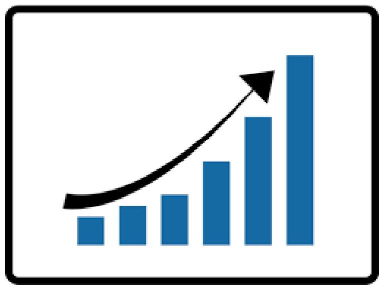
Increasing Confidence in Machine Learned (ML) Functional Behavior during Artificial Intelligence (AI) Development using Training Data Set Measurements

Presented for the Acquisition Research Symposium 2021
Special Webinar on Developing AI in Defense Programs, March 3, 2021

Presented by Bruce Nagy

Motivation: Naval Ordnance Safety and Security Activity (NOSSA) Concerned about the "Garbage In, Garbage Out" Phenomena

Quality of the Training Set (Composition)



Quantity of the Training Set (Size)

0.633666667	0.200333333	0.483666667	0.000333333	0.000333333
0.650333333	0.000333333	0.733666667	0.000333333	0.000333333
0.667	0.350333333	0.700333333	0.000333333	0.000333333
0.467	0.233666667	0.550333333	0.217	0.333666667
0.767	0.000333333	0.417	0.000333333	0.000333333
0.000333333	0.300333333	0.667	0.433666667	0.383666667
0.467	0.583666667	0.000333333	0.000333333	0.000333333
0.000333333	0.233666667	0.000333333	0.750333333	0.367
0.283666667	0.700333333	0.383666667	0.783666667	0.333666667
0.267	0.667	0.000333333	0.700333333	0.833666667
0.183666667	0.450333333	0.250333333	0.483666667	0.533666667
0.000333333	0.633666667	0.000333333	0.000333333	0.000333333
0.000333333	0.267	0.000333333	0.733666667	0.000333333
0.250333333	0.767	0.267	0.783666667	0.750333333
0.000333333	0.000333333	0.183666667	0.000333333	0.733666667
0.000333333	0.267	0.117	0.350333333	0.700333333
0.000333333	0.000333333	0.000333333	0.300333333	0.717

Results of Using Training Set To Develop AI/ML Algorithm (Array of Weights)

Operational Environment

↓

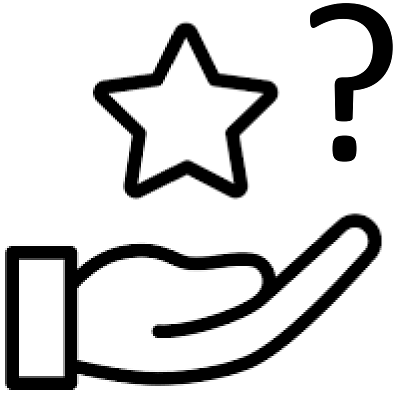
Deployed Algorithm

↓

Confidence in Functional Performance

?

Synthetic vs Real Data – the Training Conundrum



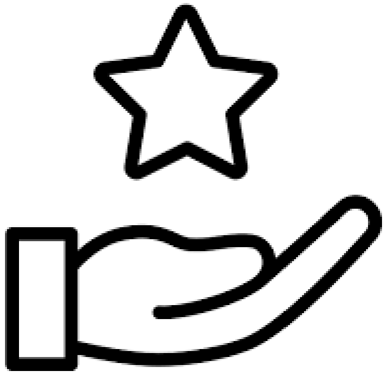
Uncertainty in Synthetic Accuracy

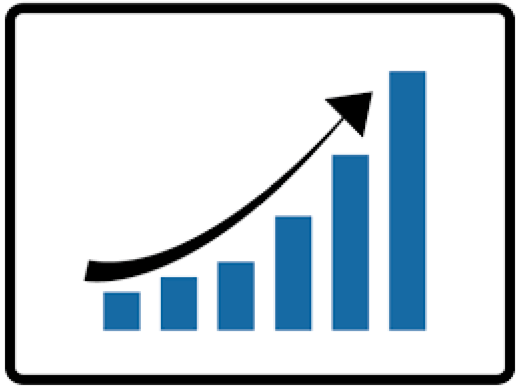
≠

Synthetic vs Real

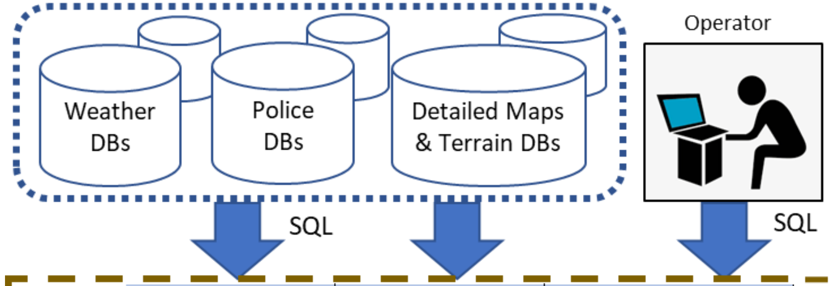
Does it adequately replicate noise?

Does it adequately replicate reality?





