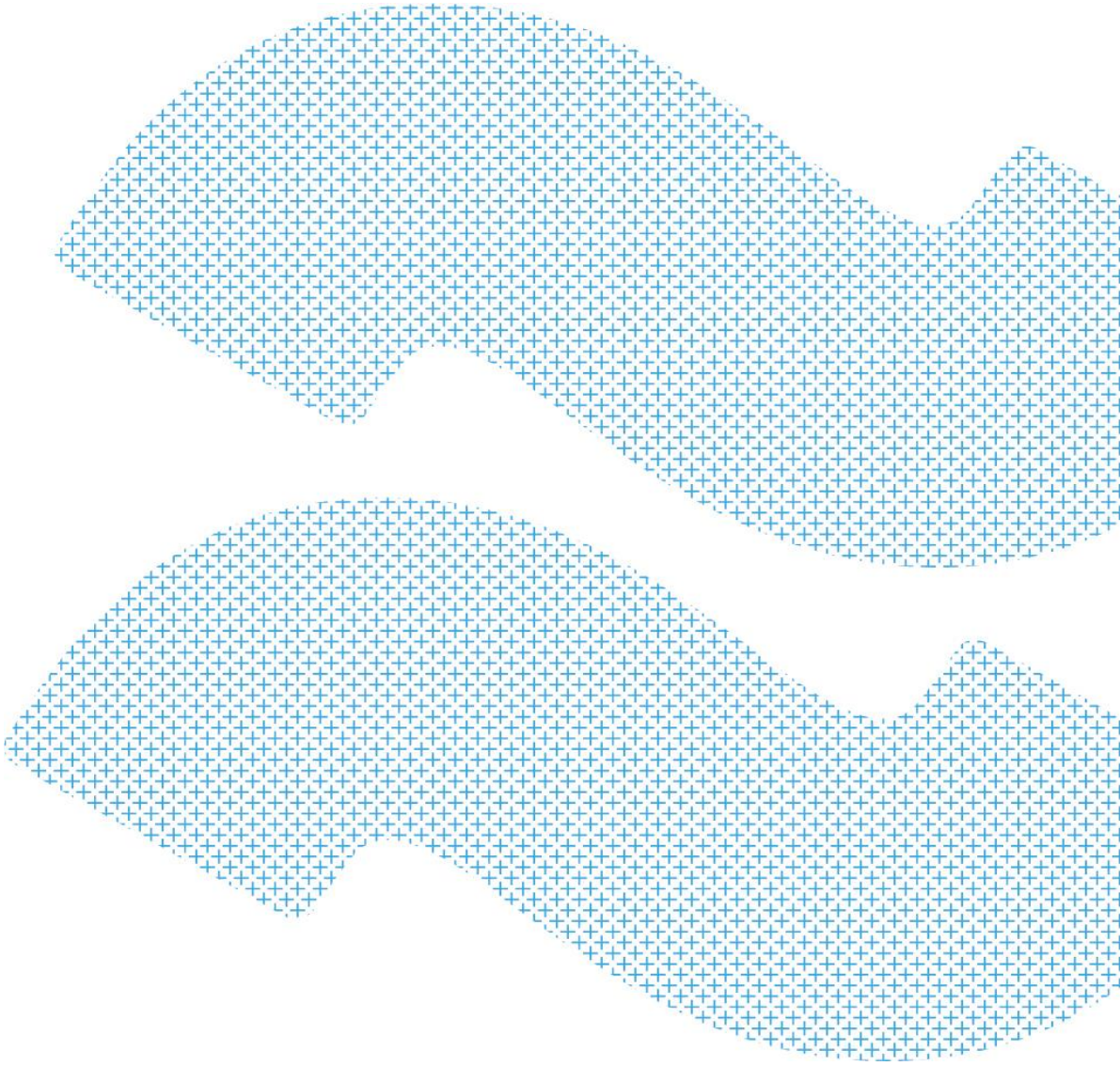
Data Science for Cyber Risk:

Measurement, Methods, and Models

drs. Scott Allen Mongeau
Data Scientist
Cyber Security
sas

February 2017



GARP | Global Association of Risk Professionals

The views expressed in the following material are the author's and do not necessarily represent the views of the Global Association of Risk Professionals (GARP), its Membership or its Management.

Scott Allen Mongeau

Scott Mongeau
Data Scientist
Cyber Security

scott.mongeau@sas.com

06 837 030 97

LinkedIn
Twitter
Blog

Experience

- **SAS Institute**
  Data Scientist
- **Deloitte**
  Manager Analytics
- **Nyenrode University**
  Lecturer Analytics
- **SARK7 Analytics**
  Owner / Principal Consultant
- **Genentech Inc. / Roche**
  Principal Analyst / Sr Manager
- **Atradius**
  Sr. R&D Engineer
- **CFSI**
  CIO

Education

- **PhD (ABD)**
- **MBA (OneMBA)**
- **Masters Financial Management**
- **Certificate Finance**
- **GD IT Management**
- **Masters Computer & Communications Technology**

YouTube

- *Introduction to Advanced Analytics*
- *Introduction to Cognitive Analytics*
- *TedX RSM: Data Analytics*

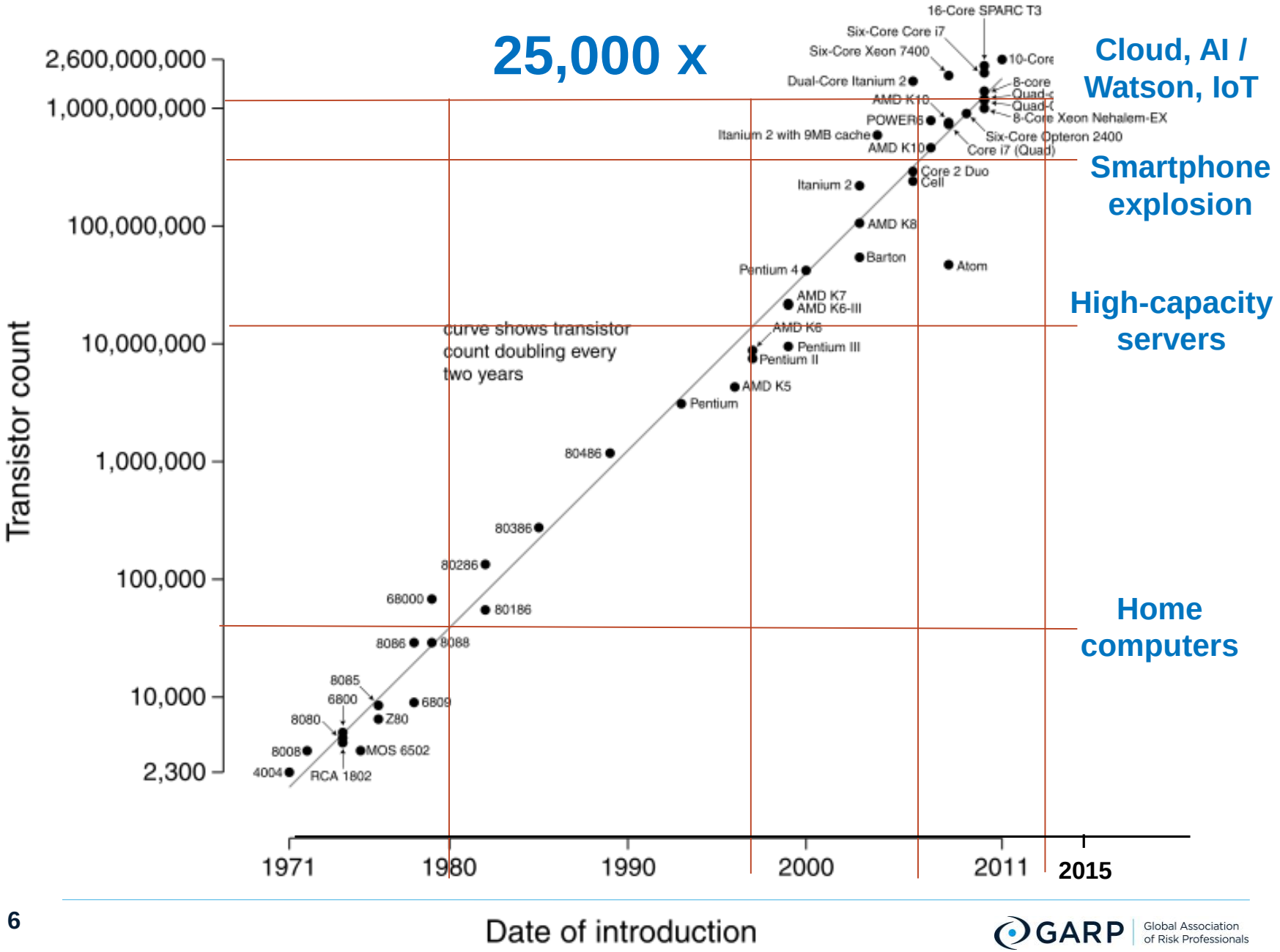
Cyber Risk: A Measurement Challenge

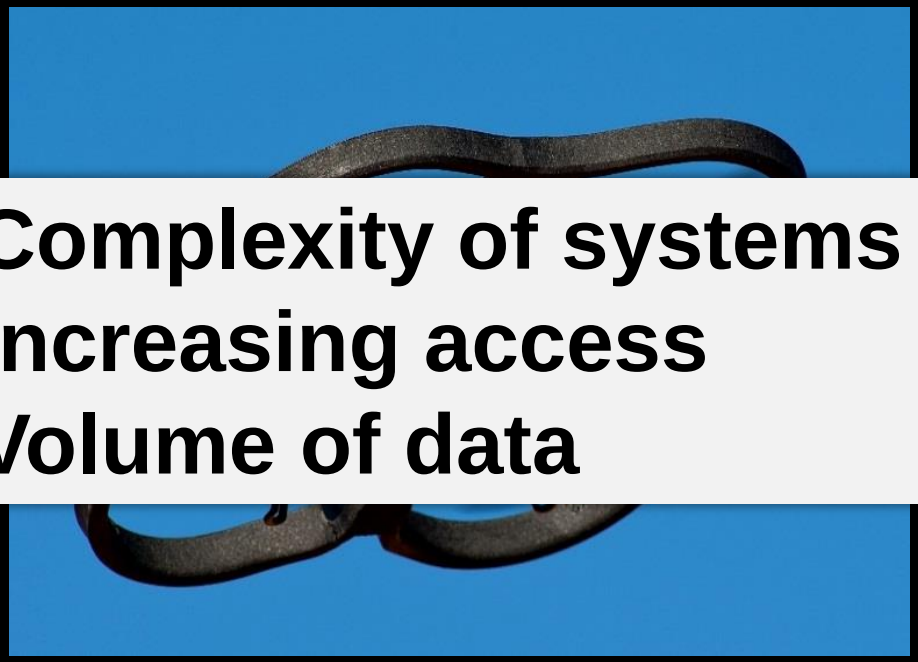
- **Context setting**
- **Cyber risk measurement**
  - What is data analytics / data science?
  - What methods and technologies are involved?
- **Experience and learnings from the field**
  - Actionable insights
  - Emerging methods
- **Trends and opportunities**

