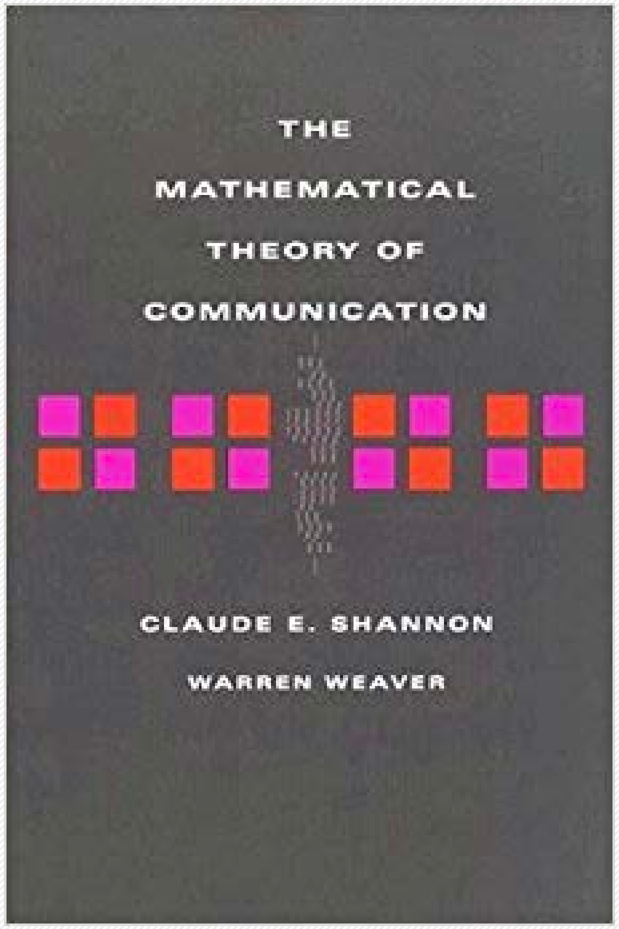
CYBER 503x
Cybersecurity Risk Management

Unit 6: Data Driven Security 1

R·I·T

Information Theory – developed in 1940s by Claude Shannon



THE MATHEMATICAL THEORY OF COMMUNICATION

CLAUDE E. SHANNON

WARREN WEAVER

- “Information” – as the amount of uncertainty reduction in a signal.

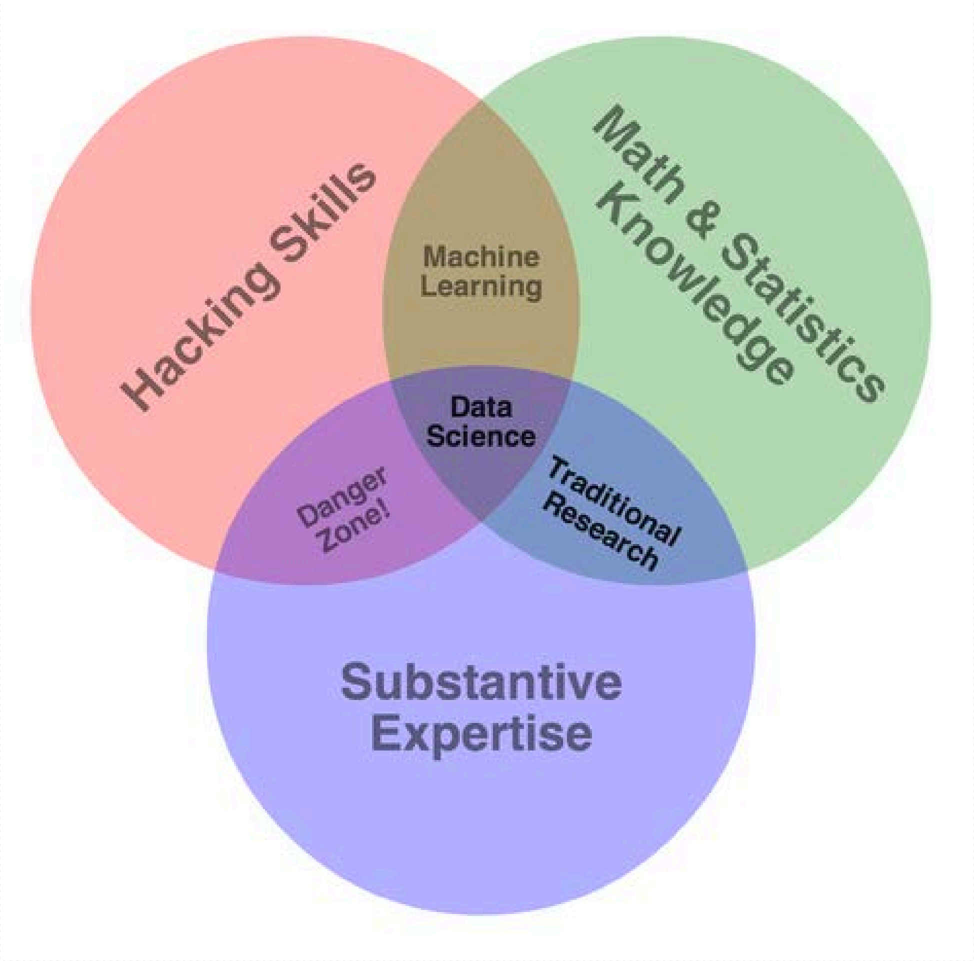
What is Data-Driven Security?

- Data: Digital Information based on empirical observations;
- Data-driven security is a discipline that
  - Leverages the pervasive security data
  - Applies statistical and machine learning analytics methods
  - Extracts actionable insights about cybersecurity risk models (e.g. threat, vulnerability, impact and cost-benefit factors) to drive decisions to assess, mitigate, and continuously evaluate risks.

What constitutes “Security Data?”

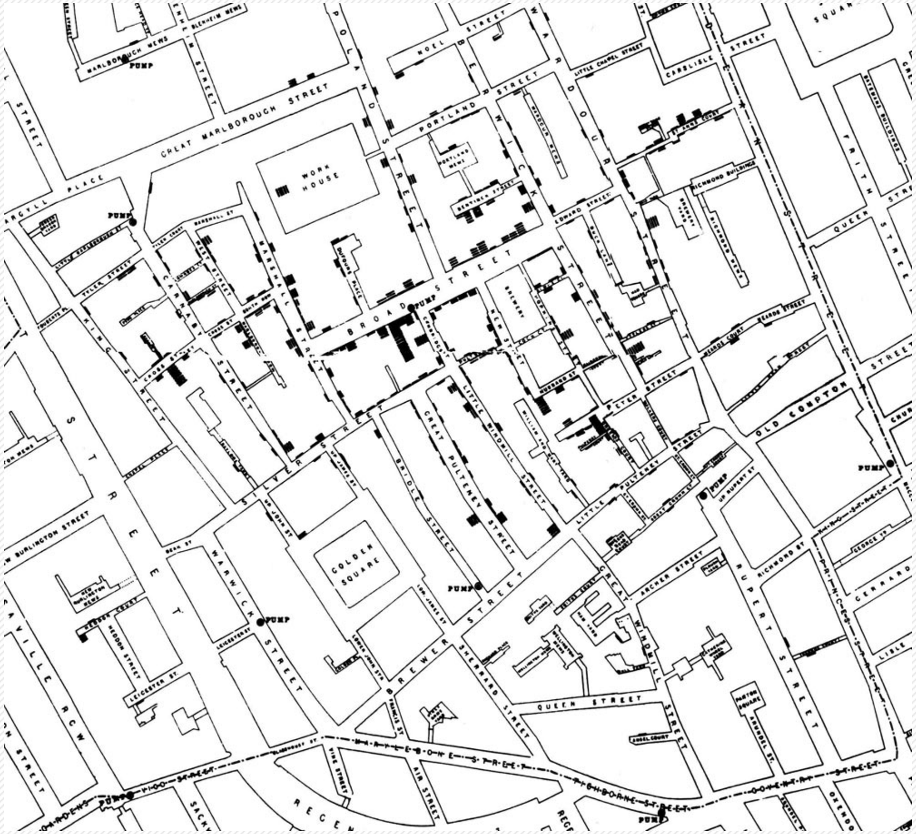
- Process, memory and system binary dumps.
- Patch release activity logs
- Network traffic data
- Web logs and click streams
- Email messages
- Video surveillance data
- Continuous multi-modal sensory data
- Real time transaction data

The Skill Sets

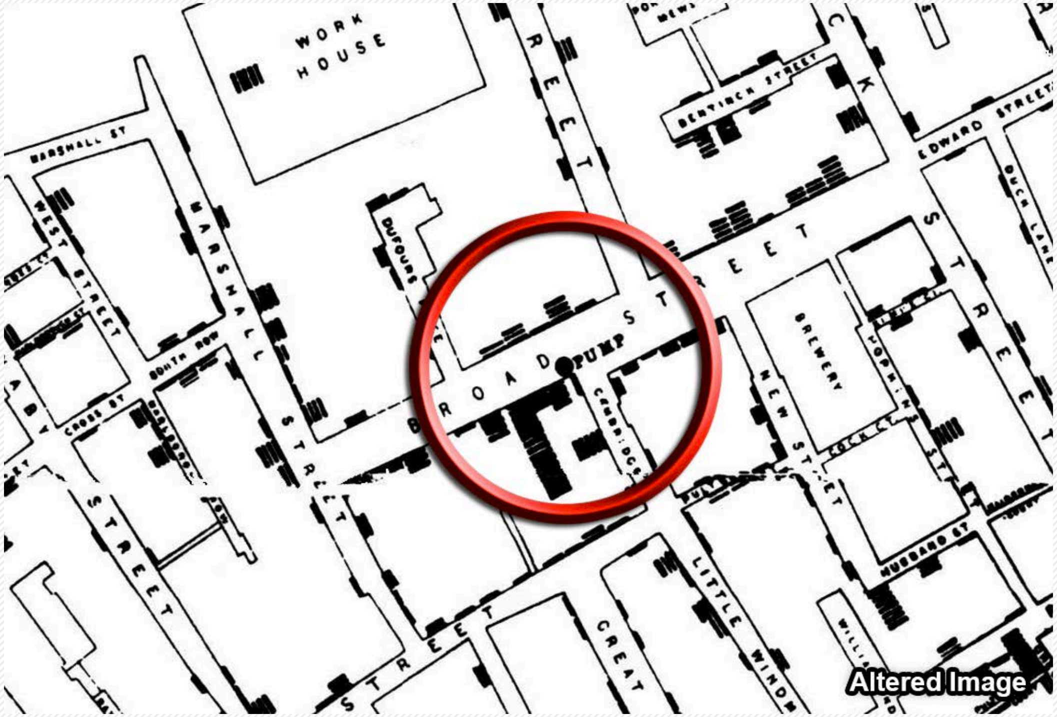


A Brief History of Learning from Data (1)

- 19th Century Data Analysis
  - Dr. John Snow’s map of the 1854 London cholera outbreak



London Cholera Map – John Snow



Altered Image

- He plotted every death on a map with ingenious mapped bar charts (see picture), and was able to show that the closer to the Broad Street water pump he plotted, the greater the number of death.
- This information helped convince the public a true sewage system was needed and spurred the city to action.

