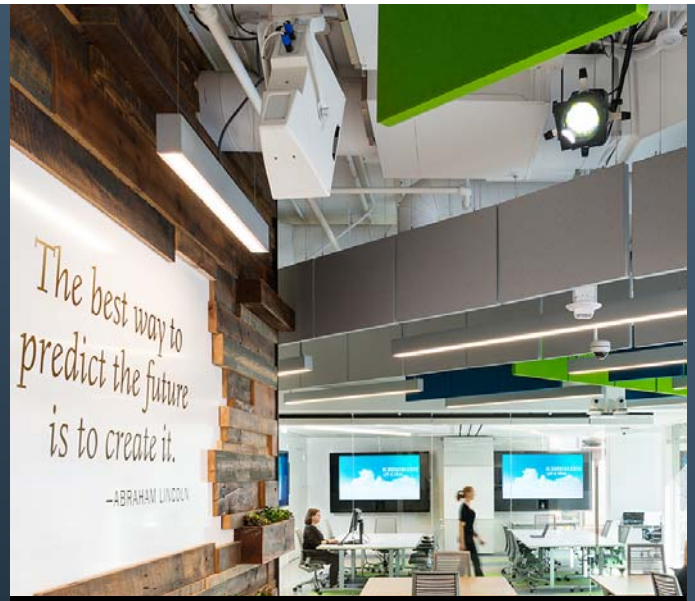
Booz | Allen | Hamilton



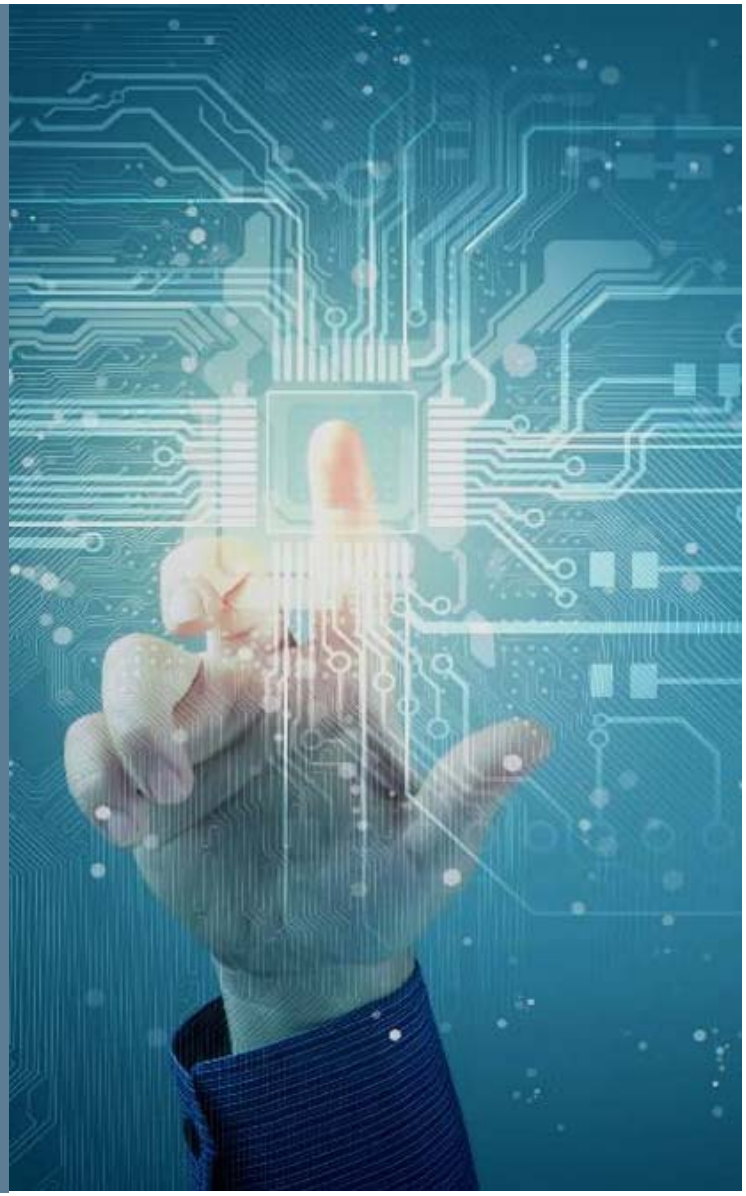
Innovation center, Washington, D.C.

