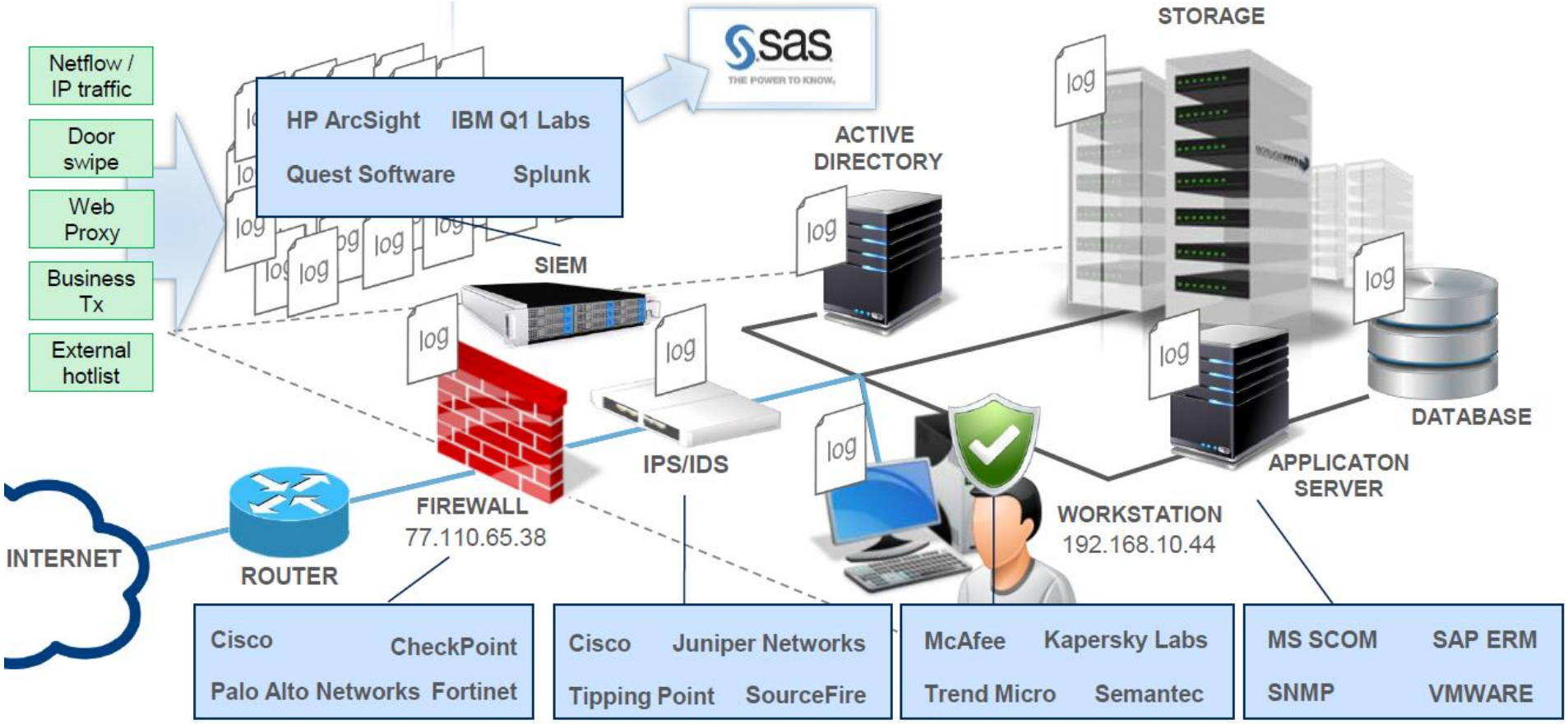
destination port


devices
source
geo location
destination IP
date stamp
destination port
IP address
threat alert
Source
location
device type
destination
geolocation

Security Data Management Challenge: Speed and Volumes



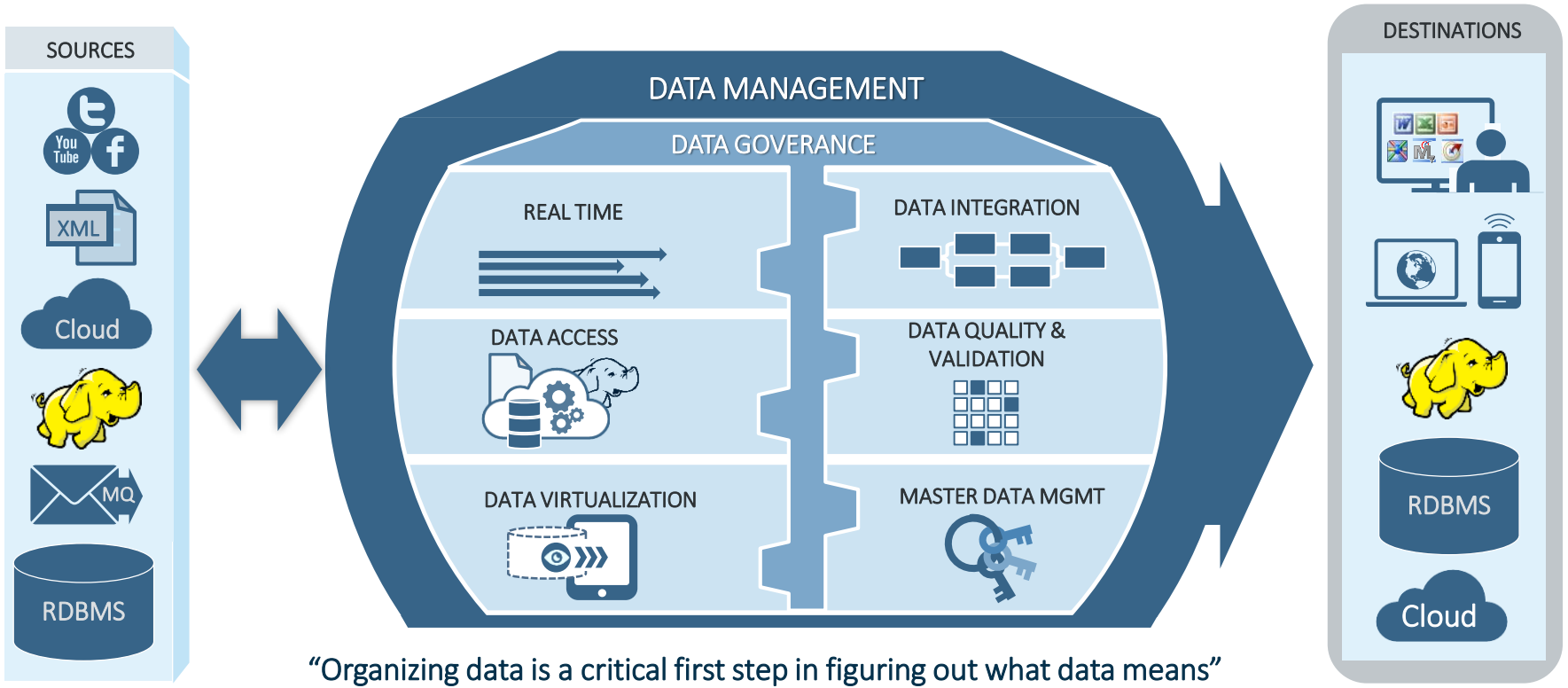
Many data sources… increasing data volumes



High false alerts… slow investigation processes


LACK OF CONTEXT
DISCONNECTED & FRAGMENTED
UNVALIDATED ALERTS
VOLUME & SPEED
MULTIPLE SYSTEMS
?
!

Data Engineering: Fusion, Quality and Delivery



"Organizing data is a critical first step in figuring out what data means"

Larry Alton, Information Management Feb 14th, 2019

Cybersecurity Frameworks & Ontologies

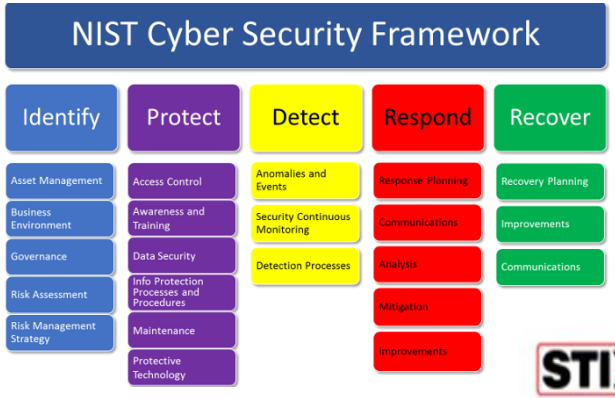
FRAMEWORKS

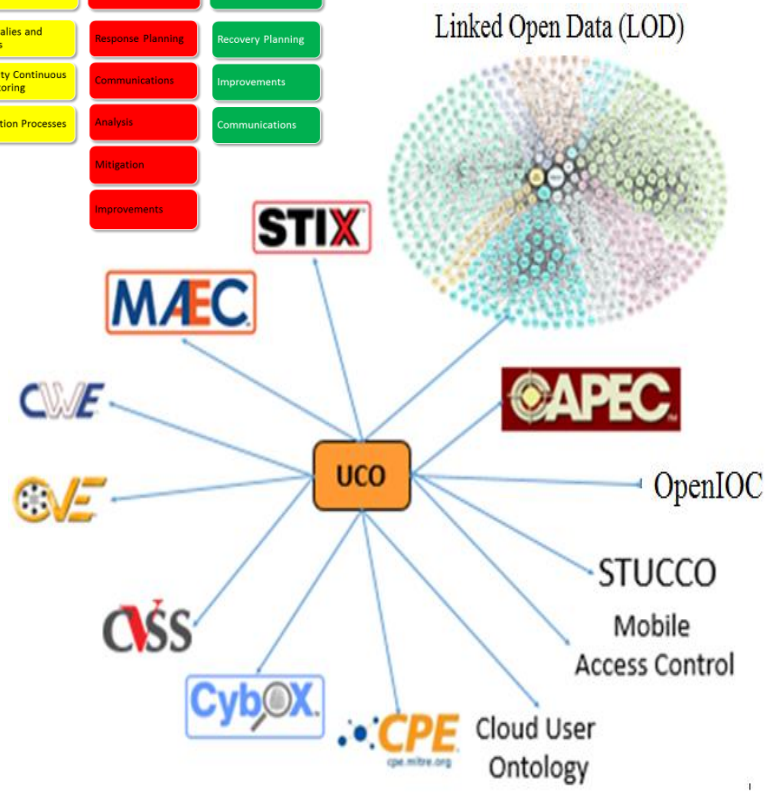
- MITRE Cyber Observable eXpression
- NIST Cybersecurity Framework
- Intrusion Kill Chain (Lockheed Martin)

ONTOLOGIES

- **DFAX** Digital Forensic Analysis eXpression
- **CVE** Cyber Intelligence Ontology
- **ICAS** Information Security (example)
- **UCO / UCO (OWL)**

  Unified Cybersecurity Ontology





U.S. Cyber Incident Data Sharing Specifications

- **Common Attack Pattern Enumeration and Classification (CAPEC):** Structured framework for describing known tactics, techniques, and procedures (TPP) applied by adversaries.
- **Cyber-investigation Analysis Standard Expression (CASE):** Open standard ontology/specification language for sharing cybersecurity case investigation information
- **Cyber Observable eXpression (CybOX):** A semantic framework for describing objects and properties in the cybersecurity domain (merged into STIX).
- **Incident Object Description Format (IODEF):** Standard data format for the exchange of incident information between security teams.
- **Integrated Cyber Analysis System (ICAS):** A U.S. DARPA initiative for documenting infrastructure to aid attack forensics and tactical cyber defense incident response
- **Malware Attribute Enumeration and Classification (MAEC):** A semantic framework for describing structured malware behavior.
- **OASIS Customer Information Quality (CIQ):** A language for representing information about individuals and organizations.
- **Structured Threat Information Expression (STIX):** A structured language specification for describing cyber threat information so it can be shared, stored, and analyzed in a consistent manner. This initiative embeds or ties to several of the other initiatives listed and is overseen by the **OASIS Cyber Threat Intelligence Technical Committee (OASIS, 2019).**
- **Unified Cyber Ontology (UCO):** A common ontology to unify and represent disparate cybersecurity domain knowledge.
- **Vocabulary for Event Recording and Incident Sharing (VERIS):** A formal metrics framework for describing security incidents and their effects in a structured manner.

https://www.us-cert.gov/Information-Sharing-Specifications-Cybersecurity


POLICE

Whitepaper: A Comprehensive Approach to Big Data Governance, Data Management and Analytics

http://sww.sas.com/cosmos/a/cosmos-images/107968_0718.pdf

SAS-Hadoop Integration Tools

SAS Interfaces to Hadoop

1) In-Database Processing with Hadoop
http://support.sas.com/documentation/cdl/en/acreldb/69580/HTML/default/viewer.htm#n0kgg6z8c14ewmn1phdwdm5cp51i.htm

2) SAS/ACCESS Interface to Hadoop
https://www.sas.com/en_ph/software/data-management/access-hadoop.html

3) SAS Data Loader for Hadoop (*Hadoop ETL*)
https://www.sas.com/en_us/software/data-loader-for-hadoop.html

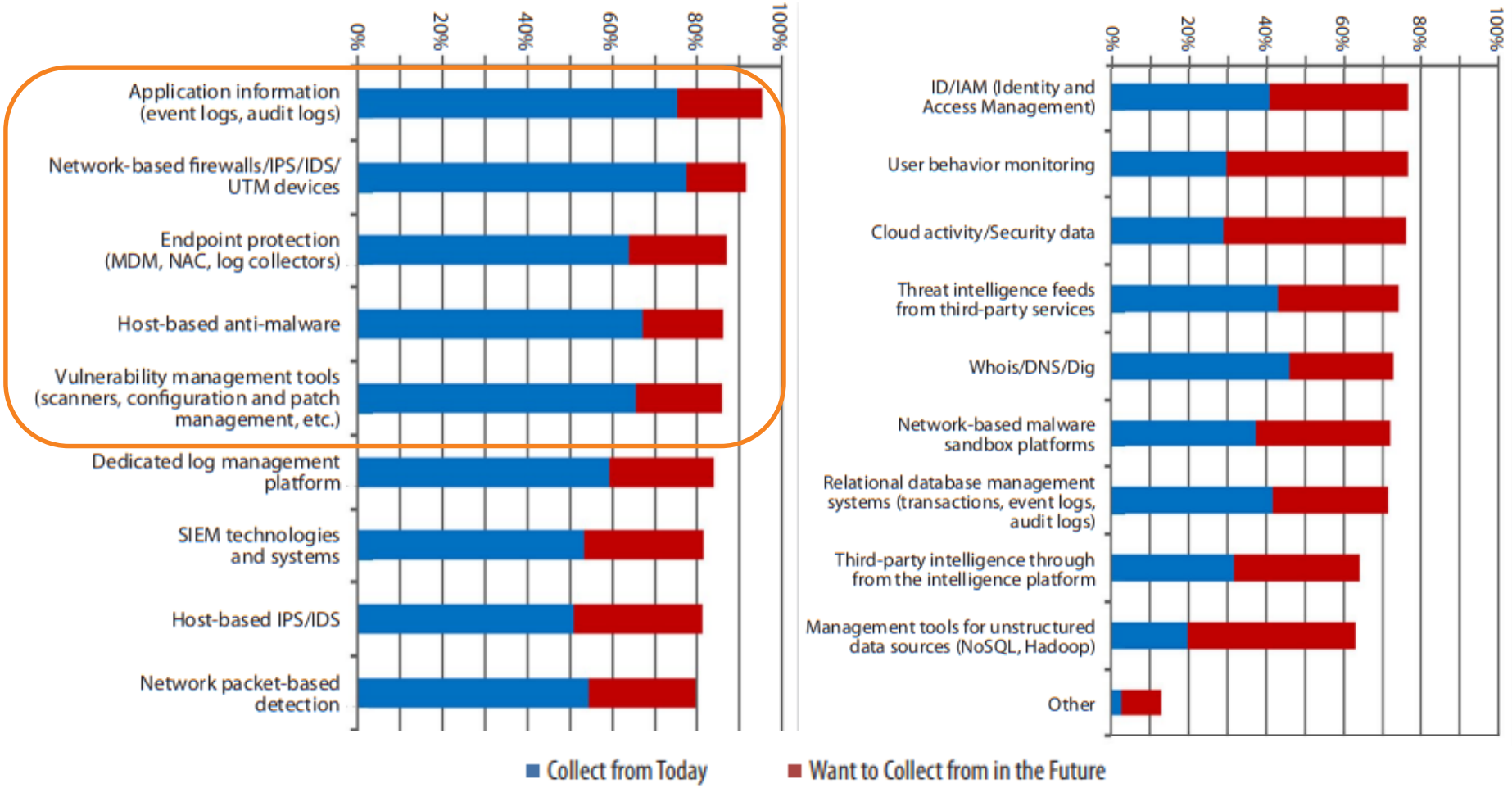
White Paper: 'Eight Considerations for utilizing Big Data Analytics with Hadoop:
https://www.sas.com/en_us/offers/sem/tdwi-8-considerations-utilizing-big-data-analytics-with-hadoop-107015.html?gclid=CjwKEAjw07nJBRDG_tvshefHhWQSJABRcE-ZvRH45VlQzQ_iUmKsrj_jF8ftzNRHjdB9WyIxr9YQYBoCNyHw_wcB

Cybersecurity Data Sources

Common Cybersecurity Data Sources

- CMDB
  (internal configuration/asset catalogs)
- Intrusion detection/prevention system
- Vulnerability scans
- SIEM-generated
- Endpoint data
- NetFlow
- Network packets
- SaaS and cloud logs

- DNS records
- Third-party reports
  (threat intelligence/feeds e.g. STIX/TAXII)
- Network component-generated logs
  (firewall, router, bridge, DHCP server, proxy server, typical device types, and so on)
- Device configurations, rules, traffic
  (firewall, router, switch)
- Contextual / 'demographic' data
  (organizational, user-demographic, and so on)
- SaaS and cloud logs


0%
20%
40%
60%
80%
100%
Application information (event logs, audit logs)
Network-based firewalls/IPS/IDS/ UTM devices
Endpoint protection (MDM, NAC, log collectors)
Host-based anti-malware
Vulnerability management tools (scanners, configuration and patch management, etc.)
Dedicated log management platform
SIEM technologies and systems
Host-based IPS/IDS
Network packet-based detection
ID/IAM (Identity and Access Management)
User behavior monitoring
Cloud activity/Security data
Threat intelligence feeds from third-party services
Whois/DNS/Dig
Network-based malware sandbox platforms
Relational database management systems (transactions, event logs, audit logs)
Third-party intelligence through from the intelligence platform
Management tools for unstructured data sources (NoSQL, Hadoop)
Other
Collect from Today
Want to Collect from in the Future

Sourcing Own Data: Open-Source Distributions / Tools

Suricata

- Network threat detection engine
- Real time intrusion detection (IDS)
- Inline intrusion prevention (IPS)
- Network security monitoring (NSM)
- Flow and malware detection probe
- Integration with external systems (i.e., SIEMs, Splunk, ELK stack)
- Supported by non-profit foundation open source project (OISF)

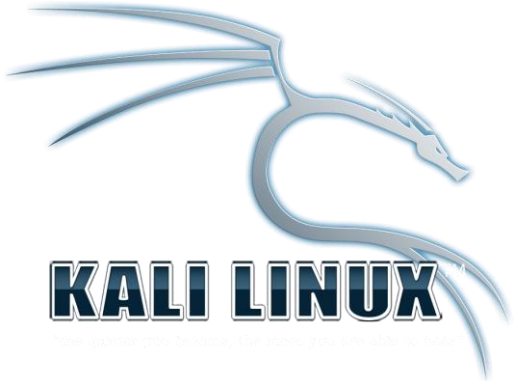
Bro

- Network security monitoring
- Intrusion detection
- Signature detection
- Network traffic analysis
- Network discovery
- Active support from academia, research labs, and open source

Sourcing Own Data: Open-Source Distributions / Tools

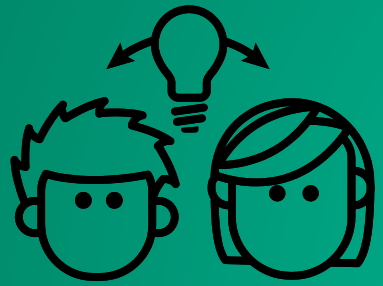
Kali Linux

- Information gathering
- Vulnerability analysis
- Wireless attacks
- Web applications
- Exploitation tools
- Forensics tools
- Stress testing
- Sniffing & spoofing
- Password attacks
- Maintaining access
- Reverse engineering
- Reporting tools
- Hardware hacking



Security Onion

- Full packet capture
  - Netsniff-ng
- NW and host-based intrusion detection
  - NIDS, Bro IDS, HIDS
- Analysis tools
  - Sguil, Squertm Enterprise Log Search & Archive



Idea Exchange

What do we hope to extract from cybersecurity data?