CYBER RISKS

GARP | Global Association of Risk Professionals

Moore's Law: Exponential growth of computing power

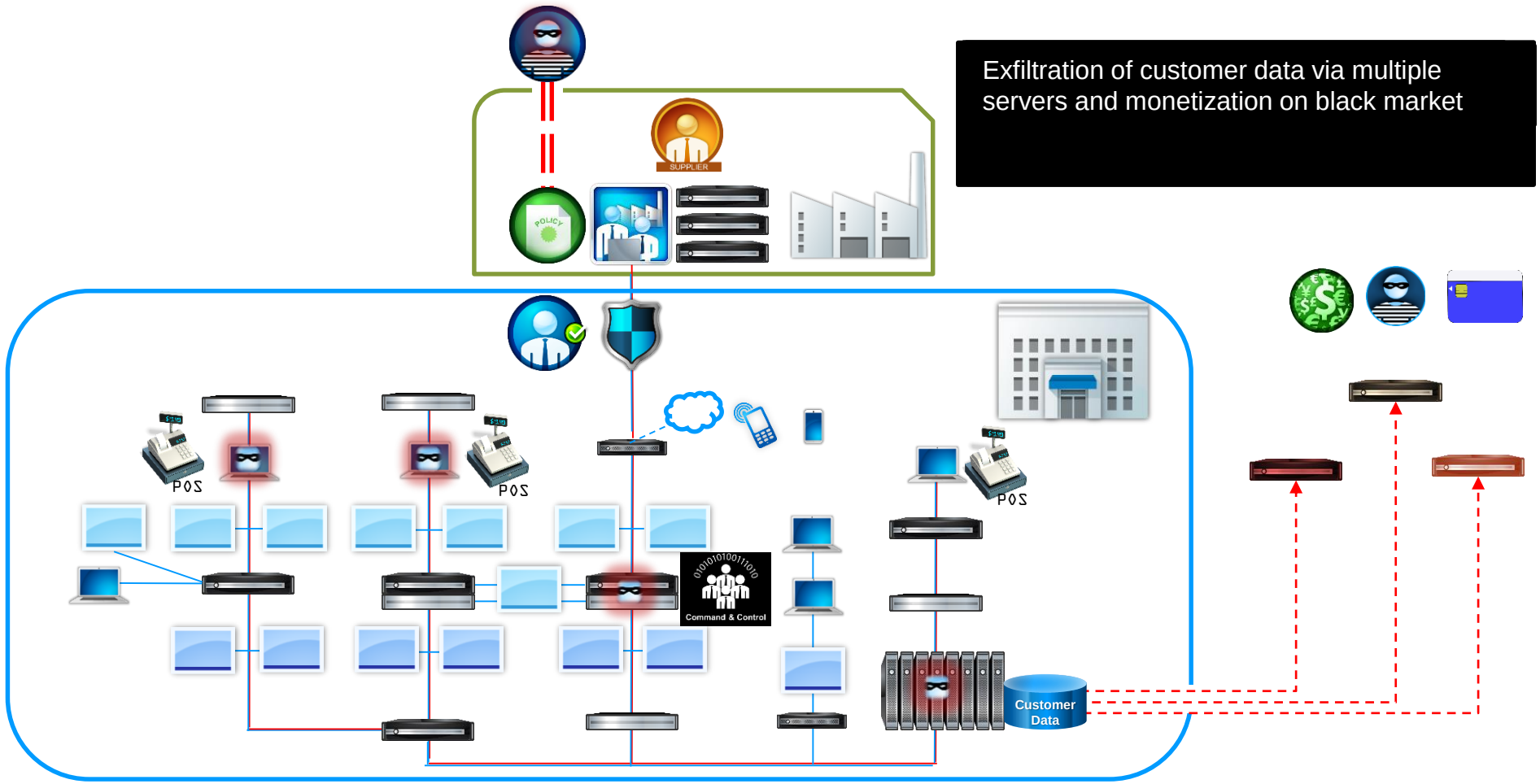




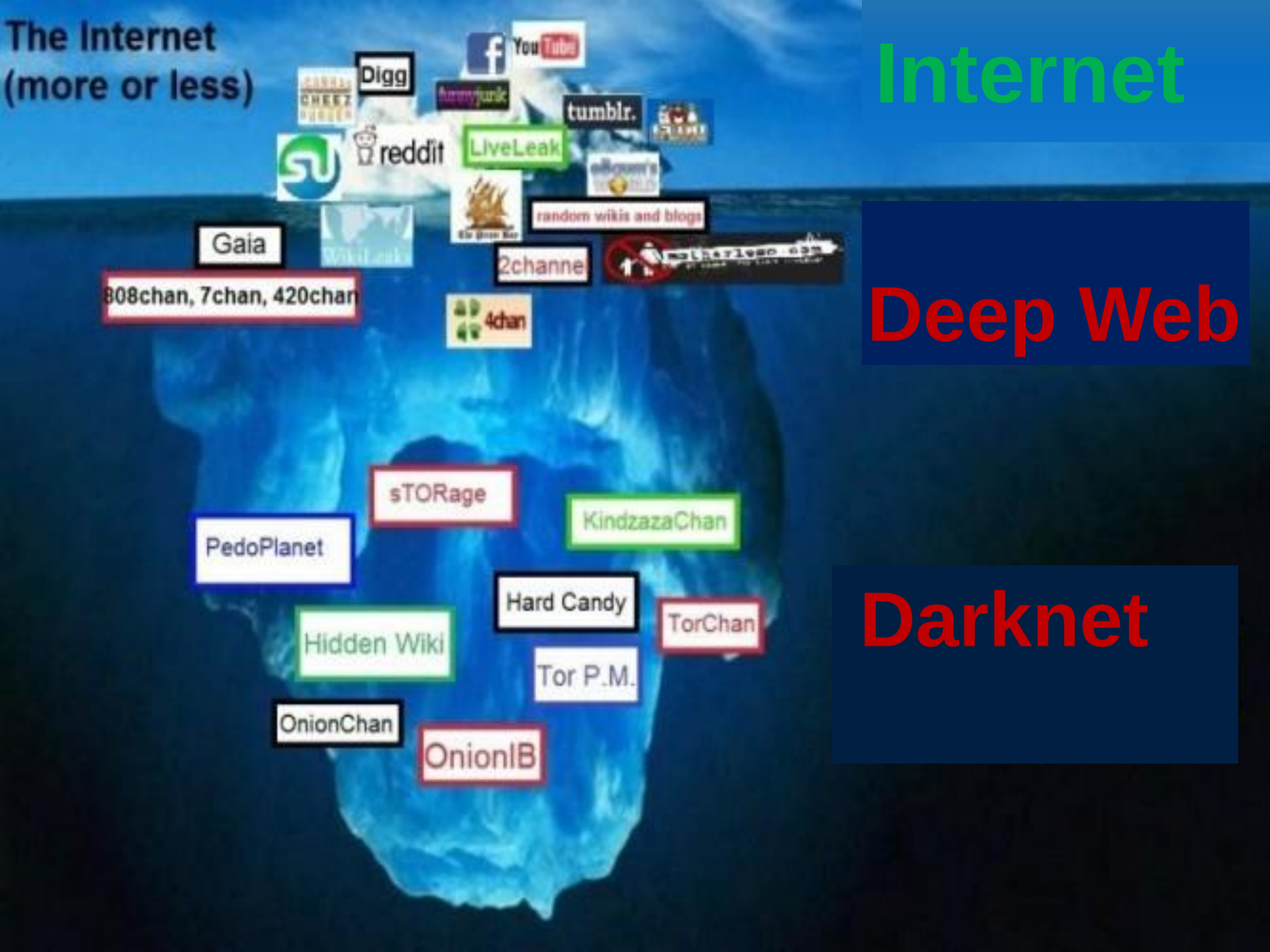
- Complexity of systems
- Increasing access
- Volume of data

- BYOD
- VMs / containers
- IoT / smart devices
- ICS SCADA

Anatomy of a sophisticated Cyber Attack



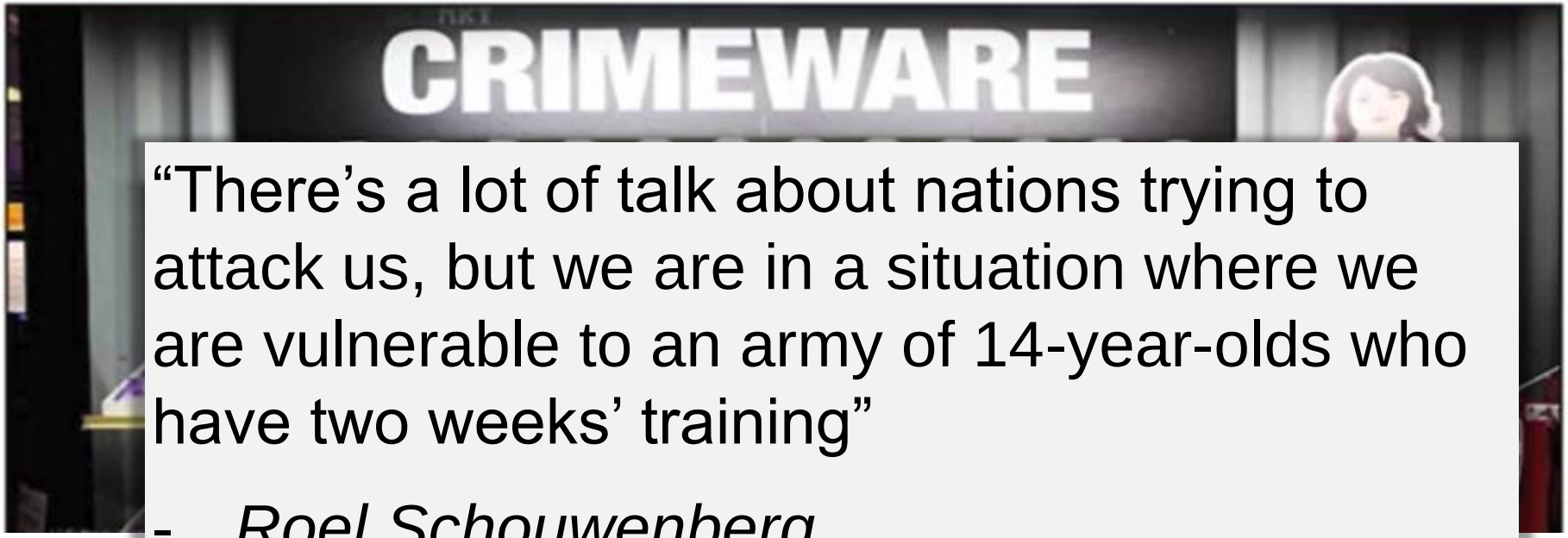
Exfiltration of customer data via multiple servers and monetization on black market


The Internet
(more or less)
Digg
YouTube
reddit
LiveLeak
tumblr.
random wikis and blogs
Gaia
808chan, 7chan, 420chan
2channe
4chan
sTORage
KindzazaChan
PedoPlanet
Hard Candy
TorChan
Hidden Wiki
Tor P.M.
OnionChan
OnionIB
Internet
Deep Web
Darknet

DISTRIBUTE, WHOLESALE, RESELLERS……

CRIMEWARE TOOLKITS

Cyber Threat Professional



"There's a lot of talk about nations trying to attack us, but we are in a situation where we are vulnerable to an army of 14-year-olds who have two weeks' training"

- *Roel Schouwenberg*
  Senior Researcher, Kaspersky Lab
  http://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet

NOTICE OF EXTORTION

Your business, [redacted], has been targeted for extortion. The selection process is random, and was not triggered by any event under your control.

Should you fail to pay the one-time monetary tribute, by the deadline provided below, your business will be **severely and irreparably damaged**. The following methods are commonly employed in cases of non-compliance:

- Negative Online Reviews
- BBB Complaints
- Harassing Telephone Calls
- Fraudulent Delivery Orders
- Telephone Denial-of-Service
- Bomb Threats
- Vandalism
- Mercury Contamination

Anonymous Reports of:

- Health Code Violations
- OSHA Violations
- Criminal Tax Evasion
- Money Laundering
- Illegal Drug Sales
- Marijuana Grow Operations
- Methamphetamine Production
- Terrorist Training Activity

The tribute price is only One Bitcoin (1 BTC), but must be paid by **August 15, 2014**. Payment is to be made to the Bitcoin Wallet Address listed below.

If payment is not received, our team will begin taking the actions listed above. Once engagement has begun, it can only be stopped for a tribute of Three Bitcoin (3 BTC). Because many of the actions we take are catastrophic and irreversible, is it advised to pay the tribute before the deadline is reached.

Payment Type: Bitcoin

Deadline: August 15, 2014

Amount Due: *(If paid before deadline)* 1 Bitcoin

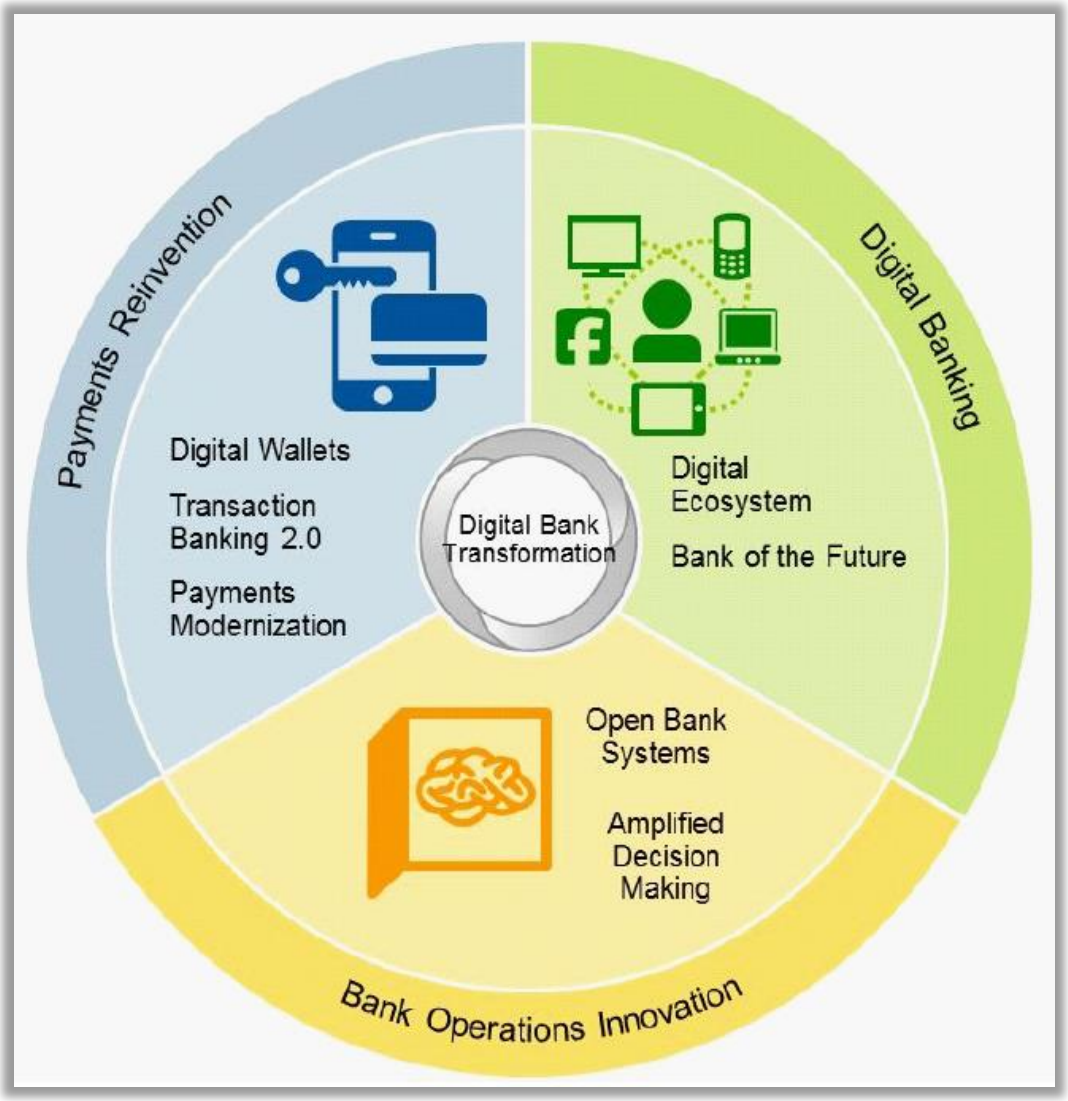
Amount Due: *(If paid after deadline)* 3 Bitcoin

Purchase Bitcoin @
https://www.coinbase.com/

17gt1BancvtnnJwy4BA41VBUH3pfbUvzE

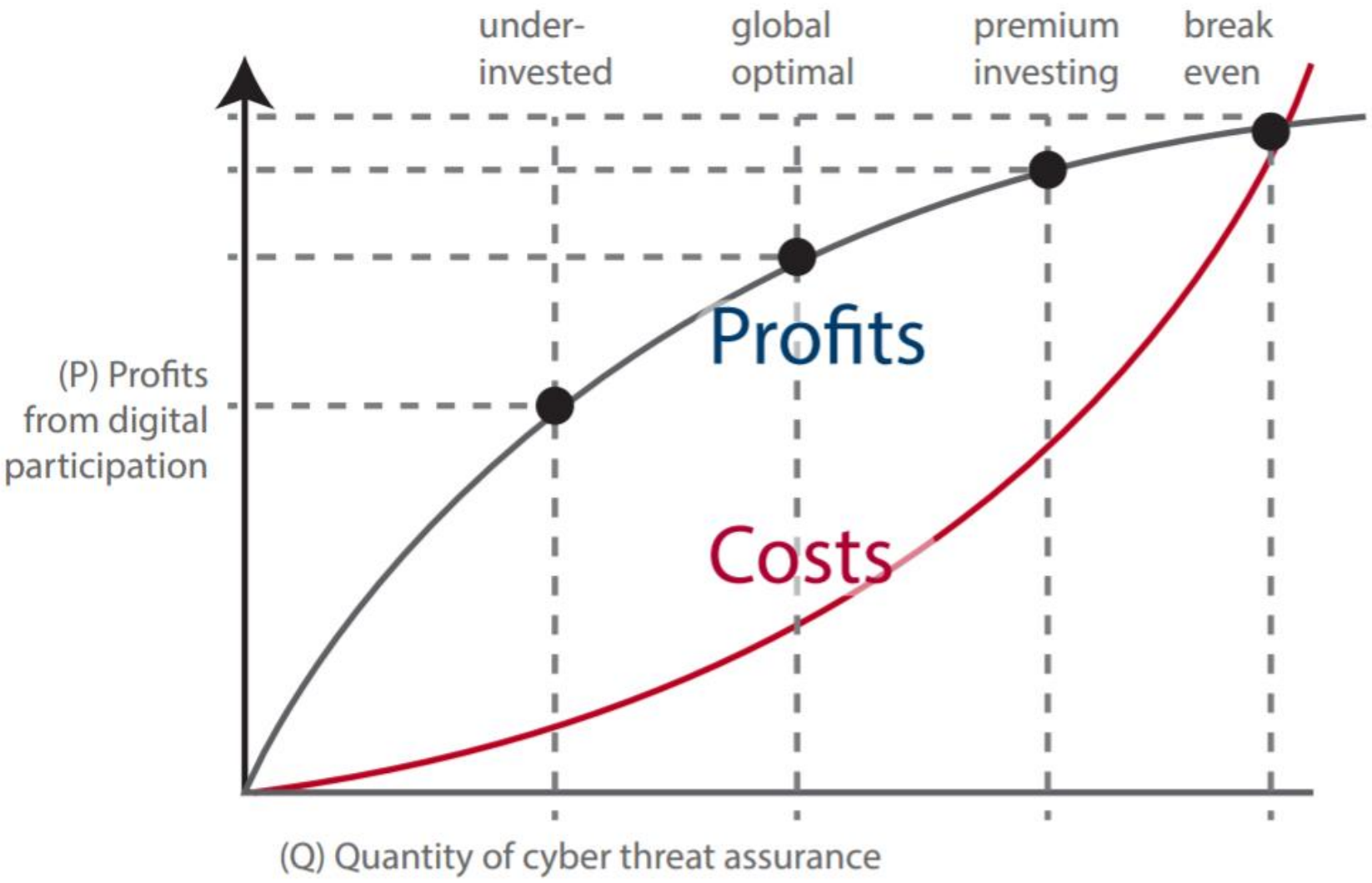
CYBER RISK MANAGEMENT

Competitive pressures… e.g. ‘digital banking’