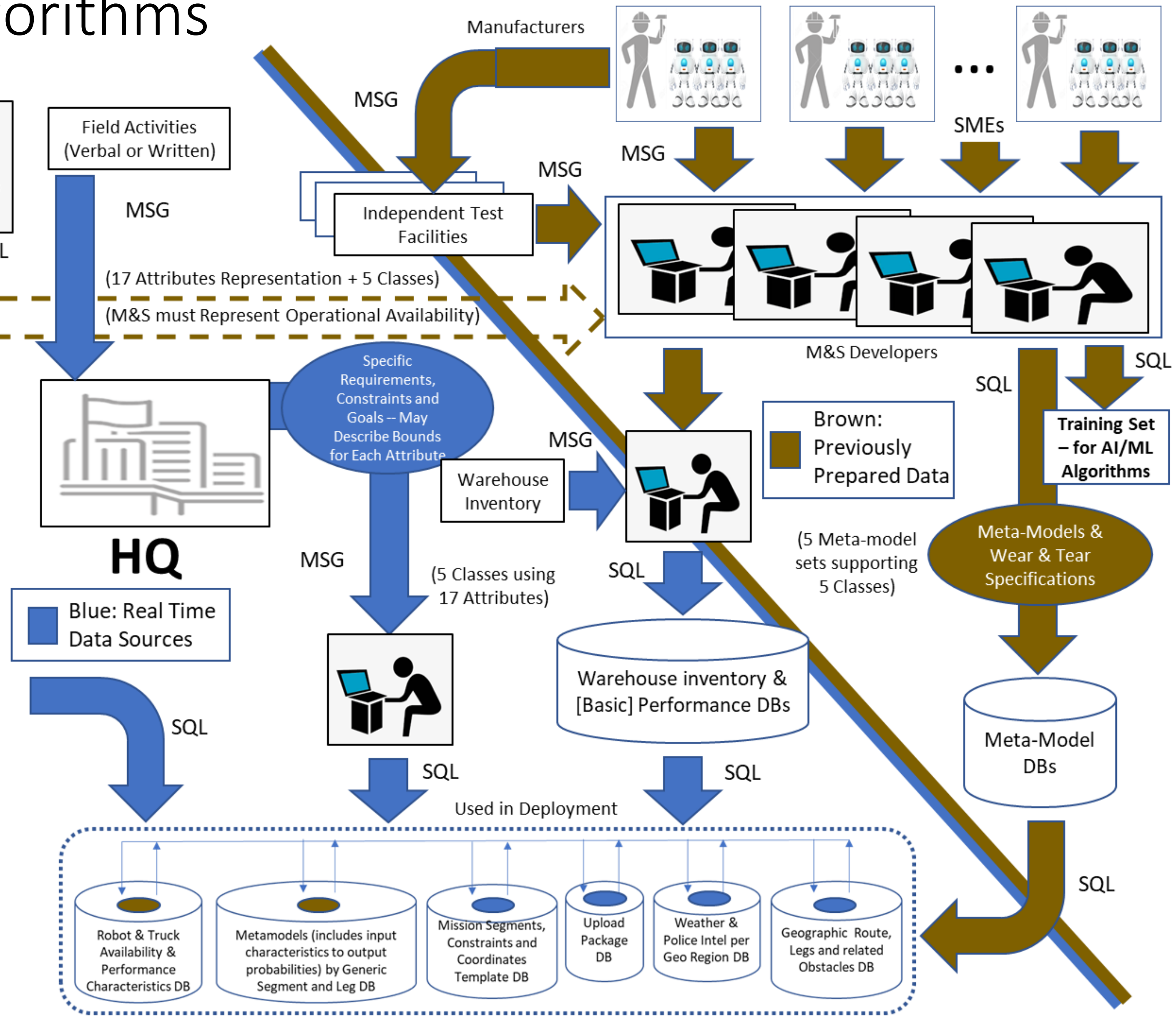
Data Sources Used by ML Algorithms



	Used to Understand Evironment	Used to Understand Route	Used by Meta-Model (Variables to Affect Robot and Route Selection)
Inputs & Attributes	Weather & Police Intel per Geo Region DB	Geographic Route, Legs and related Obstacles DB	User Expectations Input (Bascially Predicting the Future)
Experience [Operator]			Low, Sufficient, High [Expectation - Affect on Time]
Accountability [Operator]			Low, Sufficient, High [Expectation - Affect on Time]
Loader [Mechanism]			Poor, Fair, Good [Expectation - Affect on Time]
Weight [Robots]			Outside Tolerance, Within Tolerance [Expectation - Affect on Time]
Secure [Robots]			Loosely, Tightly, Firm [Expectation - Affect on Time]
Damage [Robots]			None, Minor, Significant [Expectation - Affect on Time]
Distance > 5 miles (distanceT)		X miles [Affect on Time]	
Distance <= 5 miles (distanceR)		Y miles [Affect on Time]	
Surface		Loose, Slippery, Firm [Affect on Time]	
Weather	Raining Hard, Raining Slightly, Sunny [Affect on Time]		
Incline		Steep Up, Flat [Affect on Time]	
Speed > 5 mph (propulationT)			Slow, Medium, Fast [Expectation - Affect on Time]
Speed <= 5 mph (propulationR)			Slow, Medium, Fast [Expectation - Affect on Time]
Stress [Robots]			Severe, Minor, None [Expectation - Affect on Time]
Identification [Recipient]			Unsure, Likely, Confident [Expectation - Affect on Time]
Access [to Recipient]		Highly Obstructed, Obstructed, Clear [Affect on Time]	
Mechanics [of Robot Arms]			Limited, Glitchy, Working [Expectation - Affect on Time]

Will any of this data be sparse or missing?