What Do We Hope to Gain from Cybersecurity Data?

RULES & ALERTS (SIEM)

- Real-time status
  - Exceptions
  - Known indicators
- Processes / sequences
- Forensics
  - Changing status and order of processes and events
- Aggregate understanding
  - Complex indicators

DETECT, PREDICT, OPTIMIZE (ANALYTICS)

- Probabilistic 'behavioral' understanding
  - What is 'normal' as baseline?
  - Different time ranges / processes
  - Insights into and *between* various entities, categories layers, and levels (individual, group, domain, etc.)
- Emerging anomalies (exploration)
- Complex detection (case in aggregate)
- Quantifiable risk models
- Optimization
  - Match best resource to most pressing risks

Advanced Insights

Datasets for cybersecurity training, research, and development

Sources for Cybersecurity Research Data

- HoneyPot Project: http://honeynet.org/challenges
- LANL CSR Red Teaming: https://csr.lanl.gov/data/cyber1/
- CTU-13 CTU University, Czech Rep https://mcfp.weebly.com/the-ctu-13-dataset-a-labeled-dataset-with-botnet-normal-and-background-traffic.html
- SecRepo.com: http://www.secrepo.com/
- VizSec: http://vizsec.org/data/
- Data.gov Cyber Data Sets: https://catalog.data.gov/dataset?tags=cybersecurity
- Malware Traffic Analysis: http://malware-traffic-analysis.net/
- MIT Lincoln Laboratory IDS Data Sets: https://www.ll.mit.edu/ideval/data/
- Center for Applied Internet Data Analysis (CAIDA) Data Sets: http://www.caida.org/data/overview/
- Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT): https://www.dhs.gov/publication/dhsstpia-006-protected-repository-defense-infrastructure-against-cyber-threats
- NSA Cyber Defense Exercise Data Set: https://www.iad.gov/iad/programs/cyber-defense-exercise/index.cfm

Example Research Dataset: LANL CSR Red Teaming

LANL CSR Red Teaming: https://csr.lanl.gov/data/cyber1/

- 58 consecutive days of de-identified event data collected from five sources within Los Alamos National Laboratory's corporate, internal computer network
- 12 gigabytes compressed across five data elements
  - 1,648,275,307 events in total
  - 12,425 users
  - 17,684 computers
  - 62,974 processes
- Data sources include
  - **Windows-based authentication events** from **individual computers** and **Active Directory** servers;
  - **Process start and stop events** from individual Windows computers;
  - **Domain Name Service (DNS) lookups** as collected on internal DNS servers;
  - **Network flow data** as collected on at several key router locations
  - Set of well-defined **red teaming events** that present bad behavior within the 58 days

Example Research Dataset: LANL CSR Red Teaming

"Lessons learned in cybersecurity data munging" (academic paper in progress)

- Block out some serious time…
- Don't underestimate latency (processing)
- Well, you could get a Hadoop/Elastic repository… but…
- Start small to develop an understanding: sample, sample, sample!
- Develop a 'theory' of the central entities of interest
  - Prepare to have your theory destroyed
  - i.e. "Computer" was likely an IP associated to a user and not e.g. MAC address (as a result, average of ~15 computers per user)
  - "Domain" more comprehensible (average of 3 per user)
- Importance of time epochs / time slices

Example Research Dataset: LANL – Extracting Features

1. Time
2. Source user@domain
3. Destination user@domain
4. Source computer
5. Destination computer
6. Success/failure
7. Authentication type
8. Login type
9. Authentication orientation

#	FIELD	DESCRIPTION
1	time_auth	Cumulative second epoch (0-to-N)
2	DateTimeSAS	SAS date format
3	Hour_round	Cumulative hour epoch, rounded down (0-to-N)
4	Hour	Cumulative hour epoch (0-to-N)
5	Day	Cumulative day epoch, rounded (0-to-N)
6	DateTime	Julian date/time
7	DatePart	Julian date
8	TimePart	Time
9	Time_Code	1=0:00-5:59; 2=6:00-11:59; 3=12:00-17:59; 4=18:00-23:59
10	Weekend	Weekend day (0,1)?
11	Weekday	Day of week (1=Sun; 7 = Sat)
12	WeekNum	Week number of year
13	source_userdom	Full source user@domain (user$ = comp acct; @DOM = AD)
14	source_user	Resolved source userid
15	source_user_comp	Computer source user account? (0,1)
16	source_dom	Resolved source domain
17	source_dom_AD	Source domain via AD (DOM)? (0,1)
18	source_comp	Computer/device
19	dest_userdom	Full destination user@domain (user$ = comp acct; @DOM = AD)
20	dest_user	Resolved destination userid
21	dest_user_different	Different source-dest users (0,1)
22	dest_user_comp	Computer destination user account? (0,1)
23	dest_dom	Resolved destination domain
24	dest_dom_AD	Destination domain via AD (DOM)? (0,1)
25	dest_comp	Computer/device
26	dest_comp_different	Different source-dest computers (0,1)
27	dest_comp_TGT	TGT (ticket) authentication (0,1)?
28	dest_comp_USER	User authentication (0,1)?
29	suc_fail	Authentication fail (0,1)?
30	auth_type	Authorization type (categorical)
31	authtype_?	0, 1
32	authtype_microsoft	0, 1
33	authtype_netware_a	0, 1
34	authtype_kerberos	0, 1
35	authtype_ntlm	0, 1
36	authtype_negotiate	0, 1
37	authtype_wave	0, 1
38	authtype_other	0, 1
39	logon_type	Logon type (categorical)
40	logontype_batch	0, 1
41	logontype_interactive	0, 1
42	logontype_cached	0, 1
43	logontype_reminder	0, 1
44	logontype_network	0, 1
45	logontype_nwcleartxt	0, 1
46	logontype_service	0, 1
47	logontype_unlock	0, 1
48	logontype_newcred	0, 1
49	auth_orient	Authorization orientation (categorical)
50	authorient_authmap	0, 1
51	authorient_logoff	0, 1
52	authorient_logon	0, 1
53	authorient_screenlock	0, 1
54	authorient_screenunlock	0, 1
55	authorient_tgs	0, 1
56	authorient_tgt	0, 1

Data Engineering

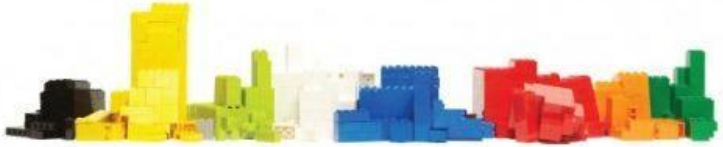
Cybersecurity Events

Irregular and Complex Events

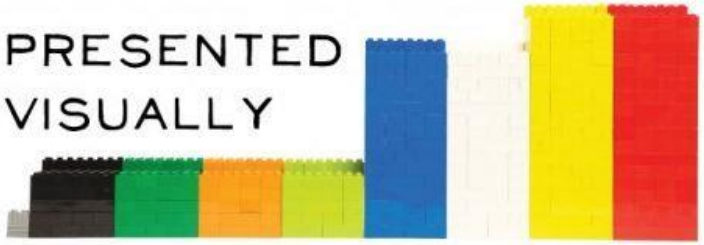
DATA

SORTED

ARRANGED

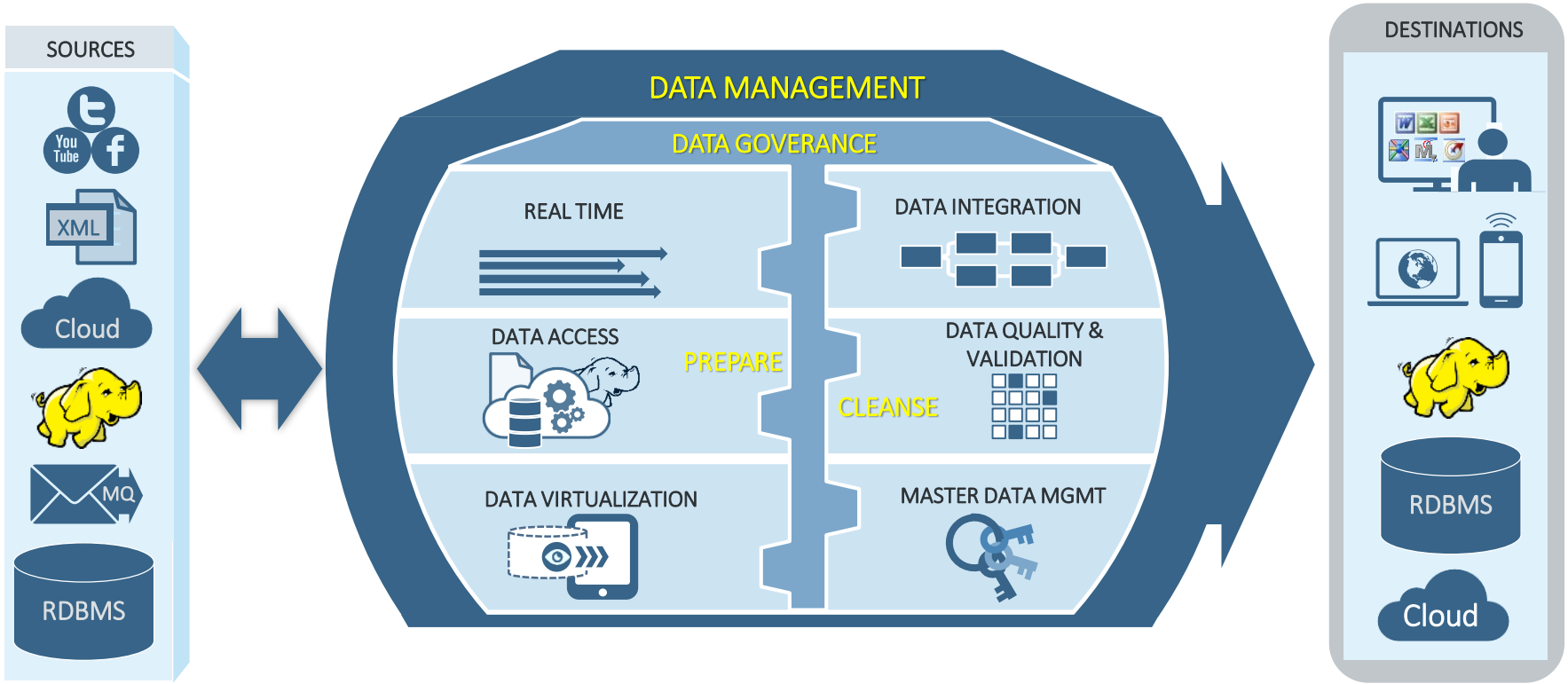


PRESENTED
VISUALLY



Self-Service Visual Analytics

Data Engineering: Data Management Context



Data Handling (aka Data Wrangling or Munging)

- Cleansing
- Filtering
- Joining
  (fusion / integration)
- Appending
- Transposing
- Transformations

- Deleting / hiding
- Interpolating
- Substituting
- Binning
- Clustering
- Reducing

SAS Data Management Taxonomy

Acquire and Discover	• Import and profile • Query or join
Transform	• Transform • Transpose
Cleanse	• Data quality transformations • Delete or hide
Integrate	• Query or join / perform merge operations • Sort and de-duplicate / cluster / collapse / bin
Deliver	• Push or make available for pull

What is ‘tidy’ data?

Data scientists spend ~80% of their time ‘cleaning’ data…

- Tidy = ‘shape’ of data matches assumptions of analytics models
- Data formats e.g. vectors, tables, cube, timeseries, graphs
- **Example**: raw network traffic is unstructured, irregular time series with complex events (multiple variables with unclear dependencies) and contextual entities (e.g. what is a ‘user’? is an IP atomic and persistent?)

ACTION	PURPOSE
Deduplication	Remove deduplicates
Extrapolation	Derive new variable (e.g. ratio)
Cast	Specify type (double, string, binary, etc.)
Binning	Reduce dimensionality via roll-up category
Imputation	Fill-in missing data
Join	Combine datasets
Aggregation	Group by (sum, count, max, min, avg)
Projection	Aggregate and reduce variables
Normalize	Reduce redundancy, linking *key entities*

Example: R Tidyverse

www.tidyverse.org

ggplot2

ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details. Learn more ...

dplyr

dplyr provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges. Learn more ...

tidyr

tidyr provides a set of functions that help you get to tidy data. Tidy data is data with a consistent form: in brief, every variable goes in a column, and every column is a variable. Learn more ...

readr

readr provides a fast and friendly way to read rectangular data (like csv, tsv, and fwf). It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes. Learn more ...

purrr

purrr enhances R's functional programming (FP) toolkit by providing a complete and consistent set of tools for working with functions and vectors. Once you master the basic concepts, purrr allows you to replace many for loops with code that is easier to write and more expressive. Learn more ...

tibble

tibble is a modern re-imagining of the data frame, keeping what time has proven to be effective, and throwing out what it has not. Tibbles are data.frames that are lazy and surly: they do less and complain more forcing you to confront problems earlier, typically leading to cleaner, more expressive code. Learn more ...

stringr

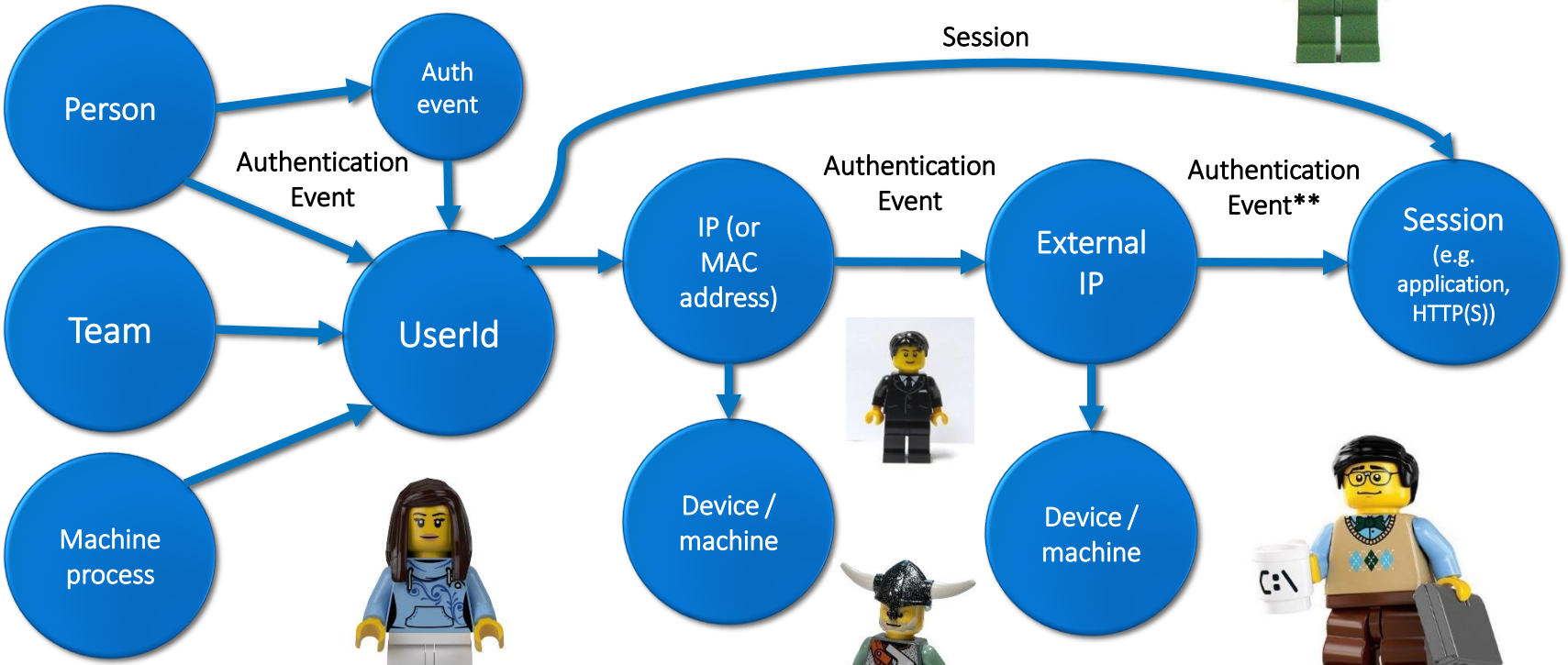
stringr provides a cohesive set of functions designed to make working with strings as easy as possible. It is built on top of stringi, which uses the ICU C library to provide fast, correct implementations of common string manipulations. Learn more ...

forcats

forcats provides a suite of useful tools that solve common problems with factors. R uses factors to handle categorical variables, variables that have a fixed and known set of possible values. Learn more ...