MULTIVARIATE ANALYSIS OF STEALTH QUANTITATES (MASQ)

Application of Machine Learning to Testing in Finance, Cyber, and Software

THE SCIENCE OF TEST WORKSHOP 2017

CONSULTING | ANALYTICS | DIGITAL SOLUTIONS | ENGINEERING | CYBER



AGENDA

INTRODUCTION TO MASQ

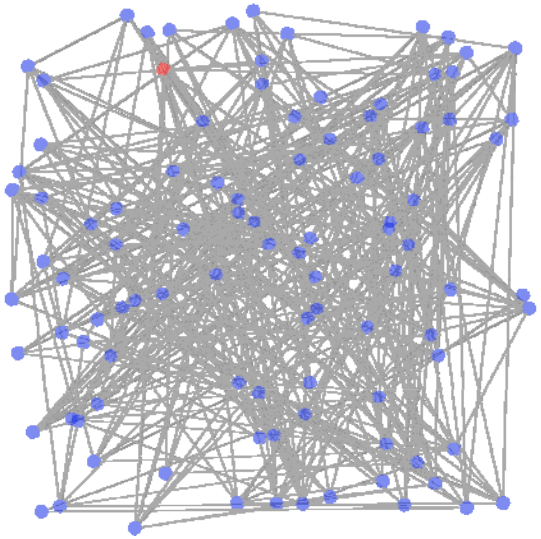
MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES (MASQ) ALGORITHM OVERVIEW

MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES APPLIED TO CYBER SECURITY (MASQ-C)

MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES APPLIED TO AUDIT TESTING (MASQ-F)

MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES APPLIED TO SOFTWARE TESTING (MASQ-S)

MASQ IS USED TO FIND RARE OBSERVATIONS IN LARGE COMPLEX DATA



"Needle in the Haystack"

STEALTH QUANTITIES ARE DANGEROUS THREATS

- Stealth Quantities are anomalies that can easily hide among multiple variables, they are anomalies that are too rare or easily modified to easily train on
- Only by applying advanced analytics to all data under investigation simultaneously will Stealth Quantities be found

AKA NEEDLE IN A HAYSTACK

- The most apt analogy for stealth quantities is the needle in a haystack. Stealth Quantities typically exist in randomly related data and defy conventional efforts of discovery.
- Stealth Quantities are hiding amongst normal observations exploiting the dimensions that exist in data to mask themselves as normal.

MASQ
MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES

WHAT IS MASQ?

- Machine Intelligence
  - MASQ is a proprietary algorithm that mines data for known and unknown signatures. Signatures are characteristic or distinctive patterns searchable within data. Signatures fall into two categories: known and unknown. Known signatures are generally well documented and have predefined properties or rules than make Supervised Learning by a machine straightforward. Unknown signatures exist within data but do not have predefined properties or rules making Supervised Learning impractical. Predicting the quantity of unknown signatures is possible, but not trivial to perform, and requires Unsupervised Learning techniques. MASQ makes use of both styles to find any anomaly in any data set.

- Discipline Agnostic Cloud Analytics
  - MASQ works on any volume, variety, velocity, or veracity of data to map all data into distinct subpopulations from which errors or malicious behavior can be flagged. MASQ was created to address the kind of multivariate systems presented by an organization's cyber, financial, or simulation enterprises of today. Multivariate acknowledges the fact that today's enterprise systems deal with hundreds to thousands of variables that characterize financial transactions, experimental data, network traffic, etc.

- Predictive Anomaly Detection
  - Stealth Quantities are unknown risks lurking within the large, complex datasets that defy many of today's methods of detection. Known risks (e.g., known viruses, past audit findings, or programming bugs) are generally well documented and have predefined properties or rules than make Supervised Learning by a machine straightforward. Unknown risks (e.g., new malware, new forms of fraud, etc.) exist within data streams but do not have predefined properties or rules making Supervised Learning, and traditional techniques, impractical or late-to-need. MASQ finds both known, current threats and new threats yet undiscovered.