Operator guesses in real time about the future of these 3 Attributes



NAWC Weapons Division NAVAL AIR WARFARE CENTER

NAVAIR

The Need to Understand Details of the Composition of the Training Data - Training Set Alignment Test (TSAT)

Procedure for calculation:

1. Determine a scale for grading from 1 to “m,” where “m” means greatest attribute priority/significance based on operational deployed needs.
2. Identify attributes a_1 to a_n to grade, such that “n” is the number of attributes being graded out of r total attributes available. Therefore n ≤ r and n ≤ m, where grading a_i with grade “m” indicates a_i(m) is the most important attribute based on operational needs. Additionally, attribute grading range is (m-n+1) to m, consecutively, where lowest grade indicates least operationally important (possibly DOE analysis and/or SME determination).
3. Identify the n attributes that occur the most times in the training data. Using the same scale “m,” grade attributes b_1 to b_n based which attribute occurred the most often within the training set (this can be a statistical number, e.g., 70% of the time b_i attribute was used in simulations or 70% of the samples/instances were collected, e.g., images, that contained attribute b_i). Again, grade “m” indicates b_i occurred the most and (m-n+1) indicates b_j occurred the least within the training set.
4. Perform $k = \sum_{i=m-n+1}^{m}(\mathrm{i})$ and $\beta = \sum_{i=1}^{n}\left(\frac{a_i(grade)}{k}\right) * bi(grade) \le m$
5. Perform $\left(\frac{\beta}{m}\right) * 100 = \alpha\% \ge 50\%$ as a constraint

DOE Signficance Ranking for LT
6
8
7
10
9
1
3
1
2
5
4
1
1
6
1
1
1

Green is 1st Order Attributes

Yellow is 2nd Order Attributes

Compare the operationally determined ranking to the simulation occurrence ranking (see circled numbers in red)

Attributes with highest significance identified in the DOE results

Green is 1st Order Attributes

Important Note: This can be used to analyze any training set, whether synthetically generated or collected from live sources

Statistic comes from how often the attribute was used in the simulation (i.e., Occurrence/Total in Data Set from our sandbox)

DOE Signficance Ranking for LT	Simulation Ranking for LT	Load Truck (LT)		Weighted Number
6	9	P(experience | LT) =	0.702	1.35
8	6	P(accountability | LT) =	0.602	1.20
7	8	P(loader | LT) =	0.602	1.40
10	10	P(weight | LT) =	0.702	2.50
9	7	P(secure | LT) =	0.602	1.58
1		P(damage | LT) =	0.002	
3		P(distanceT | LT) =	0.202	
1		P(distanceR | LT) =	0.002	
2		P(surface | LT) =	0.402	
5		P(weather | LT) =	0.302	
4		P(incline | LT) =	0.302	
1		P(propulationT | LT) =	0.002	
1		P(propulationR | LT) =	0.002	
6		P(stress | LT) =	0.402	
1		P(identification | LT) =	0.002	
1		P(access | LT) =	0.002	
1		P(mechanics | LT) =	0.002	
		Class LT Attribute Alignment Score		80%

Determined based on this statistic

This is a weighted average calculation

Yellow is 2nd Order Attributes

Is this an acceptable value? That needs to be agreed upon. Rule of thumb: below 50% means too much mismatch.

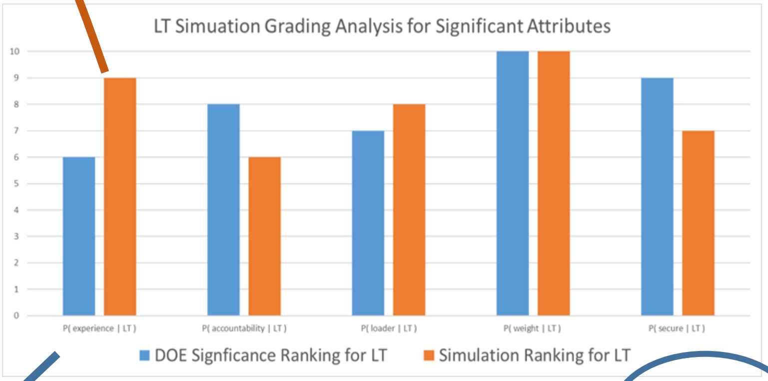
Note: This is the first measurement of the Discriminator portion of our inspired GANs

NAVAIR

How does TSAT assess Quality?

DOE Signficance Ranking for LT	Simulation Ranking for LT	P(LT) =	0.097087379	Weighted Number
6	9	P(experience | LT) =	0.702	1.35
8	6	P(accountability | LT) =	0.602	1.20
7	8	P(loader | LT) =	0.602	1.40
10	10	P(weight | LT) =	0.702	2.50
9	7	P(secure | LT) =	0.602	1.58
1		P(damage | LT) =	0.002	
3		P(distanceT | LT) =	0.202	
1		P(distanceR | LT) =	0.002	
2		P(surface | LT) =	0.402	
5		P(weather | LT) =	0.302	
4		P(incline | LT) =	0.302	
1		P(propulationT | LT) =	0.002	
1		P(propulationR | LT) =	0.002	
6		P(stress | LT) =	0.402	
1		P(identification | LT) =	0.002	
1		P(access | LT) =	0.002	
1		P(mechanics | LT) =	0.002	

This percentage describes how well what was required matches what was simulated. In this case it was 80%.