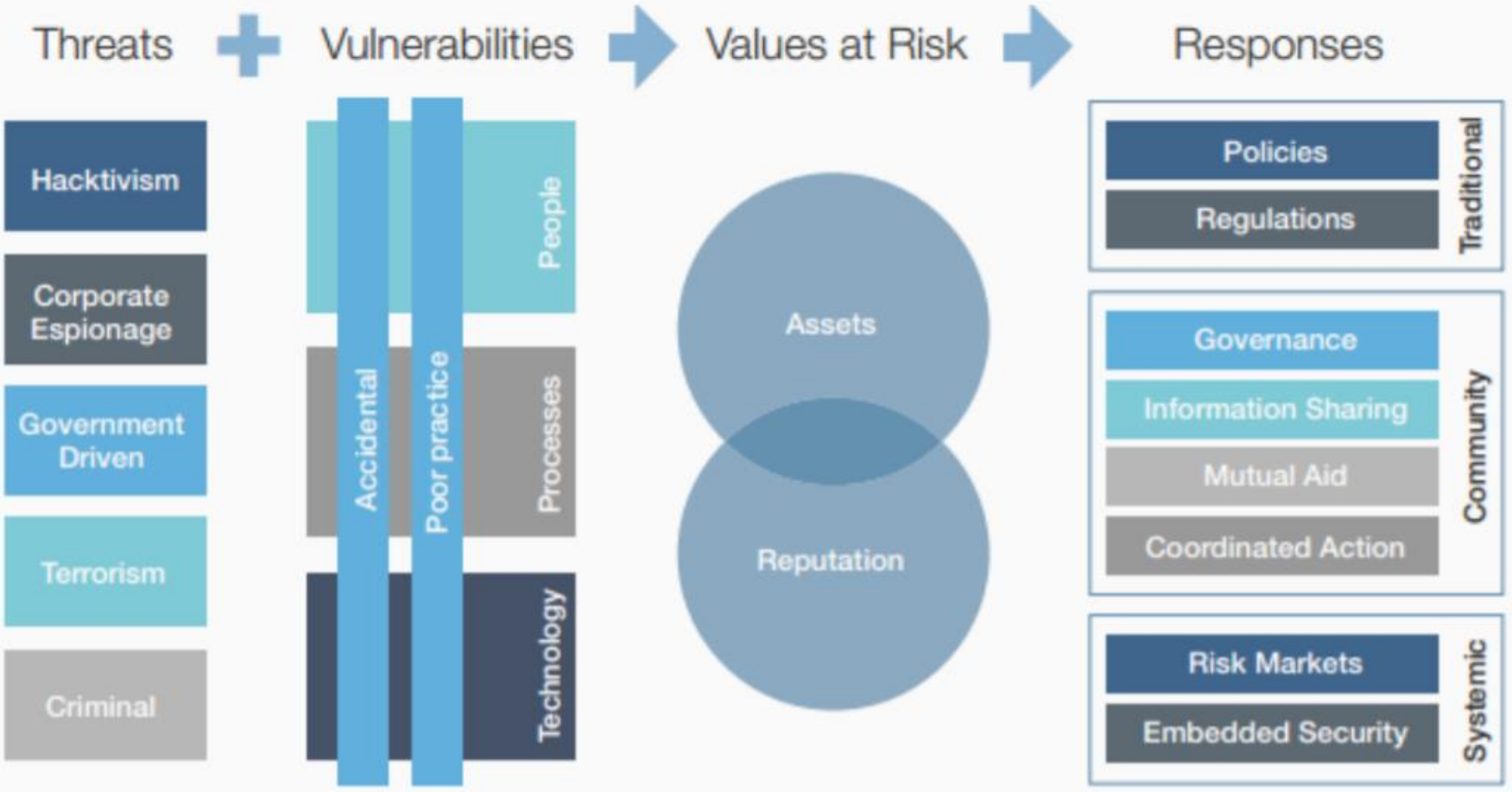
Source: Gartner. 2015. Agenda Overview for Banking and Investment Services.

Optimizing Accessibility Versus Exposure



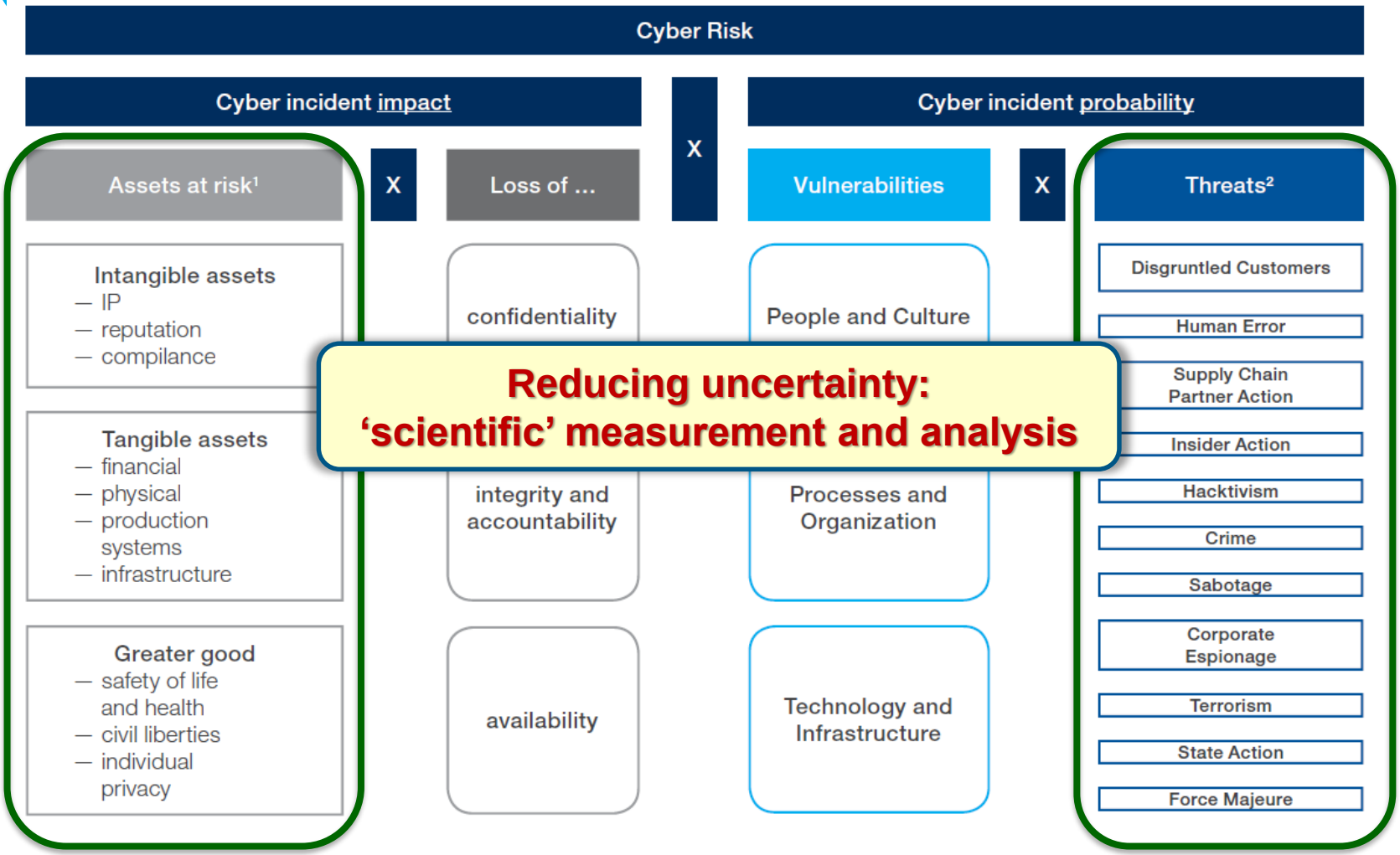
Partnering for Cyber Resilience: Towards the Quantification of Cyber Threats
WEF report in collaboration with Deloitte:
http://www3.weforum.org/docs/WEFUSA_QuantificationofCyberThreats_Report2015.pdf

Data Analytics => Measurement



Partnering for Cyber Resilience: Towards the Quantification of Cyber Threats
WEF report in collaboration with Deloitte:
http://www3.weforum.org/docs/WEFUSA_QuantificationofCyberThreats_Report2015.pdf

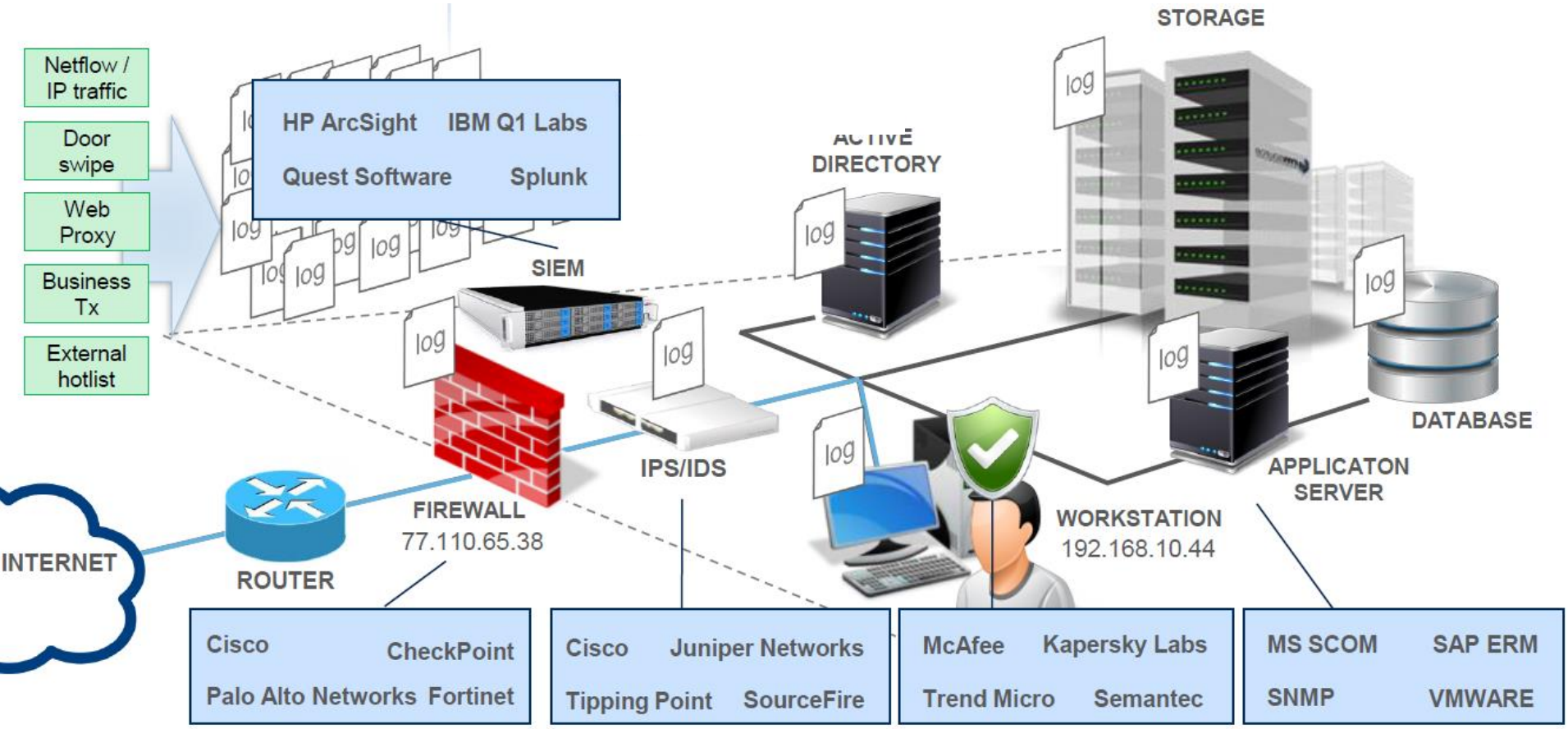
Data Science => Measurement



Advancing Cyber Resilience Principles and Tools for Boards
http://www3.weforum.org/docs/IP/2017/Adv_Cyber_Resilience_Principles-Tools.pdf