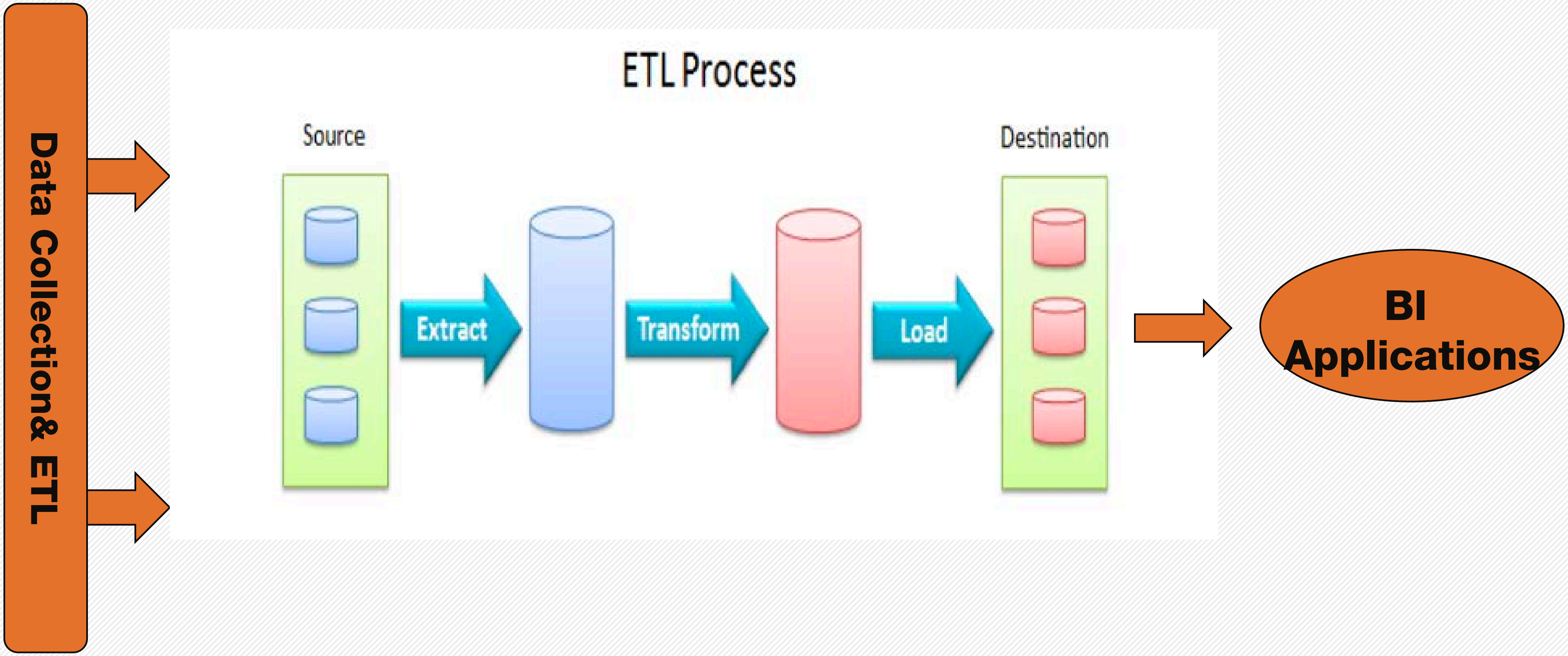
A Brief History of Learning from Data (2)

- 20th Century Data Analysis –
  - Ronal Fisher (a professor of genetics) and many of his revolutionary contributions to statistics in pioneering the statistical model in scientific inference.
  - No longer could scientists simply collect and present their data as evidence of their claim, they now had the tools to design robust experiments and techniques to model how the variables affected their experiments and observations.

A Brief History of Learning from Data (3)

- 21st Century Data Analysis
  - New visualization techniques – for describing and exploring the data
  - The rise of machine learning algorithms
    - From “defining a data model **of nature**” to “deriving an algorithmic model **from nature**”
    - The challenges of modern data: higher noise to signal ratio, large data sets
  - Deep learning & AI – using an artificial brain to protect against cyberattacks

Security Data Analytics Pipeline (1)
Data Collection & ETL



Data Collection& ETL

BI Applications

Security Data Analytics Pipeline (2) Data Exploration

Data Collection& ETL

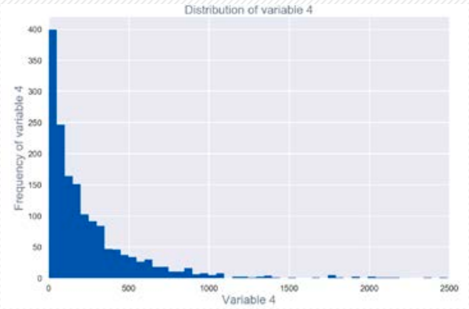
Data Exploration

Descriptive Analytics

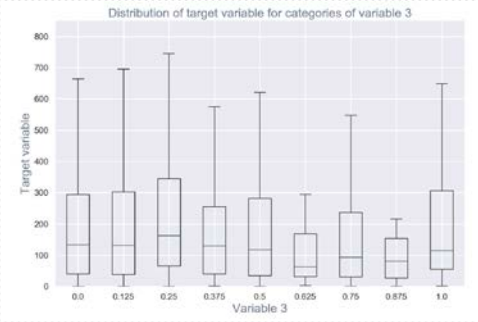
(understand data & assumptions, discover questions)

Exploration Methods:

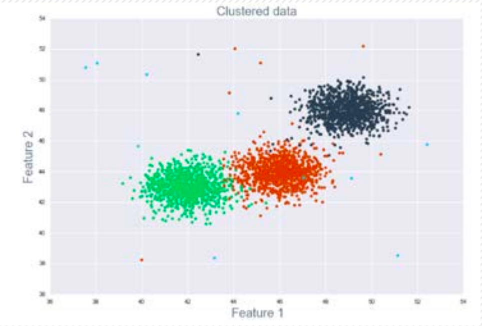
- Summary statistics & univariate visualization (e.g. histogram) & bivariate visualization (e.g. boxplot distribution graph)
- Multivariate visualizations to understand interactions between different attributes
- Dimensionality reduction
- Clustering



Univariate- histogram

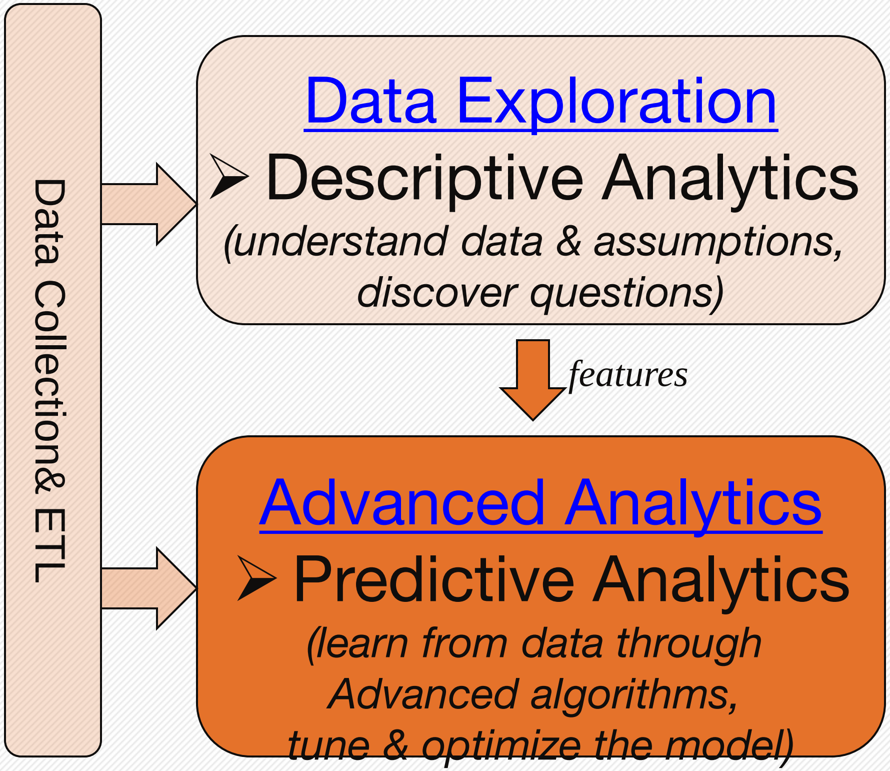


Bivariate boxplot



Data Clustering

Security Data Analytics Pipeline (3) Advanced Analytics



Data Collection& ETL

Data Exploration

➢ Descriptive Analytics

(understand data & assumptions, discover questions)

features

Advanced Analytics

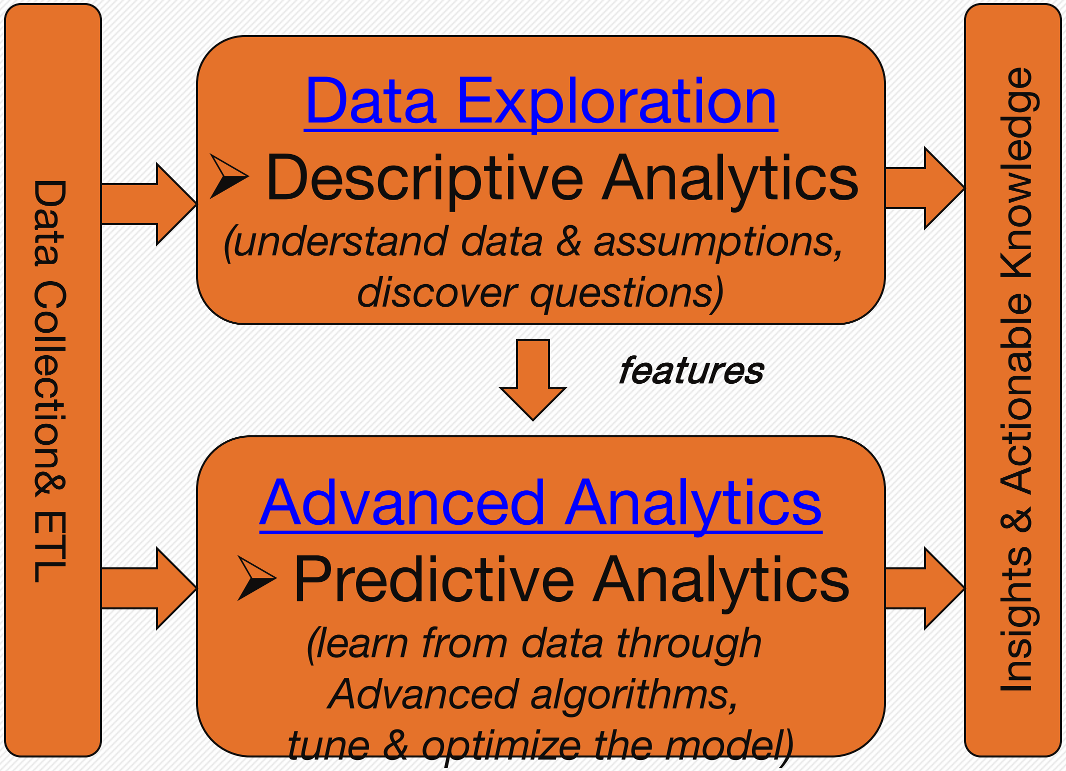
➢ Predictive Analytics

(learn from data through Advanced algorithms, tune & optimize the model)

Advanced predictive analytics for anomaly/fraud/spam detection:

- **Complex Statistical Modeling**
  - *Regression analysis*
  - *Principal Component*
  - *Bayesian modeling*
- **Machine Learning approaches**
  - *Decision tree & random forest*
  - *Support Vector Machine*
  - *Naïve Bayes*
  - *Neural Network*

Security Data Analytics Pipeline (4)
Dashboards, Reports, Decisions, Actions



Data Collection& ETL

[Data Exploration](#)

➢ Descriptive Analytics

(understand data & assumptions, discover questions)

features

[Advanced Analytics](#)

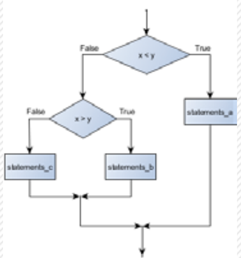
➢ Predictive Analytics

(learn from data through Advanced algorithms, tune & optimize the model)

Insights & Actionable Knowledge

Dashboards & Reports

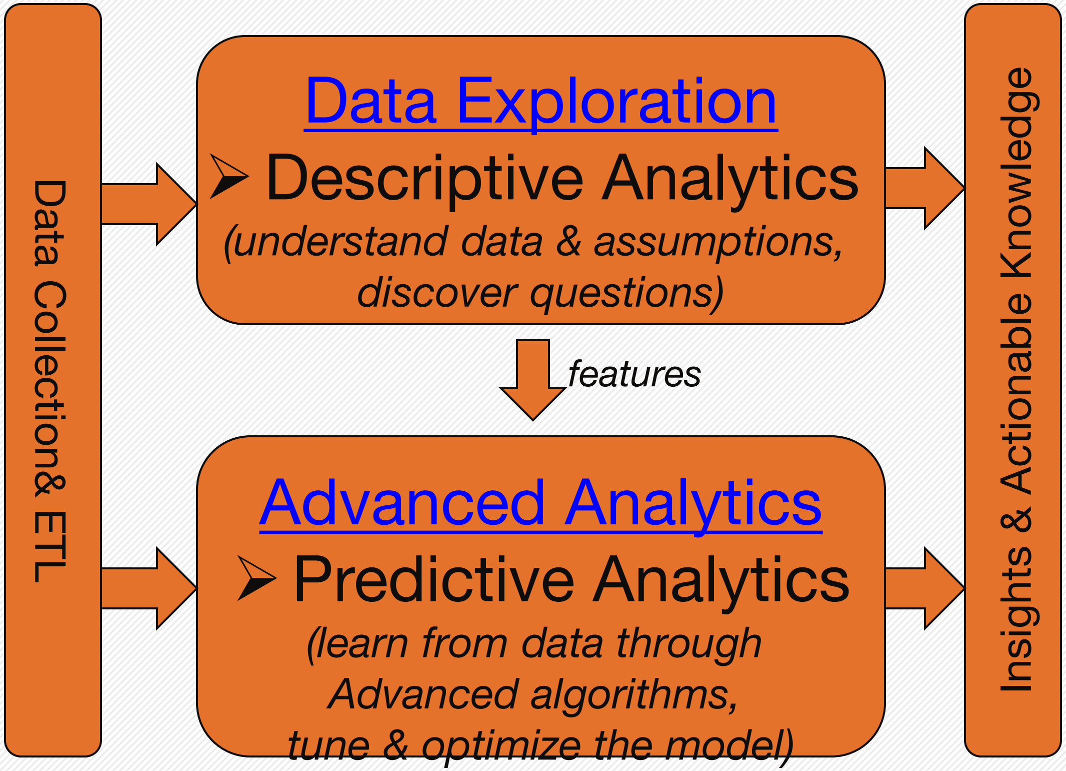
- ◆ Threat
- ◆ Vulnerability
- ◆ Impact
- ◆ Control effectiveness