Class LT Attribute Alignment Score	80%

From the simulation results that created the training data, was what you expected in terms of precedence of data source available for an attribute represented in the simulation?

NAVAIR

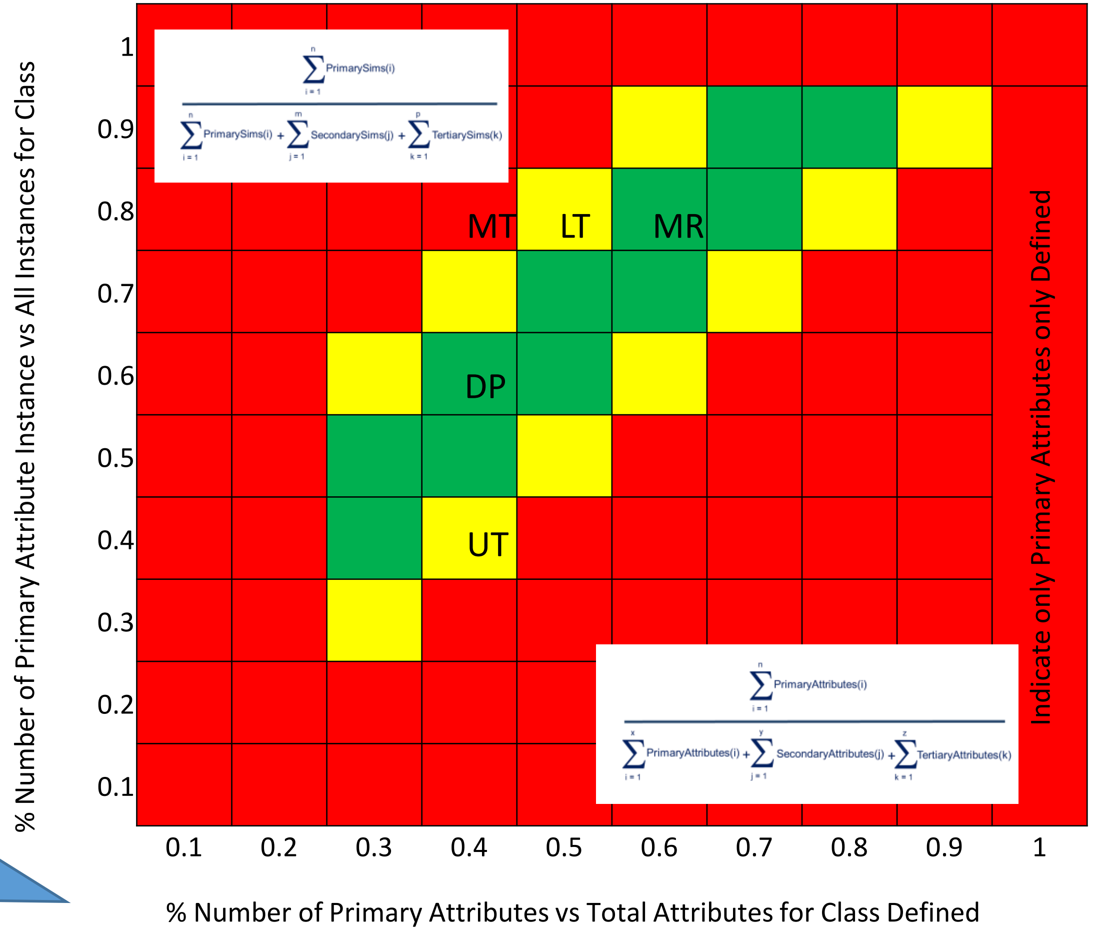
Source to Attribute Ratios for 2nd (StAR-n) Order Grouping Matrix

(Defined based on Deployed Operational Conditions transcribed into Requirements)

Instances from simulations or collected live data that created quantity of training data containing primary attributes

This matrix supports primary and secondary analysis only – used as an example

This matrix chart is for analyzing quantity regarding primary attributes to others



DOE Focus on Creating Sufficient Training Data by asking the question: How much on the Training Data should consist of Primary vs Secondary Attributes depends on Data Sources

Evidence of Simulations

+ Justification to Handle Unexpected

+ External Source to Monitor & Intercede

Note: Matrices can be created for Primary, Secondary and Tertiary attributes, not just Primary!

NAWC Weapons Division NAVAL AIR WARFARE CENTER

NAVAIR

At Requirements stage and checked during Architecture review:

- **First Step**: Create a ten by ten matrix, labeling each axis from zero to 1.
- **Second Step**: Label the horizontal axis "% Number of Primary Attributes vs Total Attributes for Class" and the vertical axis "% Number of Primary Attribute Instances vs All Instances for Class"
- **Third Step**: Determine a three-color zone scheme (see example), where green indicates that the ratio fell within acceptable limits, yellow indicates ratio is boarder line acceptable, and red color zone indicated ration is outside expected limits. Color of the zone should how well training data reflects operational environment. Based on color zone, determine evidence justification. Examples (used for guidance only) are described below:
  - Zone Green: Evidence of data by showing appropriate n-th order groups of training sets collected or generated by the simulations, including success rates as well as the TSAT results.
  - Zone Yellow: Zone Green evidence plus justification on why n-th group precedence can still handle the unexpected and provide acceptable success rates.
  - Zone Red: Zone Green and Yellow evidence as to how this algorithm is going to be supervised or monitored when operationally unexpected events occur.