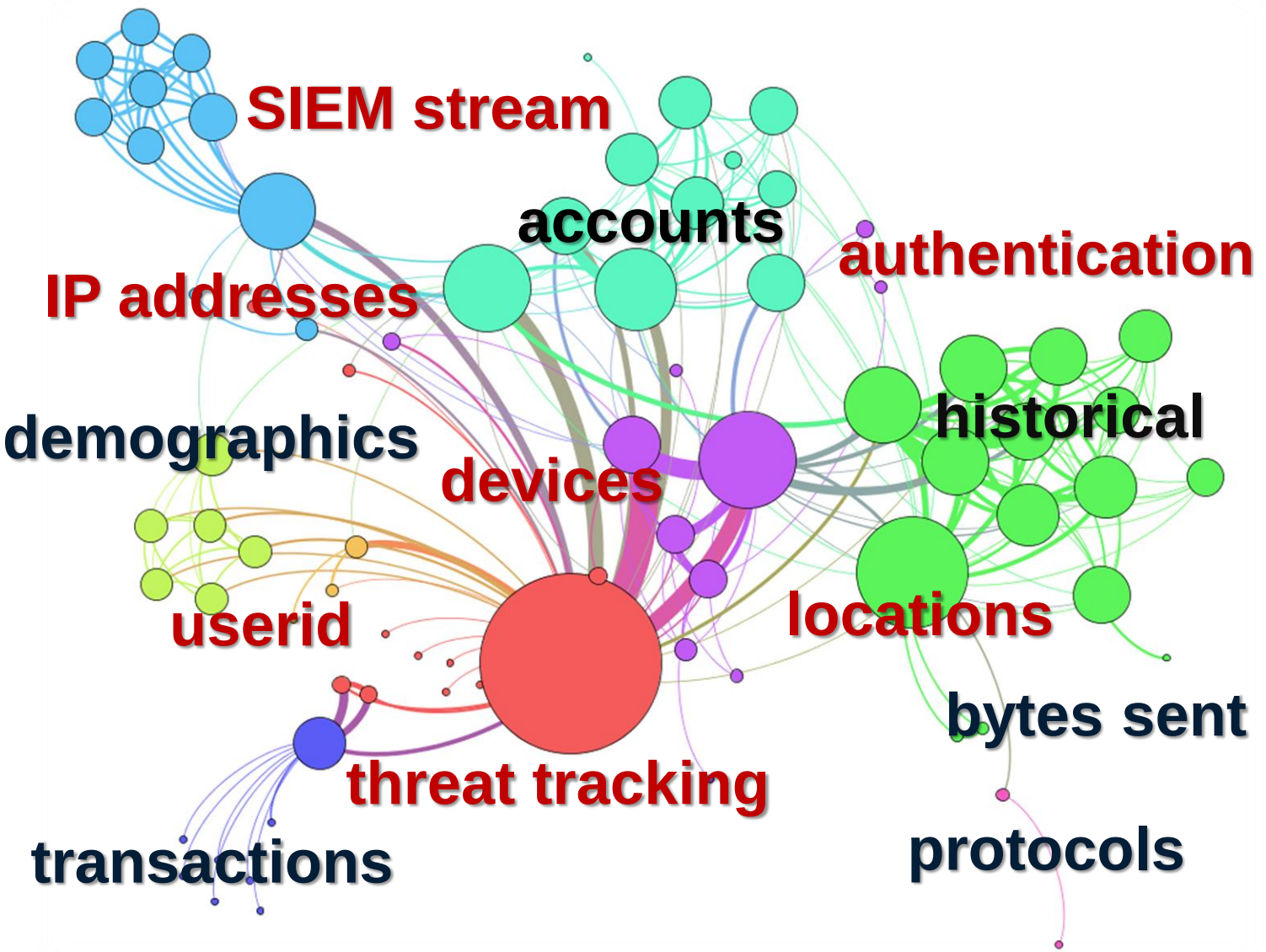
CYBER RISK MEASUREMENT

Many data sources… increasing data volume

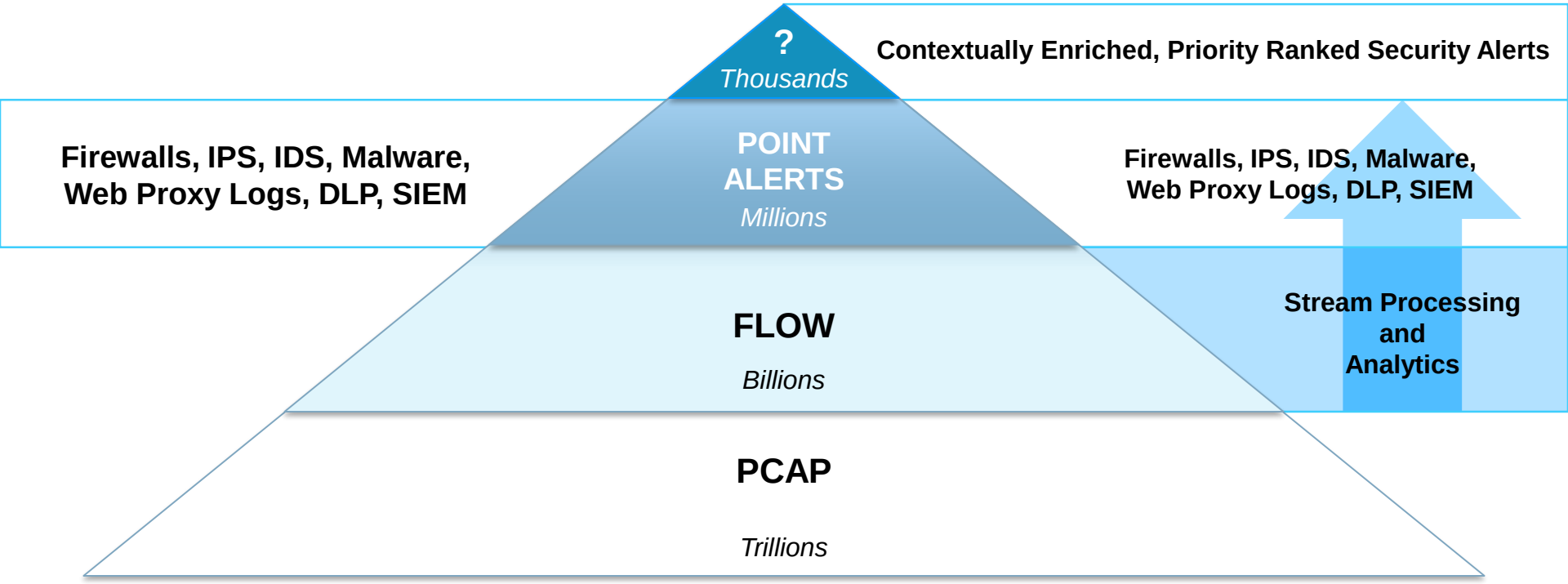


***Source*: Cyber Security Solutions, 2014.**

Linking and Managing 'Big' Cyber Data



Cyber Data Types and Monthly Volumes



In Search of: Targeted, Relevant, Actionable Alerts…

Addressing Challenges: Cyber Risk Analytics

CHALLENGES

Data overload

Disconnected & low quality data

High false positive alerts

Unknown unknowns – no baseline

Slow and manual investigation processes

CYBER DATA SCIENCE

Focused insights from Big Data

Managing and rationalizing data

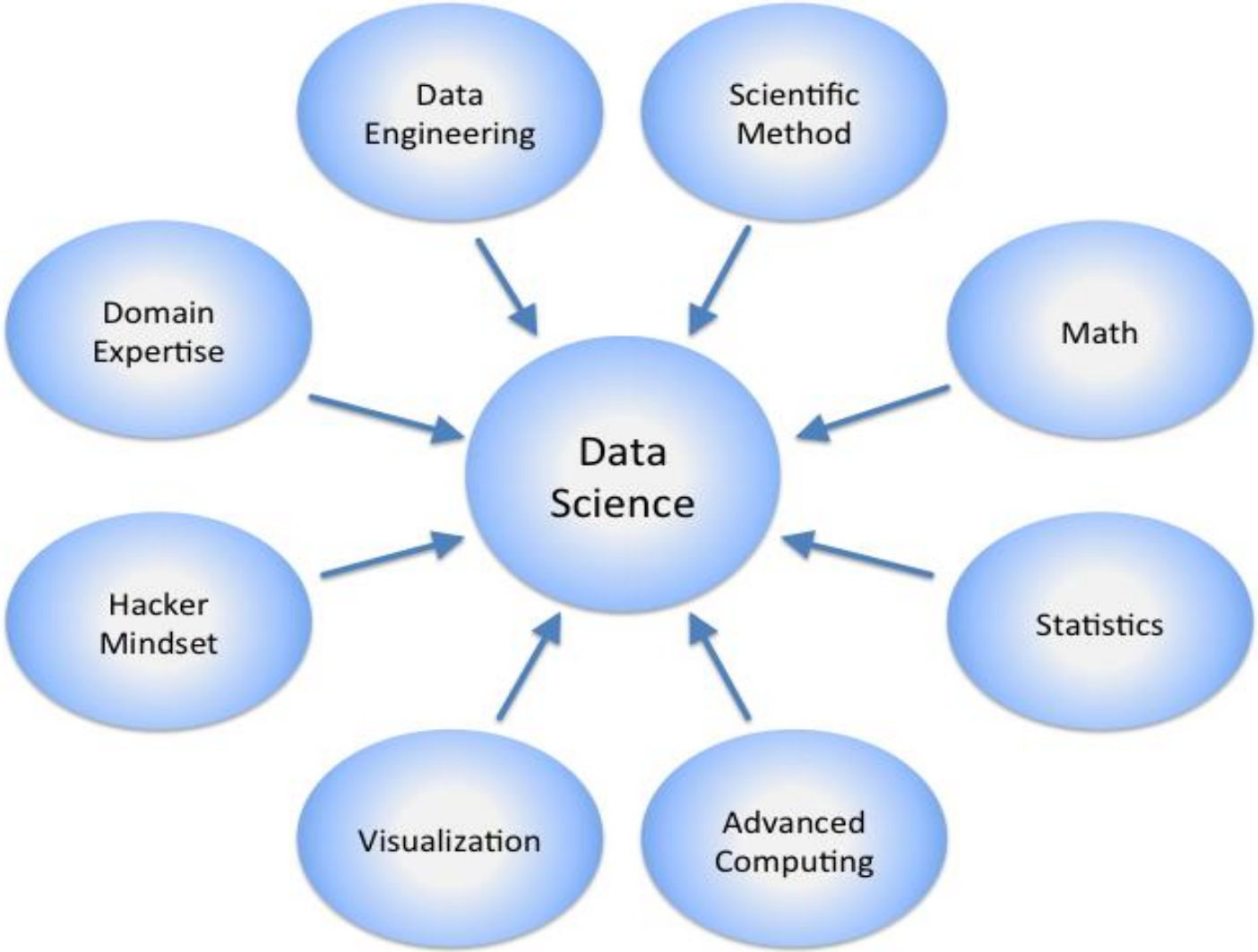
Machine learning identifies hidden patterns

Diagnostics for understanding ‘normal’

Targeted alerts based on anomalies

Data Science?

Data Science / Data Analytics: An Interdisciplinary Practitioner Field

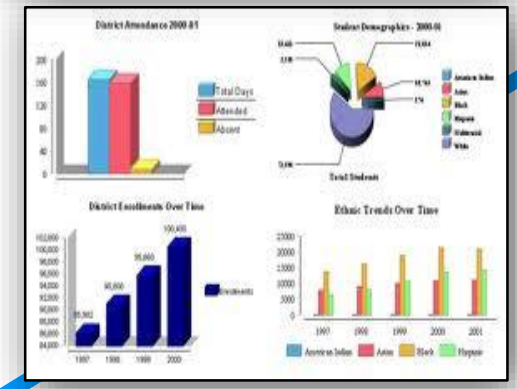


SOPHISTICATION

VALUE

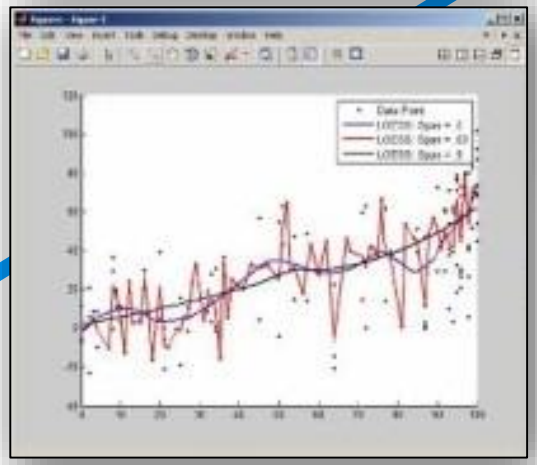
What has happened?

DESCRIPTIVE



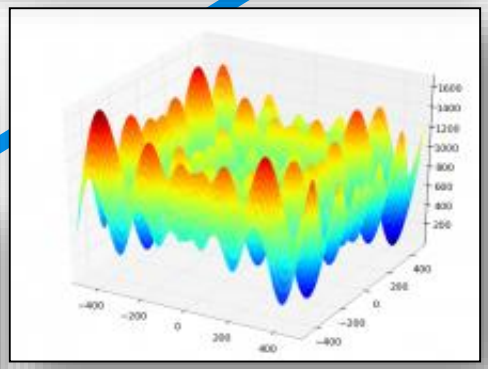
What will happen?

PREDICTIVE



How to optimize?

PRESCRIPTIVE

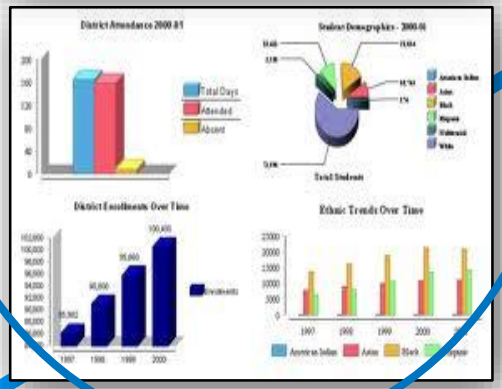


SOPHISTICATION

VALUE

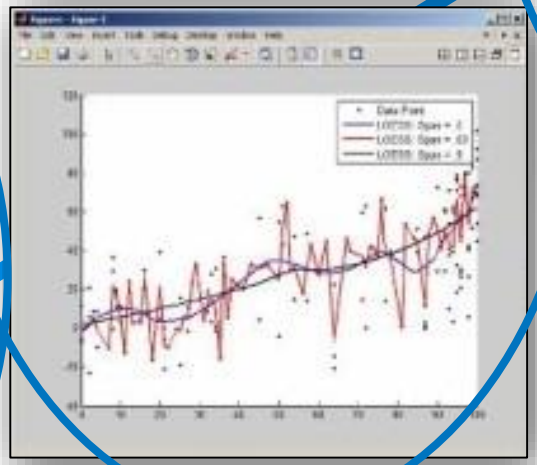
Business Intelligence (BI)

DESCRIPTIVE



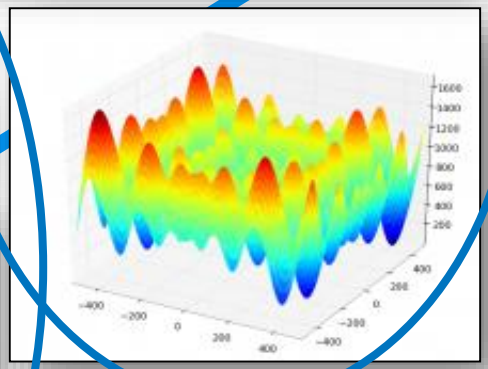
Econometrics Financial Analysis Machine Learning