MASQ
MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES

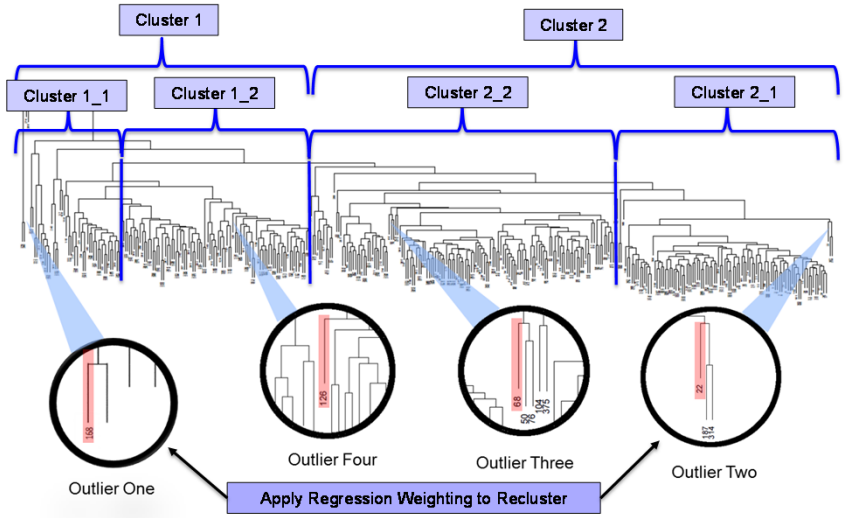
HOW MASQ WORKS

MASQ is a process that leverages machine learning techniques by combining both supervised and unsupervised machine learning techniques (hybrid machine learning)

- ***Unsupervised Learning*** means that you do not need to "tell" the program what to look for prior to analyzing the data; every time you examine data in MASQ it can find new types of rare events
- ***Supervised Learning*** captures SME feedback regarding the "goodness" or "badness" of rare event groups. MASQ then learns what rare events are acceptable and those that are not which streamlines repeated application of the software to additional data

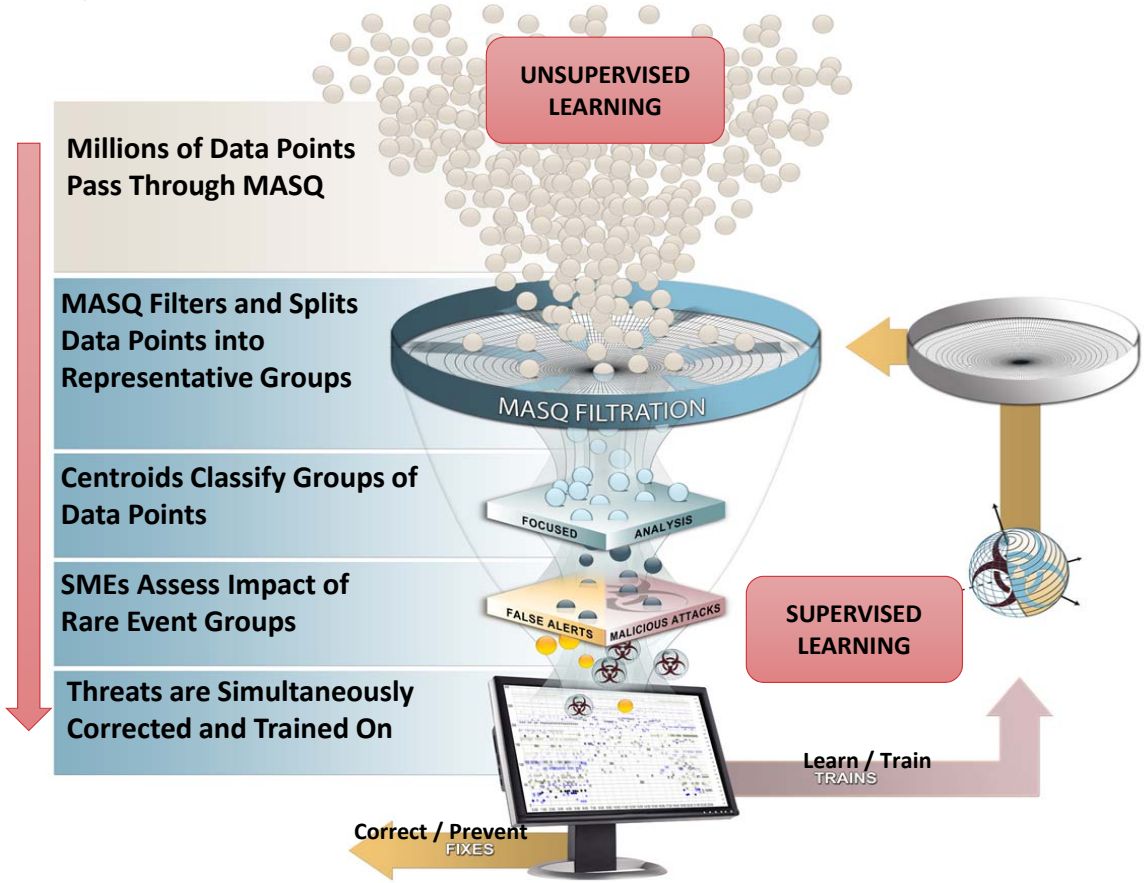
MASQ finds rare events through highly advanced clustering techniques