When training set is produced during Algorithm code review:

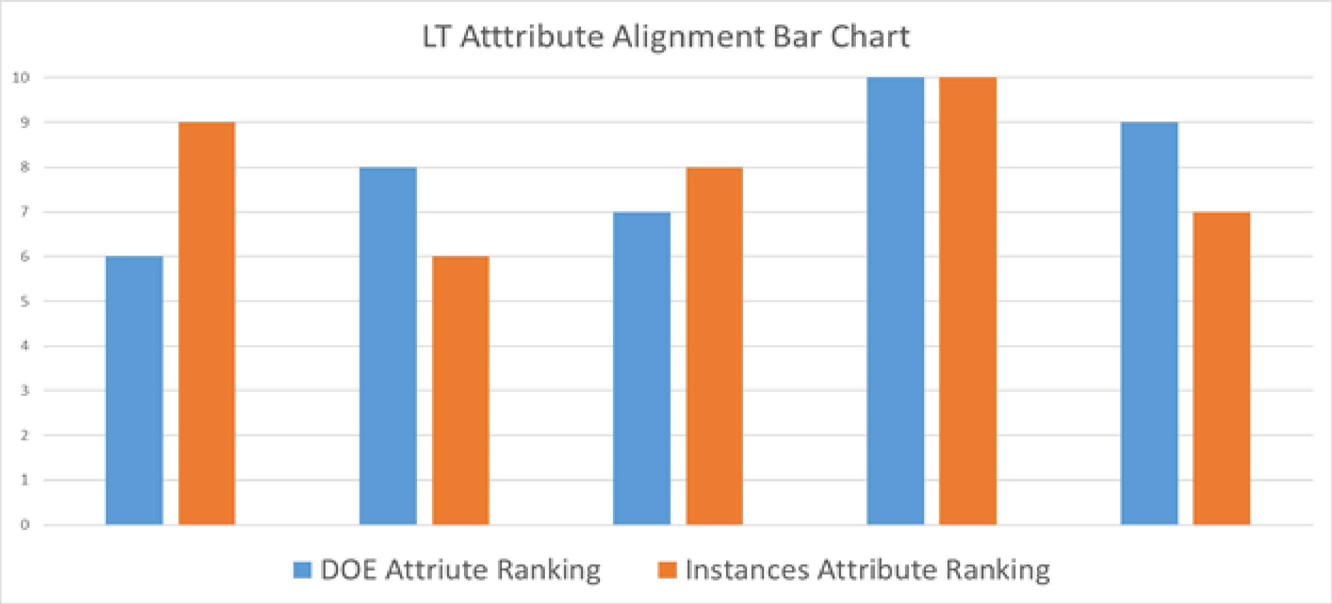
- **Fourth Step**: Calculate the σ and δ (see Figure 6 as an example) ratios. Each ratio should be less than 1. The example below is for primary attributes, but can be done for any n-th order attributes:
  - σ (by Class) = (Number of Primary Attributes / Number of All Attributes) ≤ 1.
  - δ (by Class) = (Number of all Primary Instances / Number of All Instances) ≤ 1.
- **Fifth Step**: Plot (x, y) using (σ, δ) pair of numbers and assess where the pair fall within the color zones to determine support action. See example.
  - Zone Green: Evidence of data by showing appropriate n-th order groups of training sets collected or generated by the simulations, including success rates as well as the TSAT results.
  - Zone Yellow: Zone Green evidence plus justification on why n-th group precedence can still handle the unexpected and provide acceptable success rates.
  - Zone Red: Zone Green and Yellow evidence as to how this algorithm is going to be supervised or monitored when operationally unexpected events occur.

Example of for Load Truck (LT)

TSAT

StAR-n

Acceptance Threshold Defined in Requirements



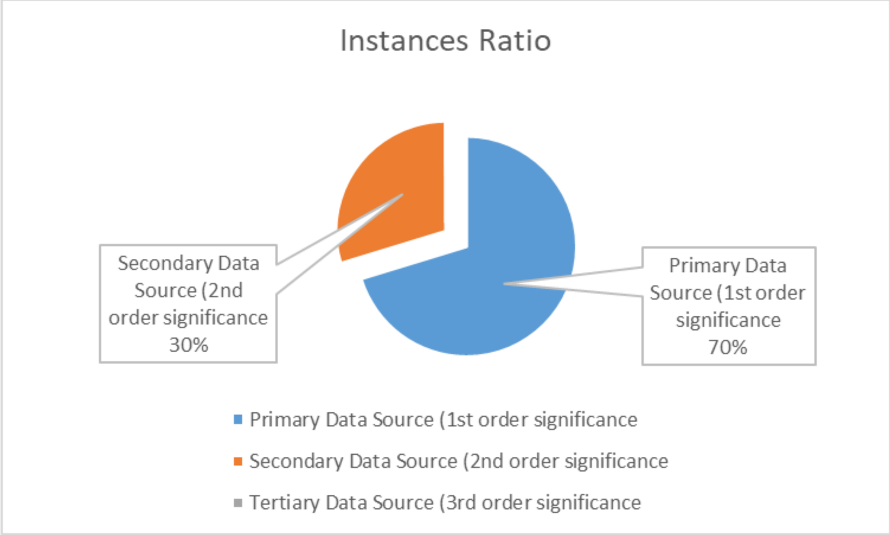
Note: Expected from DOE compared to Actuals from measuring how often the attributes were simulated