PREDICTIVE



Operations Management

PRESCRIPTIVE

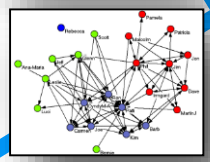


Optimizing Systems

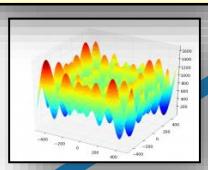
Understanding Patterns

Data visualization

SEMANTIC

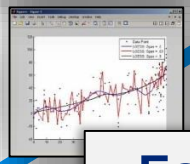


PRESCRIPTIVE



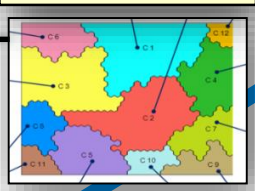
Understanding Social Context & Meaning

PREDICTIVE



Identifying Factors & Causes

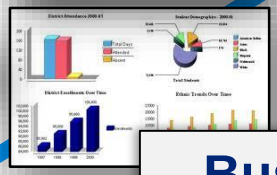
DIAGNOSTICS



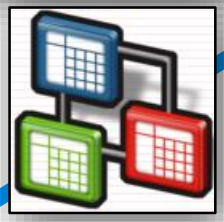
Forecasting & Probabilities

Traditional BI

DESCRIPTIVE



DATA QUALITY



Business Intelligence

Transformed

analytics maturity

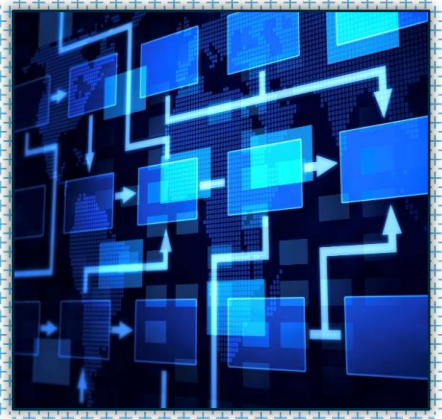
Aspirational

Transactional

business value

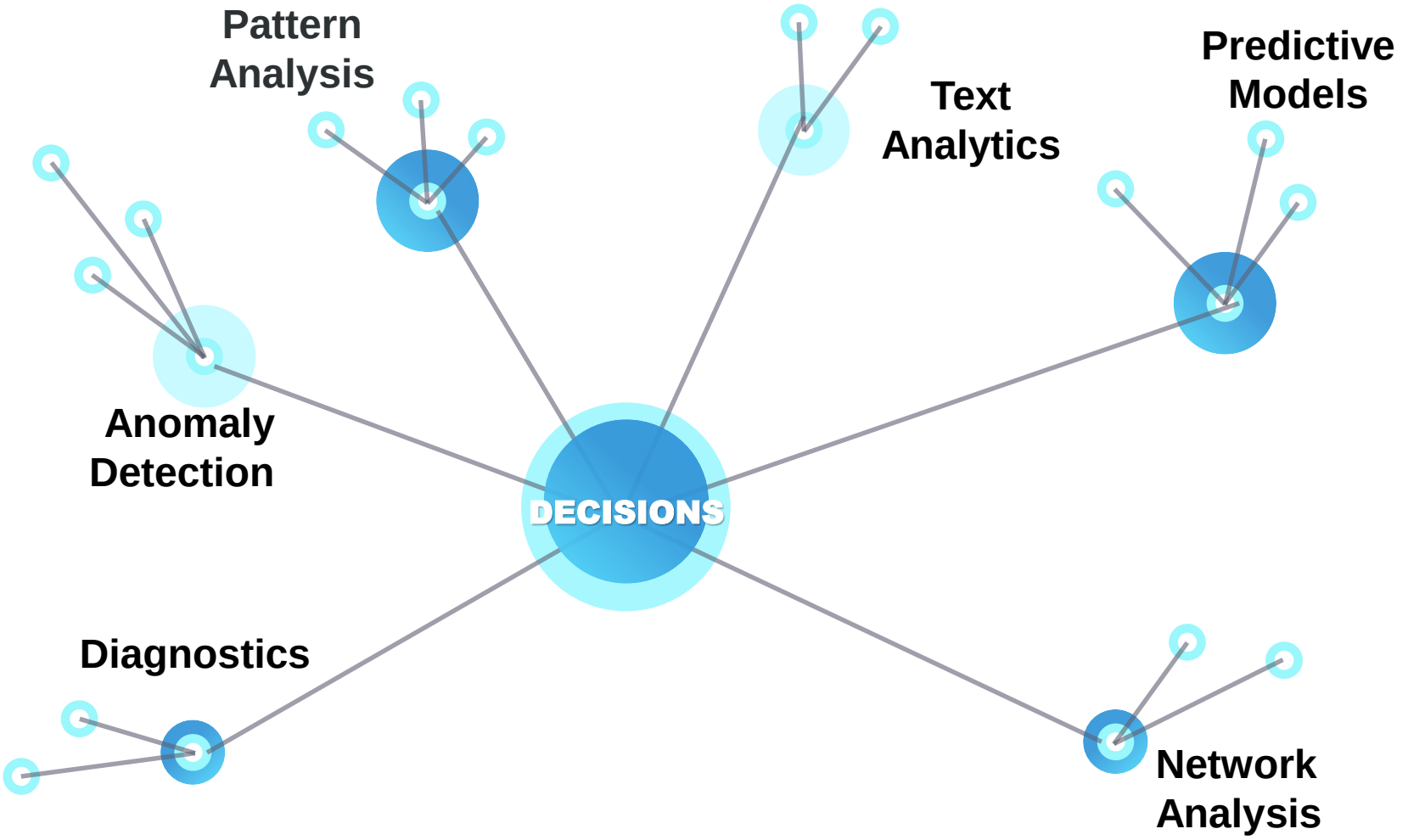
Strategic

Data Science for Cyber Risk



GARP | Global Association of Risk Professionals

Fraud Analytics: A Mature, Adjacent Domain

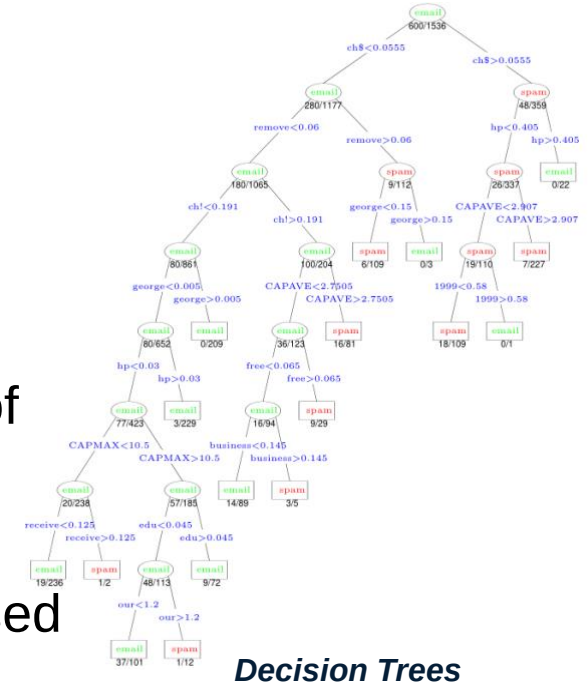


Two Major Approaches

Prediction:

Supervised learning

- You have a baseline: a dataset with examples of what you are attempting to predict or classify (random forests, boosted trees)
- *Example*: known examples of cyber attacks based on Net Flow data

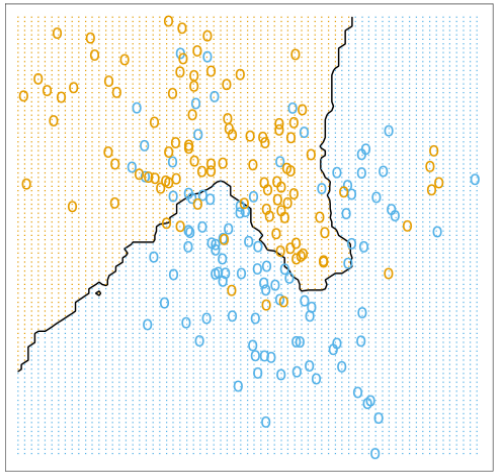


Decision Trees

Discovering Patterns:

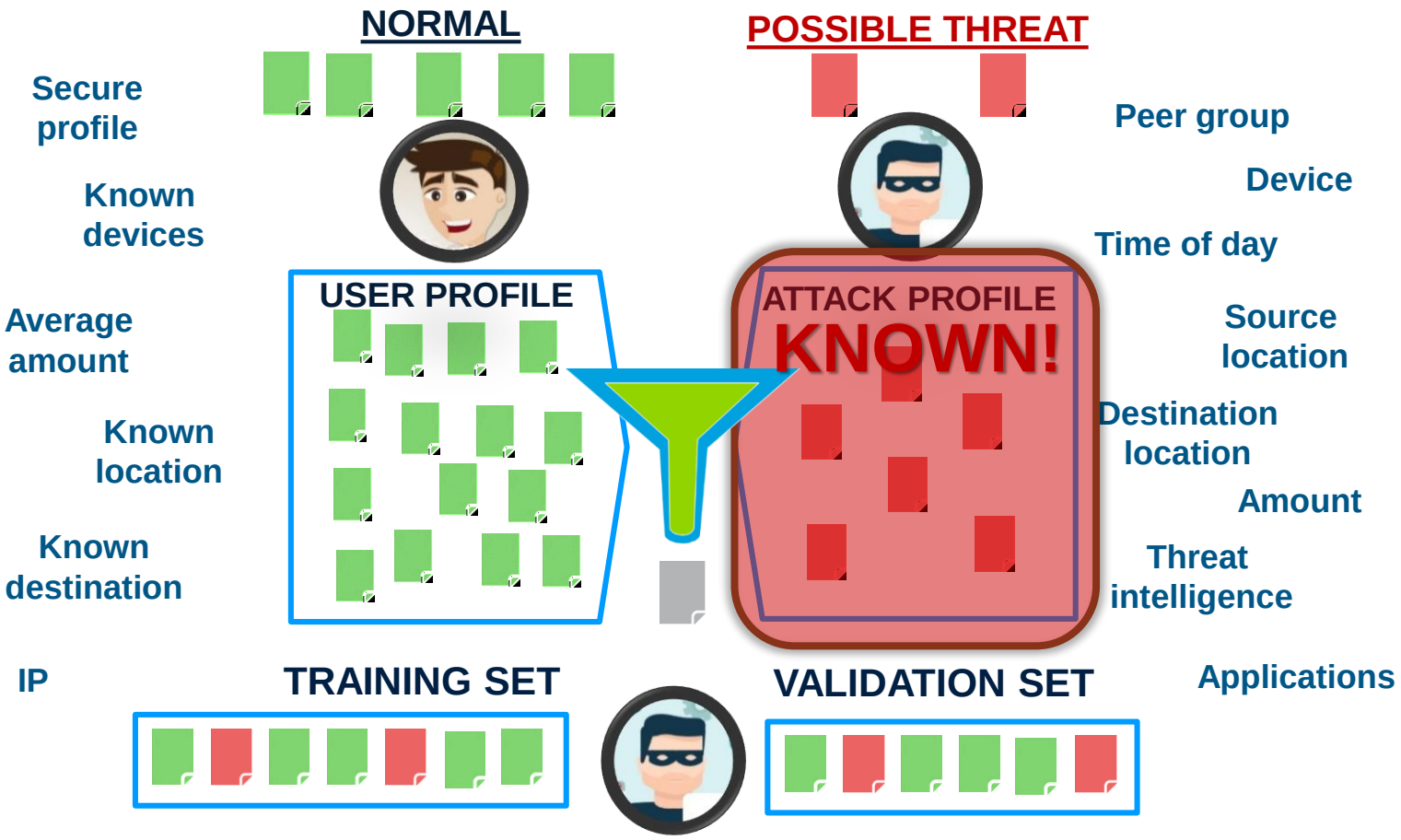
Unsupervised learning

- You have a dataset, but little idea concerning the patterns and categories
- *Example*: your have a large set of Net Flow data, but do not know patterns



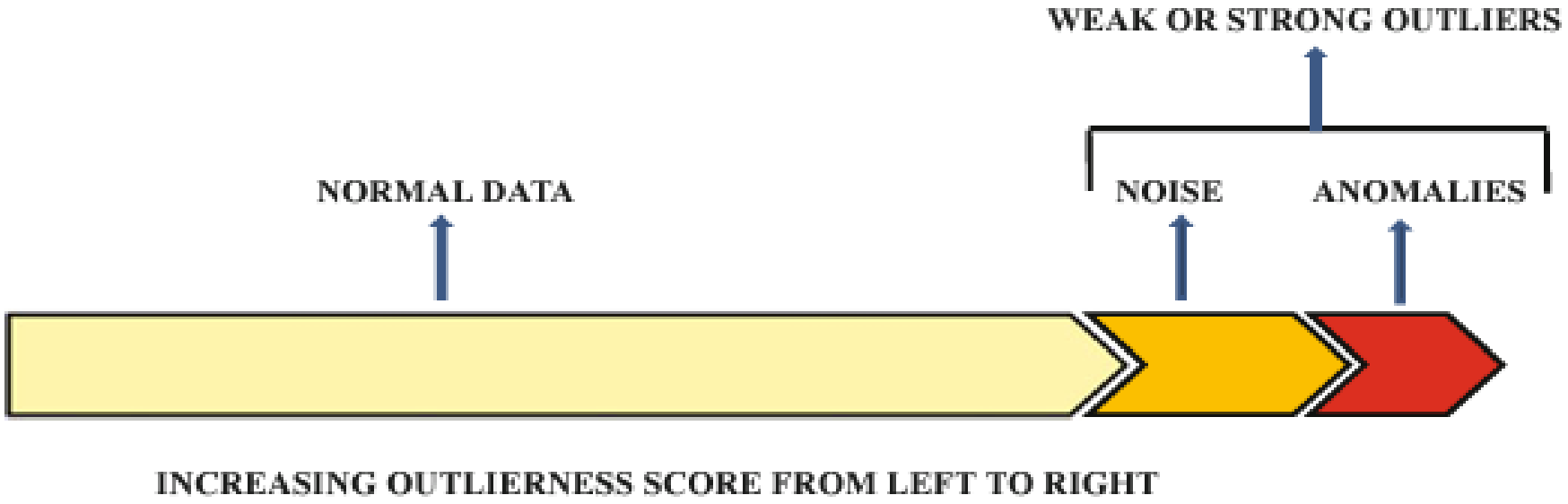
Cluster Analysis

Supervised: Machine Learning



Unsupervised: Pattern & Anomaly Detection

- **Understand normal: ‘normal is crazy enough!’**
- **Identify outliers on the basis of deviations**



Data Analytics Technologies

Data management

- *Relational databases*: Oracle, IBM DB2, MS SQL Server, data warehouses
- *NOSQL*: graph databases, document stores, key-value, column family
- *Mass storage*: Hadoop clusters, cloud approaches, SAP Hana
- *Data management*: SAS, many 3rd party products

Statistical analysis software

- *Math-focused: Matlab*, Mathematica
- *Programmatic / scripting*: Python, PERL, Java, Haskell
- *Packaged tools*: SAS, SAS JMP, SPSS, R, SAP Infinite Insight

Machine learning

- SAS Enterprise Miner
- SAP Predictive Analysis
- R, MatLab, Python, etc.
- Visualization / dashboards
- Tableau, QlikView, SAS Data Visualization

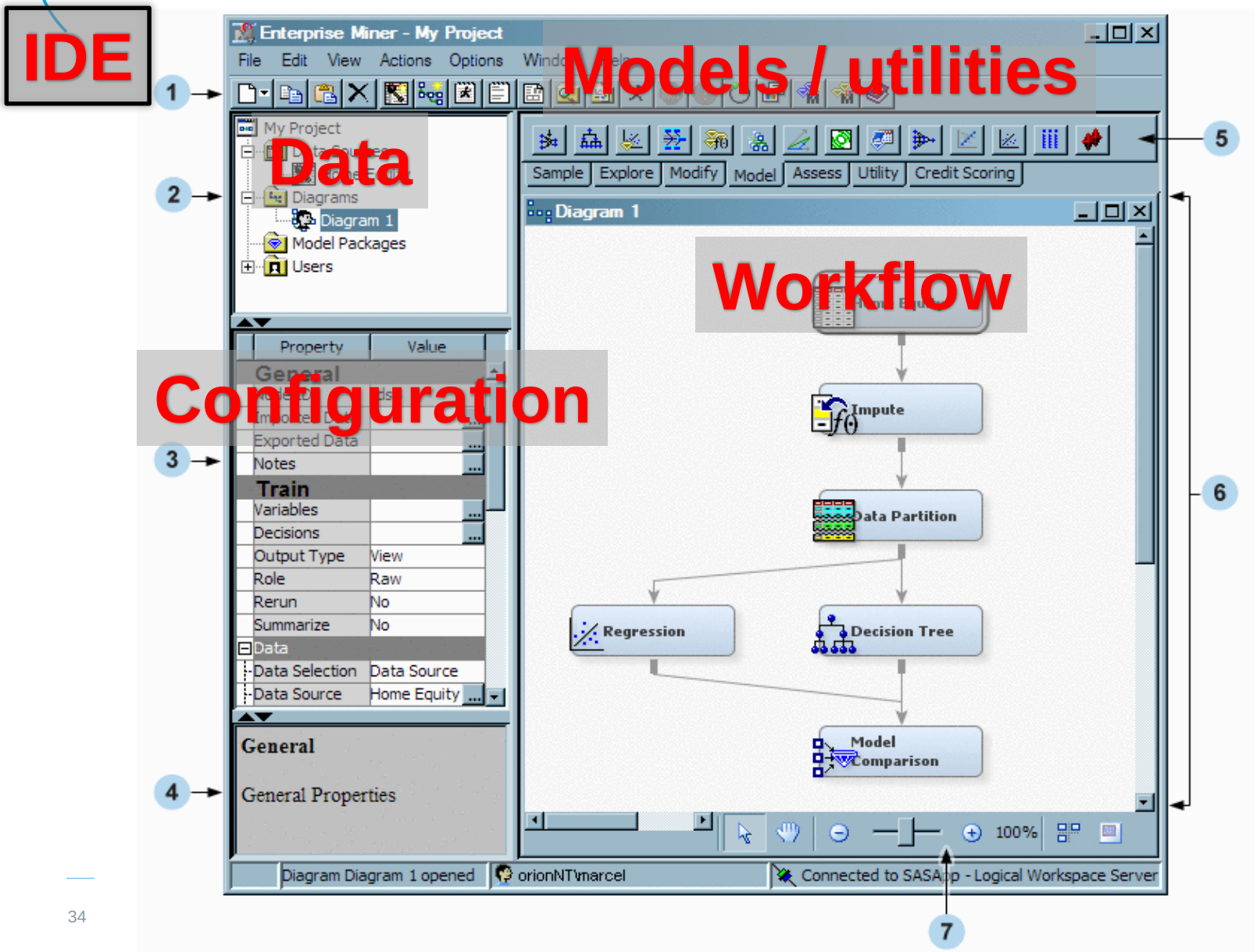
Semantic

- Text mining, sentiment analysis, (i.e. SAS Contextual Analysis)

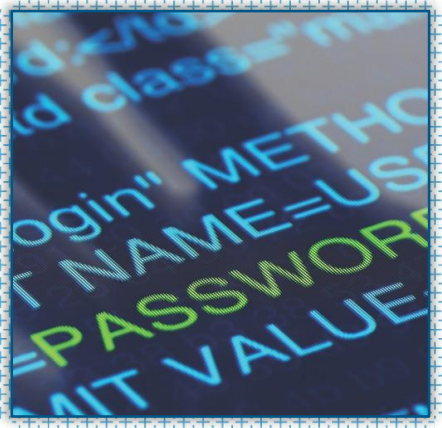
Big Data

- Hadoop, Map Reduce, Hive, Spark, etc. etc.