Feature Extraction

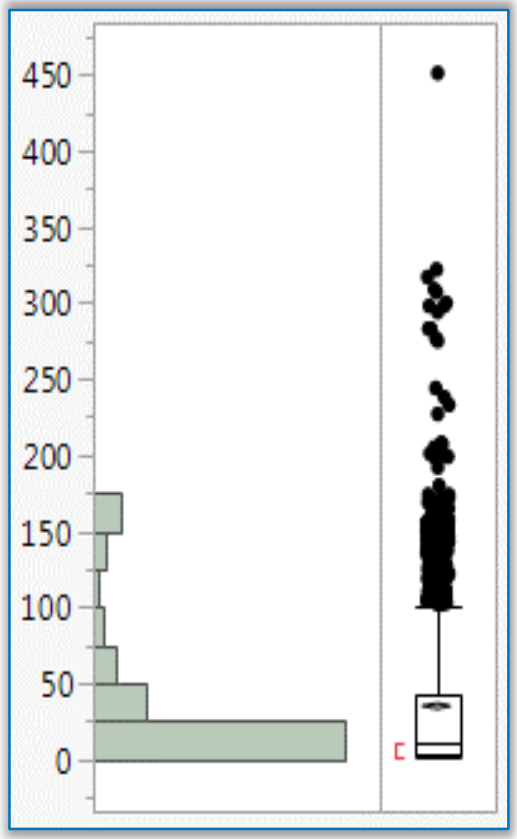
What is a User, Anyway?



behavioral profile

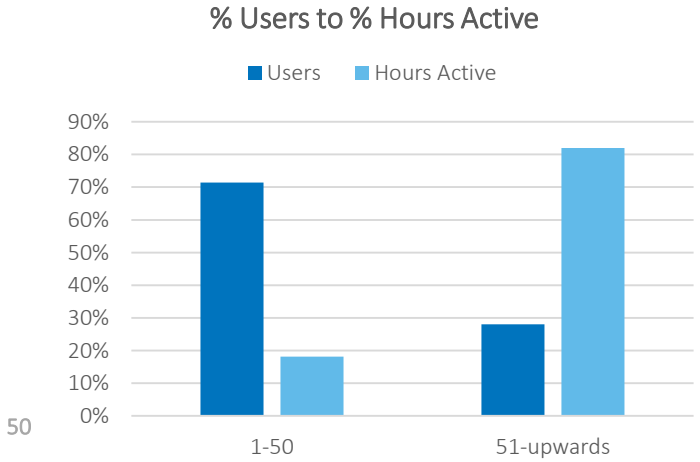
Feature Selection / Extraction

Understanding Network Behavioral Patterns



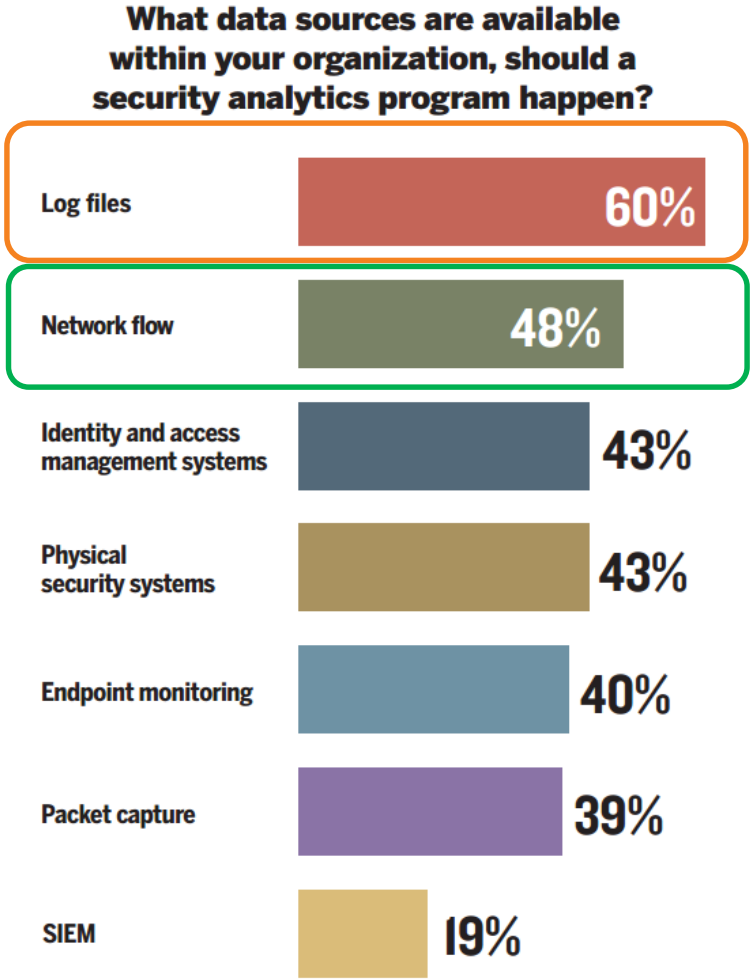
Pareto Principle

- **80/20%** pattern in network-usage
- *Outliers*: multiple devices 24 hours online
- High correlation: hrs online and breadth of activities
- Pattern observed across multiple networks



50

Focused Data Source: NetFlow



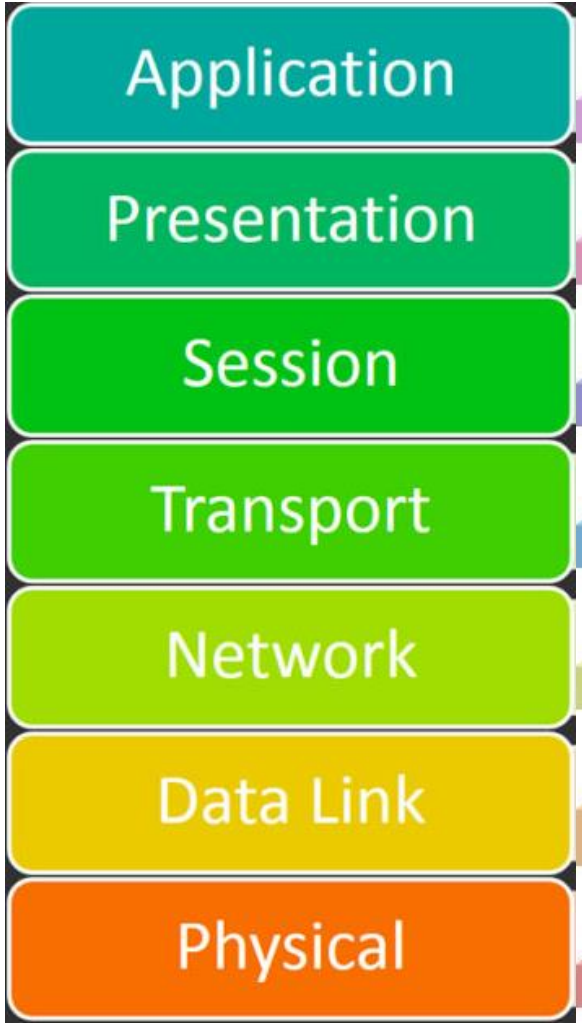
SOURCE
Security Brief Magazine. (2016). "Analyze This! Who's Implementing Security Analytics Now?" Available at https://www.sas.com/en_th/whitepapers/analyze-this-108217.html

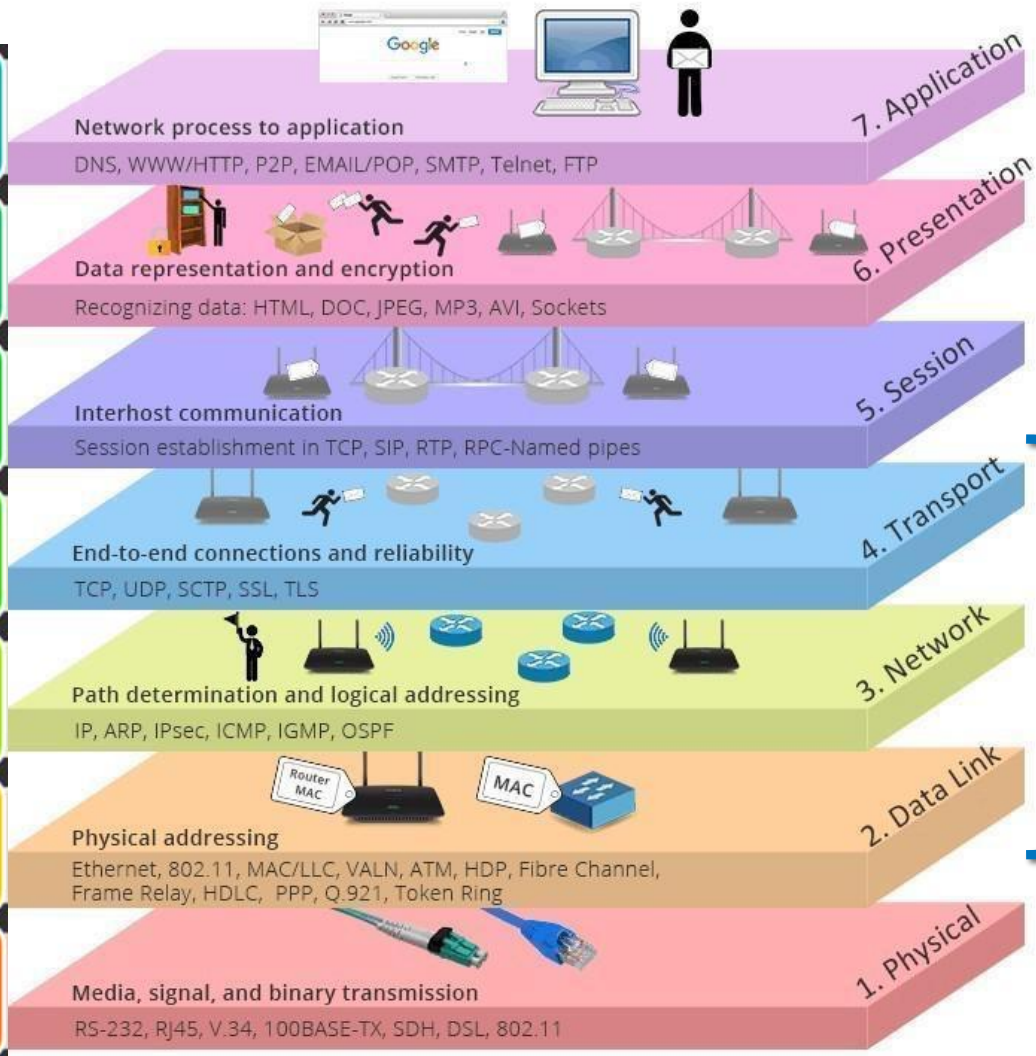
NetFlow Data

Date flow start Duration Proto Src IP Addr:Port Dst IP Addr:Port Flags Tos Packets Bytes Flows
2005-08-30 06:53:53.370 63.545 TCP 113.138.32.152:25 -> 222.33.70.124:3575 .AP.SF 0 62 3512 1
2005-08-30 06:53:53.370 63.545 TCP 222.33.70.124:3575 -> 113.138.32.152:25 .AP.SF 0 58 3300 1

Top 10 flows ordered by bytes:
Date flow start Duration Proto Src IP Addr:Port Dst IP Addr:Port Flags Tos Packets Bytes pps bps Bpp Flows
2005-08-30 06:50:11.218 700.352 TCP 126.52.54.27:47303 -> 42.90.25.218:435 ..... 0 1.4 M 2.0 G 2023 5.6 M 1498 1
2005-08-30 06:47:06.504 904.128 TCP 198.100.18.123:54945 -> 126.52.57.13:119 ..... 0 567732 795.1 M 627 2.5 M 1468 1
2005-08-30 06:47:06.310 904.384 TCP 126.52.57.13:45633 -> 91.127.227.206:119 ..... 0 321148 456.5 M 355 4.0 M 1490 1
2005-08-30 06:47:14.315 904.448 TCP 126.52.57.13:45598 -> 91.127.227.206:119 ..... 0 320710 455.9 M 354 4.0 M 1490 1
2005-08-30 06:47:14.316 904.448 TCP 126.52.57.13:45629 -> 91.127.227.206:119 ..... 0 317764 451.5 M 351 4.0 M 1489 1
2005-08-30 06:47:14.315 904.448 TCP 126.52.57.13:45634 -> 91.127.227.206:119 ..... 0 317611 451.2 M 351 4.0 M 1489 1
2005-08-30 06:47:06.313 904.384 TCP 126.52.57.13:45675 -> 91.127.227.206:119 ..... 0 317319 451.0 M 350 4.0 M 1490 1
2005-08-30 06:47:06.313 904.384 TCP 126.52.57.13:45619 -> 91.127.227.206:119 ..... 0 314199 446.5 M 347 3.9 M 1490 1
2005-08-30 06:47:06.321 790.976 TCP 126.52.54.35:59898 -> 132.94.115.59:2466 ..... 0 254717 362.4 M 322 3.7 M 1491 1
2005-08-30 06:47:14.316 904.384 TCP 126.52.54.35:59773 -> 55.107.224.187:11709 ..... 0 272710 348.5 M 301 3.1 M 1340 1

NFDUMP https://github.com/phaag/nfdump



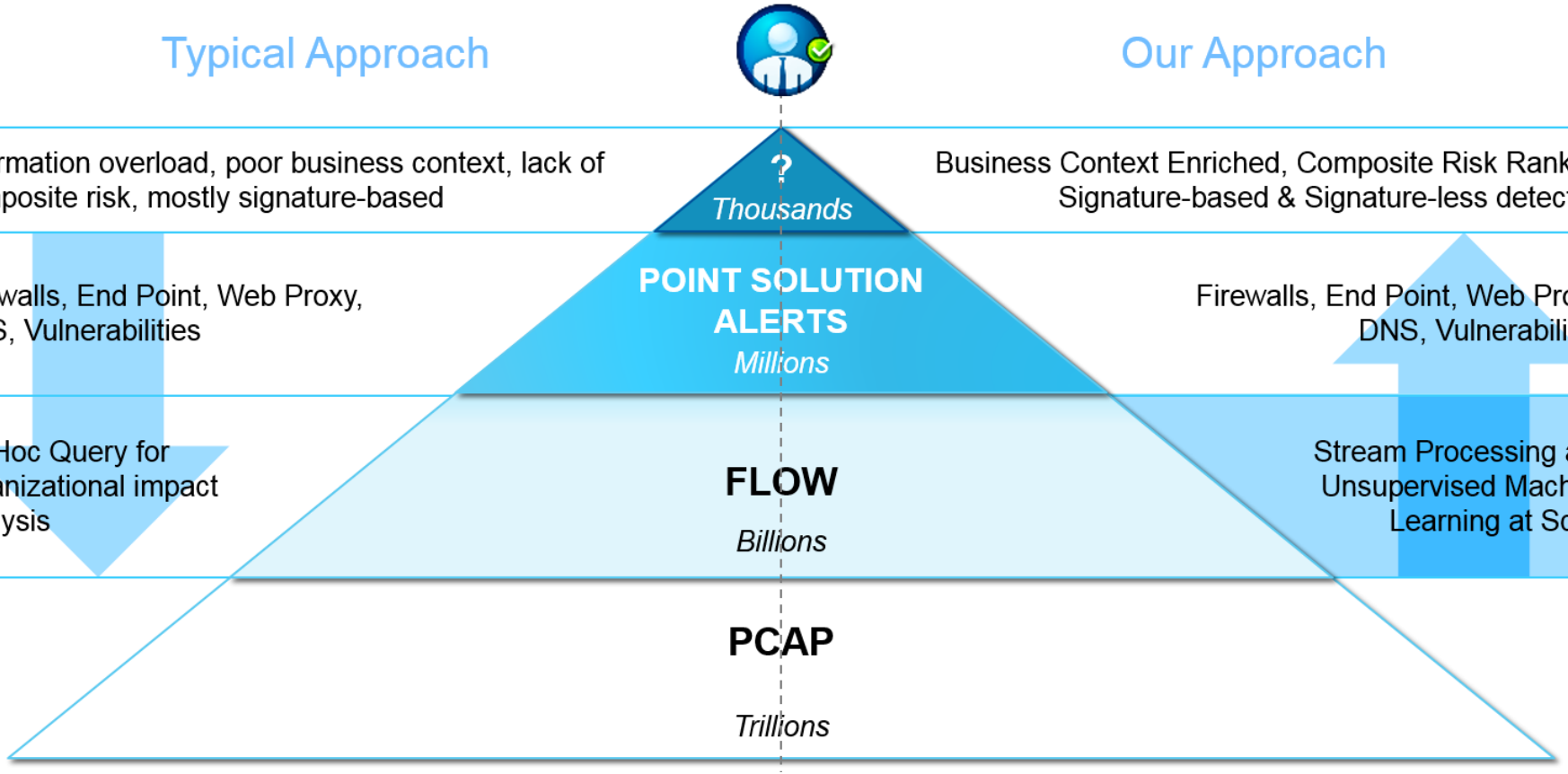


NetFlow

Key Flow Data Supplied in Records

1. Number of flows
2. IP protocol
3. Type of service
4. TCP/UDP source & dest port
5. IPv4/IPv6 source & dest
6. BGP Autonomous System
7. MPLS TOP
8. MAC source & dest
9. Min & max time-to-live

Data Volumes and Security Challenge

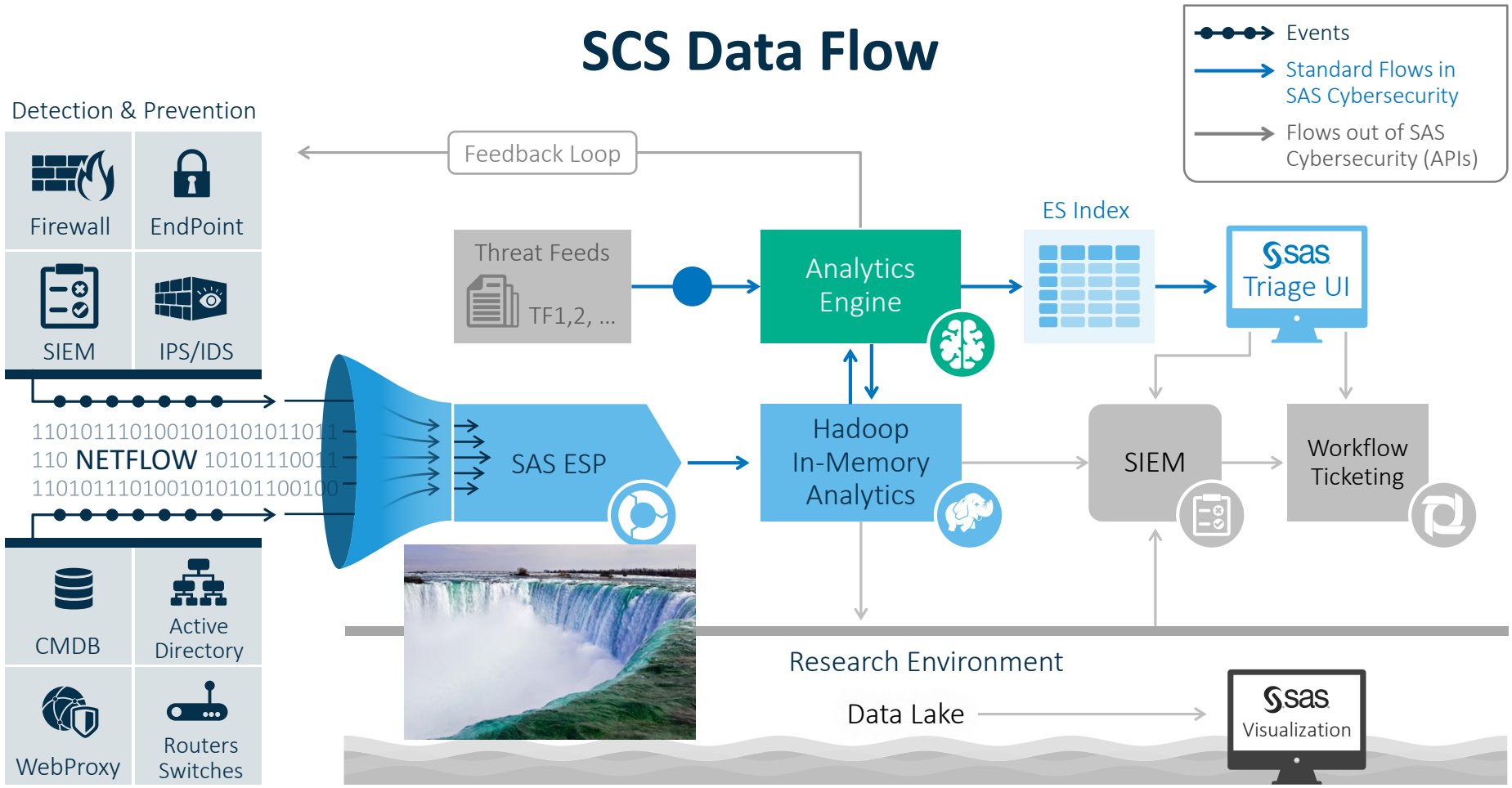


Network Flow (NetFlow) Analysis Tools

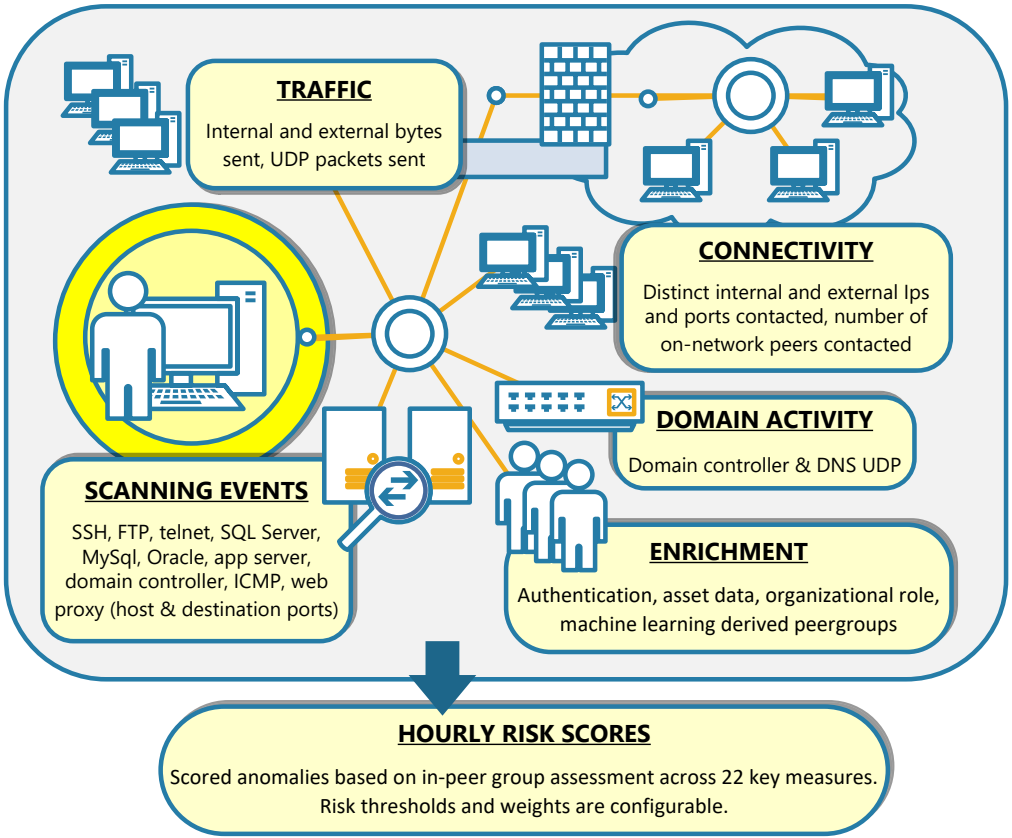
- Silk (CERT)
- NFDUMP
- Scrutinizer
- Cisco NTA
- Bro (complement)
- SAS Cybersecurity