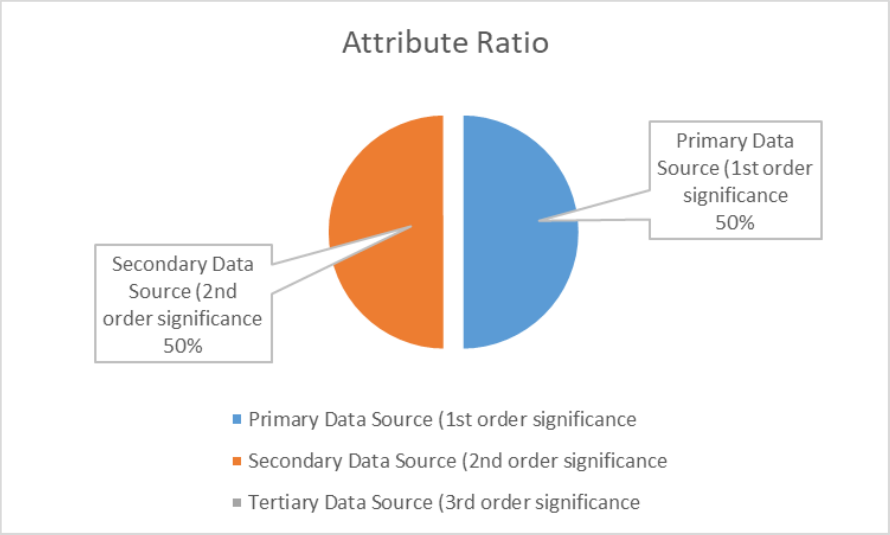
Another graph could be used for less significant variables?

What rigor score is acceptable?

Class LT Attribute Alignment Score	80%

Scale: 100% (highest level of attribute alignment between operations * training set)





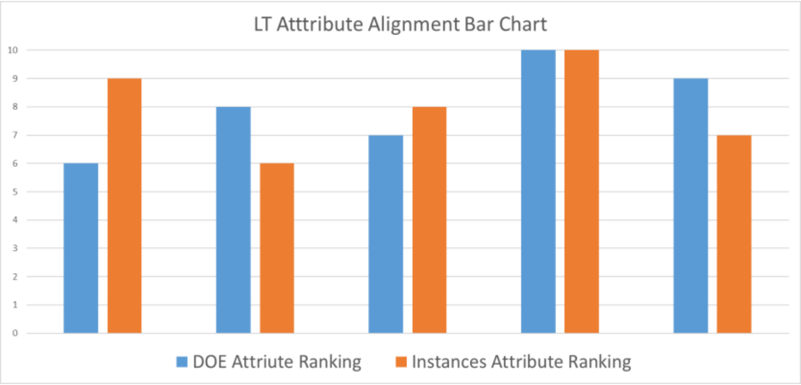
These ratios need to be defined in requirements and then compared after training set is generated

TSAT (Quality) and STAR-n (Quantity) Analysis of M&S

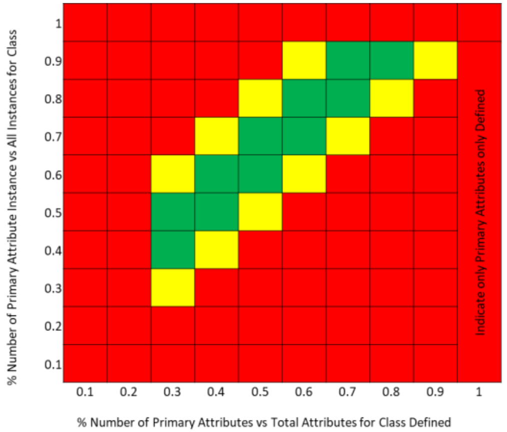
These examples would be modified to suite operational deployment conditions and then defined in requirements. Once defined, the requirements would be compared to simulation generated data set.

Both TSAT and STAR-n are used to describe how well the training data was organized using quality and quantity measurements

Does it represent the primary, secondary and tertiary data sources adequately?