Decisions & Actions

- ◆ Automated Detection (fraud, spam, anomaly)
- ◆ Threat Intelligence
- ◆ Intelligent defenses

Security Data Analytics Pipeline



Data Collection& ETL

[Data Exploration]

➢ Descriptive Analytics

(understand data & assumptions, discover questions)

features

[Advanced Analytics]

➢ Predictive Analytics

(learn from data through Advanced algorithms, tune & optimize the model)

Insights & Actionable Knowledge

Dashboards & Reports

- ◆ Threat
- ◆ Vulnerability
- ◆ Impact
- ◆ Control effectiveness

Decisions & Actions

- ◆ Automated Detection (fraud, spam, anomaly)
- ◆ Threat Intelligence
- ◆ Intelligent defenses

Common Descriptive Analytics for Security Data

- Grouping and aggregation through summary statistics, e.g. average, median, standard deviation
- Time series analysis
- Cross-sectional analysis
- Quartile analysis
- Correlation & association analysis

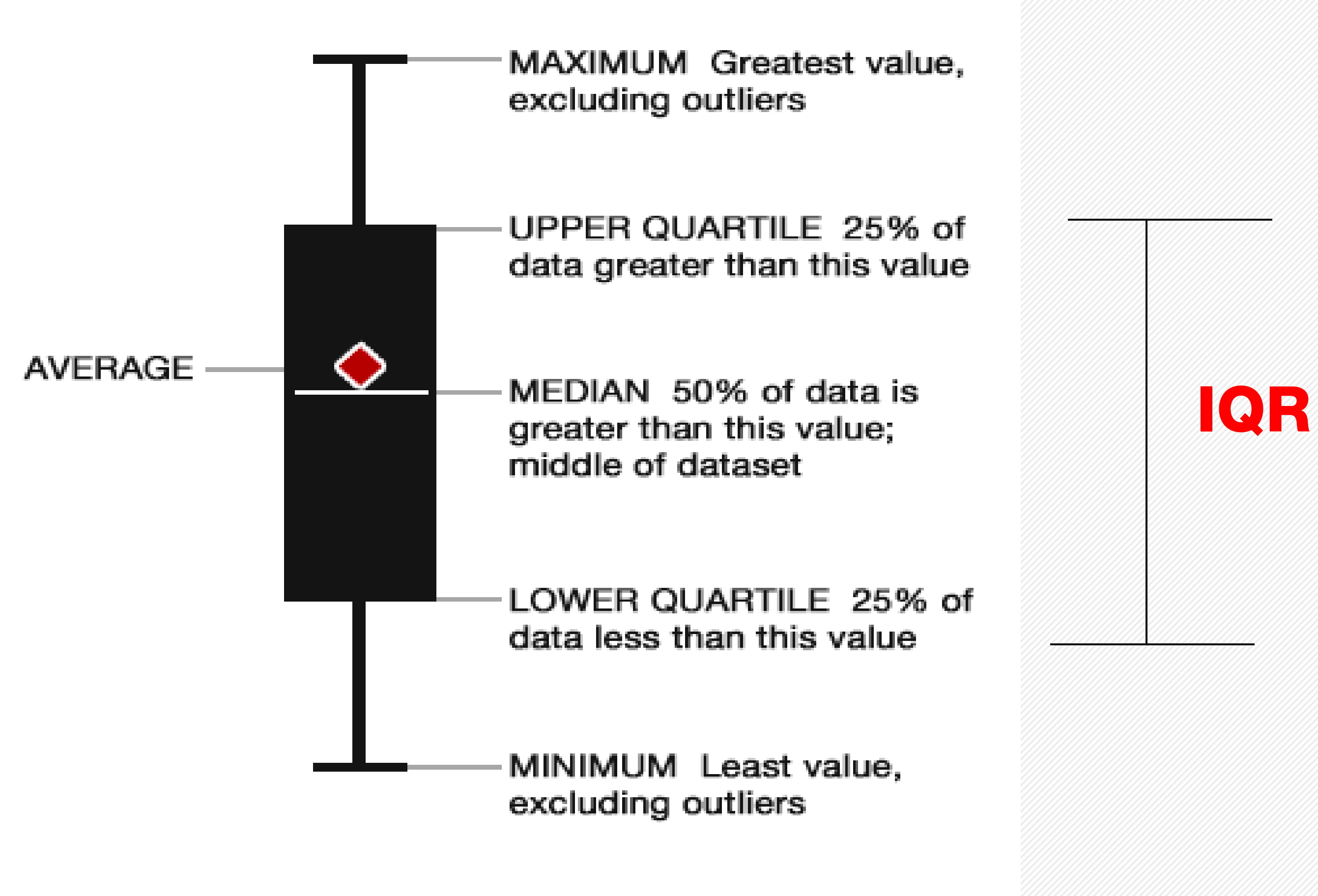
Summary Statistics

- Arithmetic Mean
  - It is the sum of a collection of numbers divided by the number of numbers in the collection.
  - Not a robust statistics, often influenced by outliers.
- Median
  - It is the number that separates the top 50% of elements from the bottom, when sorted in order.
  - More robust statistics when the data set has outlier(s).
- Example:
  - Data set: {9, 25, 3, 17, 6, 13, 2, 5, 19};
    first sorted: {2,3,5,6,9,13,17,19,25}
  - Arithmetic mean = 11
  - Median = 9

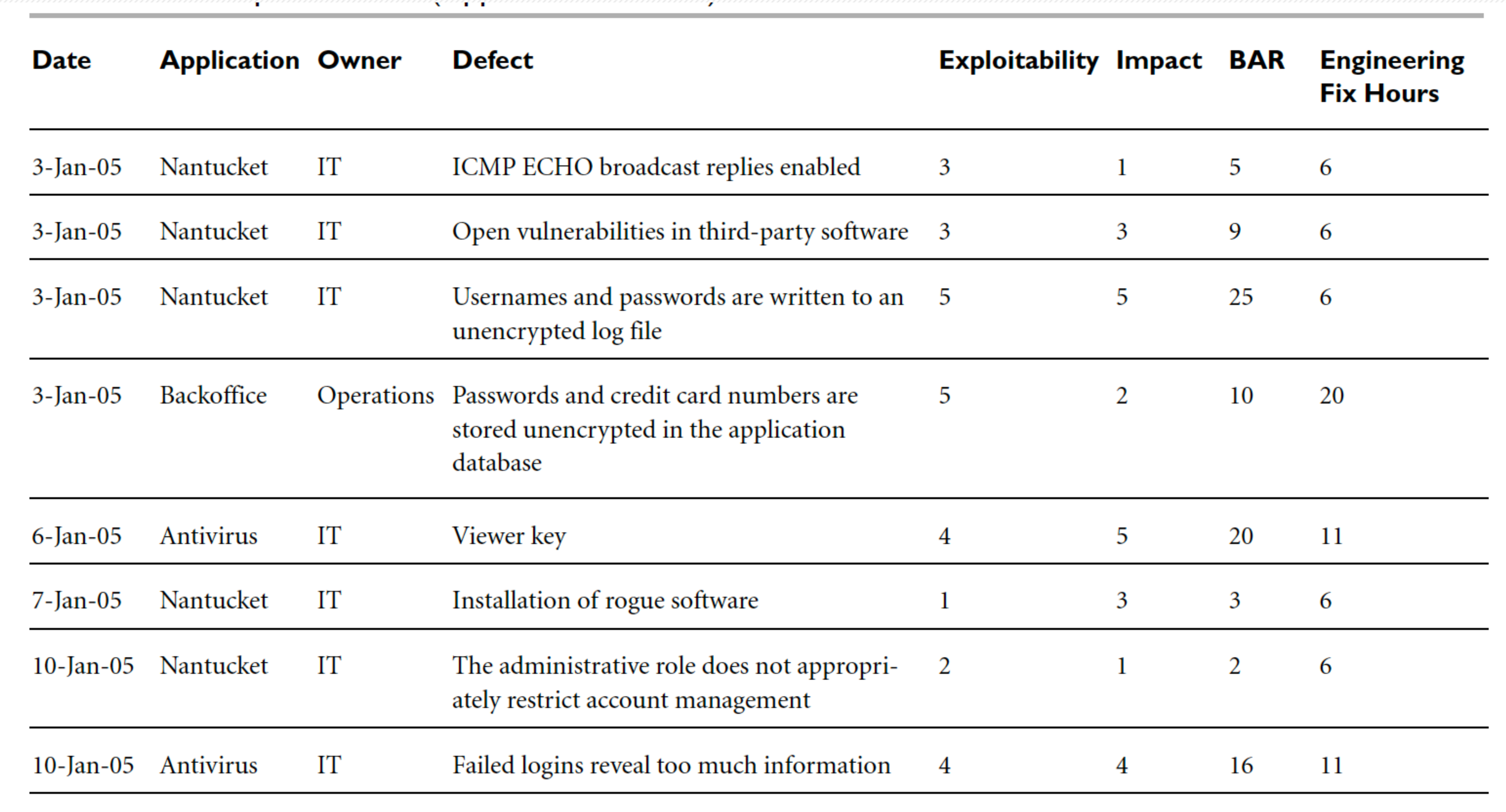
Variance & Standard Deviation

- Variance
  - The average of the **squared** differences from the mean.
- Standard Deviation (σ)
  - It is a measure of how spread out numbers are.
  - It is the square root of the variance.

Summary Statistics in Boxplot

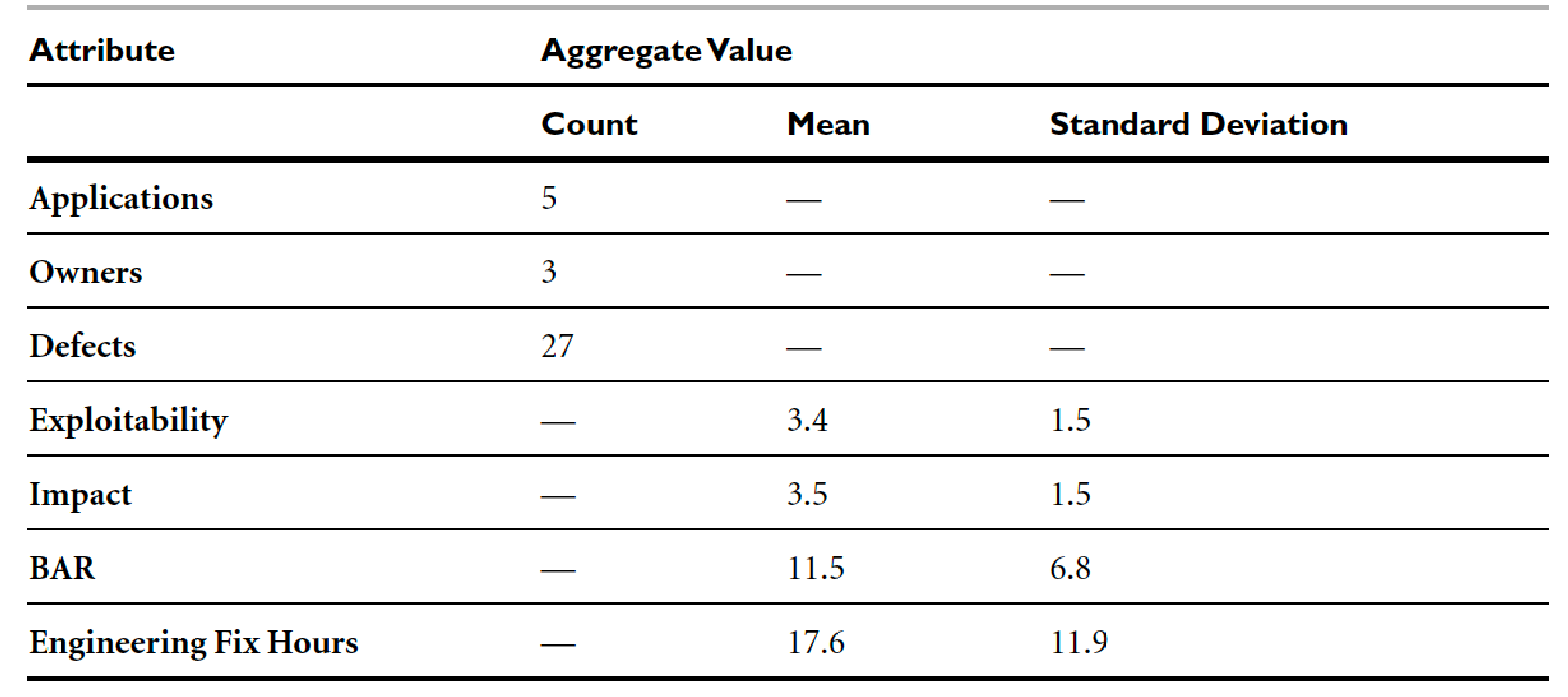

MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
AVERAGE
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
IQR