Date flow start	Duration	Proto	Src IP Addr:Port		Dst IP Addr:Port	Flags	Tos	Packets	Bytes	pps	bps	Bpp	Flows
2005-08-30 06:50:11.218	700.352	TCP	126.52.54.27:47303	->	42.90.25.218:435	.....	0	1.4 M	2.0 G	2023	5.6 M	1498	1
2005-08-30 06:47:06.504	904.128	TCP	198.100.18.123:54945	->	126.52.57.13:119	.....	0	567732	795.1 M	627	2.5 M	1468	1
2005-08-30 06:47:06.310	904.384	TCP	126.52.57.13:45633	->	91.127.227.206:119	.....	0	321148	456.5 M	355	4.0 M	1490	1
2005-08-30 06:47:14.315	904.448	TCP	126.52.57.13:45598	->	91.127.227.206:119	.....	0	320710	455.9 M	354	4.0 M	1490	1
2005-08-30 06:47:14.316	904.448	TCP	126.52.57.13:45629	->	91.127.227.206:119	.....	0	317764	451.5 M	351	4.0 M	1489	1
2005-08-30 06:47:14.315	904.448	TCP	126.52.57.13:45634	->	91.127.227.206:119	.....	0	317611	451.2 M	351	4.0 M	1489	1
2005-08-30 06:47:06.313	904.384	TCP	126.52.57.13:45675	->	91.127.227.206:119	.....	0	317319	451.0 M	350	4.0 M	1490	1
2005-08-30 06:47:06.313	904.384	TCP	126.52.57.13:45619	->	91.127.227.206:119	.....	0	314199	446.5 M	347	3.9 M	1490	1
2005-08-30 06:47:06.321	790.976	TCP	126.52.54.35:59898	->	132.94.115.59:2466	.....	0	254717	362.4 M	322	3.7 M	1491	1
2005-08-30 06:47:06.504	904.128	TCP	198.100.18.123:54945	->	126.52.57.13:119	.....	0	567732	795.1 M	627	2.5 M	1468	1
2005-08-30 06:47:06.310	904.384	TCP	126.52.57.13:45633	->	91.127.227.206:119	.....	0	321148	456.5 M	355	4.0 M	1490	1
2005-08-30 06:47:14.315	904.448	TCP	126.52.57.13:45598	->	91.127.227.206:119	.....	0	320710	455.9 M	354	4.0 M	1490	1
2005-08-30 06:47:14.316	904.448	TCP	126.52.57.13:45629	->	91.127.227.206:119	.....	0	317764	451.5 M	351	4.0 M	1489	1
2005-08-30 06:47:14.315	904.448	TCP	126.52.57.13:45634	->	91.127.227.206:119	.....	0	317611	451.2 M	351	4.0 M	1489	1
2005-08-30 06:47:06.313	904.384	TCP	126.52.57.13:45675	->	91.127.227.206:119	.....	0	317319	451.0 M	350	4.0 M	1490	1
2005-08-30 06:47:06.313	904.384	TCP	126.52.57.13:45619	->	91.127.227.206:119	.....	0	314199	446.5 M	347	3.9 M	1490	1
2005-08-30 06:47:06.321	790.976	TCP	126.52.54.35:59898	->	132.94.115.59:2466	.....	0	254717	362.4 M	322	3.7 M	1491	1

SCS Data Flow



SAS Cybersecurity - Summary Data



Exploring and Extracting Data

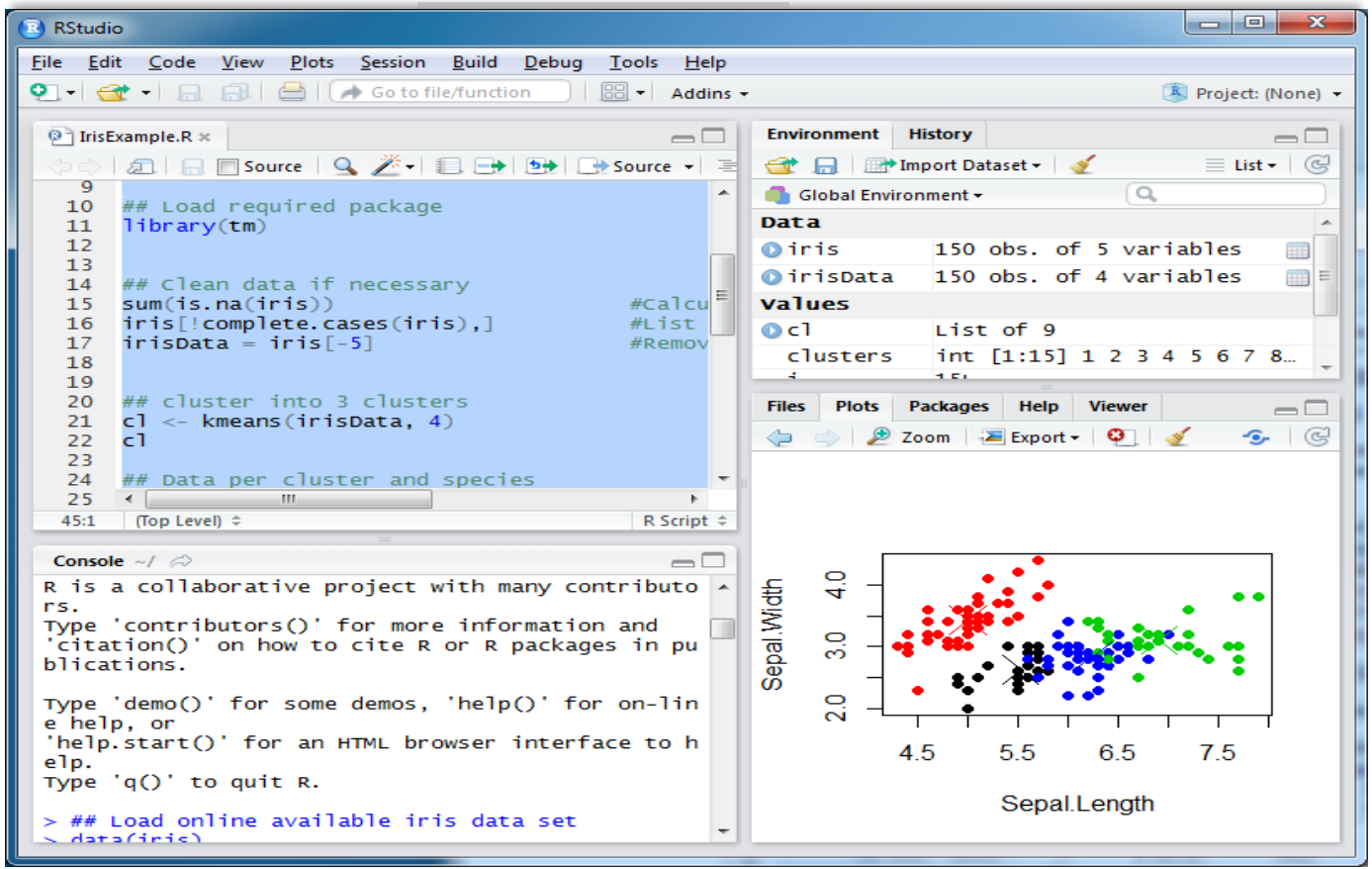
Tooling

Introducing tools for practical exercise

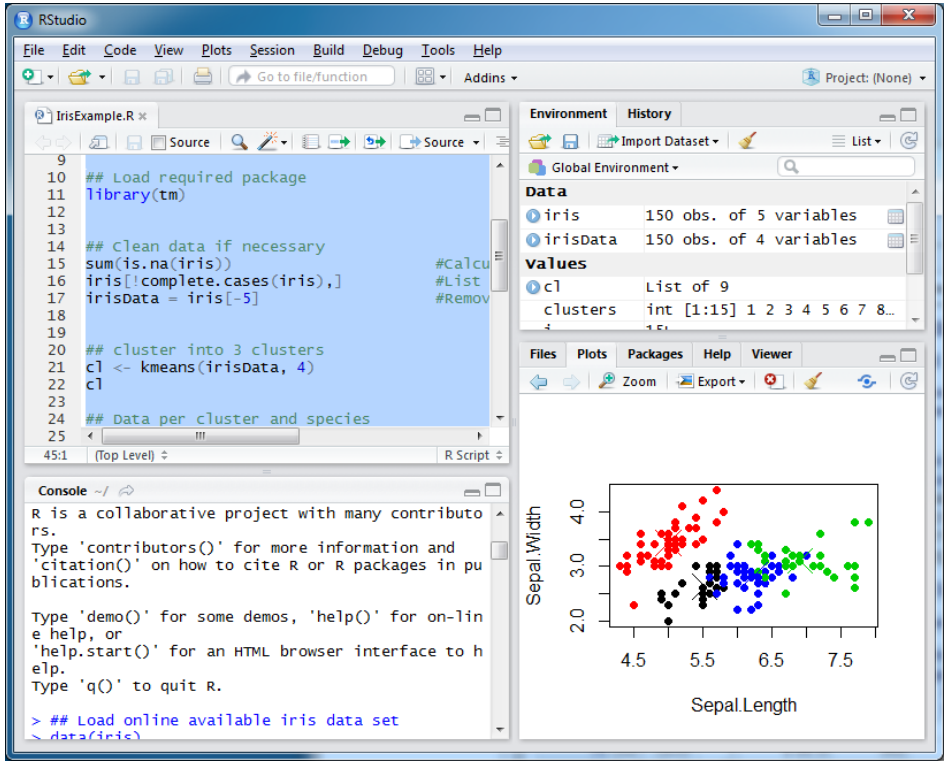
Data Scientist

Data Analysis and Exploration

R / R Studio



Example: Feature Engineering - PCA and Clustering with R



http://www.idvbook.com/teaching-aid/data-sets/the-iris-data-set/

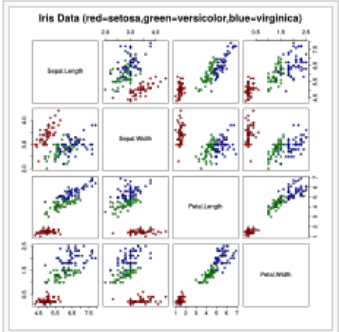
https://en.wikipedia.org/wiki/Iris_flower_data_set

Iris flower data set

From Wikipedia, the free encyclopedia

The ***Iris* flower data set** or **Fisher's *Iris* data set** is a multivariate data set introduced by Ronald Fisher in his 1936 paper *The use of multiple measurements in taxonomic problems* as an example of linear discriminant analysis.[1] It is sometimes called **Anderson's *Iris* data set** because Edgar Anderson collected the data to quantify the morphologic variation of *Iris* flowers of three related species.[2] Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".[3]

The data set consists of 50 samples from each of three species of *Iris* (*Iris setosa*, *Iris virginica* and *Iris versicolor*). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

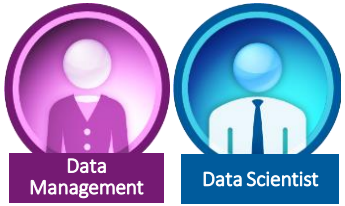


Scatterplot of the data set

Ronald Fisher

Tooling

Introducing tools for practical exercise

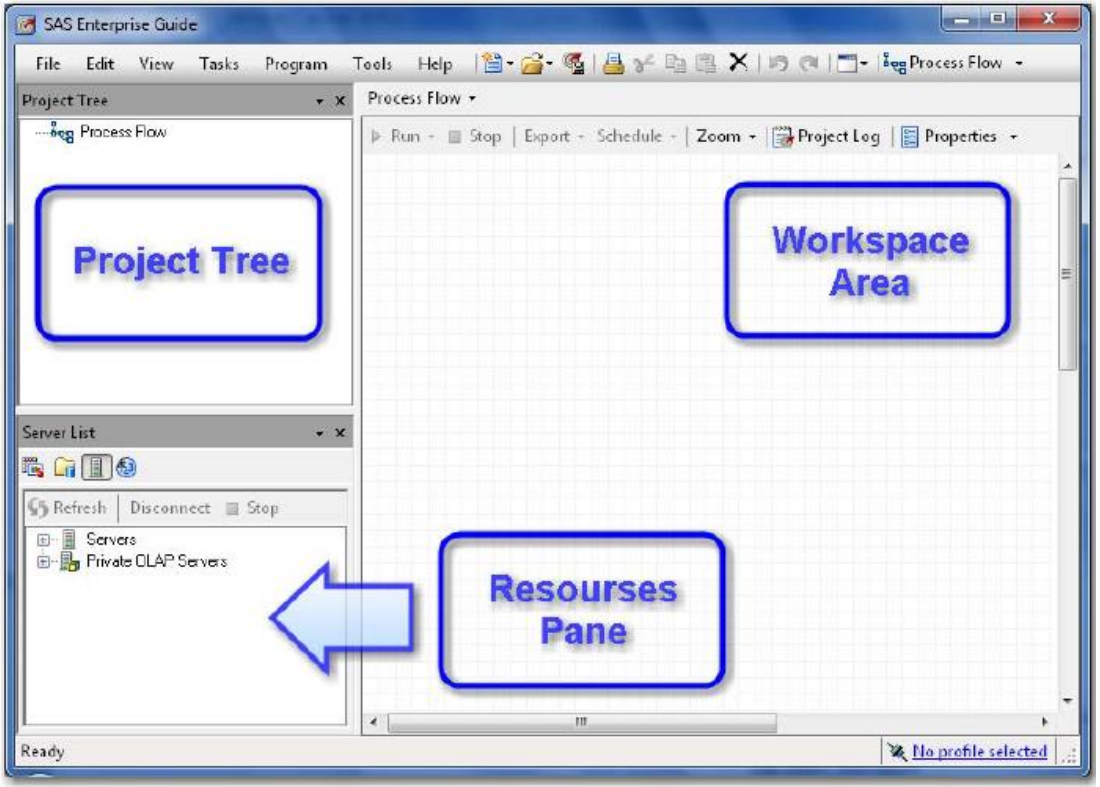


Data Analysis, Transformation, Analytics

SAS Enterprise Guide

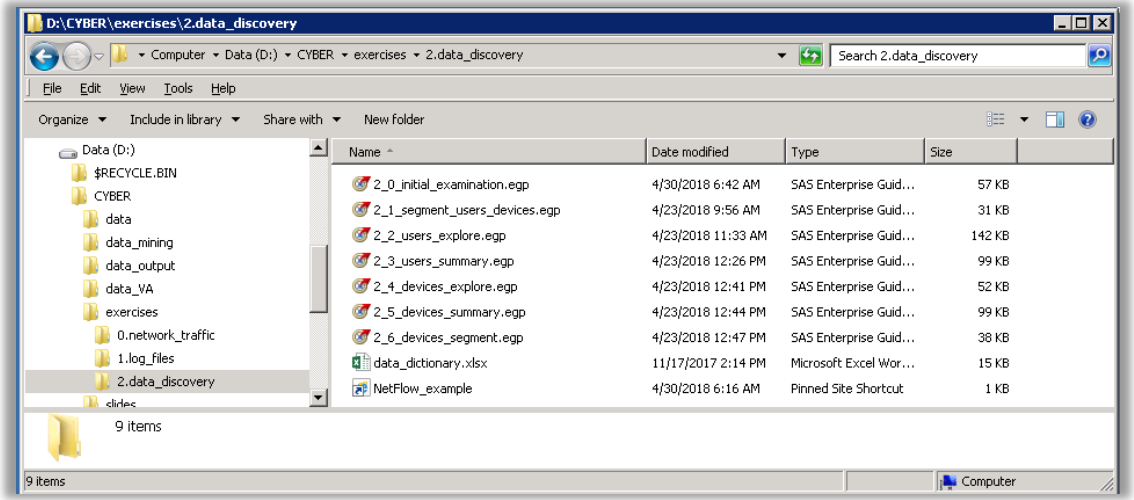
SAS Enterprise Guide – user-friendly interface to SAS Analytics:

- Preliminary data analysis
- Converting data into analytics-ready variables
- Creating workflows that structure and automate a complex set of procedures
- Performing statistical analysis, analytics, and machine learning
- Integrating SAS code



Hands-on NetFlow Data Exploration / Extraction

- Data
  - ~1 million hourly records over 24 days (March-April 2017)
  - Hourly summarized NetFlow measures by device IP address (some with bound UserId)
- Tool
  - SAS Enterprise Guide
- Goals
  - Profile dataset
  - Exploratory analysis
  - Segment records
  - Extract (meta-) measures
  - Roll-up a ‘meaningful’ summarized data set