- Algorithms split Data into Groups using Hierarchical Clustering, Applies Weights, Splits again, until the Data is Separated into Distinct Populations



MASQ
MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES

THE MASQ PROCESS



MASQ
MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES

Use Case: Detecting Cyber Anomalies

TITLE

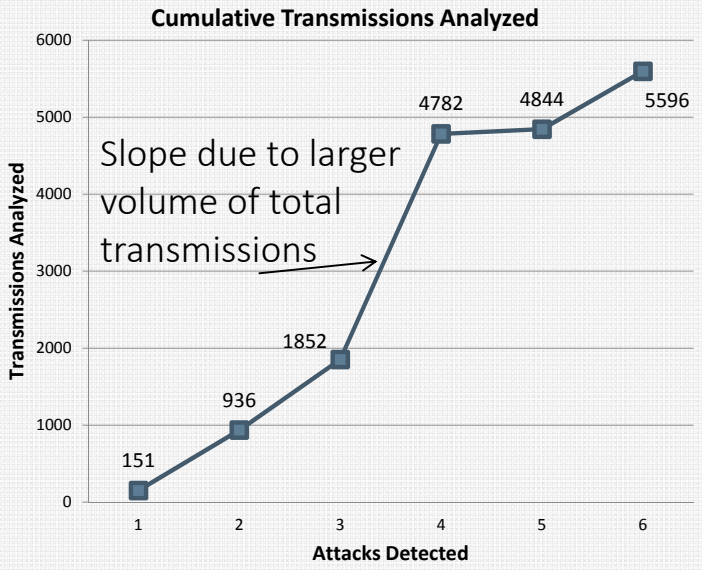
Detecting Cyber Attacks within Millions of Netflow Transmissions

THE CHALLENGE

Naval commands are often targets of sophisticated new forms of cyber attacks that traditional cyber security efforts are much more likely to misidentify as acceptable behavior. Six known attacks were injected into six portions of the client's network data to test MASQ. SQL Injection and Pass the Hash were considered difficult to detect by SMEs.