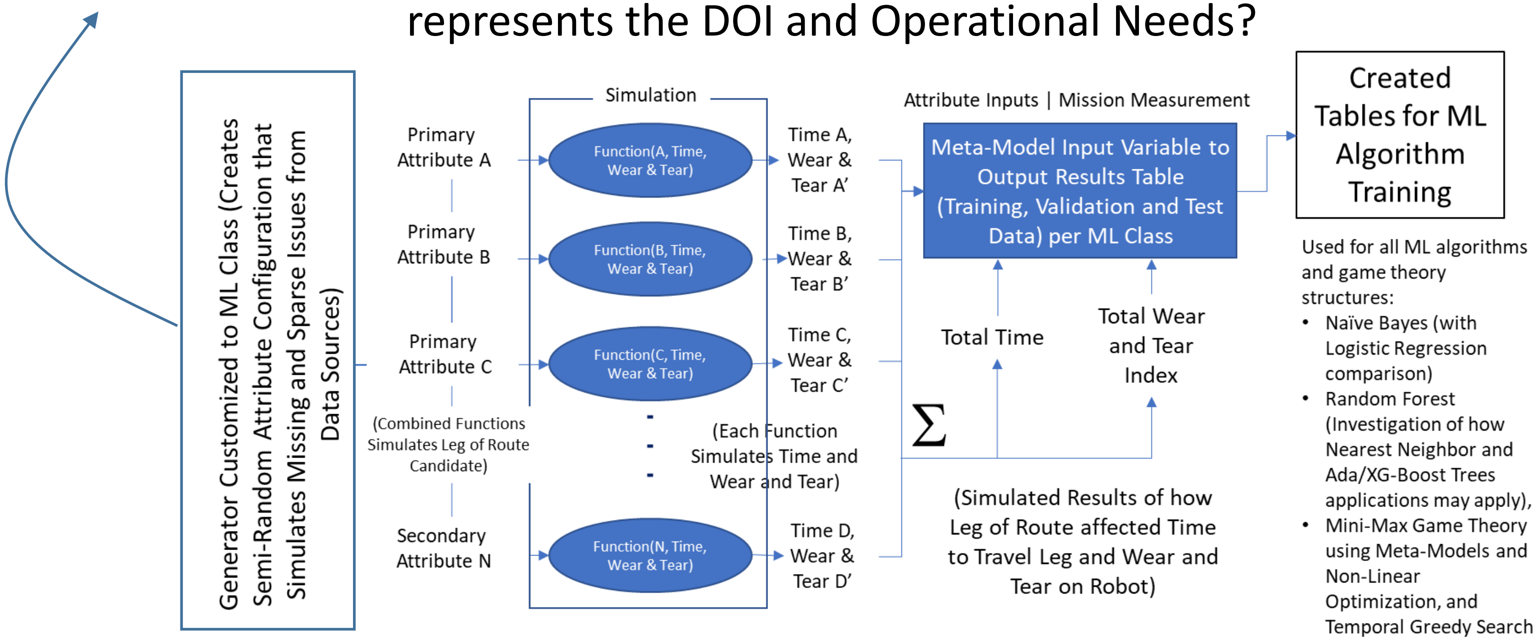


Class LT Attribute Alignment Score	80%



Does it seem reasonable with regard to ratios associated with the data sources?

Am I producing the right noise output needed to create training data that represents the DOI and Operational Needs?



Used for all ML algorithms and game theory structures:
- Naïve Bayes (with Logistic Regression comparison)
- Random Forest (Investigation of how Nearest Neighbor and Ada/XG-Boost Trees applications may apply),
- Mini-Max Game Theory using Meta-Models and Non-Linear Optimization, and Temporal Greedy Search

15,000 lines of simulated data to understand

NAWC Weapons Division NAVAL AIR WARFARE CENTER

NAVAIR

TSAT and StAR-n Concerns (Remember this is about your Training Data)…

TSAT

Example 1 (Poor focus)

Class LT Attribute Alignment Score	32%

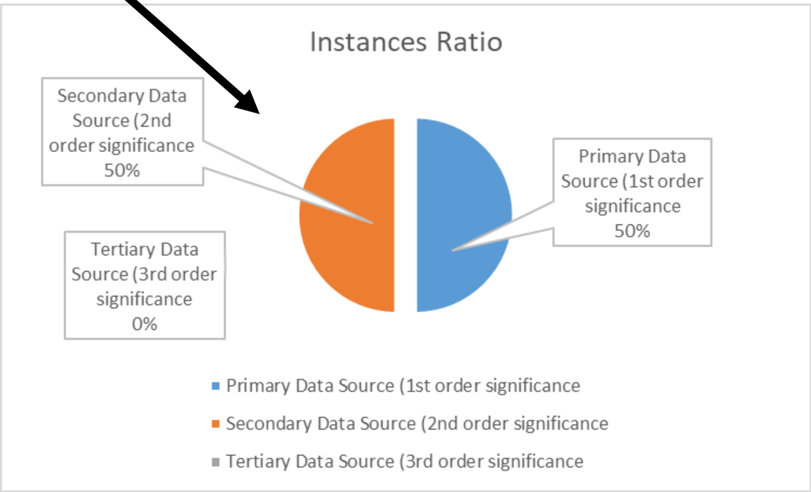
Example 2 (Marginal focus)

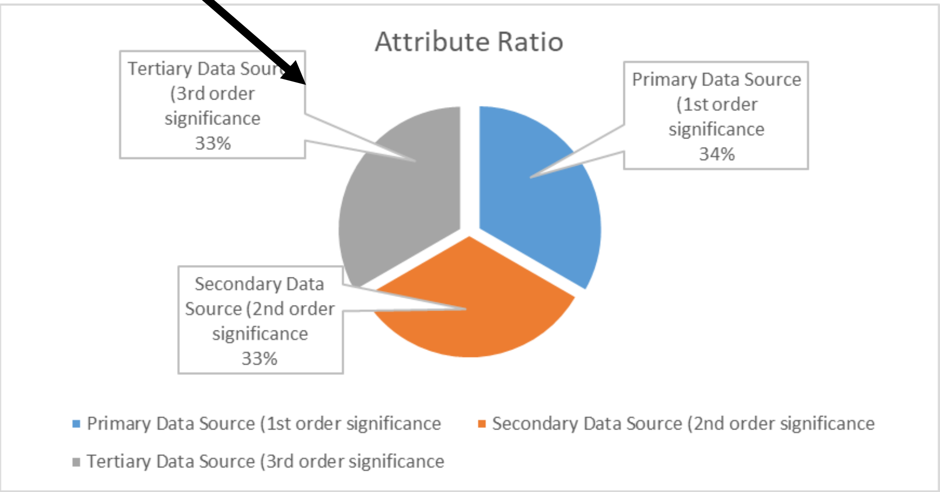
Class DP Attribute Alignment Score	53%

Issue: What DOE determined to be significant attributes is NOT what this M&S developer is focusing within his simulation model in generating a data set

StAR-n

Why are there NO simulations being preformed on 3rd order attributes when there are 3rd order attributes?