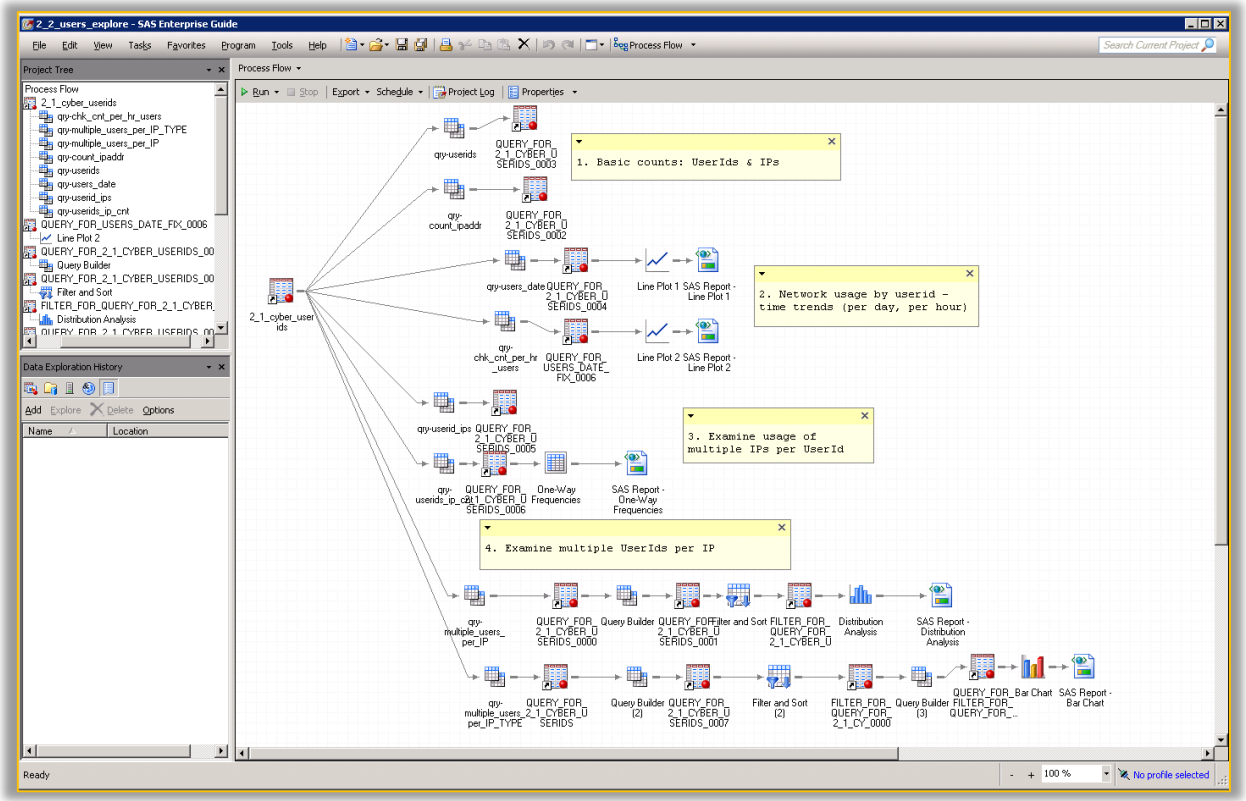
Exercise

Exploring and extracting cybersecurity data

Data Management

Data Quality as Foundation for Analytics

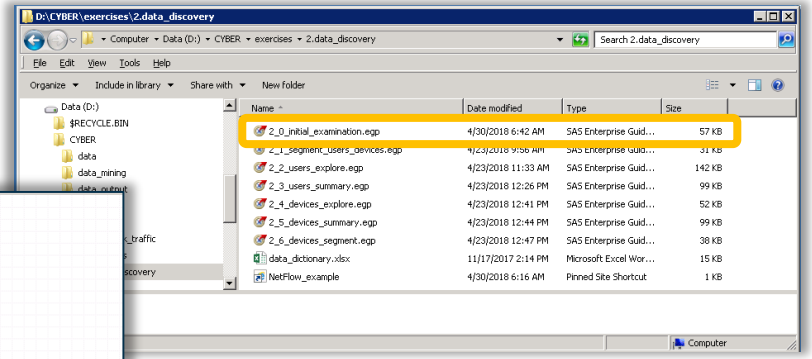
Data Handling for Cybersecurity Data

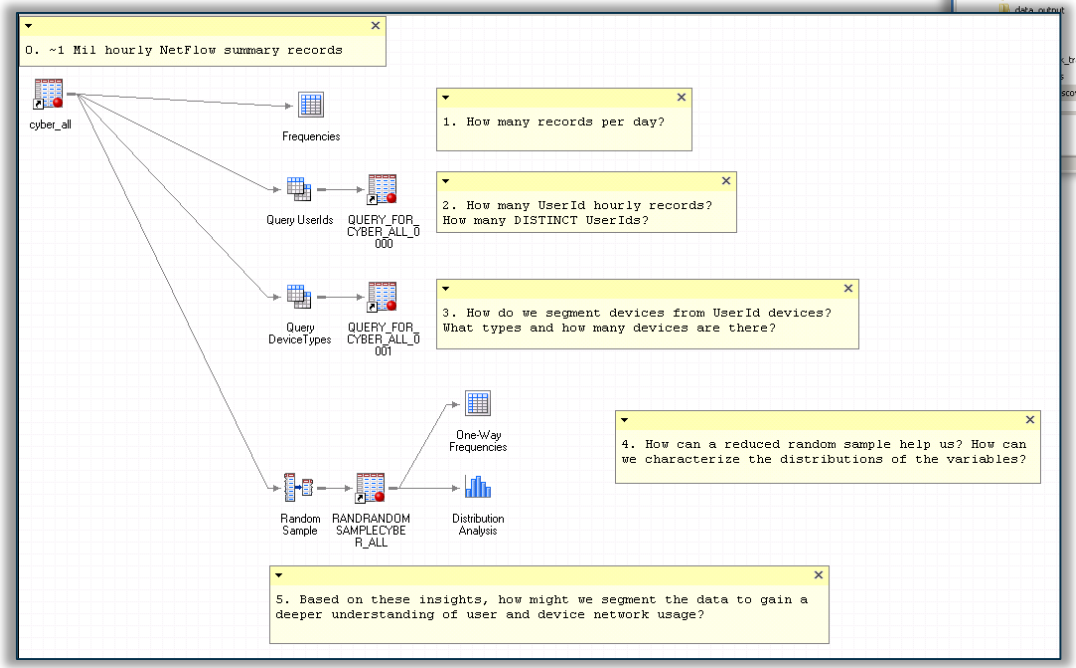


Hands-on NetFlow Data Exploration / Extraction

1. ‘2_0_initial_examinination.egp

   D:\@CYBER\2.DATA\data_explore

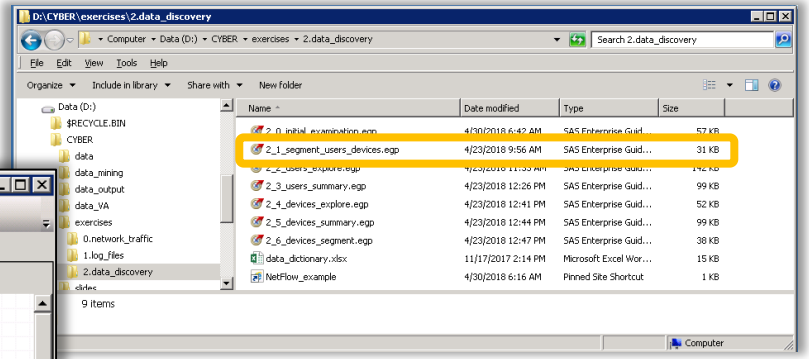


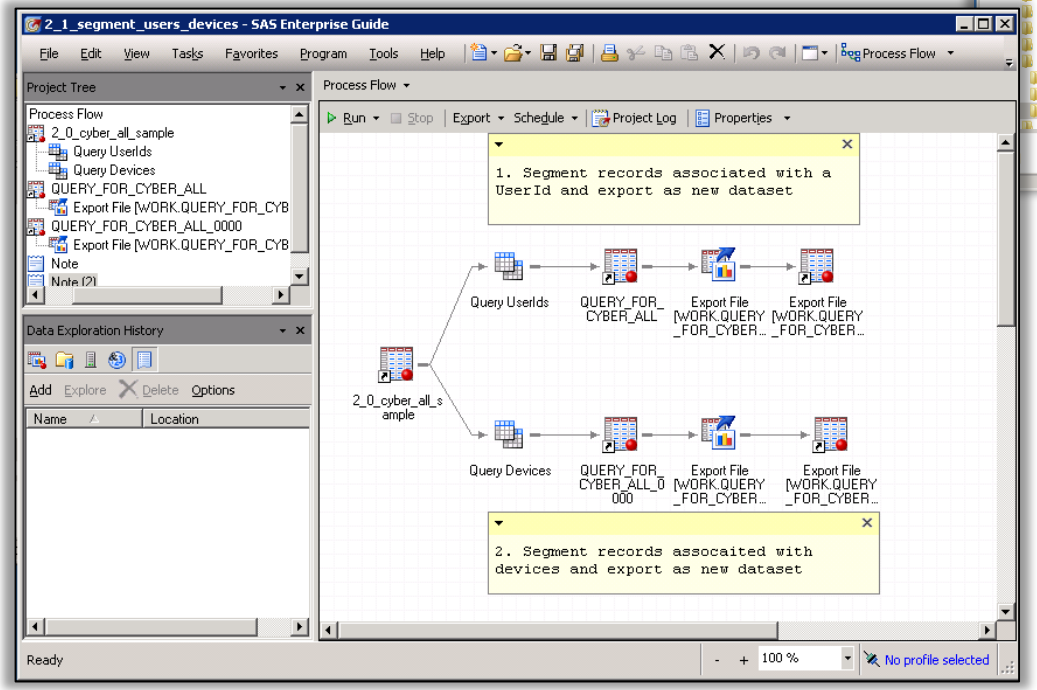


Hands-on NetFlow Data Exploration / Extraction

2. Open '**2_1_segment_user_devices.egp**'

   D:\@CYBER\2.DATA\data_explore

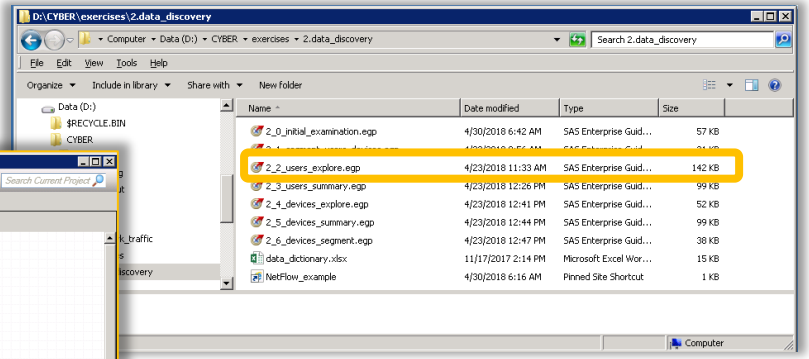


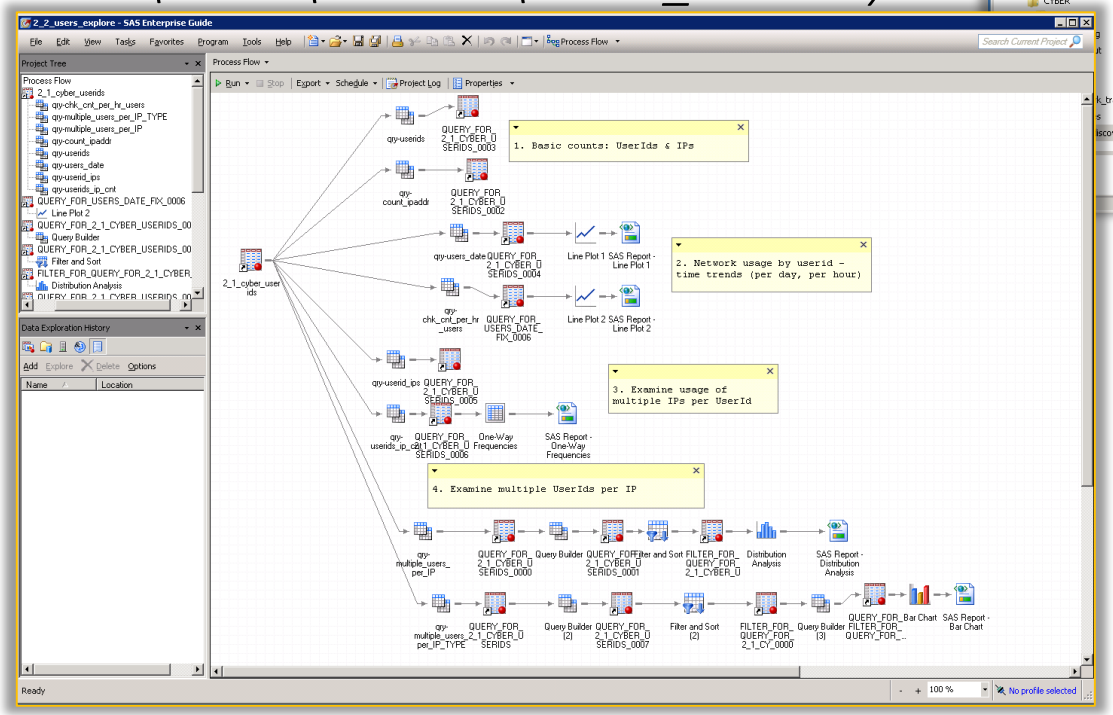


Hands-on NetFlow Data Exploration / Extraction

3. Open '**2_2_user_explore.egp**'

   D:\CYBER\exercises\2.data_discovery

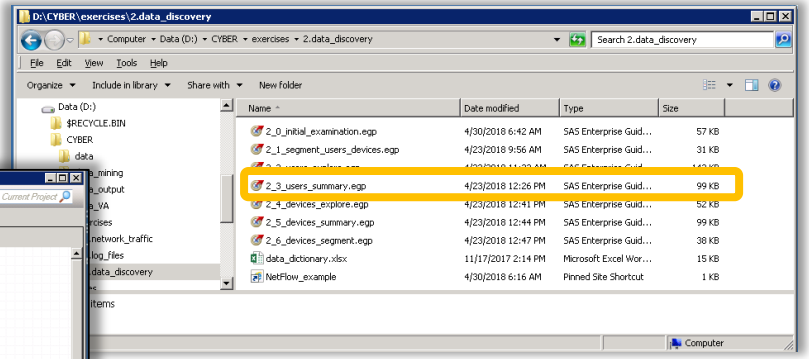


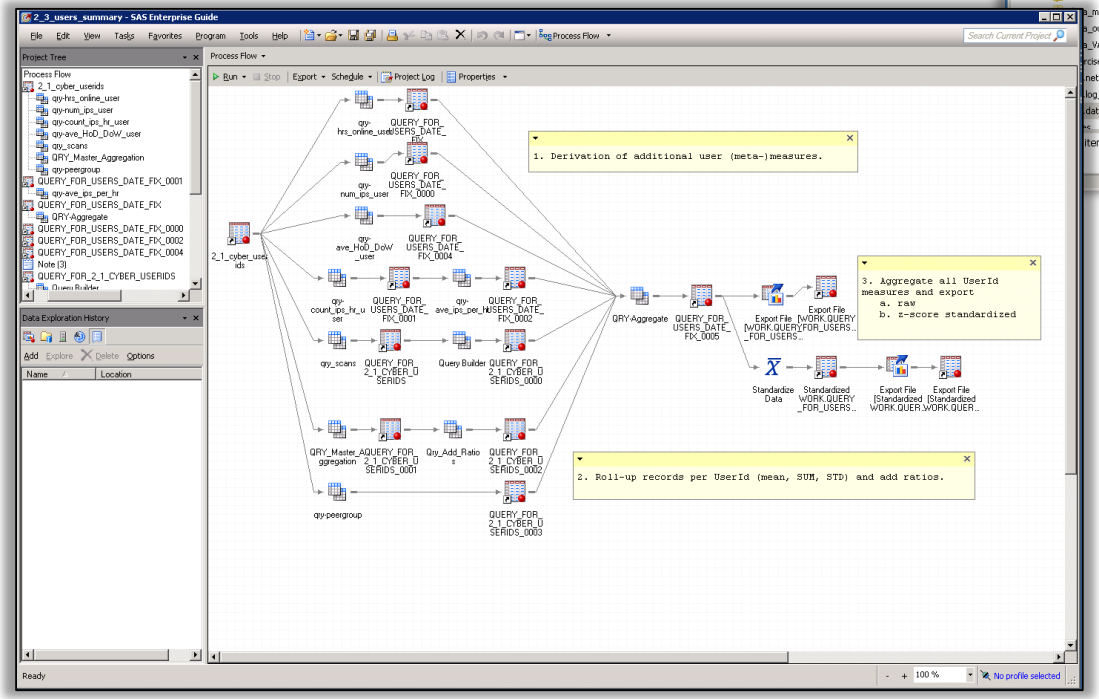


Hands-on NetFlow Data Exploration / Extraction

4. Open '**2_3_user_summary.egp**'

   D:\CYBER\exercises\2.data_discovery

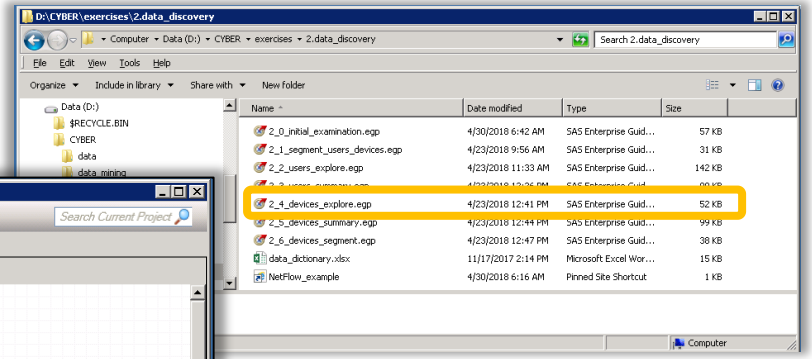


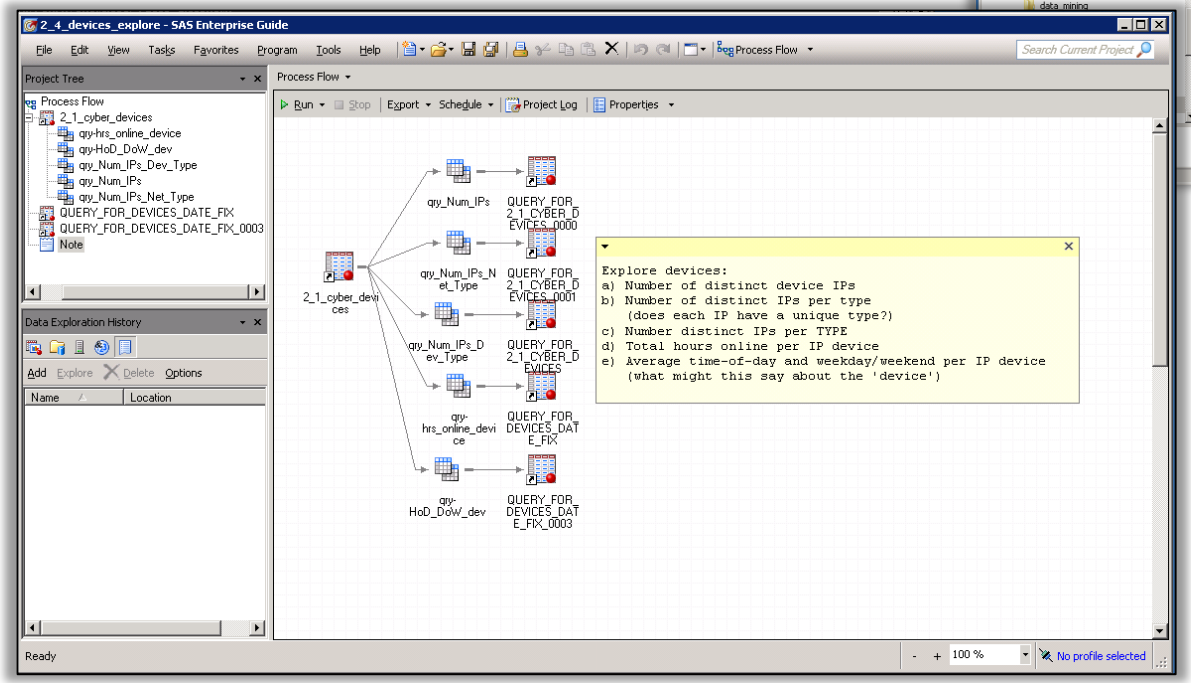


Hands-on NetFlow Data Exploration / Extraction

5. Open '**2_4_devices_explore.egp**'

   D:\CYBER\exercises\2.data_discovery

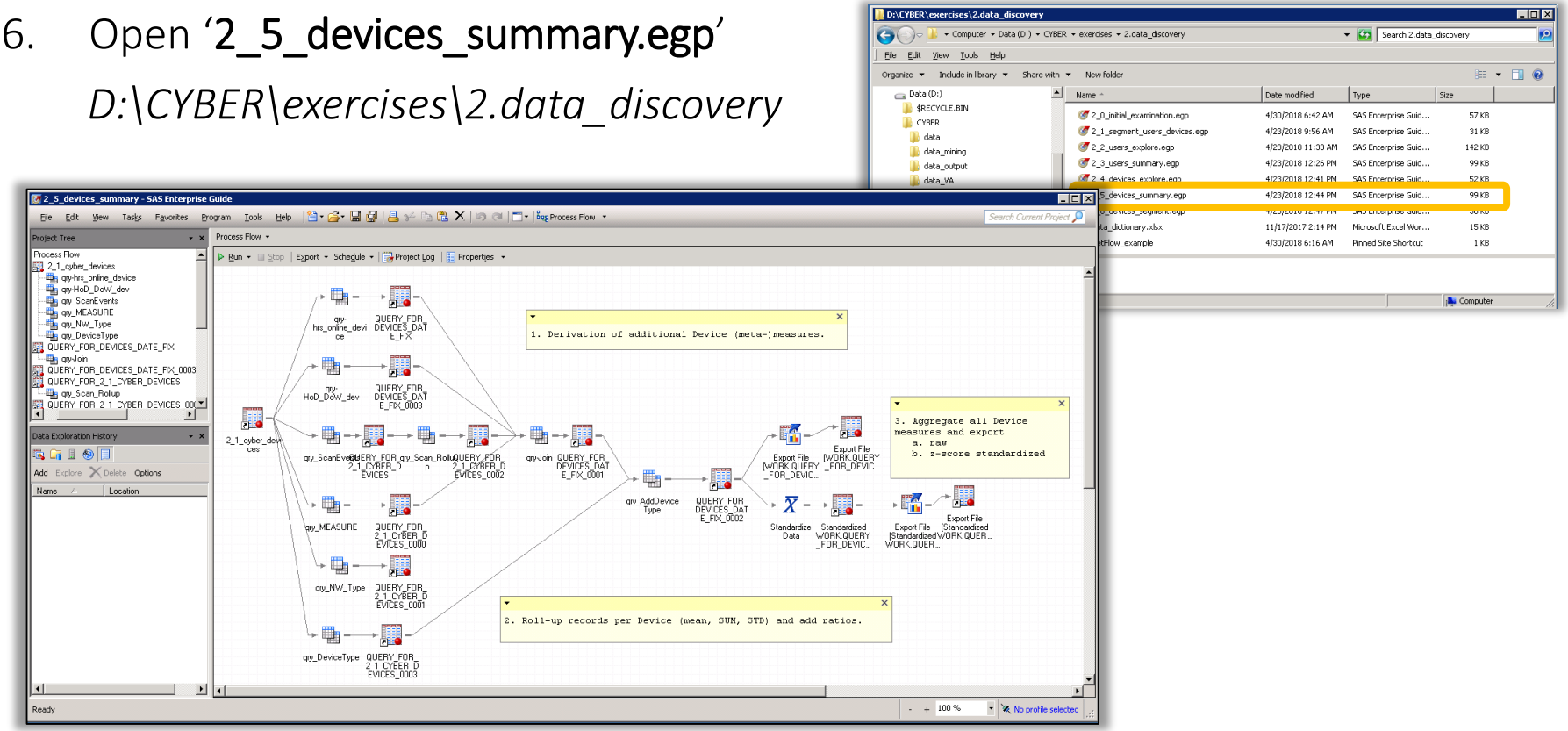




Hands-on NetFlow Data Exploration / Extraction

6. Open '**2_5_devices_summary.egp**'

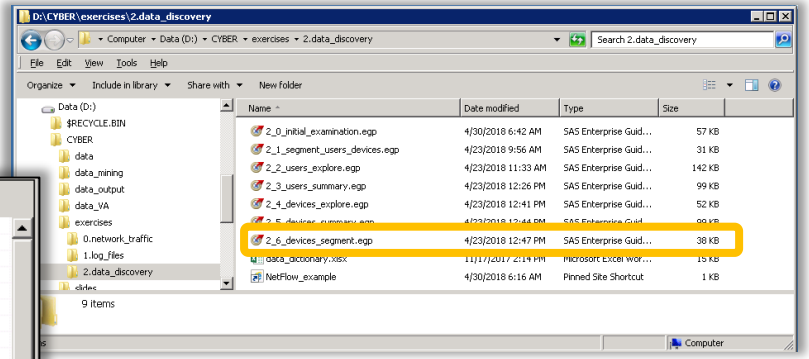
   D:\CYBER\exercises\2.data_discovery

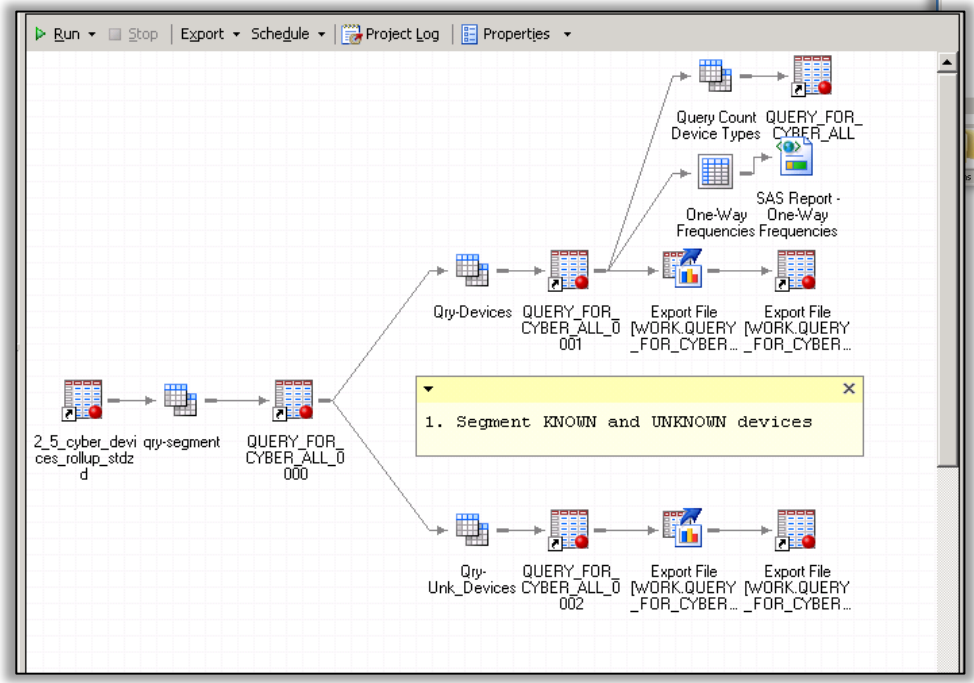


Hands-on NetFlow Data Exploration / Extraction

7. Open '**2_6_devices_segment.egp**'

   D:\CYBER\exercises\2.data_discovery