Enterprise Miner

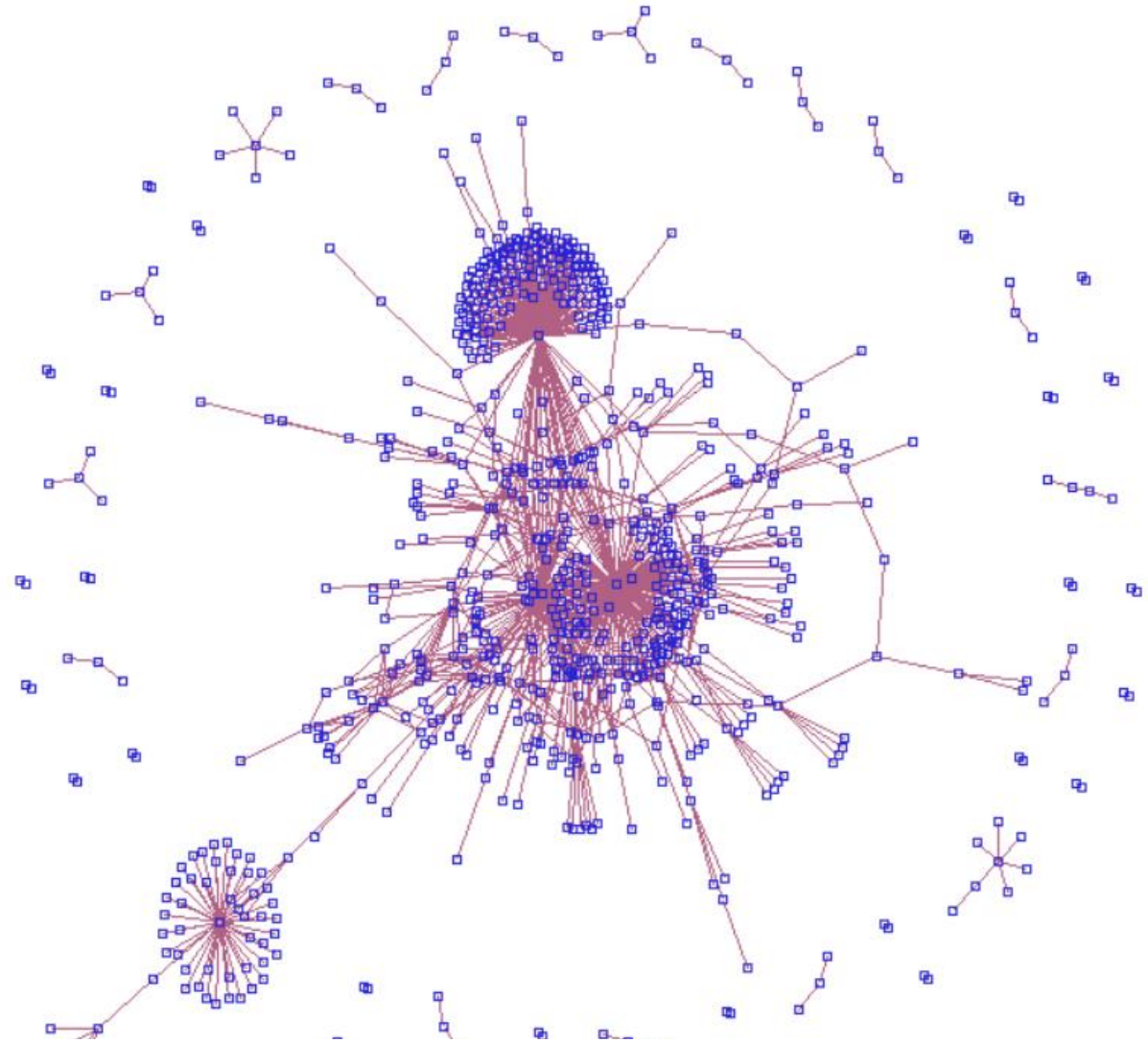


Learnings from the Field



GARP Global Association of Risk Professionals

Learnings for the Field: Network Discovery



Example Measures

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

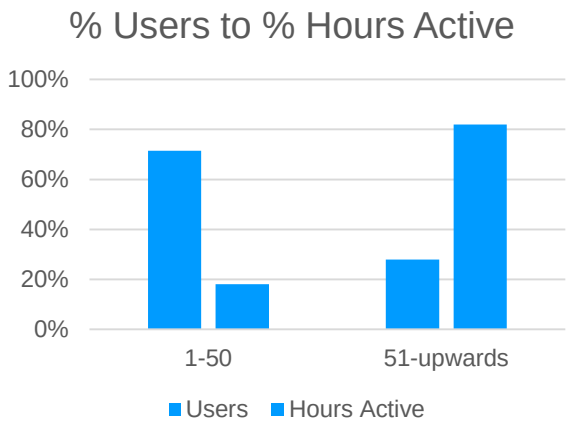
Learnings from the Field: Derive and Amplify

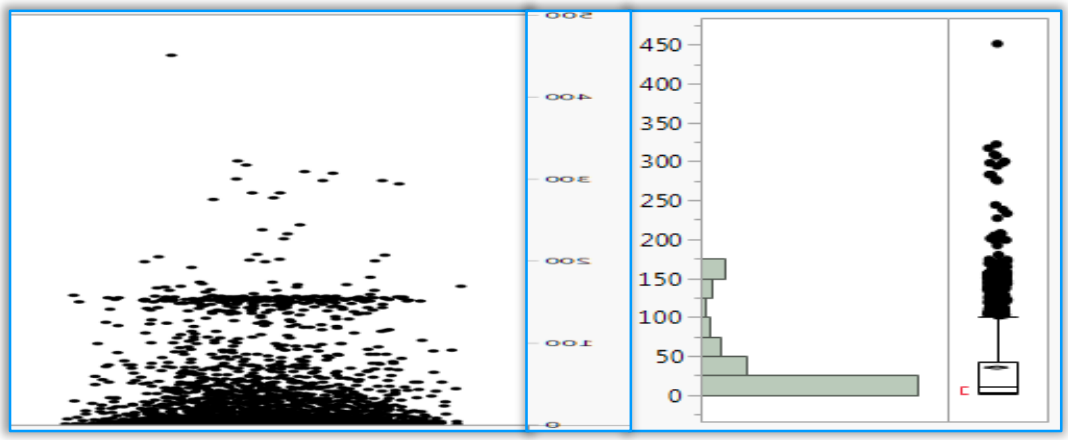
- Average total # hours online per period (userid or device)
- Average # IPs active per hour per userid
- Propensity to be active on network after work hours
- Propensity to be active on network on weekends and holidays

Learnings from the Field: User Patterns

Pareto Principle

- **80/20%** pattern in network-usage (user hours online)
  - *Outliers*: multiple devices 24 hours online
  - High correlation (80-90%) between hours online and propensity to align with multiple usage patterns…
- Pattern has been observed across multiple samples





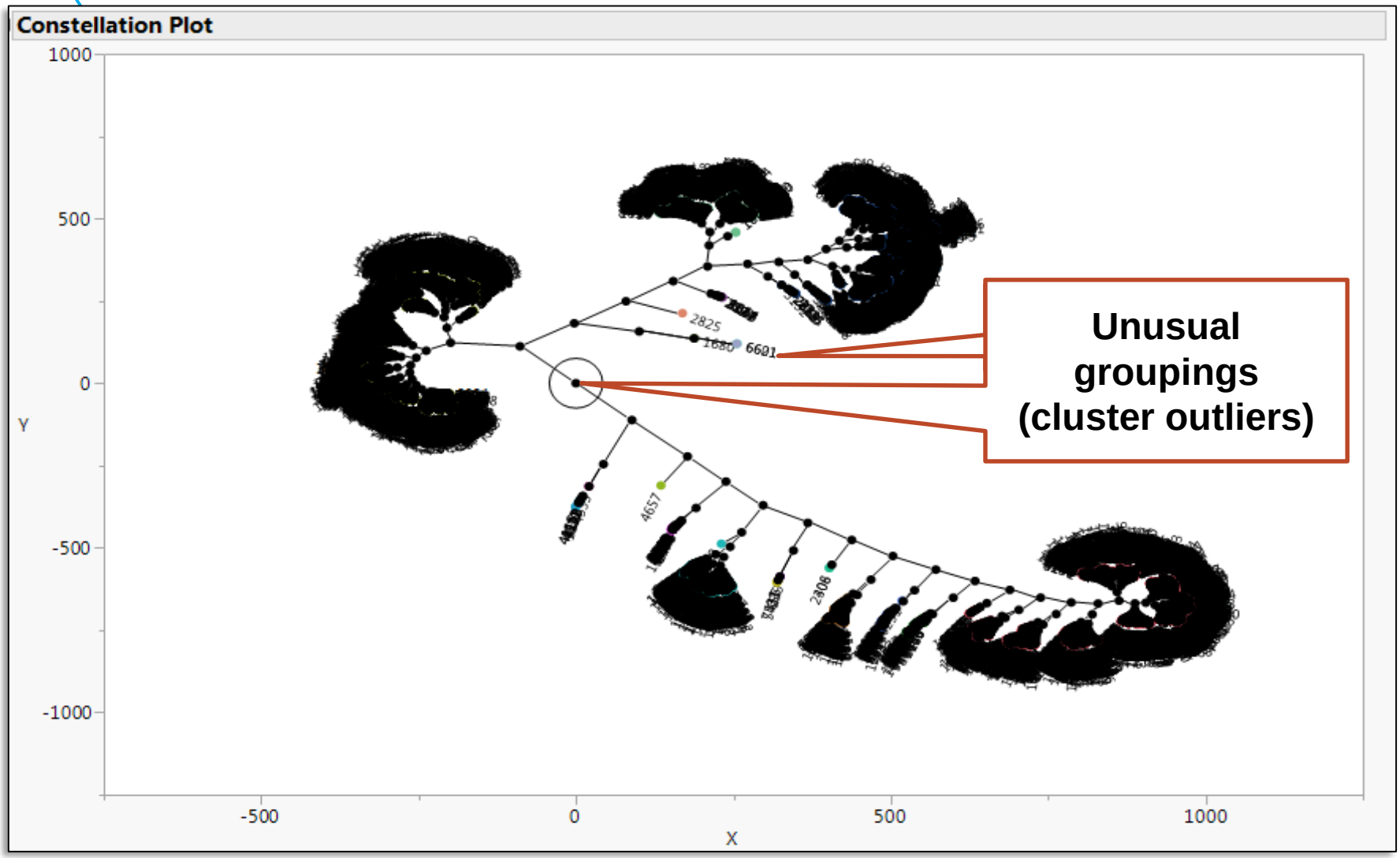
Learnings from the Field: Power of ratios…

Similar to the efficacy of financial ratios, ratios of key security measures may be more indicative of threats than single point measures.

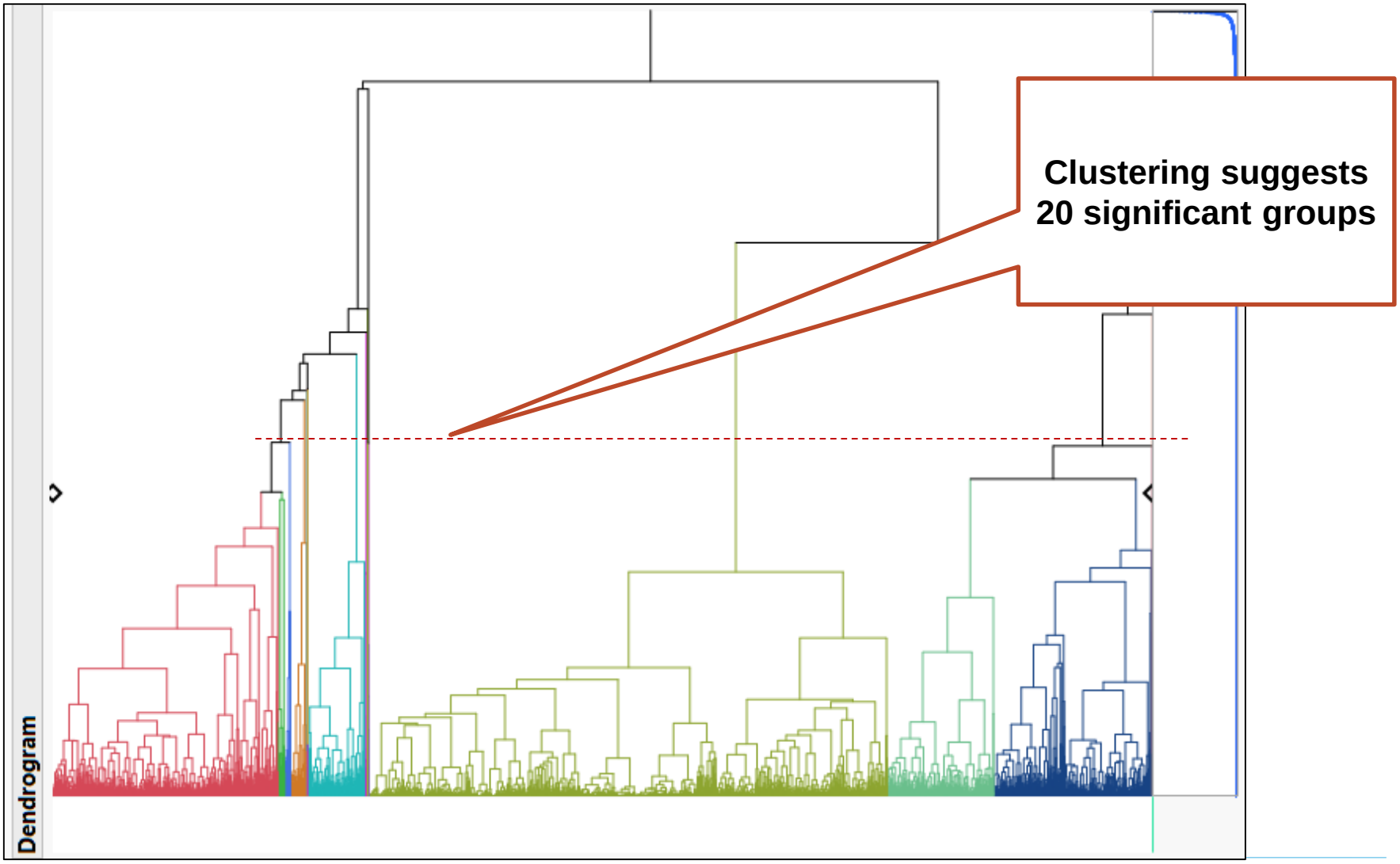
For instance:

- *Ratio of total 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

Learnings for the Field: Not All Users are Alike…

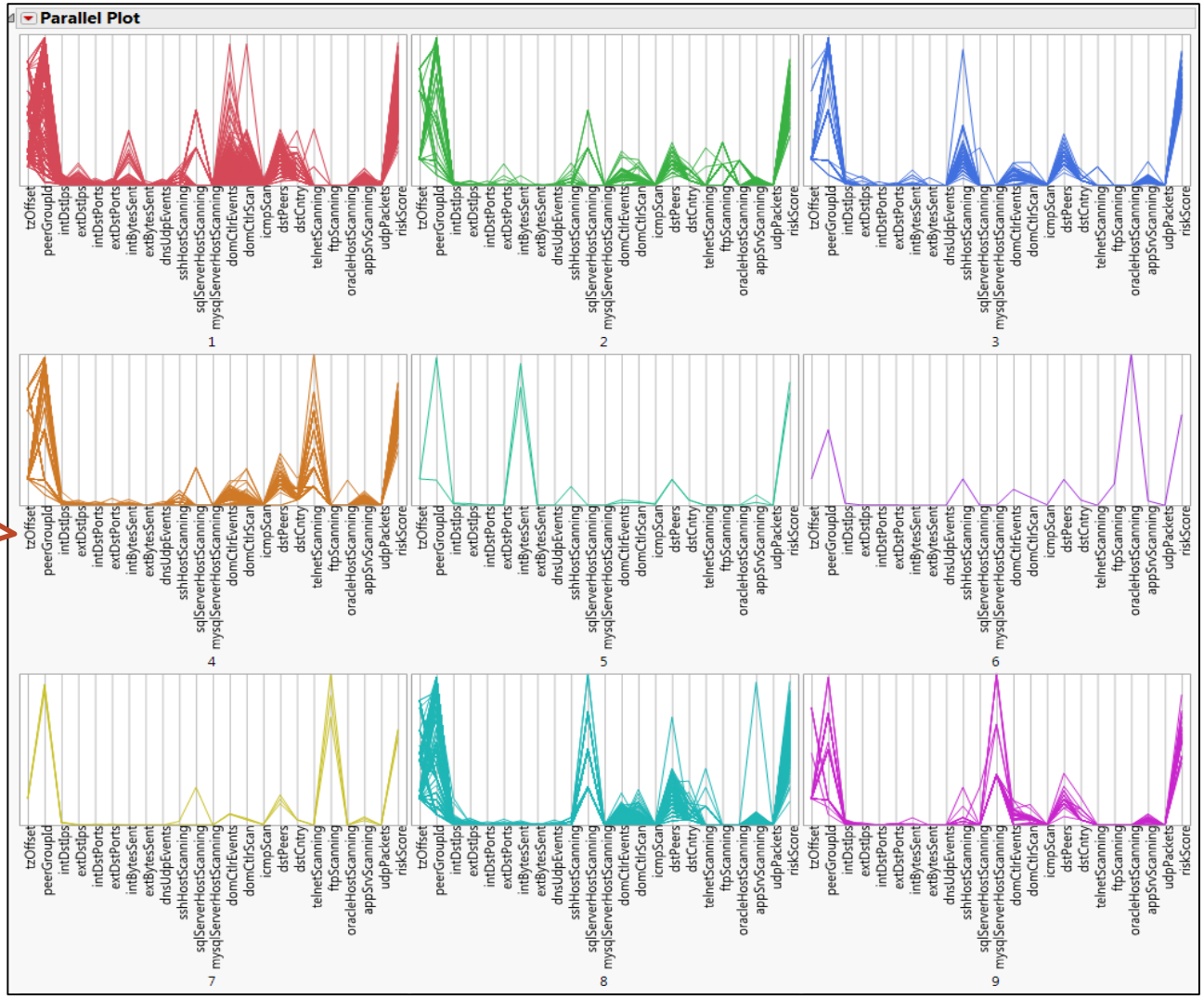


Learnings for the Field: Not All Users are Alike…

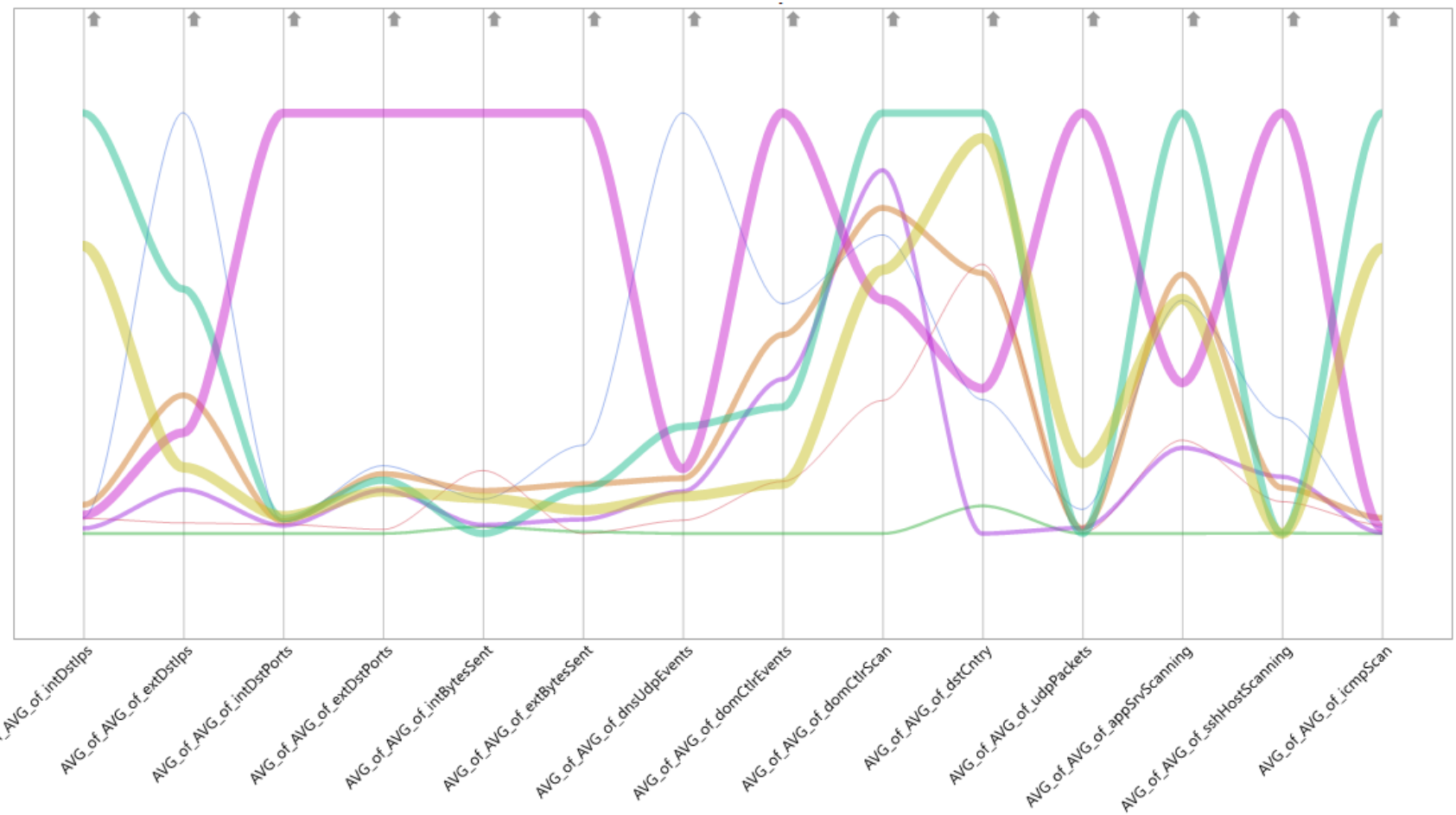


Patterns in Complexity: Cluster Analysis

Each cluster has a signature pattern of 22 measures (high and low)

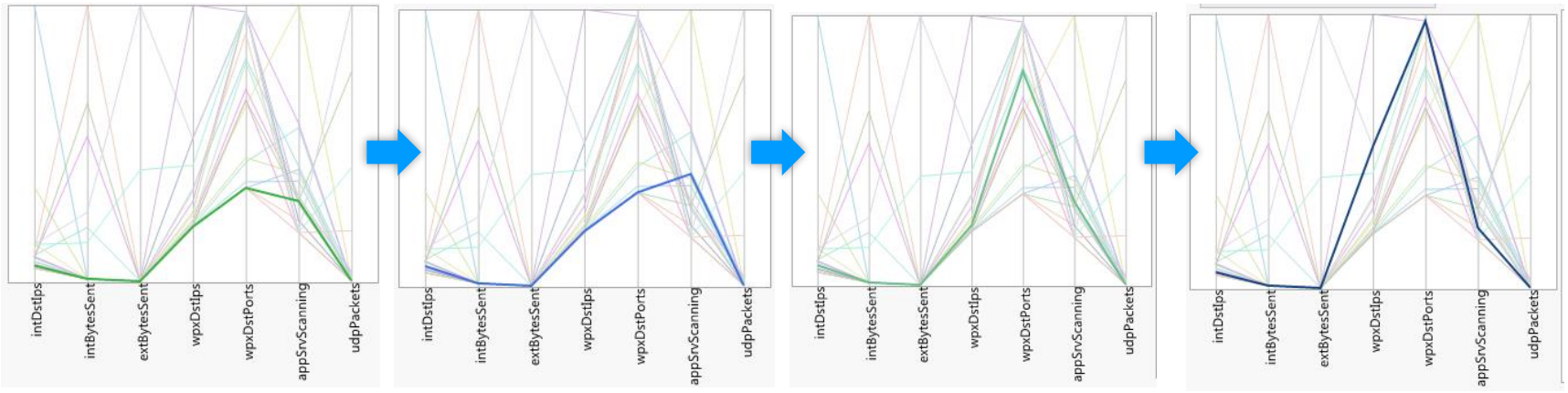


Patterns in Complexity: Cluster Analysis



Attack Pattern Identification Example

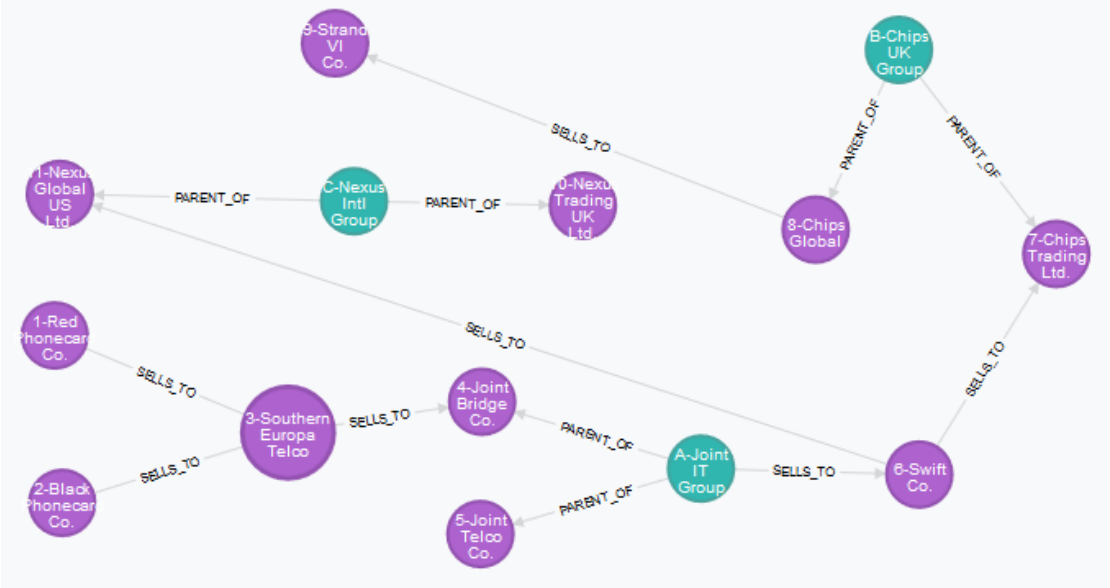
SIGNATURE PATTERN FOR IDENTIFIED INFECTED IP



Web Proxy Host Scanning Analysis	Devices on the network that are anomalously scanning for external devices via the Web Proxy server
Web Proxy Destination Port Scanning Analysis	Devices on the network that are anomalously scanning for external devices via the Web Proxy server
Application Server Host Scanning Analysis	Identify devices on the network that are anomalously scanning for devices hosting an http or application server

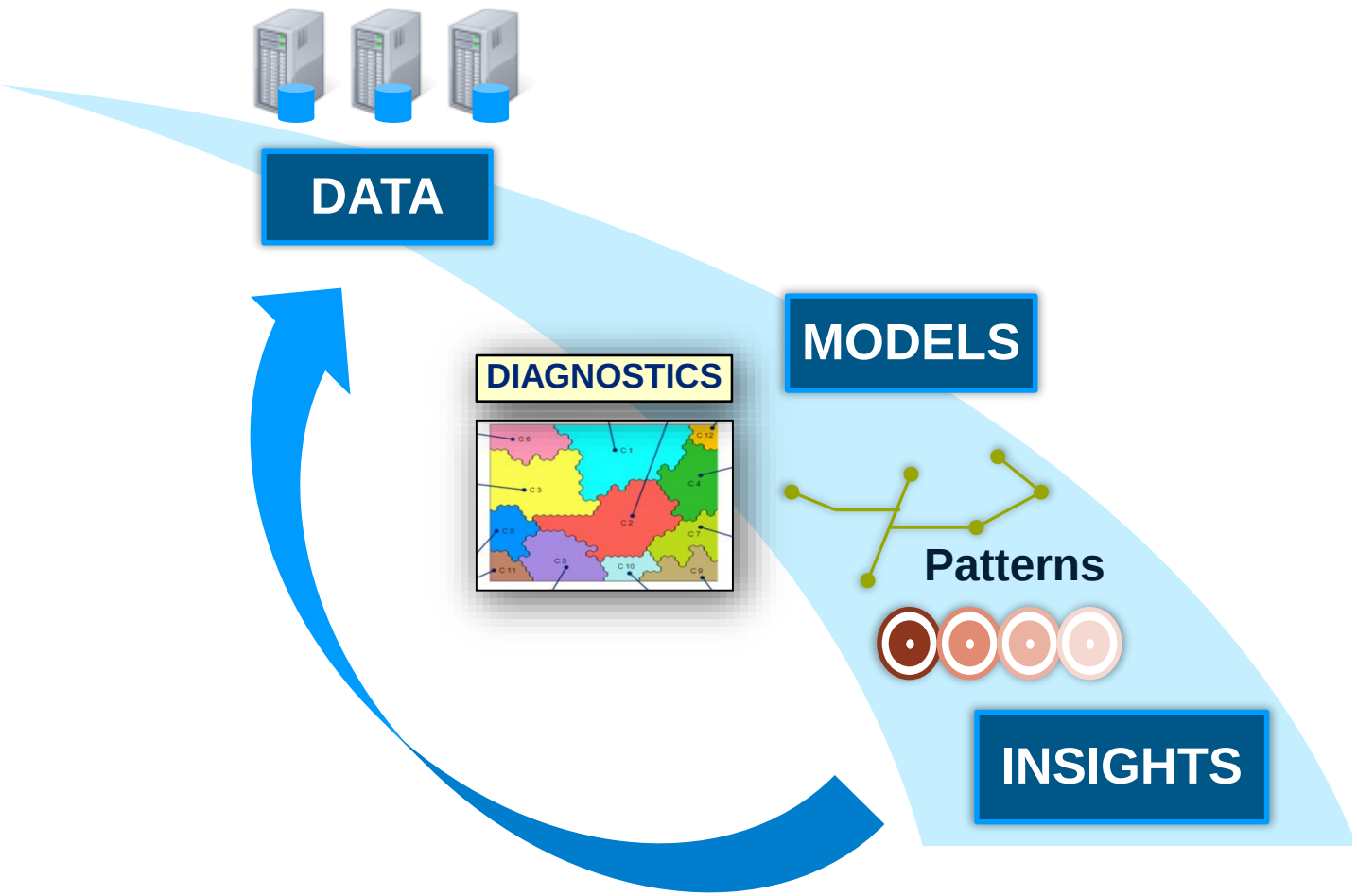
Graph data storage / network analytics for cyber attack vector pattern capture

- NOSQL graph database ‘network pattern’ storage & retrieval
- Building a cyber attack pattern ‘library’
- Identifying suspicious patterns in large & complex datasets