Example Security Data Set



Date	Application	Owner	Defect	Exploitability	Impact	BAR	Engineering Fix Hours
3-Jan-05	Nantucket	IT	ICMP ECHO broadcast replies enabled	3	1	5	6
3-Jan-05	Nantucket	IT	Open vulnerabilities in third-party software	3	3	9	6
3-Jan-05	Nantucket	IT	Usernames and passwords are written to an unencrypted log file	5	5	25	6
3-Jan-05	Backoffice	Operations	Passwords and credit card numbers are stored unencrypted in the application database	5	2	10	20
6-Jan-05	Antivirus	IT	Viewer key	4	5	20	11
7-Jan-05	Nantucket	IT	Installation of rogue software	1	3	3	6
10-Jan-05	Nantucket	IT	The administrative role does not appropriately restrict account management	2	1	2	6
10-Jan-05	Antivirus	IT	Failed logins reveal too much information	4	4	16	11

Grouping/Aggregation Example



Attribute	Aggregate Value		
	Count	Mean	Standard Deviation
Applications	5	—	—
Owners	3	—	—
Defects	27	—	—
Exploitability	—	3.4	1.5
Impact	—	3.5	1.5
BAR	—	11.5	6.8
Engineering Fix Hours	—	17.6	11.9

Time Series Analysis

- A time series contains:
  - A series of observation for a particular attribute
  - Measured at regular intervals
- Analysis interval:
  - It should be sufficiently precise that it lends insight, but not so detail that overwhelms the reader.
  - A typical analysis interval for most security metrics worth measuring is monthly or quarterly intervals.

Time Series Analysis

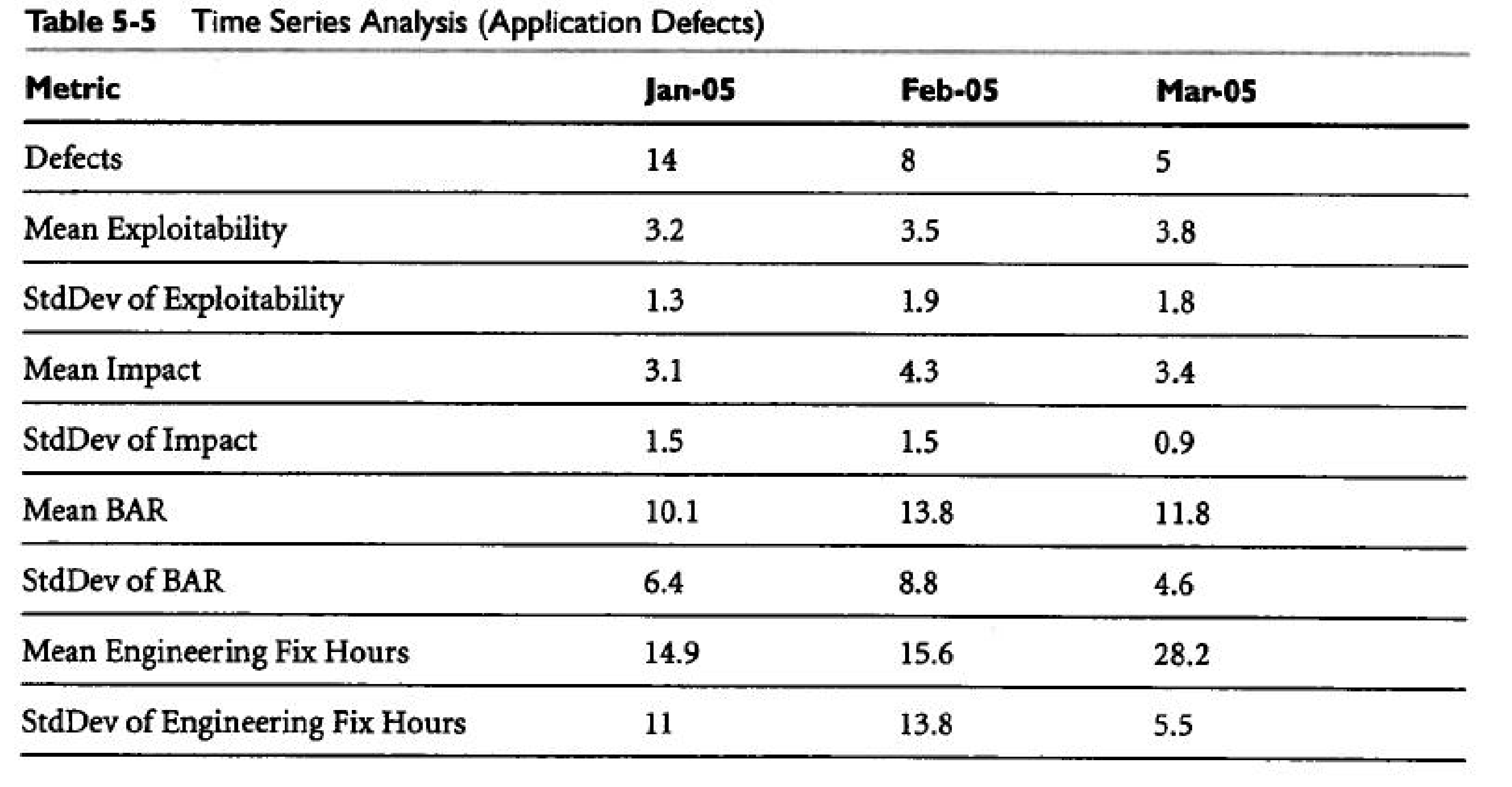


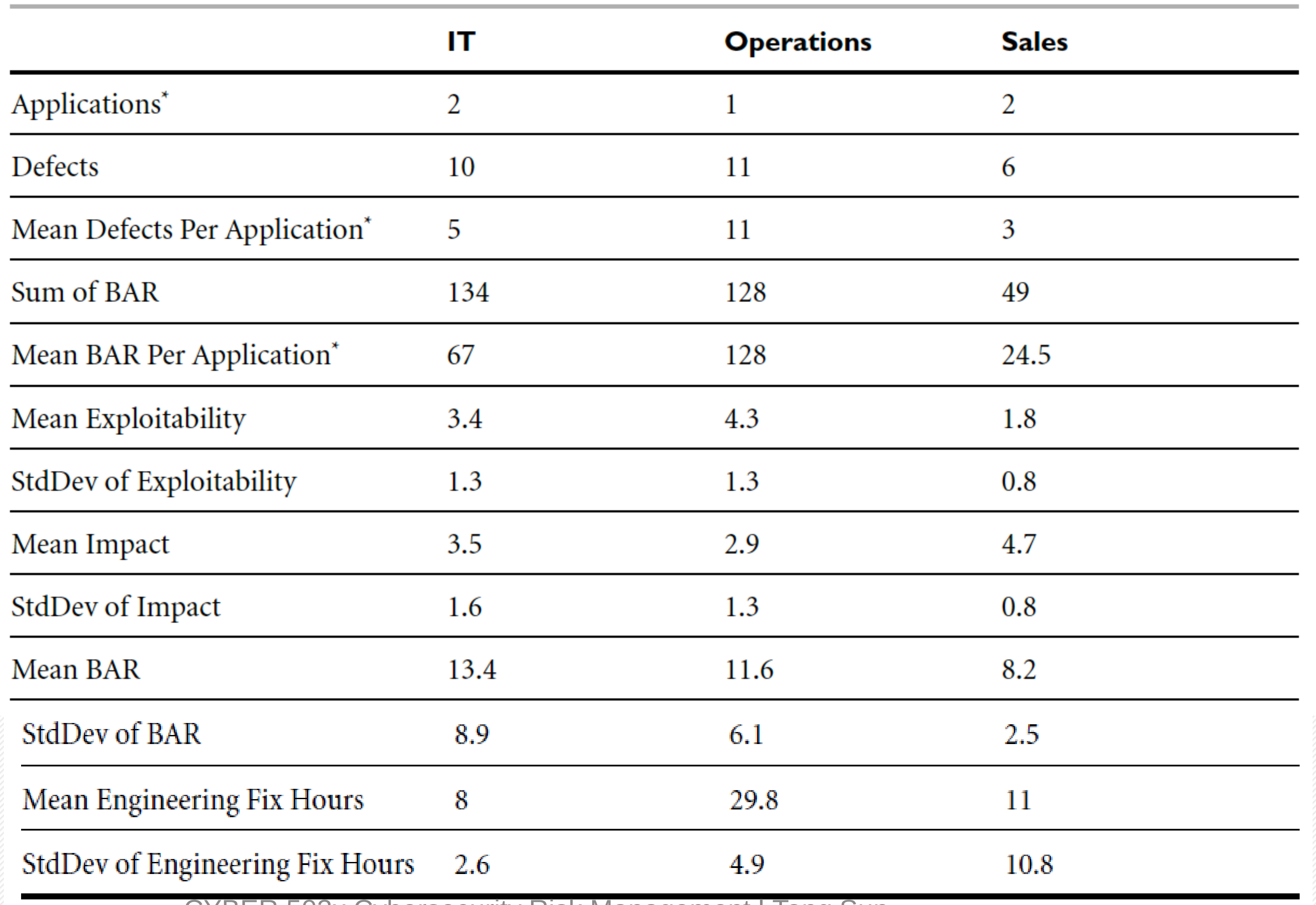
Table 5-5 Time Series Analysis (Application Defects)

Metric	Jan-05	Feb-05	Mar-05
Defects	14	8	5
Mean Exploitability	3.2	3.5	3.8
StdDev of Exploitability	1.3	1.9	1.8
Mean Impact	3.1	4.3	3.4
StdDev of Impact	1.5	1.5	0.9
Mean BAR	10.1	13.8	11.8
StdDev of BAR	6.4	8.8	4.6
Mean Engineering Fix Hours	14.9	15.6	28.2
StdDev of Engineering Fix Hours	11	13.8	5.5

Cross-Sectional Analysis

- Step 1: the analyst selects an attribute to use for creating the cross section – that is, an attribute to slice with.
  - Typically, textual attributes such as department, industry or categories.
- Step 2: The analyst groups and aggregates the data
- Step 3: analyze the results.

Cross-Sectional Analysis

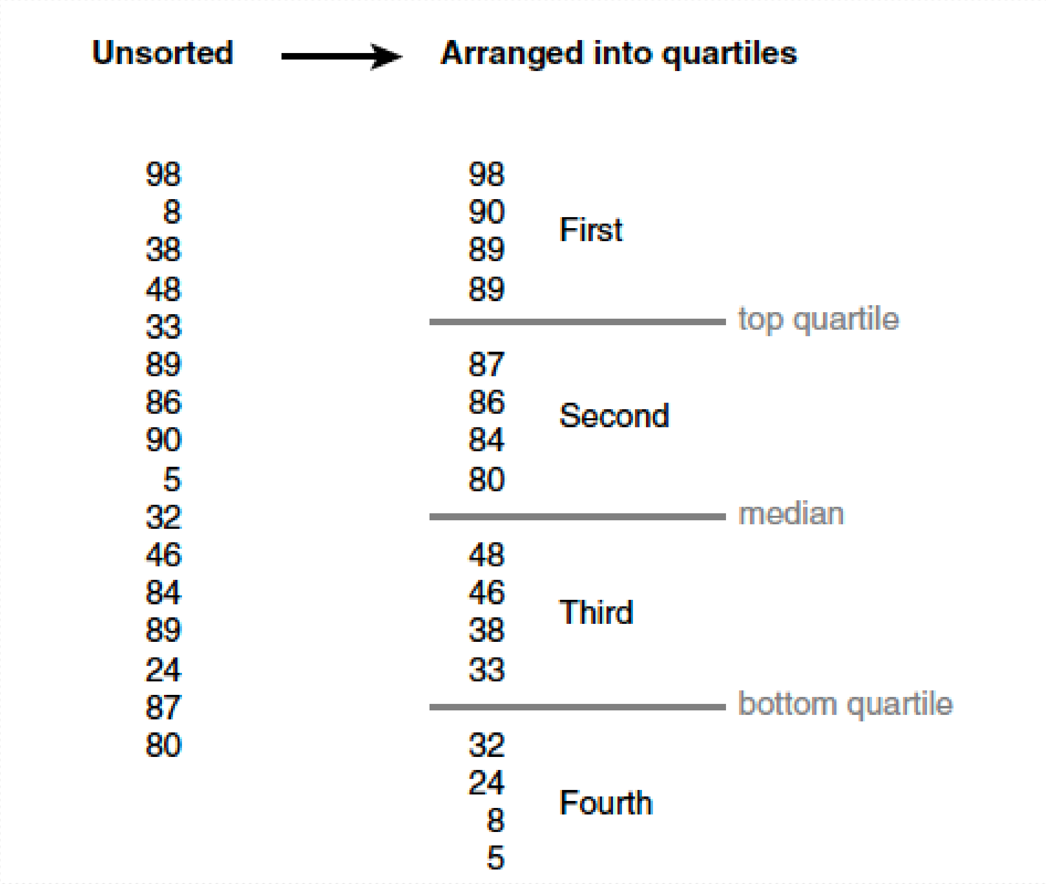


	IT	Operations	Sales
Applications*	2	1	2
Defects	10	11	6
Mean Defects Per Application*	5	11	3
Sum of BAR	134	128	49
Mean BAR Per Application*	67	128	24.5
Mean Exploitability	3.4	4.3	1.8
StdDev of Exploitability	1.3	1.3	0.8
Mean Impact	3.5	2.9	4.7
StdDev of Impact	1.6	1.3	0.8
Mean BAR	13.4	11.6	8.2
StdDev of BAR	8.9	6.1	2.5
Mean Engineering Fix Hours	8	29.8	11
StdDev of Engineering Fix Hours	2.6	4.9	10.8