SAS Data Management

Overview of SAS data management approaches

Traditional Data Management

Traditional Data Management - Highly Managed

Users: ETL Developers, IT Users, Data Stewards, ….

Example blog post – structured data management challenges
https://sctr7.com/2014/06/27/the-cutting-edge-network-analytics-for-financial-fraud-detection-and-mitigation/

Traditional Data Management

Data Integration & Quality

Extract, Transform, Load

Managed

Monitored

Data Governance

Glossary

Auditing

Lineage

Traditional Data Management - Highly Managed

Users: ETL Developers, IT Users, Data Stewards, ….

Ad-hoc Data Preparation

Data Preparation

Self-service

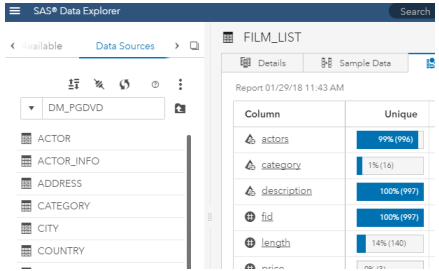
Wrangling

Blending

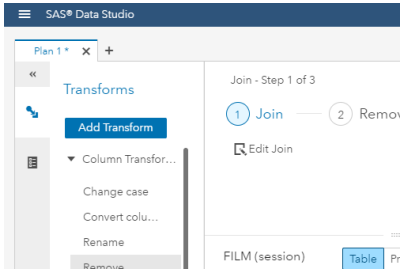
Ad-hoc Data Prep - Very Flexible

Users: Data Scientists & Business Users/Analysts

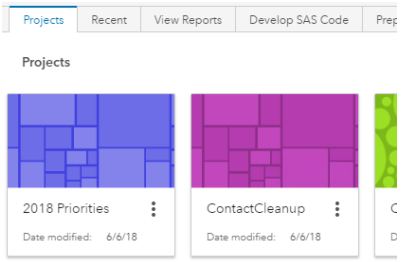
SAS Data Preparation Suite



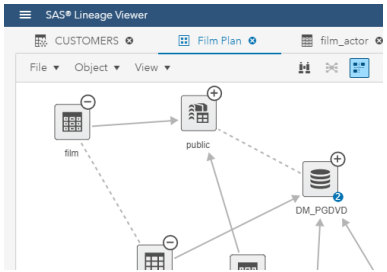
Manage Data



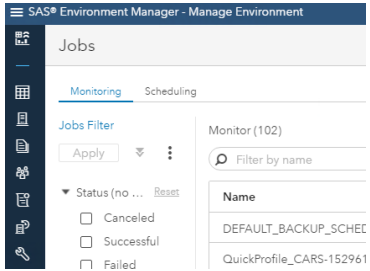
Prepare Data



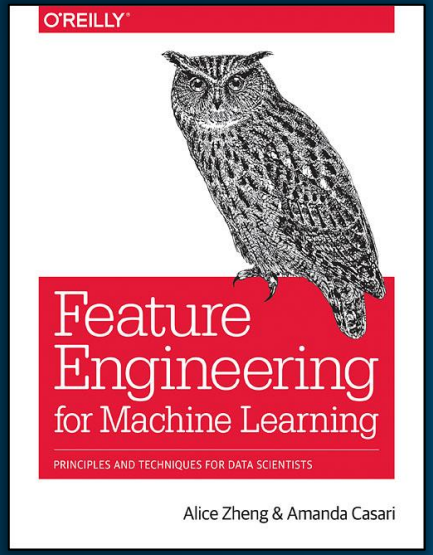
Organize Data Projects



Explore Lineage

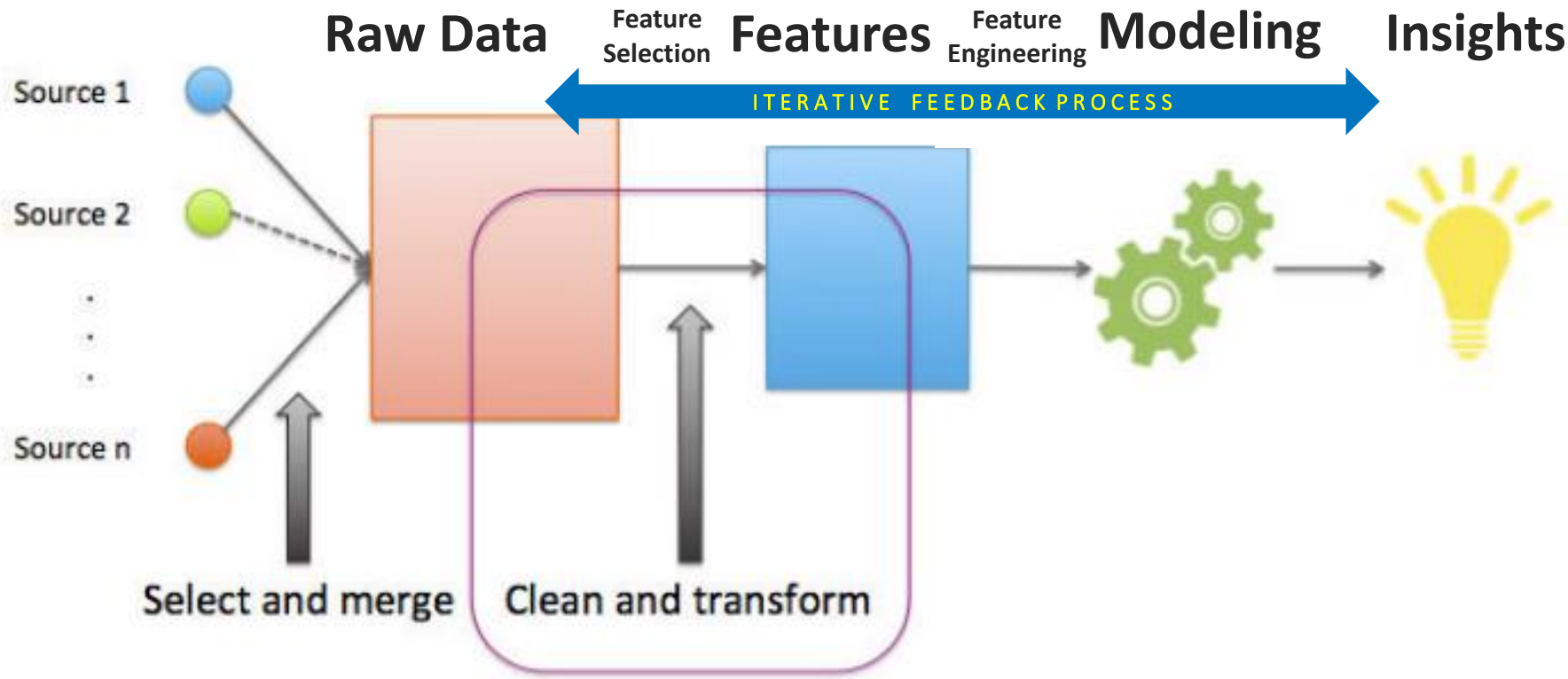


Monitor Jobs

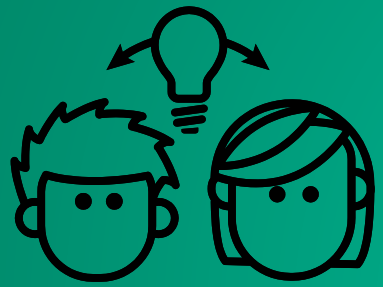


Feature Engineering

Feature Selection, Extraction, Engineering



SOURCE: Alice Zheng, Amanda Casari. 2016. Feature Engineering for Machine Learning Models. O'Reilly Media.

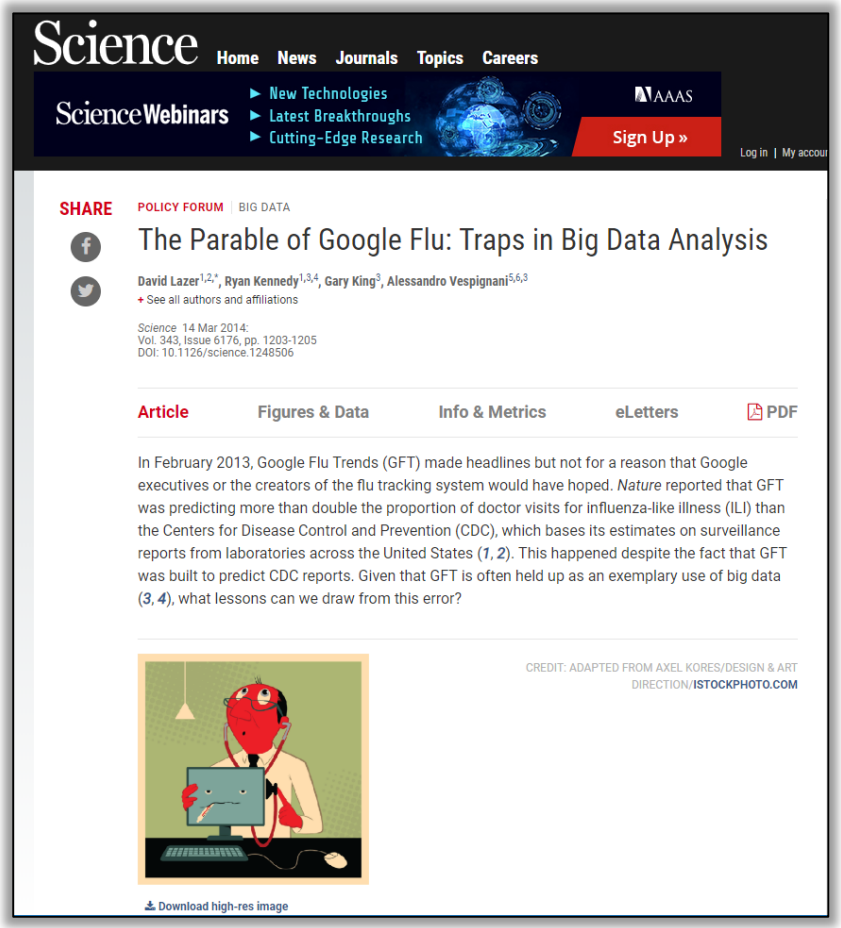


Key Concepts

Why invest the effort in feature extraction and selection for cybersecurity data?

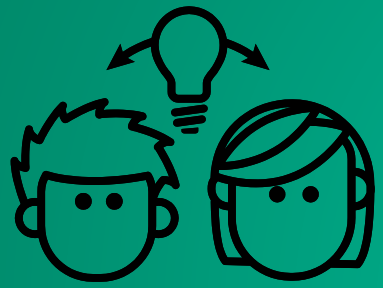
Why Feature Extraction / Selection?