OUR SOLUTION

+ **Mine ALL of the Data:** 13.8 million total transmissions were analyzed in minutes. MASQ operates by conducting multidimensional analysis across hundreds of variables simultaneously.
+ **Improve Understanding:** MASQ summarized all packets into 21,600 groups containing between 500 and 1,500 transmissions each
    + **SQL Injections** found in clusters 22, 113, and 7
    + **Pass the Hash** found in clusters 45, 147, and 76
+ **Improve Efficiency:** MASQ breaks data into clusters. This means that MASQ uses a small subset of data to identify thousands of problem points. Attacks were identified after analyzing less than 0.05% of the 13.8 million transmissions



Timestamp	Attack	Number of Transmissions	Transmissions Analyzed
23 Jul 1000	SQL Injection	1,616,130	151
23 Jul 1100	Pass the Hash	2,912,504	785
24 Jul 1300	SQL Injection	3,699,647	916
25 Jul 1200	Pass the Hash	3,359,869	2930
24 Jul 0100	SQL Injection	1,112,536	62
24 Jul 0100	Pass the Hash	1,119,467	752
Total		13,820,153	5,596

MASQ
MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES

Use Case: Find Accounting Anomalies for Audit Readiness

TITLE

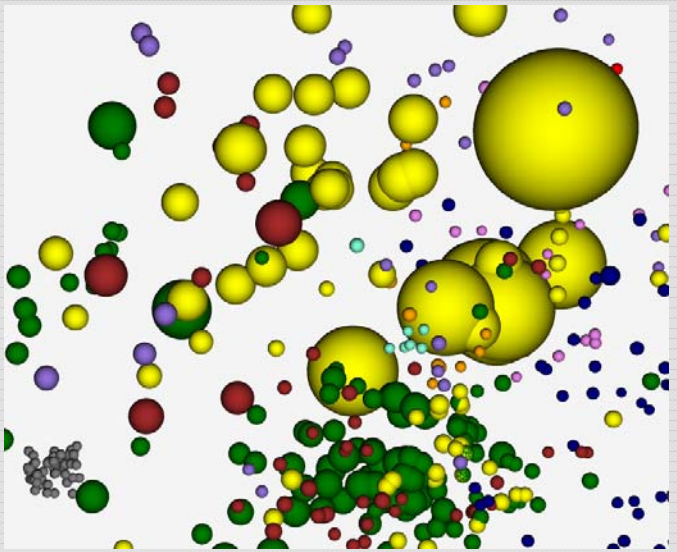
Identification of Accounting Anomalies Within Financial Transactions

THE CHALLENGE

The Navy processes millions of transactions quarterly. The sheer volume of data makes it impossible for a human to determine anomalies leading to audit failure. The client was interested to determine if MASQ could find accounting behavior not known to SMEs and improve audit readiness.

OUR SOLUTION

+ **Mine ALL of the Data:** The MASQ team ran analysis on the STARS-FL segment MILPAY and MILSTRIP independently of each other. The team also processed ERP general ledger data as well for FY15.
+ **Improve Understanding:** We built profiles of anomalous and normal transactions across all business lines so that SMEs are able to identify ALL existing anomalies and predict new anomalies. Key findings included:
    + A large number of credits and debits were discovered to have both positive negative and values.
    + Missing obligation document for extended period of time (over 30 days)
    + Document type mismatches. Quantity and Dollar mismatches, etc.
+ **Improve Efficiency:** MASQ breaks data into clusters. This means that MASQ uses a small subset of data to identify thousands of problem points.



Visualization of Data clusters including credit/debit mismatches (dark red), missing debits (purple), document type mismatches (yellow), normal data (green)

	MILPAY		MILSTRIP
Time Period	FYXX	FYYY Q1	FYXX Q1
Transactions	778,807		7,536,054
Documents	37,684	54,210	1,288,224
Anomalies Found	87,824	3,984	336,712
Erorr Percentage	11.90%	7.30%	4.50%