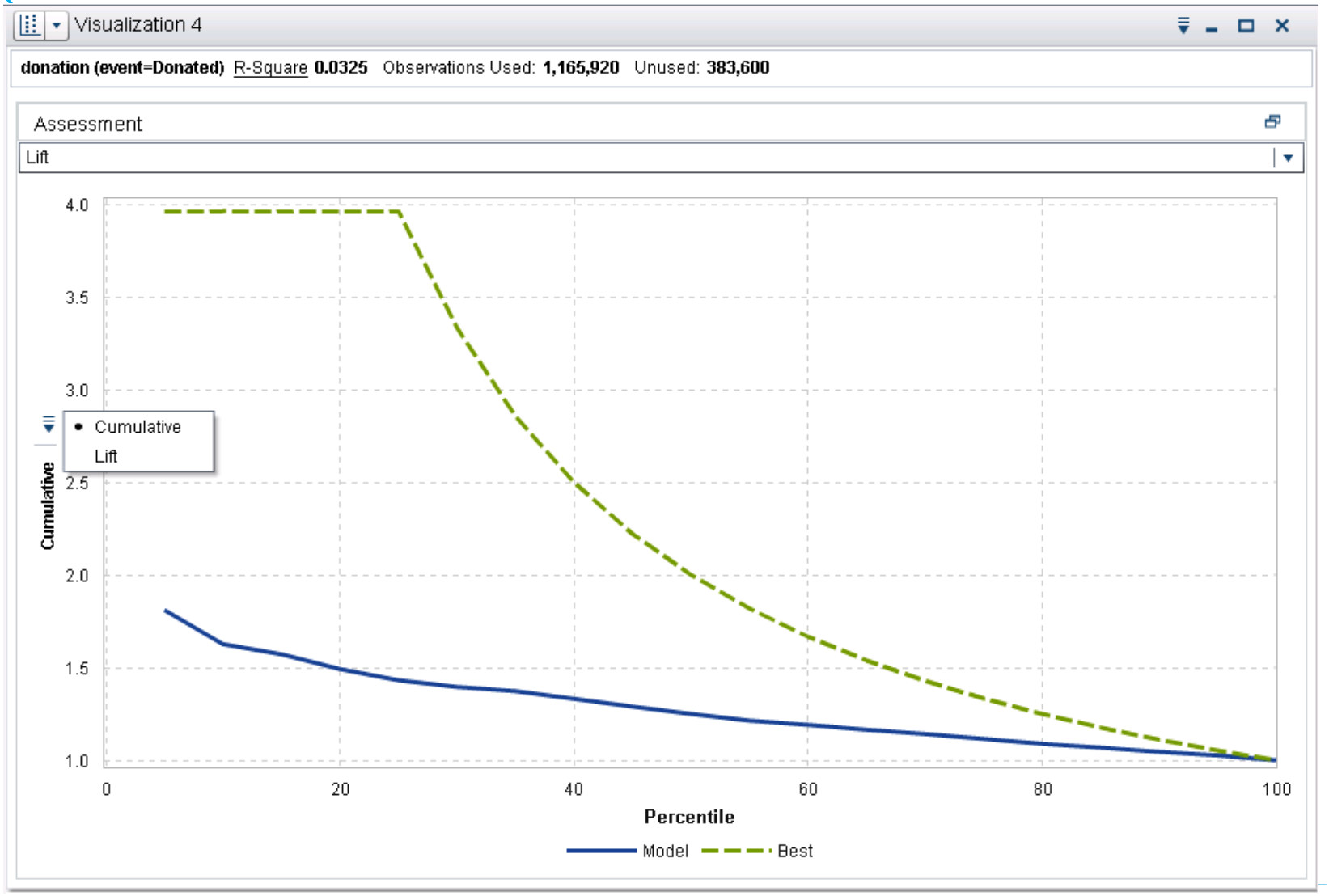


Conclusion: Managing Models

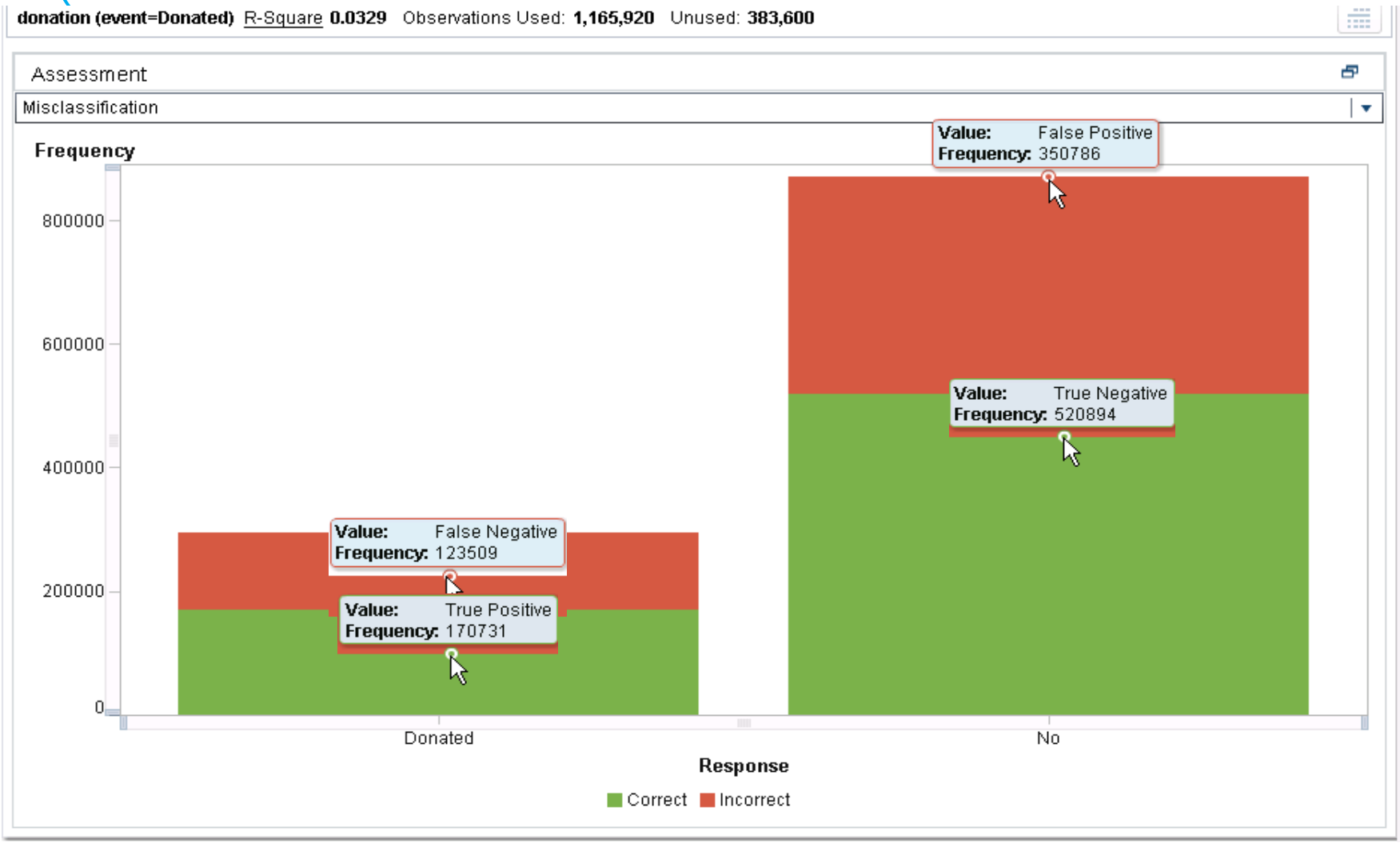
GARP | Global Association of Risk Professionals


DATA
DIAGNOSTICS
MODELS
Patterns
INSIGHTS

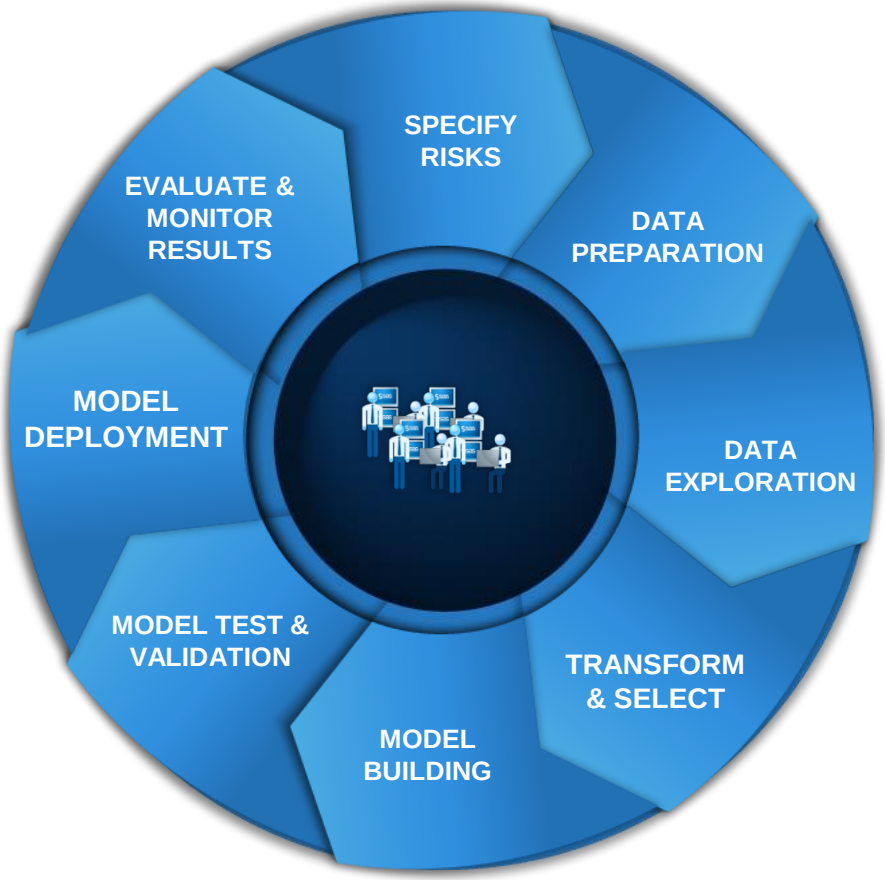
Model Diagnostics: Lift



Model Diagnostics: Misclassification Rate



Cyber Data Analytics Model Management

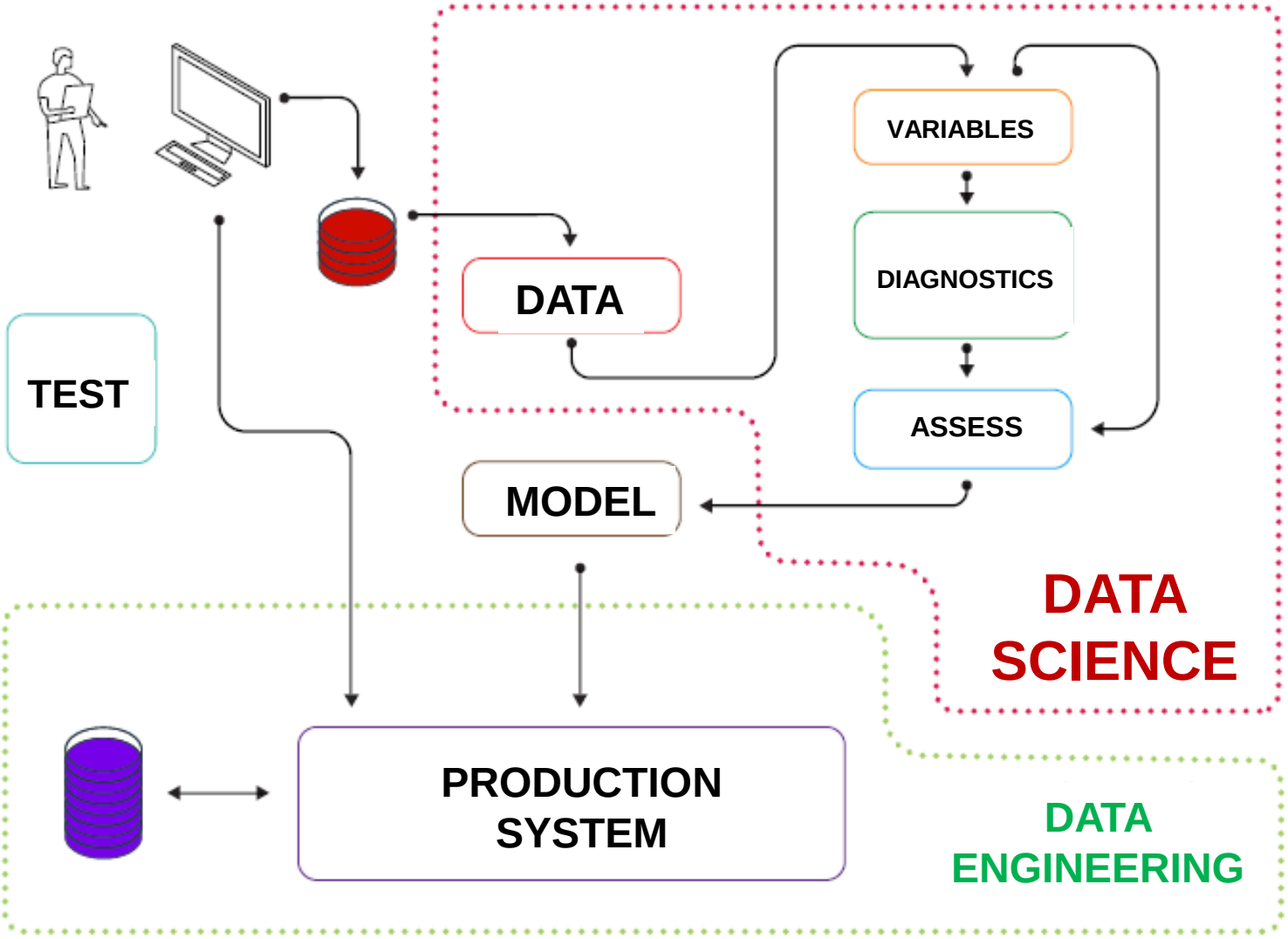


PATTERN IDENTIFICATION

DETECTION OPTIMIZATION

Objective | ETL | Develop | Validate | Deploy | Monitor | Retire

Production Analytics



Cyber Risk Model Management

- Models should be treated as enterprise assets
- Model management requires collaboration between personas
- Models need to be efficiently deployed into production
- Models need to be effectively deployed into production

*"**Model risk management** begins with robust model development, implementation, and use.*
Another essential element is a sound model validation process.
A third element is governance, which sets an effective framework with defined roles and responsibilities for clear communication of model limitations and assumptions, as well as the authority to restrict model usage".

Source: Supervisory Guidance on Model Risk Management, April 2011, The Federal Reserve System

Challenges:
Data Science in Commercial Organizations?

EXPERIMENTATION

- Most organizations have limited appetite for conducting experimentation / trial-and-error…
- But it is rare that a data scientist will get a model / framework right on the first try
- This is a new realm – it is essential to perform diagnostic tests and to adopt a mindset that allows for exploration of emerging phenomenon

“If I had six hours to chop down a tree, I’d spend the first four hours sharpening my axe.’

- Abraham Lincoln

Creating a culture of risk awareness®

Global Association of Risk Professionals

111 Town Square Place
14th Floor
Jersey City, New Jersey 07310
U.S.A.
+ 1 201.719.7210

2nd Floor
Bengal Wing
9A Devonshire Square
London, EC2M 4YN
U.K.
+ 44 (0) 20 7397 9630

www.garp.org

About GARP | *The Global Association of Risk Professionals (GARP) is a not-for-profit global membership organization dedicated to preparing professionals and organizations to make better informed risk decisions. Membership represents over 150,000 risk management practitioners and researchers from banks, investment management firms, government agencies, academic institutions, and corporations from more than 195 countries and territories. GARP administers the Financial Risk Manager (FRM®) and the Energy Risk Professional (ERP®) Exams; certifications recognized by risk professionals worldwide. GARP also helps advance the role of risk management via comprehensive professional education and training for professionals of all levels. www.garp.org.*

GARP Global Association of Risk Professionals