- Data overload -> reduction of dataset
- Poor data quality -> refinement
- Unlinked data -> associations
- Lack of context -> link assumptions
- Model efficacy -> target to phenomenon



https://science.sciencemag.org/content/343/6176/1203.full

85

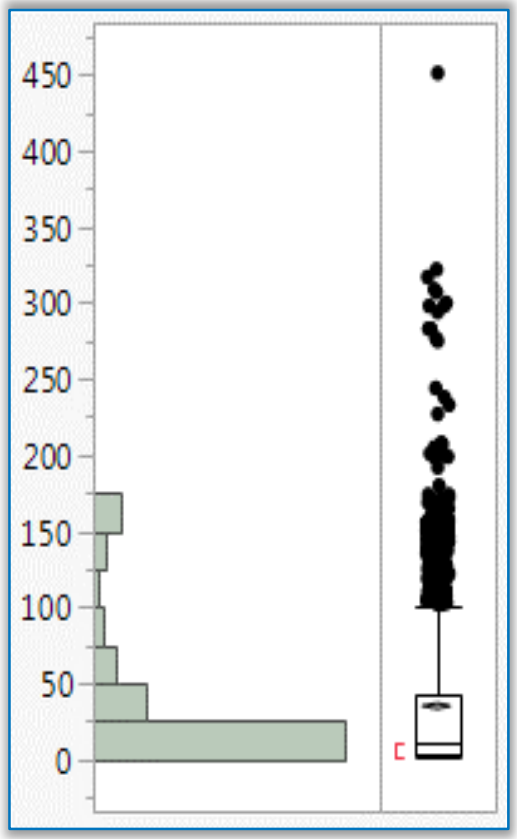


Learnings from the Field

Guidance based on practical learnings

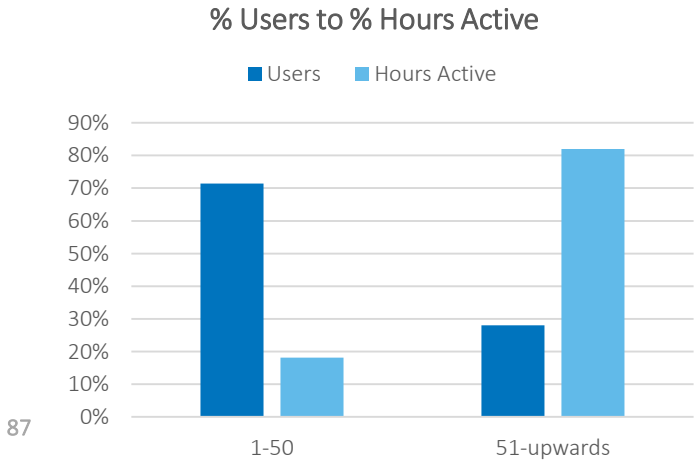
Feature Selection / Extraction

Understanding Network Behavioral Patterns



Pareto Principle

- **80/20%** pattern in network-usage
- *Outliers*: multiple devices 24 hours online
- High correlation: hrs online and breadth of activities
- Pattern observed across multiple networks



Feature Extraction

Ratios as Key Measures

Ratios **of key are possibly more indicative of threats than single point measures…**

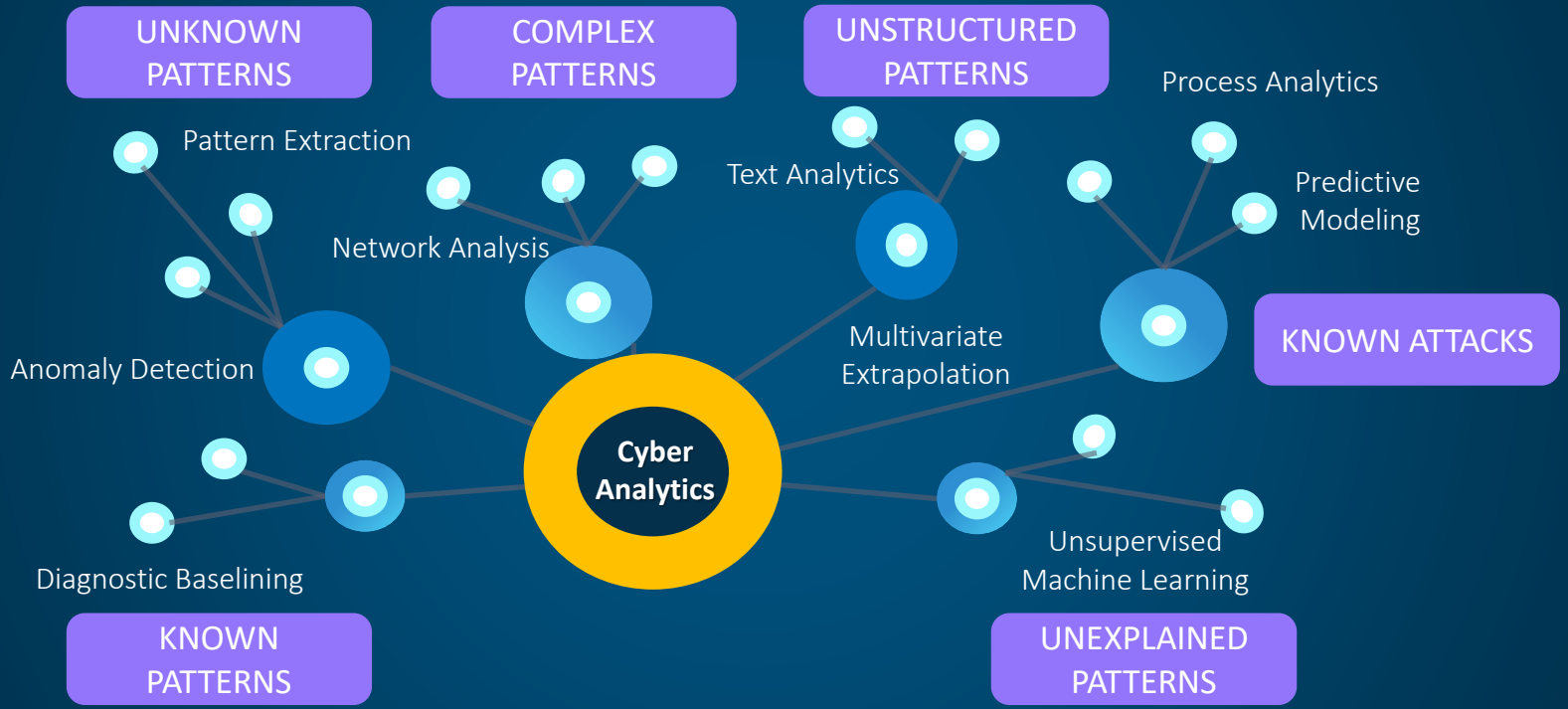
Examples

- *Ratio of total data flows per hour TO unique destination IPs*
  - Measures nearing high of 1:1 would be threat indicator of scanning activities.
- *Ratio of unique internal destination IPs TO unique external IPs*
  - Low might be threat indicator, perhaps bot net data exfiltration.
- *Ratio of unique destination ports TO unique source ports*
  - Low would generally be considered a threat, as might indicate a compromised system engaging in vulnerability surveillance across a range of outgoing ports to compromise a new system at a particular port.

Advanced Insights

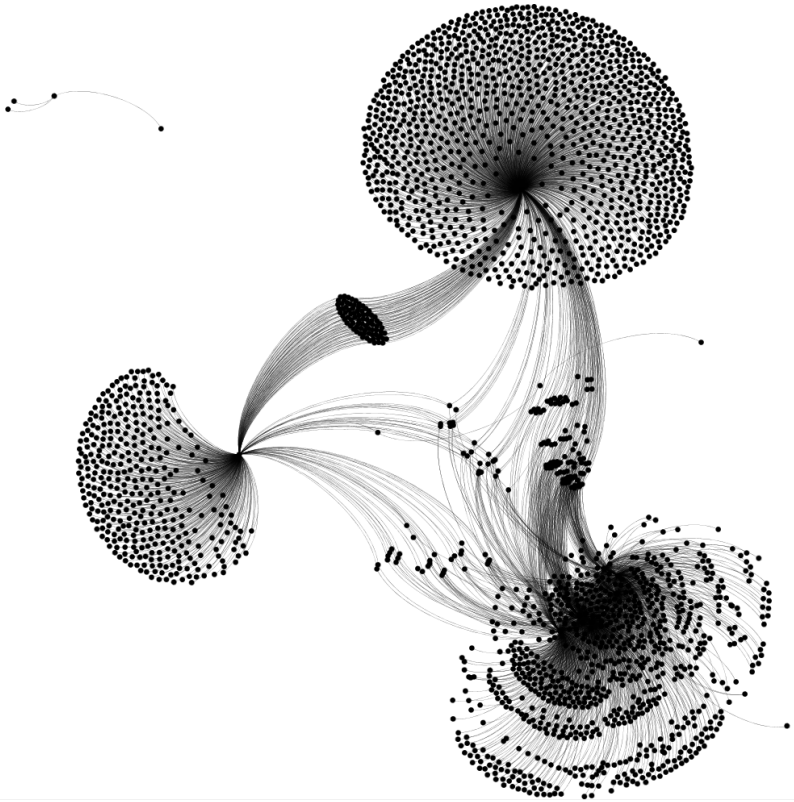
Insights into feature engineering

CSDS: Diverse Analytics Toolkit



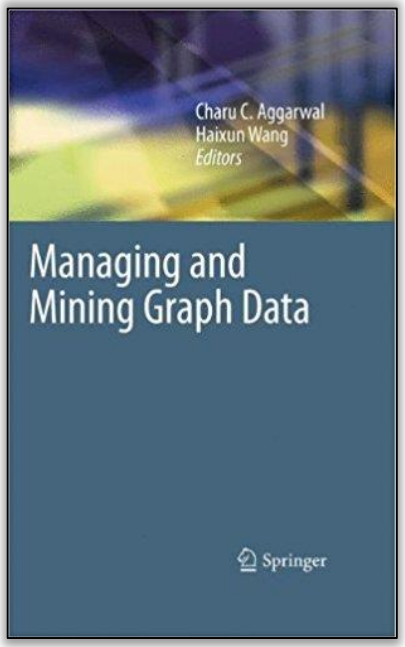
Feature Extraction: Advanced Measures

Network 'Graph' Measures



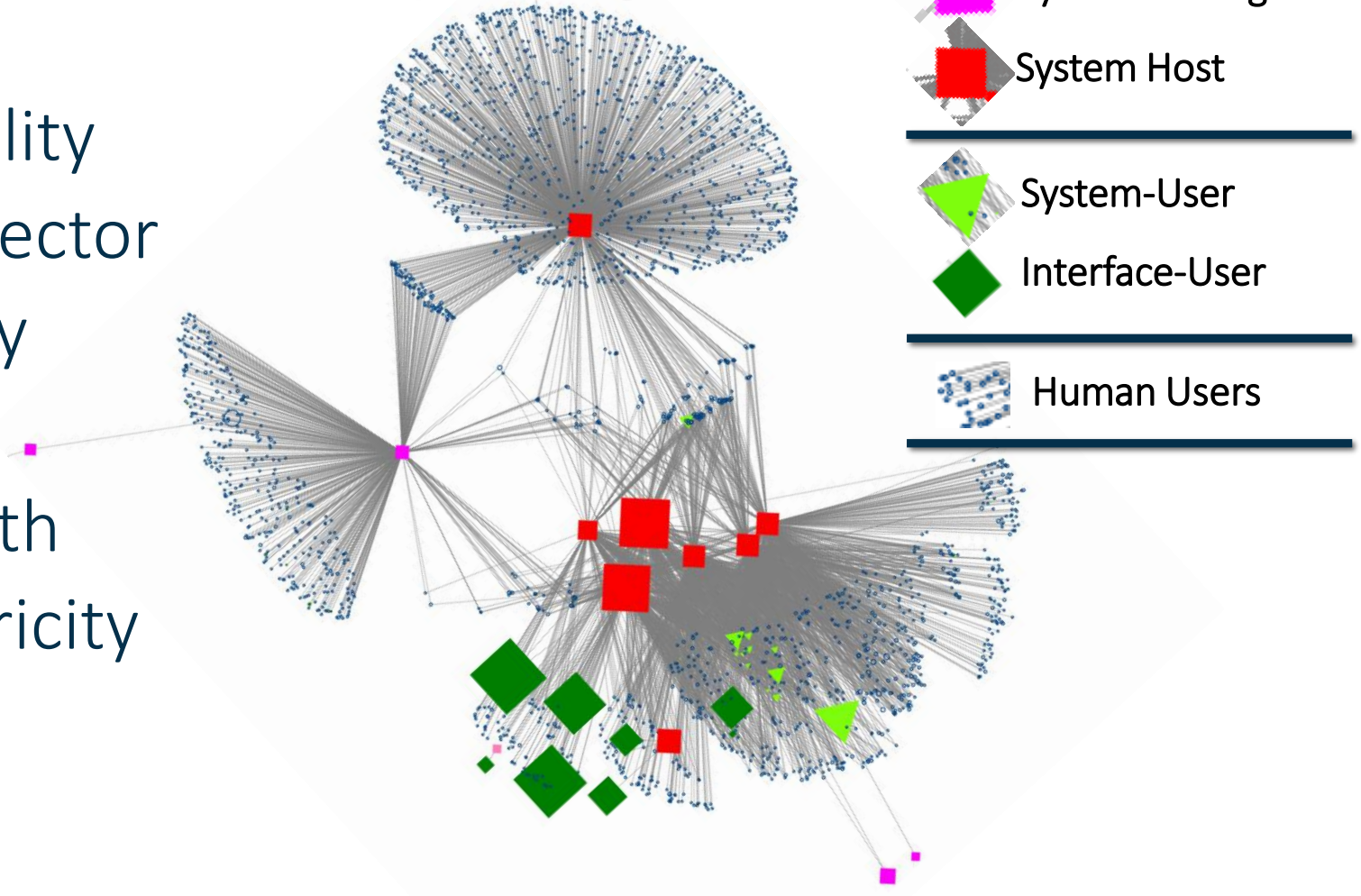
Example 'Graph' Measures

- Centrality
- Eigenvector
- Density
- : Reach
- Strength
- Recopricity

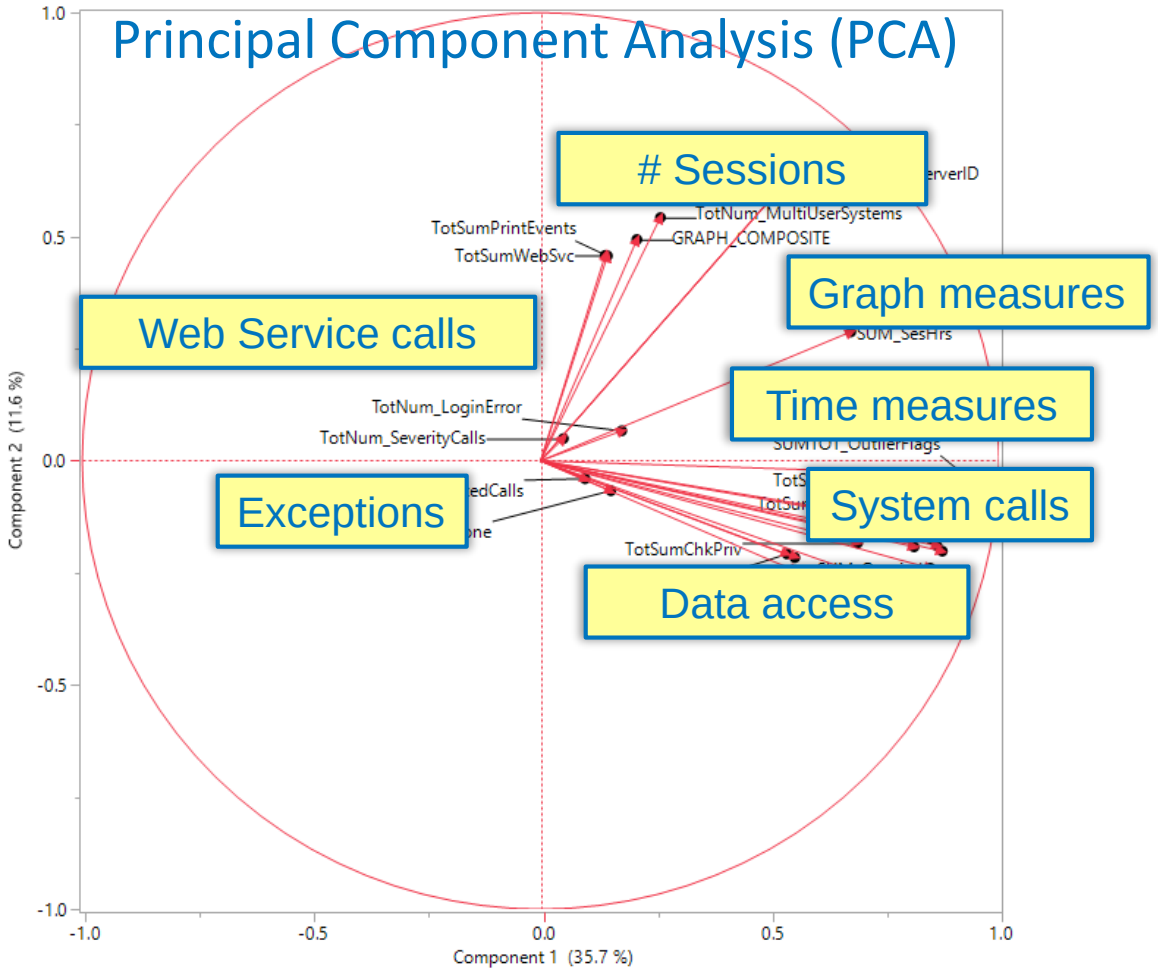


Network Graph Analytics

- Centrality
- Eigenvector
- Density
- Reach
- Strength
- Recopricity



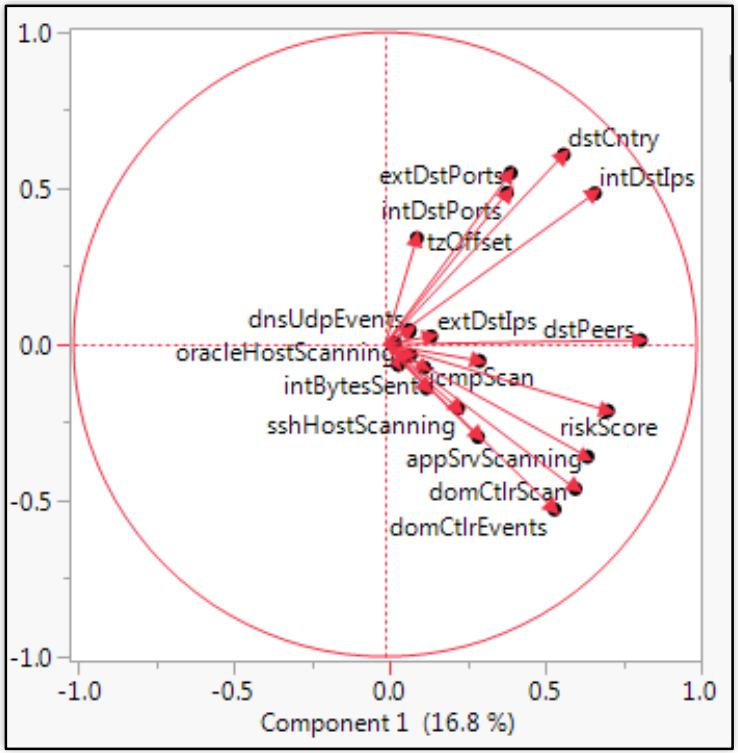
Dimensionality Reduction

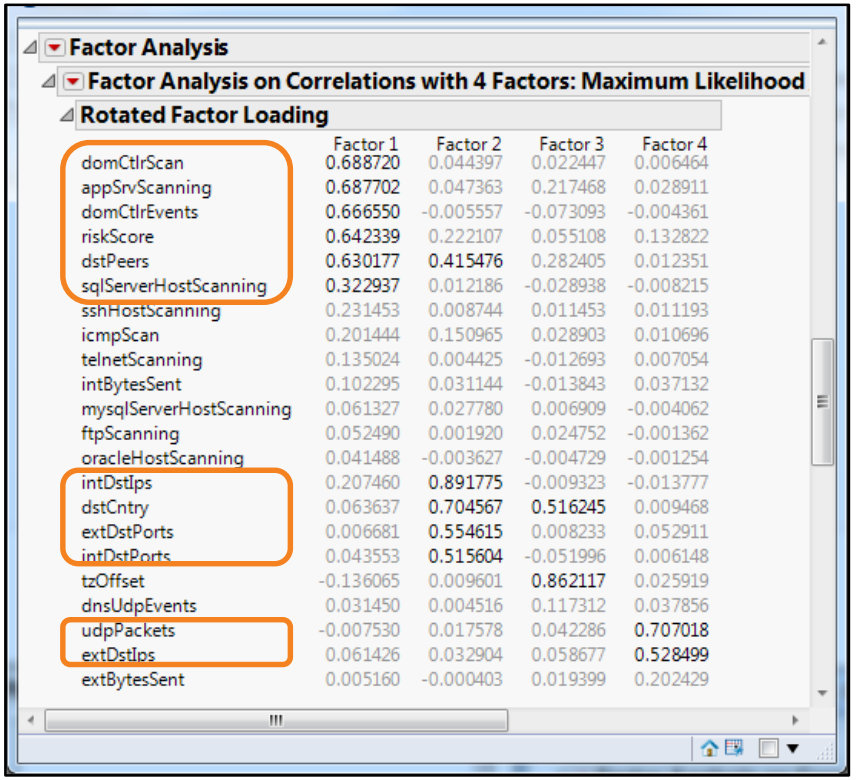


Feature Selection: Advanced Methods

Seeking Connections Amongst Variables

- Examining relationships between variables -> potentially aggregating measures
- Example: correlation, binning, Variable Clustering, Principal Component Analysis (PCA)