MASQ
MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES

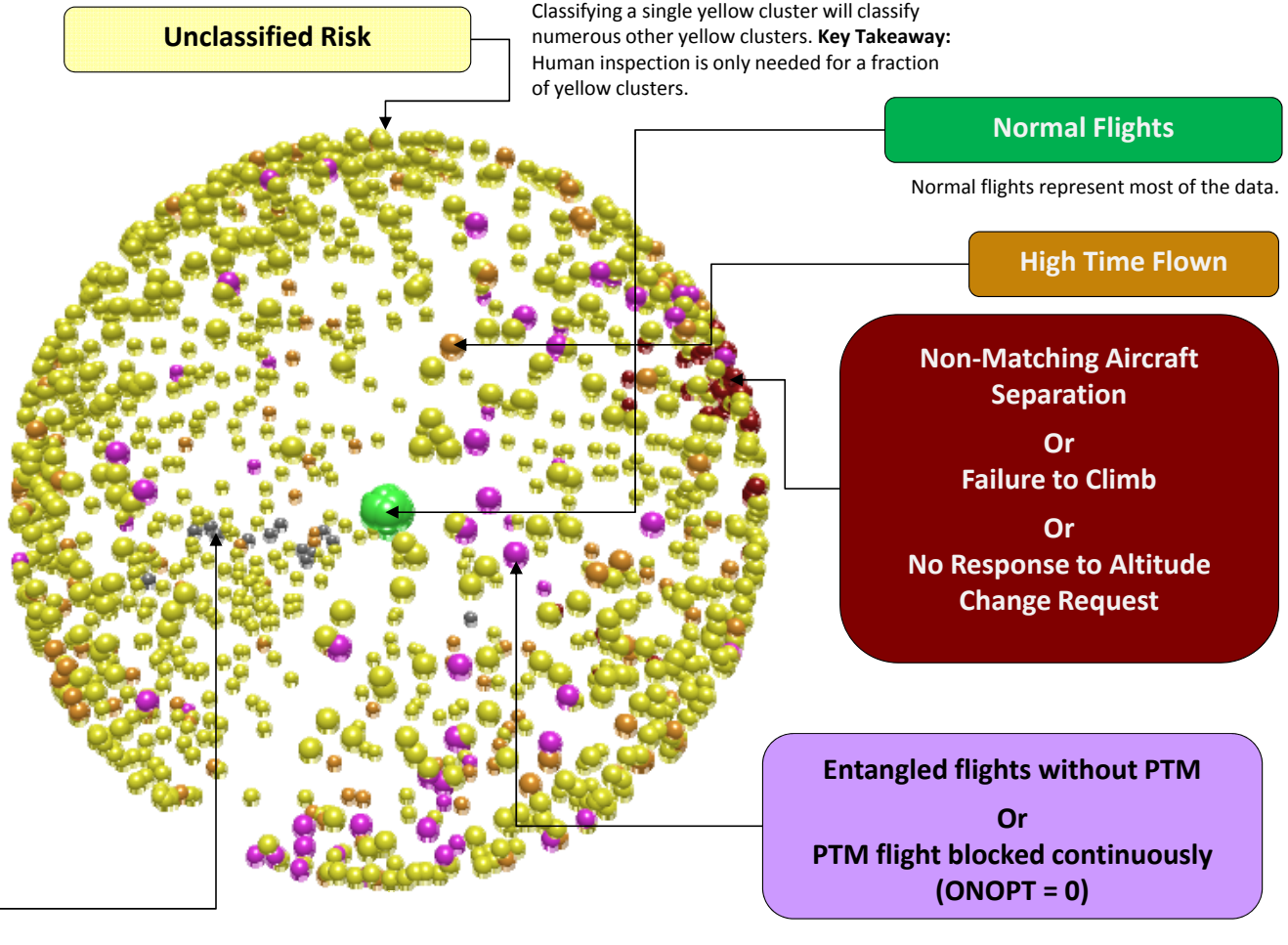
Use Case: Find Simulation Anomalies for Software Validation

THE CHALLENGE

Batch simulations are highly complex computer software that create an operational environment to systematically test NextGen concepts. The volume, veracity, and variety of data produced by batch simulations makes the Verification & Validation (V&V) process very challenging. NASA needs a V&V process for batch simulation that will ensure success in rapid fashion.

OUR SOLUTION

NASA provided the MASQ team with 1.2 TB of PTM Batch Simulation. The data contained three errors known to the customer. The MASQ team isolated the three known errors within the data and found an additional six risks and ~40 high-priority anomalous patterns in the data. Reconciling these ~50 anomalies will give NASA assurance that experiment analysis and results publication can proceed. Further, because the anomalies are sub settable from the data, analysis teams are able to predict the power of test results and proceed with limited hypothesis testing.



MASQ
MULTIVARIATE ANALYSIS OF STEALTH QUANTITIES