Quartile Analysis

- Instead of considering all records in the aggregation equally, Quartile analysis takes extra steps by ranking each aggregated result (e.g. risk score) into four “quartiles”.
  - The first quartile represents the best 25%
  - The second quartile cuts off the best 50%
  - The third quartile cuts off the top 75%
  - The forth quartile represents the worst 25%
- Very powerful, and easy-to-understand.
- Aggregated statistics between the 1st and 4th quartile are usually dramatic and revealing.

Sample Quartile Data



Unsorted → Arranged into quartiles

Unsorted: 98, 8, 38, 48, 33, 89, 86, 90, 5, 32, 46, 84, 89, 24, 87, 80

First: 98, 90, 89, 89

top quartile

Second: 87, 86, 84, 80

median

Third: 48, 46, 38, 33

bottom quartile

Fourth: 32, 24, 8, 5

Example: Quartile Analysis of Network Vulnerability Data

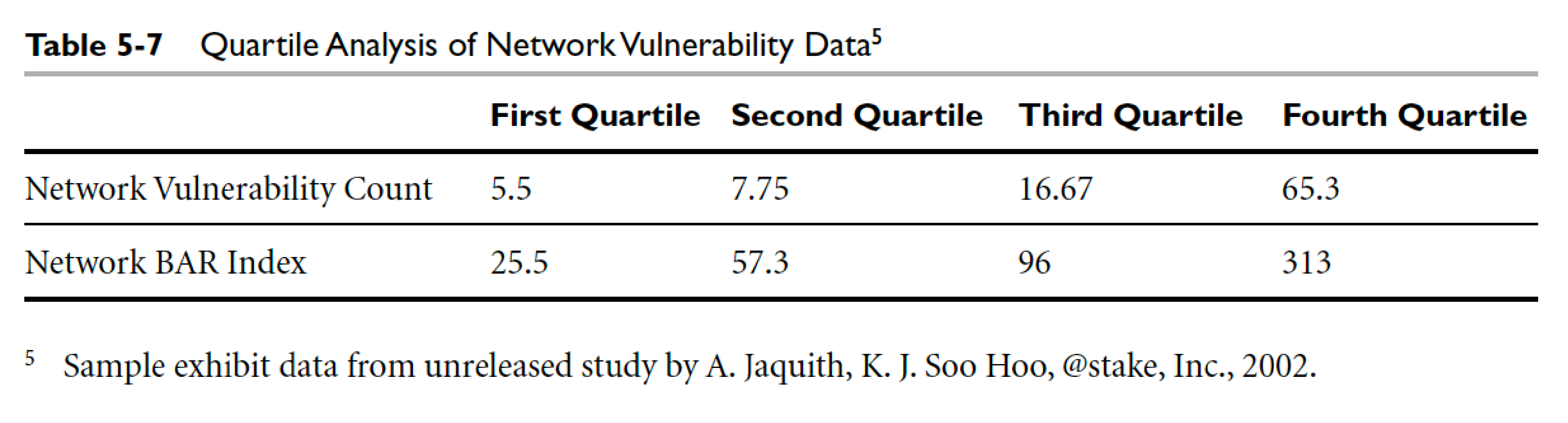


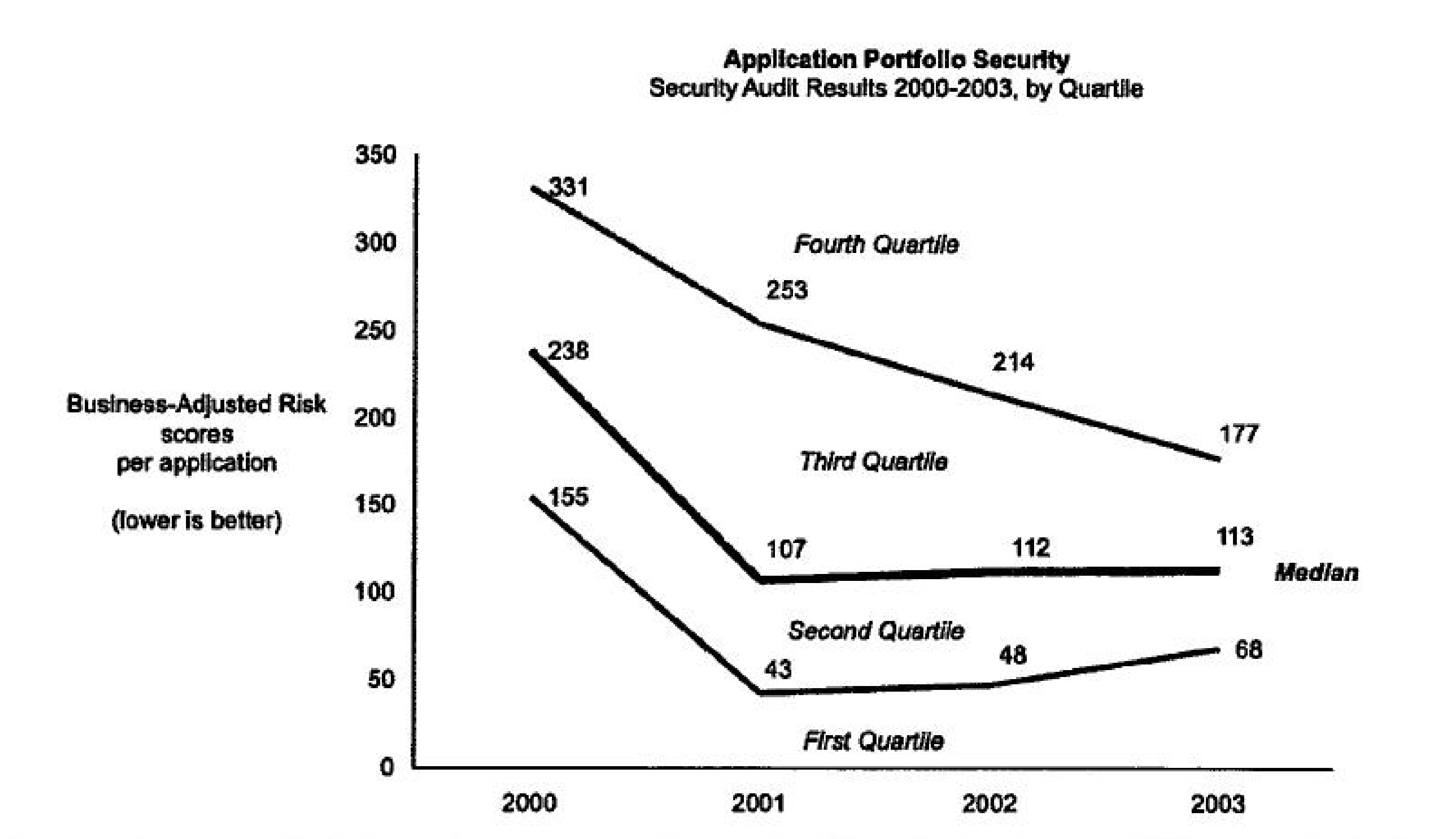
Table 5-7 Quartile Analysis of Network Vulnerability Data[5]

	First Quartile	Second Quartile	Third Quartile	Fourth Quartile
Network Vulnerability Count	5.5	7.75	16.67	65.3
Network BAR Index	25.5	57.3	96	313

[5] Sample exhibit data from unreleased study by A. Jaquith, K. J. Soo Hoo, @stake, Inc., 2002.

- Quartile summary statistics
  - More macro-level behavior
  - Another level of refinement
  - First-vs-Forth Analysis

Quartile Time Series Chart



Correlation Analysis

- Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two attributes in a data set.
  - *A **Positive** correlation is a relationship between two or more attributes whereby their values increase or decrease together.*
  - *Similarly, a negative correlation is a negative relationship, whereby when one attribute increases, the other will decrease, and vice versa.*
  - *If there is no consistent linear pattern in the change between attributes, they are said to be uncorrelated.*
- Correlation is not “Causation”

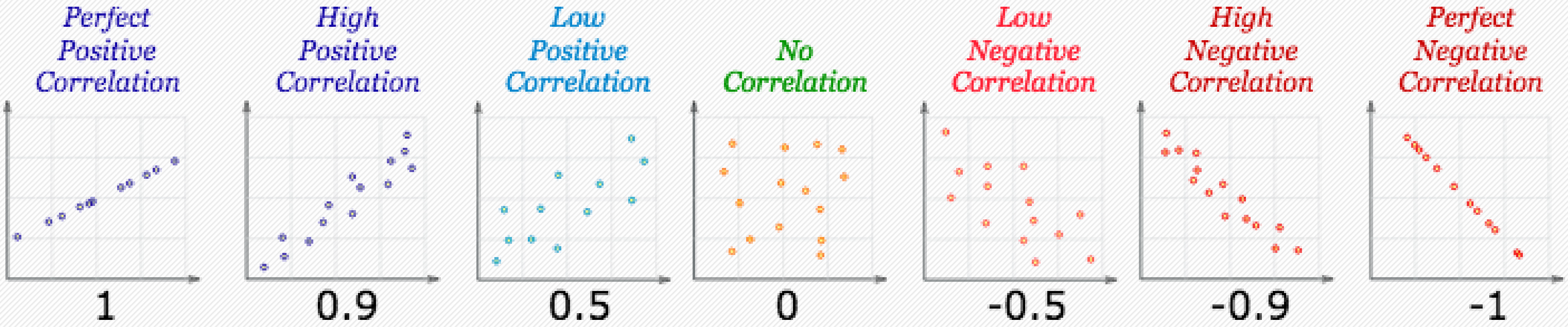
Covariance & Correlation Coefficiency

- Covariance
  Refers to the tendency of one set of values to move with another set.
- Correlation coefficiency
  Normalizes the covariance number to a scale ranging from -1 (opposite direction) to +1 (the sets move perfectly together). 0 indicates that the two data sets have no apparent relationship.