Factor Analysis

Factor Analysis on Correlations with 4 Factors: Maximum Likelihood

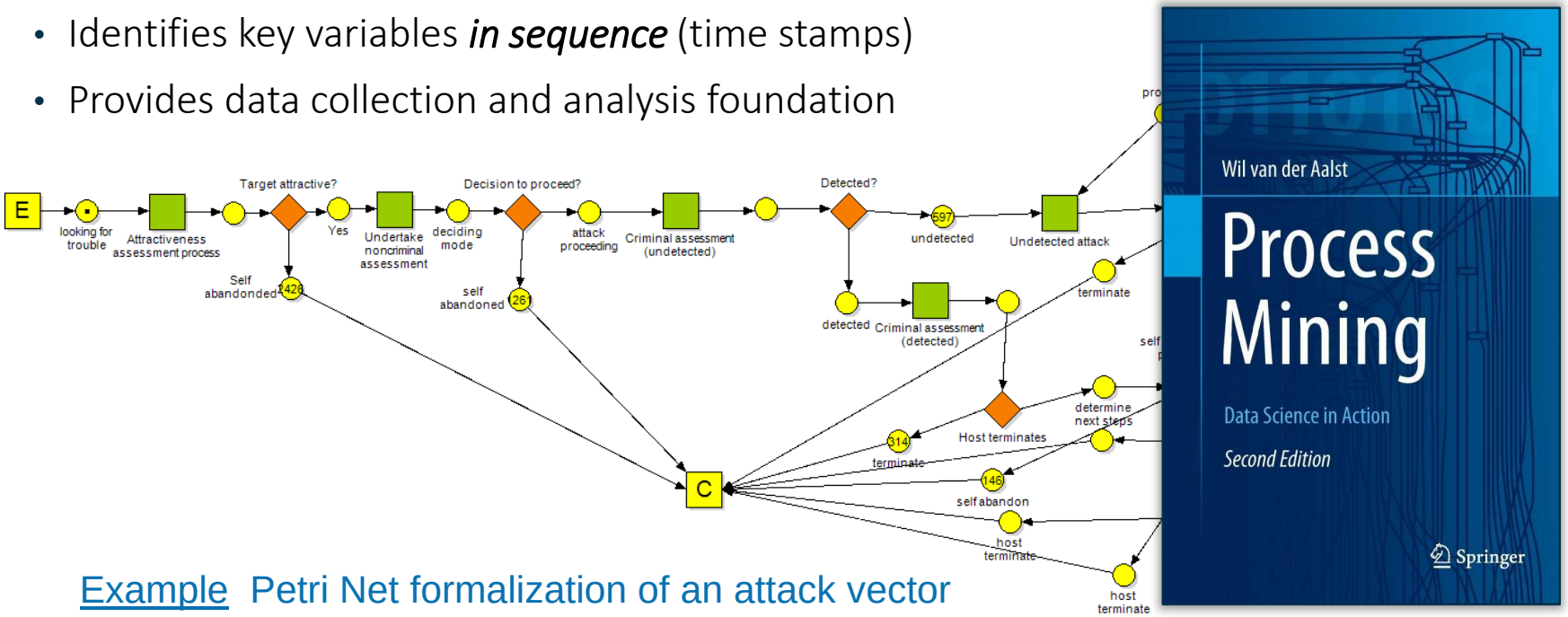
Rotated Factor Loading

	Factor 1	Factor 2	Factor 3	Factor 4
domCtlrScan	0.688720	0.044397	0.022447	0.006464
appSrvScanning	0.687702	0.047363	0.217468	0.028911
domCtlrEvents	0.666550	-0.005557	-0.073093	-0.004361
riskScore	0.642339	0.222107	0.055108	0.132822
dstPeers	0.630177	0.415476	0.282405	0.012351
sqlServerHostScanning	0.322937	0.012186	-0.028938	-0.008215
sshHostScanning	0.231453	0.008744	0.011453	0.011193
icmpScan	0.201444	0.150965	0.028903	0.010696
telnetScanning	0.135024	0.004425	-0.012693	0.007054
intBytesSent	0.102295	0.031144	-0.013843	0.037132
mysqlServerHostScanning	0.061327	0.027780	0.006909	-0.004062
ftpScanning	0.052490	0.001920	0.024752	-0.001362
oracleHostScanning	0.041488	-0.003627	-0.004729	-0.001254
intDstIps	0.207460	0.891775	-0.009323	-0.013777
dstCntry	0.063637	0.704567	0.516245	0.009468
extDstPorts	0.006681	0.554615	0.008233	0.052911
intDstPorts	0.043553	0.515604	-0.051996	0.006148
tzOffset	-0.136065	0.009601	0.862117	0.025919
dnsUdpEvents	0.031450	0.004516	0.117312	0.037856
udpPackets	-0.007530	0.017578	0.042286	0.707018
extDstIps	0.061426	0.032904	0.058677	0.528499
extBytesSent	0.005160	-0.000403	0.019399	0.202429

Feature Extraction: Advanced Measures

Process Mining / Analytics

- Formal specification of common process vectors (behavioral decision trees)
  - Identifies key variables *in sequence* (time stamps)
  - Provides data collection and analysis foundation

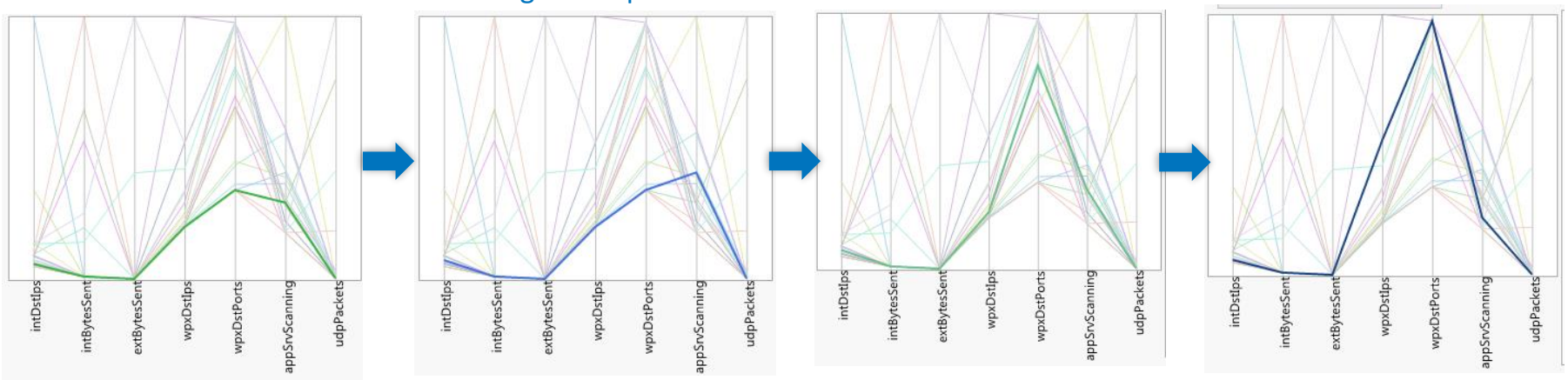


Example Petri Net formalization of an attack vector

Feature Extraction: Advanced Measures

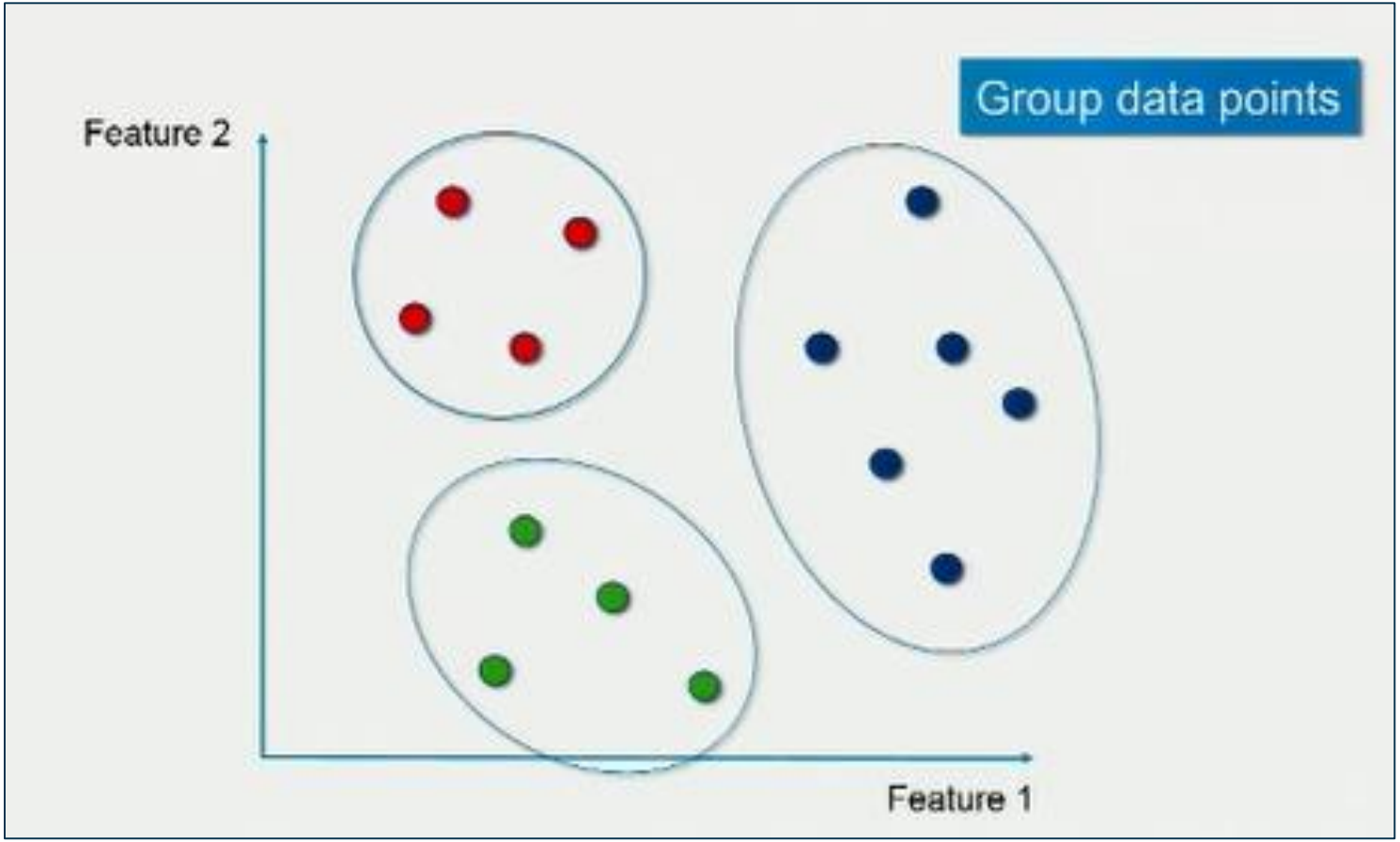
Process Analytics Example

Signature pattern for identified INFECTED IP

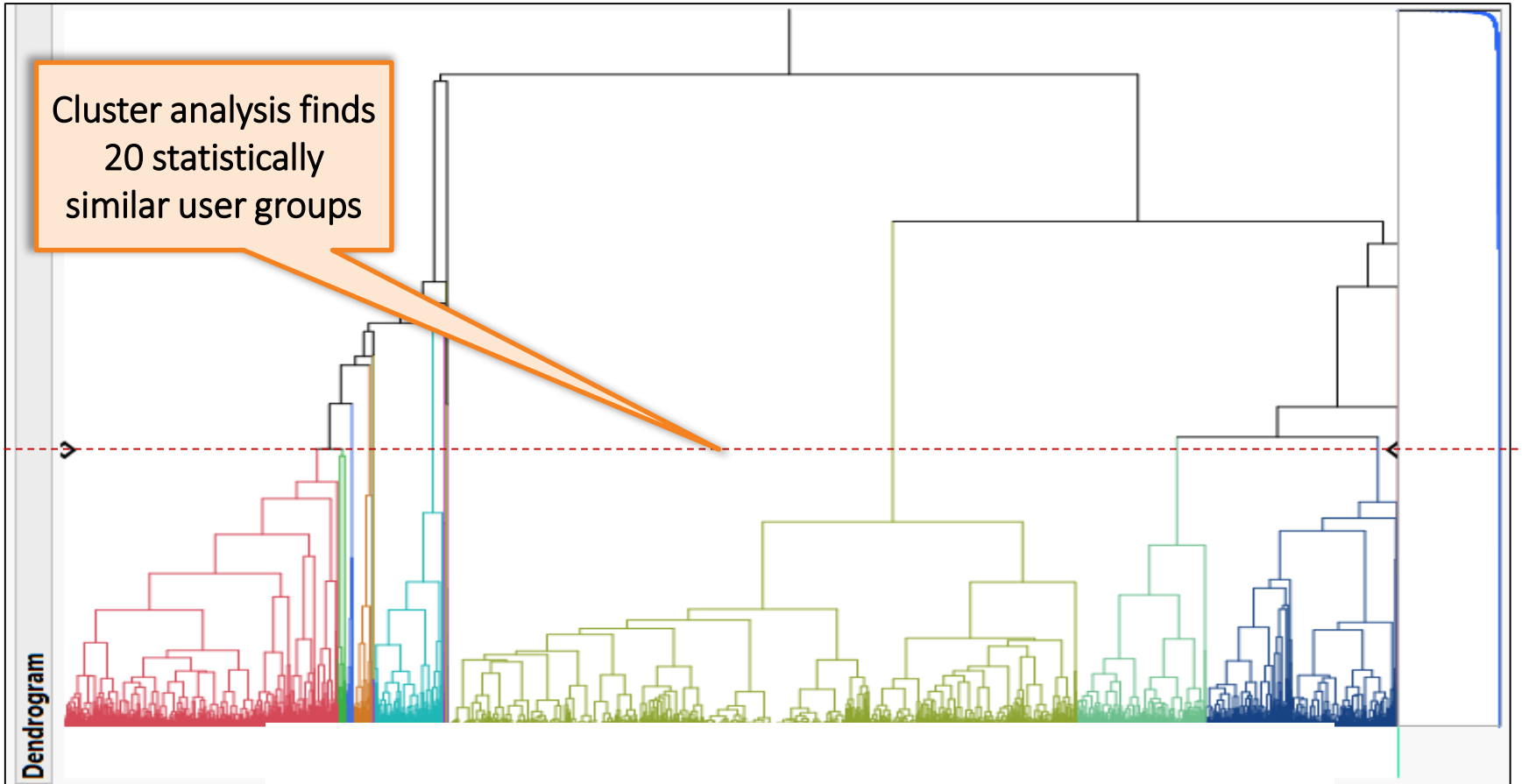


WpxDstIps: **Web Proxy Host Scanning Analysis**	Number distinct external hosts attempting to reach through the web proxy
WpxDstPorts: **Web Proxy Destination Port Scanning Analysis**	Maximum number of distinct external destination ports connected
AppSrvScanning: **Application Server Host Scanning Analysis**	Scanning for devices hosting an http or application server

Unsupervised Machine Learning : Cluster Analysis



Pattern Extrapolation Machine Learning (Unsupervised)

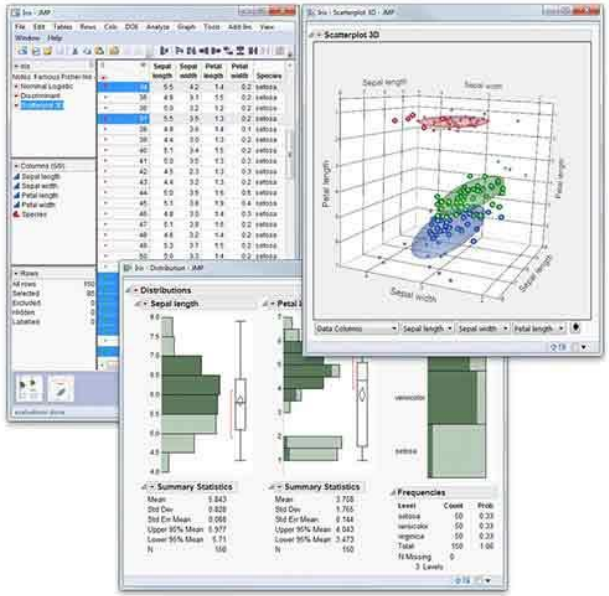


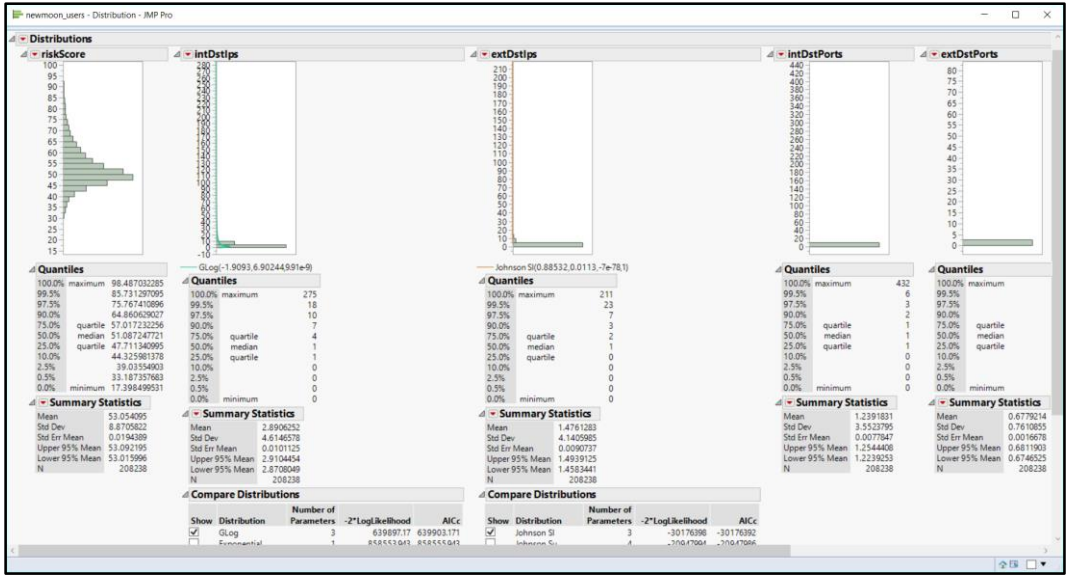
Data Exploration and Cluster Analysis of Users

This exercise illustrates how to explore statistical factors associated with user network behavior and to generate statistically self-similar groups using cluster analysis.

Descriptive / Diagnostic Data Exploration

SAS JMP





SAS JMP Professional Desktop Analytics Tool

Large, disparate or messy data from lots of sources...

Data in databases

ORACLE DB2
SQL SERVER TERADATA

Data cleansing
Data integration
Data sampling
Data security
Data access
Analytics

Use JMP as a SAS client, write SAS code in JMP, submit and view results. Utilize interactive JMP graphics, visualization and analysis tools.

Wrap-Up

Section Review

Cybersecurity Analytics Maturity

Anomaly Detection	Data-aware Investigations	Predictive Detection	Risk Awareness / Resource Optimization

- Big data management
- Flags, rules, and alerts

- Structured data
- Counts of key measures
- Foundation for comparisons across entities and over time








Cybersecurity Analytics Maturity

Anomaly Detection	Data-aware Investigations	Predictive Detection	Risk Awareness / Resource Optimization
• Big data management • Flags, rules, and alerts • Multivariate statistics, inference & unsupervised machine learning • Segments extracted as baselines	Understanding • Cleansing • Diagnostics • Labeling • Feature engineering	?	?

Cybersecurity Data Science (CSDS) Lifecycle