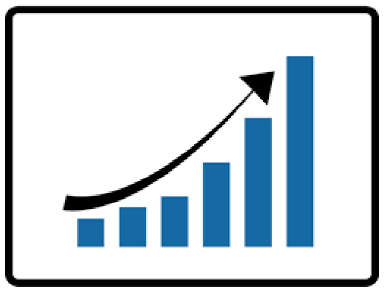




As a reminder: The order significance sources relates to primary, secondary and tertiary data sources providing the related attributes in the algorithm. In the above example, it indicates that the operational environment will have 3rd order attributes to support noisy environments, yet the M&S is not modeling that situation. Therefore, given these graphs, the ML algorithm will not be trained properly.

Naval Ordnance Safety and Security Activity (NOSSA) Addressing the “Garbage In, Garbage Out” Concern

Quality of the Training Set
(Composition)



Quantity of the Training Set
(Size)

0.633666667	0.200333333	0.483666667	0.000333333	0.000333333
0.650333333	0.000333333	0.733666667	0.000333333	0.000333333
0.667	0.350333333	0.700333333	0.000333333	0.000333333
0.467	0.233666667	0.550333333	0.217	0.333666667
0.767	0.000333333	0.417	0.000333333	0.000333333
0.000333333	0.300333333	0.667	0.433666667	0.383666667
0.467	0.583666667	0.000333333	0.000333333	0.000333333
0.000333333	0.233666667	0.000333333	0.750333333	0.367
0.283666667	0.700333333	0.383666667	0.783666667	0.333666667
0.267	0.667	0.000333333	0.700333333	0.833666667
0.183666667	0.450333333	0.250333333	0.483666667	0.533666667
0.000333333	0.633666667	0.000333333	0.000333333	0.000333333
0.000333333	0.267	0.000333333	0.733666667	0.000333333
0.250333333	0.767	0.267	0.783666667	0.750333333
0.000333333	0.000333333	0.183666667	0.000333333	0.733666667
0.000333333	0.267	0.117	0.350333333	0.700333333
0.000333333	0.000333333	0.000333333	0.300333333	0.717

Results of Using Training Set
To Develop AI/ML Algorithm
(Array of Weights)

Operational Environment

Deployed Algorithm

Confidence in Functional Performance

?

TSAT & StAR-n

References

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. arXiv:1406.2661v1.
- Goodfellow, I.J. (2017). NIPS 2016 Tutorial: Generative Adversarial Networks. ArXiv, abs/1701.00160.
- Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B.. and A Bharath, A. (2017) Generative Adversarial Networks An overview. DOI 10.1109/MSP.2017.2765202
- Foody, G., McCulloch, M. & Yate,s W. (1995) The effect of training set size and composition on artificial neural network classification, International Journal of Remote Sensing, 16:9, 1707-1723, DOI: 10.1080/01431169508954507.
- Kim, Y., Sidney, J, Buus, S., Sette1 A., Nielsen M., and Peters B. (2014) Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions, BMC Bioinformatics 2014, 15:241.
- Pei, K., Cao, Y., Yang, J., Jana, S. (2017) DeepXplore: Automated Whitebox Testing of Deep Learning Systems, ACM ISBN 978-1-4503-5085-3/17/10.
- Rodríguez-Pérez, R., Vogt M., and Bajorath, J. (2017) Influence of Varying Training Set Composition and Size on Support Vector Machine-Based Prediction of Active Compounds, Journal of Chemical Information and Modeling 2017 57 (4), 710-716 DOI: 10.1021/acs.jcim.7b00088
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P.F., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. ArXiv, abs/1606.06565.
- Everitt, T. (2018) Towards Safe Artificial General Intelligence. PhD thesis, Australian National University, pages 20-21.
- National Institute of Standards and Technology. (2019). U.S. LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools (Response to Executive Order 13859). Washington D.C.: US Department of Commerce.
- Defense Standardization Program Office. (2012) System Safety (MIL-STD 882E). Washington, D.C.: Pentagon.
- Radio Technical Commission for Aeronautics. (2012). Software Considerations in Airborne Systems and Equipment Certificatio. (DO-178C). Washington D.C.: Federal Aviation Administration.
- Joint Software Systems Safety Engineering Workgroup. (2010) Joint Software Systems Safety Engineering Handbook (SSSEH v1.0). Washington, D.C.: Pentagon.
- Wiggers, K. (2020) Waymo’s driverless cars were involved in 18 accidents over 20 months VentureBeat. https://venturebeat.com/2020/10/30/waymos-driverless-cars-were-involved-in-18-accidents-over-20-month/.
- Joint Software Systems Safety Engineering Workgroup. (2017) Software System Safety Implementing Process and Tasks Supporting MIL-STD-882E (JS-SSA-IF Rev. A). Washington, D.C.: Pentagon.
- Shneiderman, B., (2016) Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight, Proceedings of the National Academy of Sciences of the United States of America. https://doi.org/10.1073/pnas.1618211113