Algorithms for Correlation Analysis

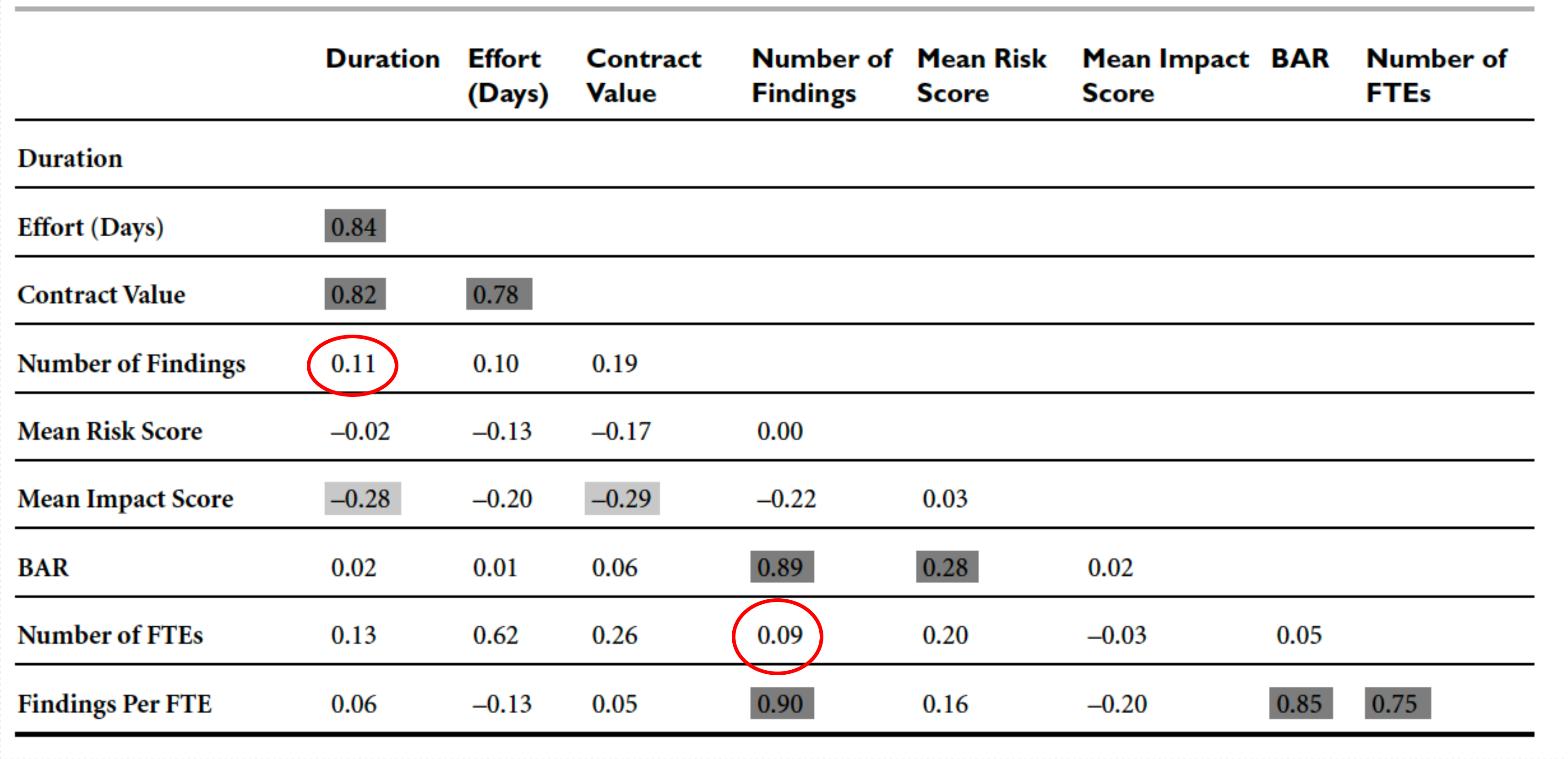
- Pearson r correlation
- Spearman rank correlation
- Kendall rank correlation
- Use the Statisticssolutions.com website to learn more http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

Correlation ScatterPlot Examples: Strong, Weak and Negative Correlation



https://www.mathsisfun.com/data/scatter-xy-plots.html

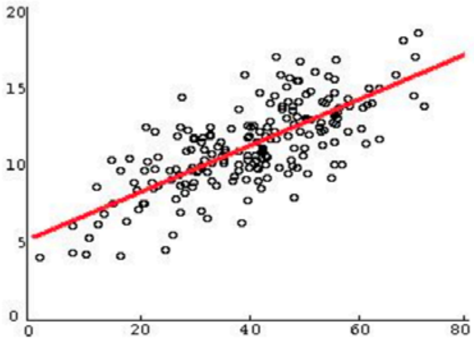
Correlation Matrix



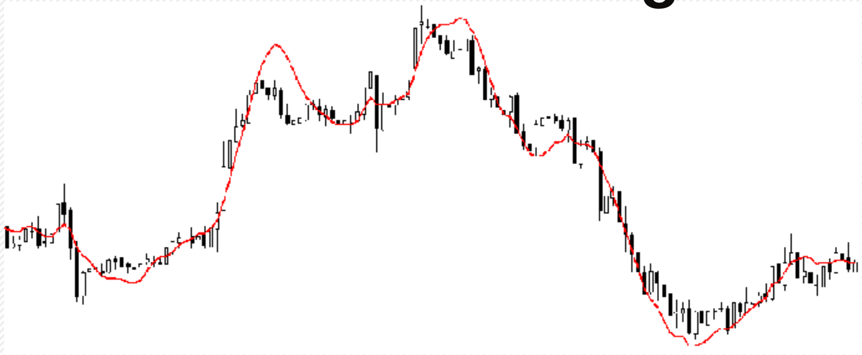
	Duration	Effort (Days)	Contract Value	Number of Findings	Mean Risk Score	Mean Impact Score	BAR	Number of FTEs
Duration								
Effort (Days)	0.84							
Contract Value	0.82	0.78						
Number of Findings	0.11	0.10	0.19					
Mean Risk Score	–0.02	–0.13	–0.17	0.00				
Mean Impact Score	–0.28	–0.20	–0.29	–0.22	0.03			
BAR	0.02	0.01	0.06	0.89	0.28	0.02		
Number of FTEs	0.13	0.62	0.26	0.09	0.20	–0.03	0.05	
Findings Per FTE	0.06	–0.13	0.05	0.90	0.16	–0.20	0.85	0.75

Regression Analysis

Linear Regression



Non-Linear Regression



- Determine the best-fitting equation f :
  - Y = f (X), where X are input(s), Y is output
- Minimize prediction errors (e.g. Root Mean Square Error – EMSE)
- Not descriptive, but inferential analysis (still not direct causal inference)
  - Estimate how different observable inputs contribute to an observable output
  - Given a specific X, estimate or predict what the output Y is

Correlation vs. Regression

- Similarities
  - For standard linear regression coefficient, it can be computed use the same algorithm such as Pearson’s correlation as in correlation analysis, although the coefficient’s meanings are different.
  - Both can be linear or non-linear relationship
  - Neither simple linear regression nor correlation answer questions of causality directly.
- Differences
  - The regression equation Y=f(X) can be used to make predictions on Y based on the value of X
  - Correlation quantifies the degree to which two attributes are related, but it does not fit a line through the data.
  - Correlation coefficient indicates the extent to which two variables move together, while regression coefficient indicates the impact of a unit change in the known variable (X) on the estimated variable (Y).

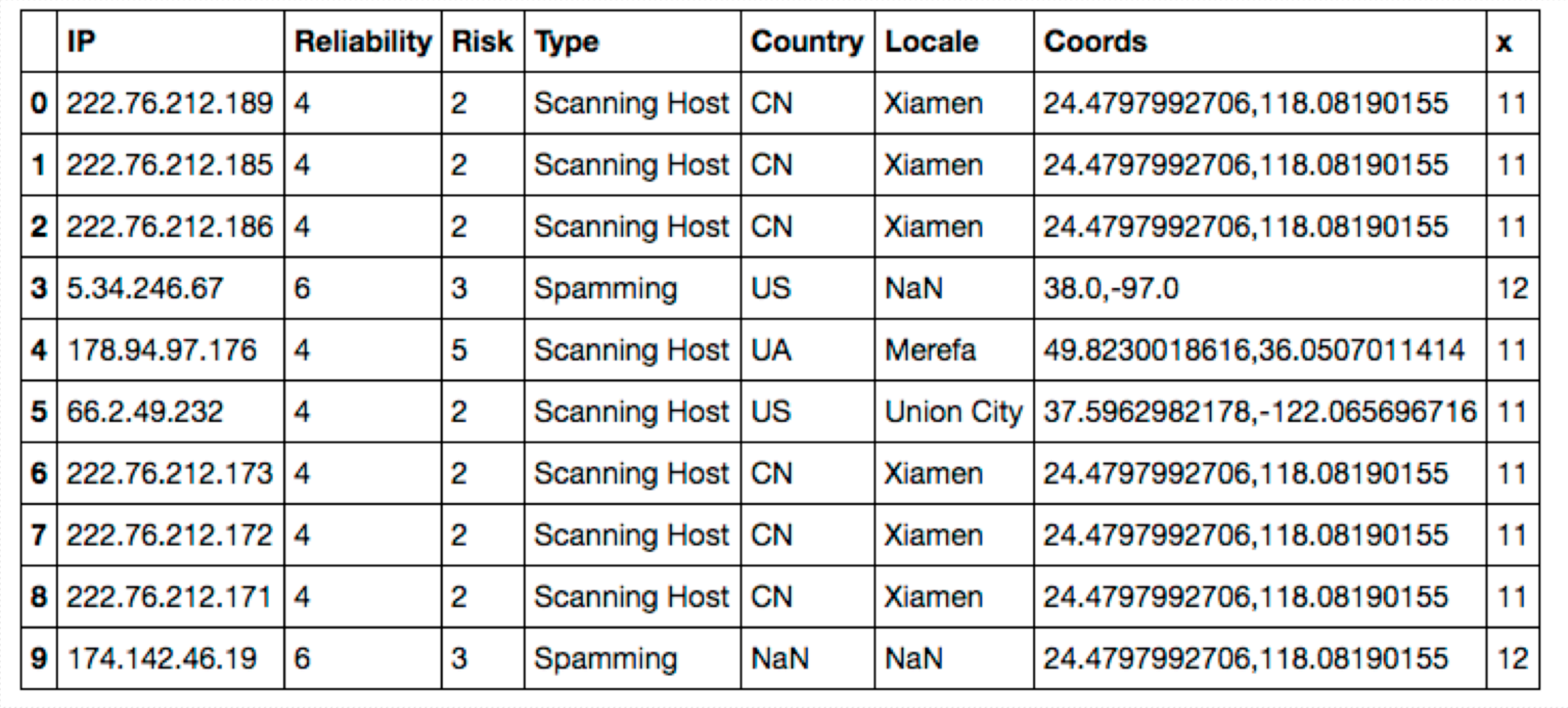
Building Analytics Toolbox

- Python (www.python.org)
  - Easy to learn for people with an existing programming background
  - Flexibility & extensibility
- R (www.r-project.org)
  - It was created by Statisticians with extensive statistical analysis packages
  - For those who are new to statistical languages, becoming proficient in R may pose more of a challenge

An Example Problem

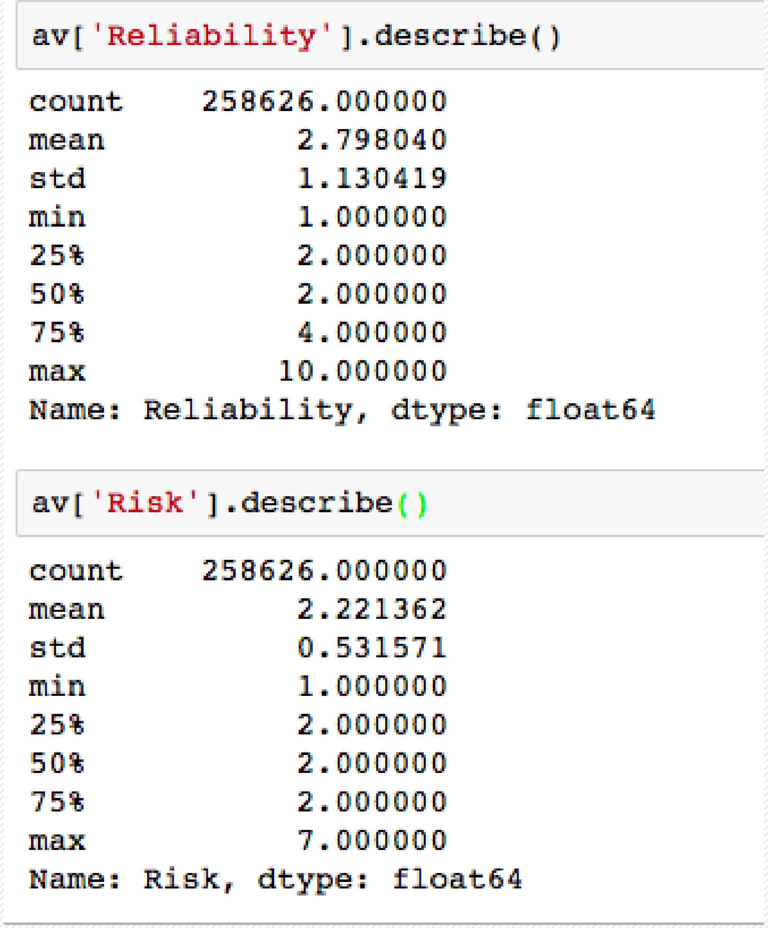
- Research Question:
  - To reduce the number of “trivial” alerts without sacrificing visibility
- AlienVault IP Reputation Database https://www.alienvault.com/
  - It is freely available data set that contains information on various
  - Types of “badness” across the Internet

Example: AlienVault's IP Reputation Database



	IP	Reliability	Risk	Type	Country	Locale	Coords	x
0	222.76.212.189	4	2	Scanning Host	CN	Xiamen	24.4797992706,118.08190155	11
1	222.76.212.185	4	2	Scanning Host	CN	Xiamen	24.4797992706,118.08190155	11
2	222.76.212.186	4	2	Scanning Host	CN	Xiamen	24.4797992706,118.08190155	11
3	5.34.246.67	6	3	Spamming	US	NaN	38.0,-97.0	12
4	178.94.97.176	4	5	Scanning Host	UA	Merefa	49.8230018616,36.0507011414	11
5	66.2.49.232	4	2	Scanning Host	US	Union City	37.5962982178,-122.065696716	11
6	222.76.212.173	4	2	Scanning Host	CN	Xiamen	24.4797992706,118.08190155	11
7	222.76.212.172	4	2	Scanning Host	CN	Xiamen	24.4797992706,118.08190155	11
8	222.76.212.171	4	2	Scanning Host	CN	Xiamen	24.4797992706,118.08190155	11
9	174.142.46.19	6	3	Spamming	NaN	NaN	24.4797992706,118.08190155	12

Exploring Data: Descriptive Statistics



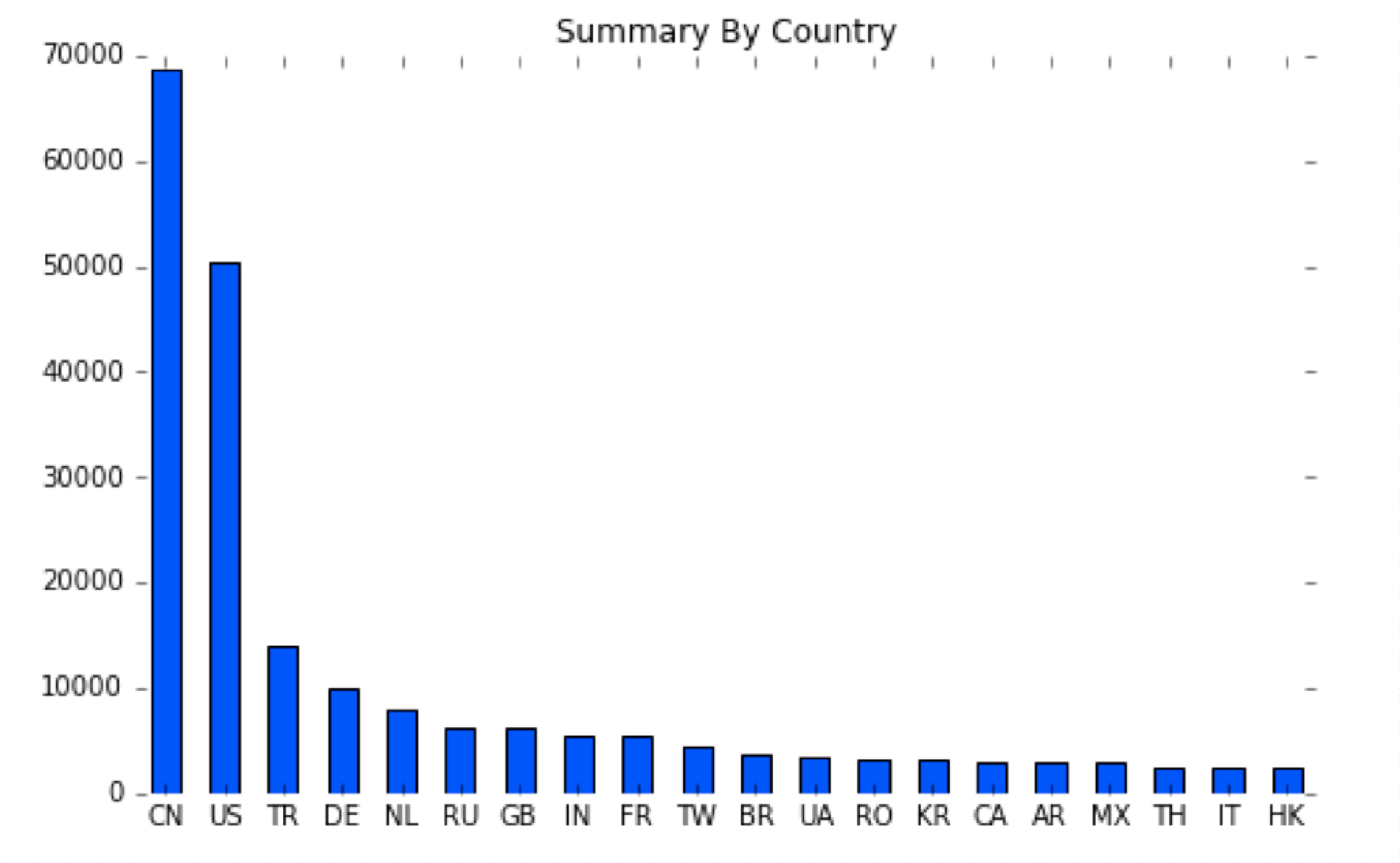
```
av['Reliability'].describe()
```

```
count    258626.000000
mean          2.798040
std           1.130419
min           1.000000
25%           2.000000
50%           2.000000
75%           4.000000
max          10.000000
Name: Reliability, dtype: float64
```

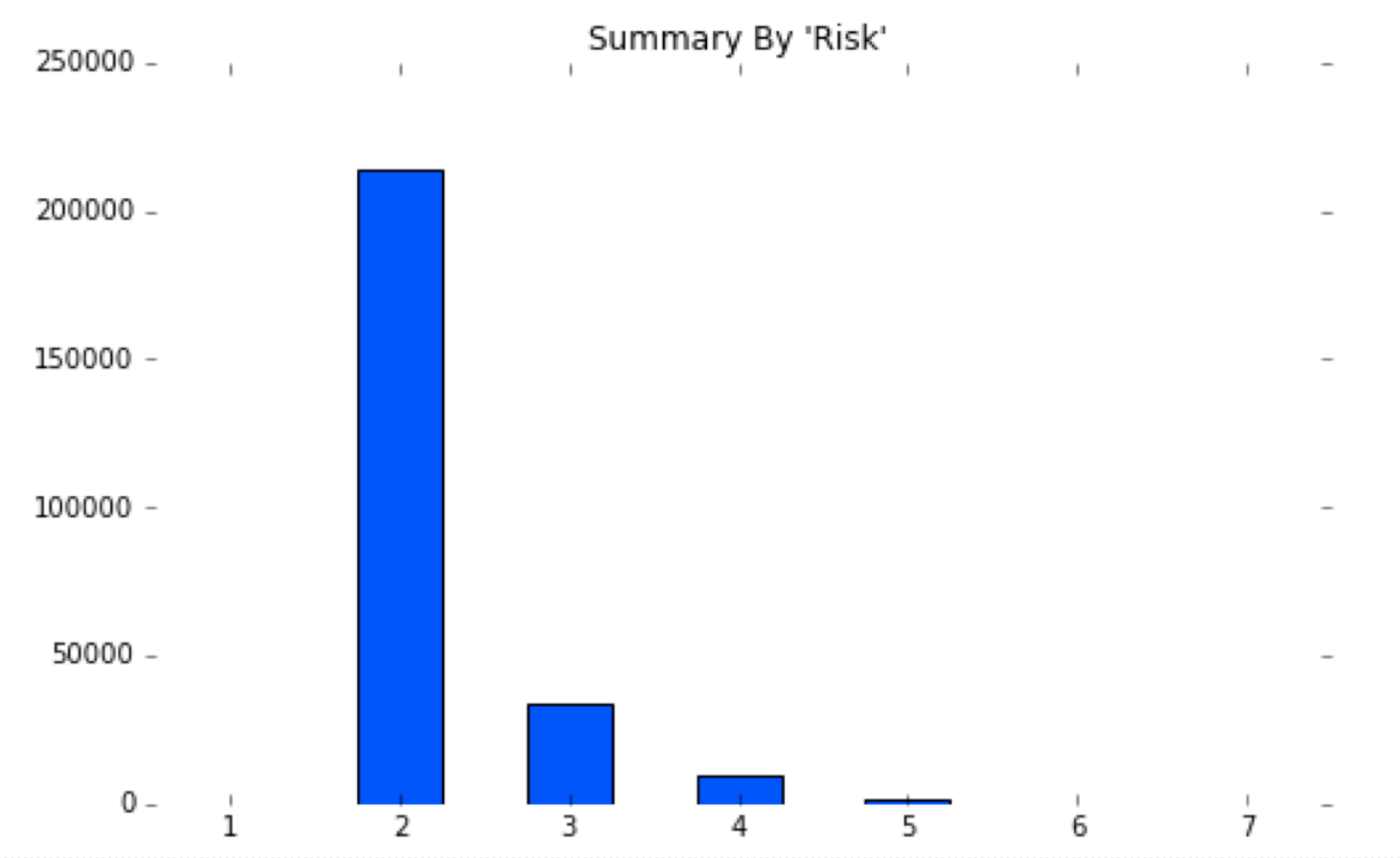
```
av['Risk'].describe()
```

```
count    258626.000000
mean          2.221362
std           0.531571
min           1.000000
25%           2.000000
50%           2.000000
75%           2.000000
max           7.000000
Name: Risk, dtype: float64
```

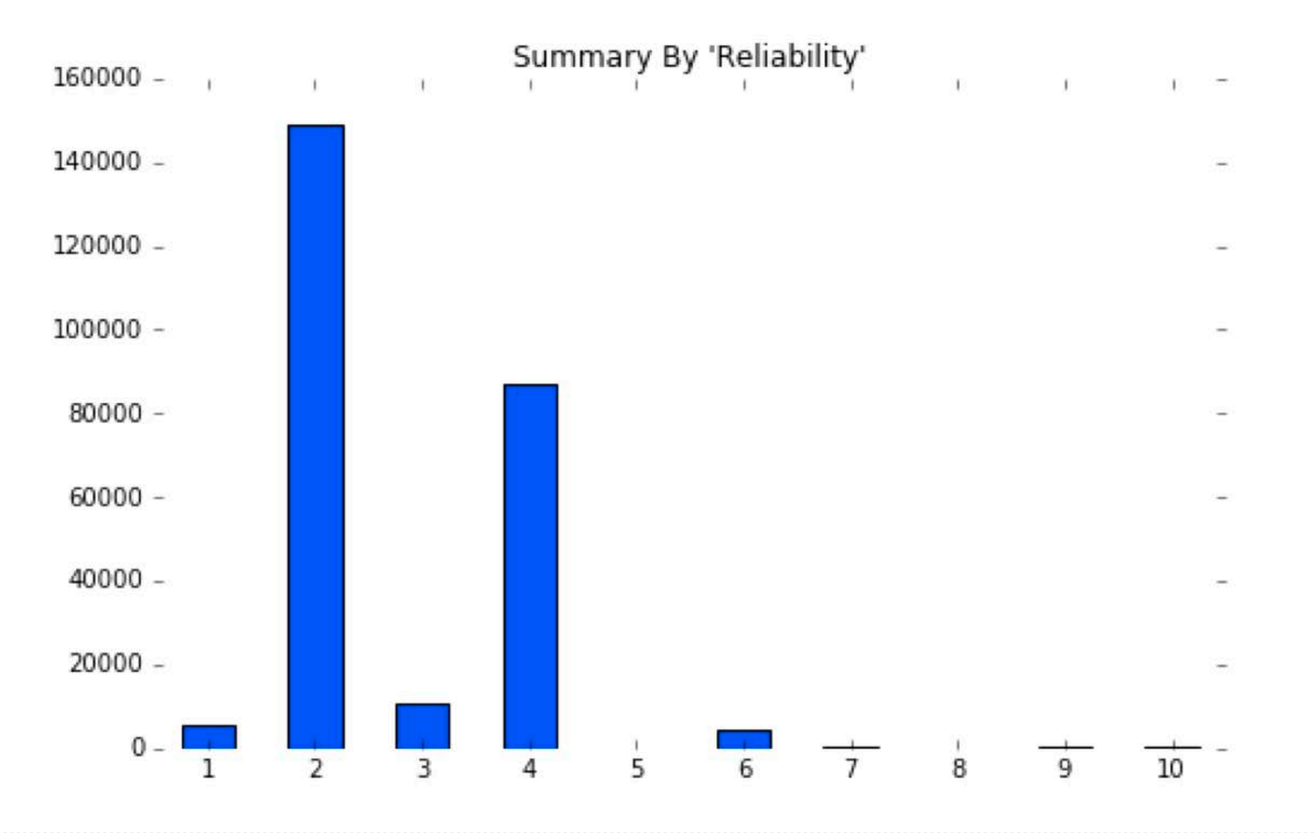
Exploring Data: Distribution Graph (1)



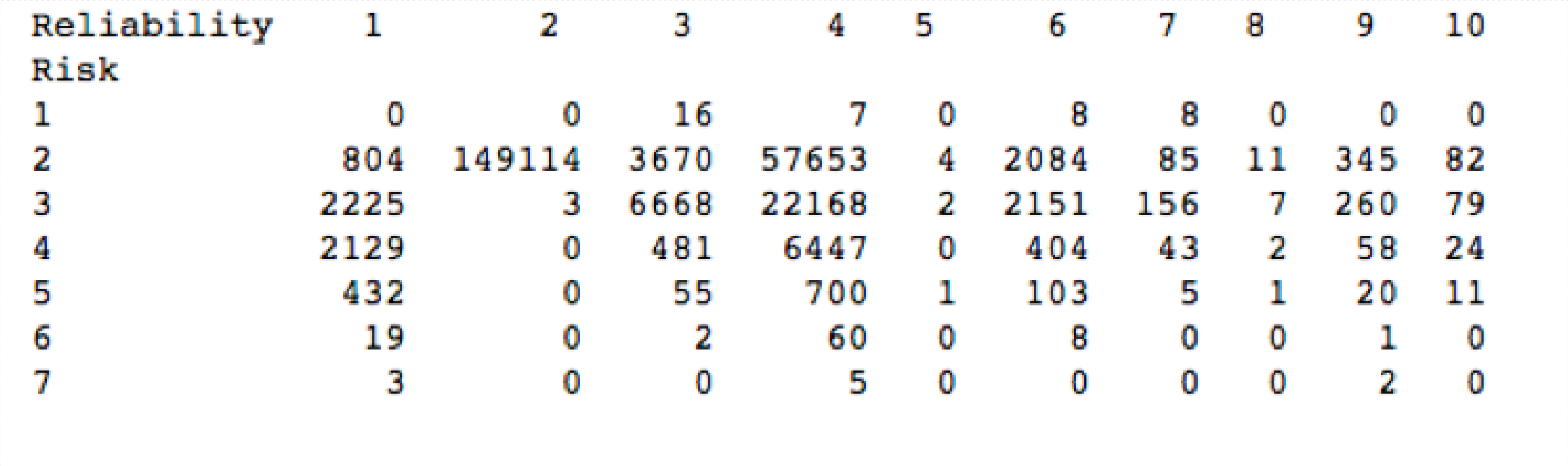
Exploring Data: Distribution Graph (2)



Exploring Data: Distribution Graph (3)



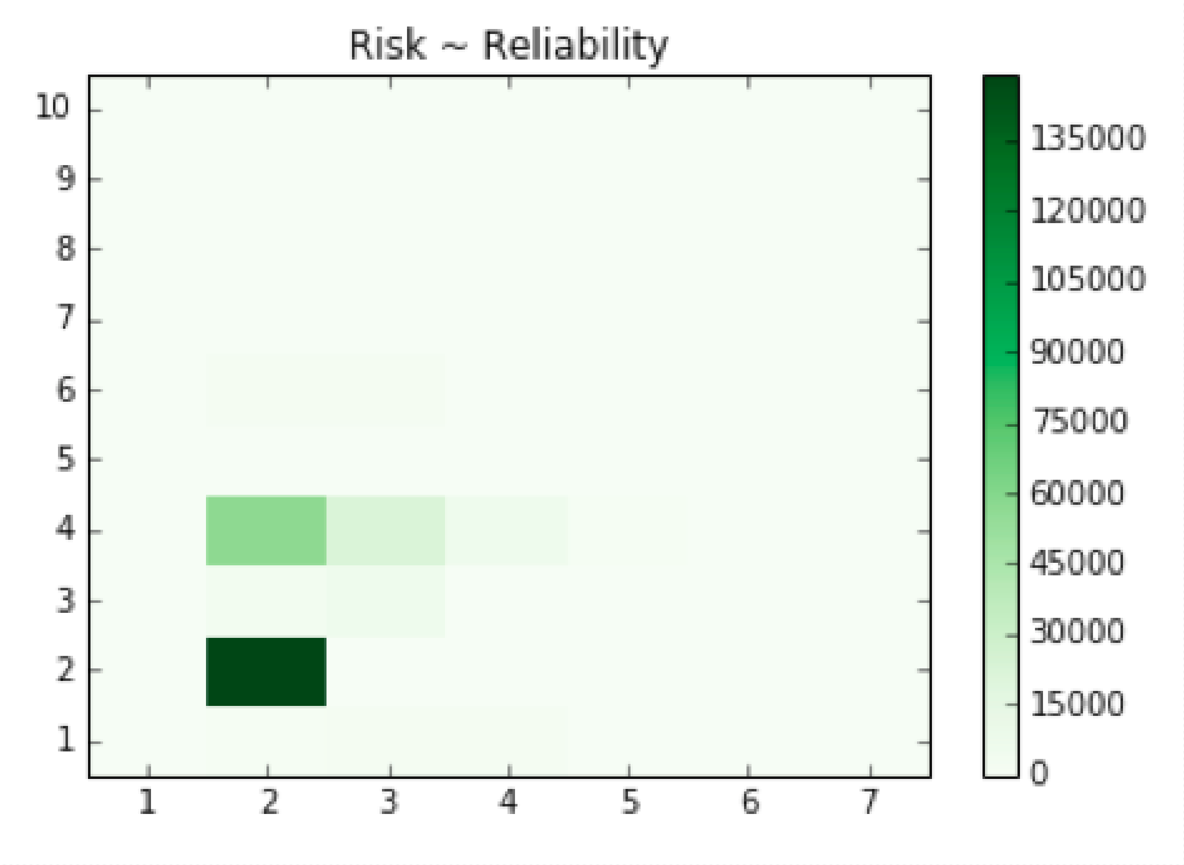
Homing In on a Question



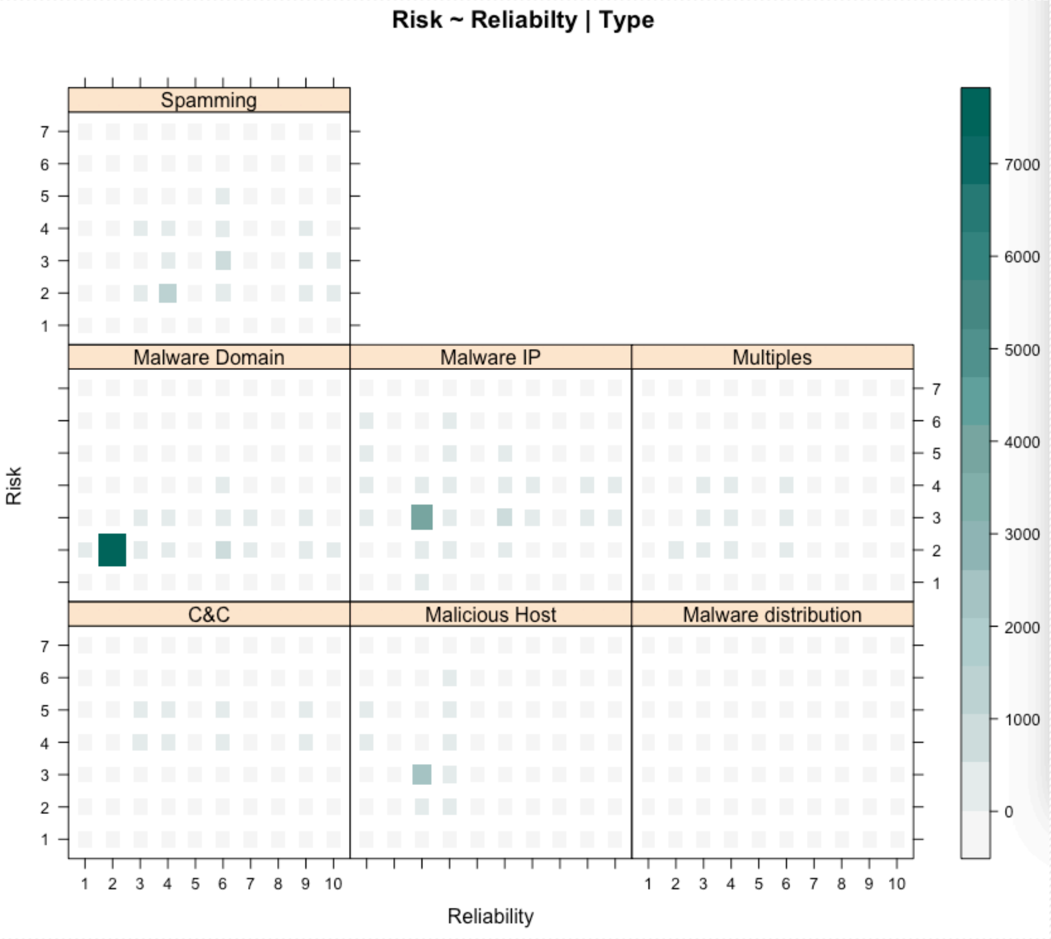
```
Reliability    1      2    3     4 5    6   7  8   9 10
Risk
1              0      0   16     7 0    8   8  0   0  0
2            804 149114 3670 57653 4 2084  85 11 345 82
3           2225      3 6668 22168 2 2151 156  7 260 79
4           2129      0  481  6447 0  404  43  2  58 24
5            432      0   55   700 1  103   5  1  20 11
6             19      0    2    60 0    8   0  0   1  0
7              3      0    0     5 0    0   0  0   2  0
```

“Reliability” & “Risk” Contingency Table

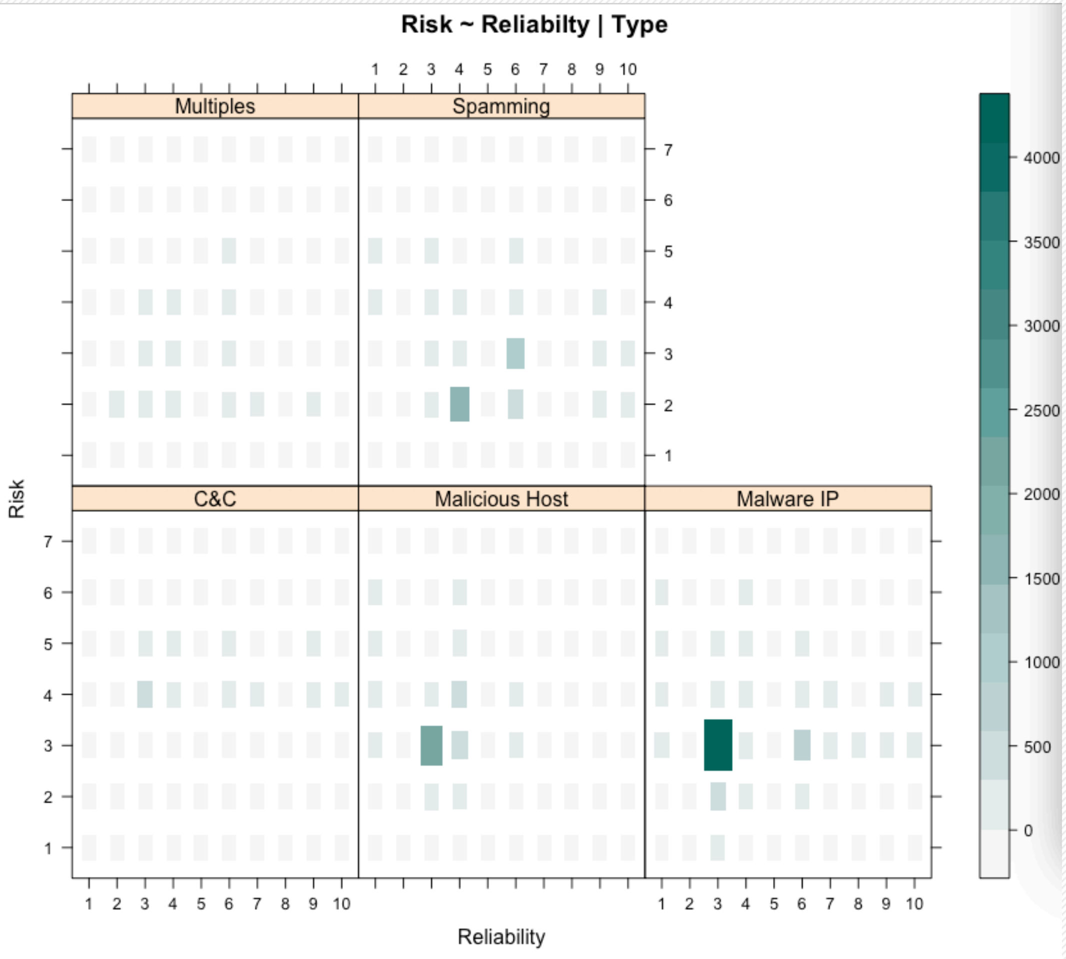
Another View: Risk/Reliability Correlation



Risk/Reliability & Type Correlation (1)



Risk/Reliability & Type Correlation (2)


Risk ~ Reliabilty | Type
Multiples
Spamming
C&C
Malicious Host
Malware IP
Risk
Reliability
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7
0
500
1000
1500
2000
2500
3000
3500
4000