Tips for CDRLs/requirements when acquiring/developing AI-enabled systems

Presented for the Acquisition Research Symposium, May 11-12, 2022

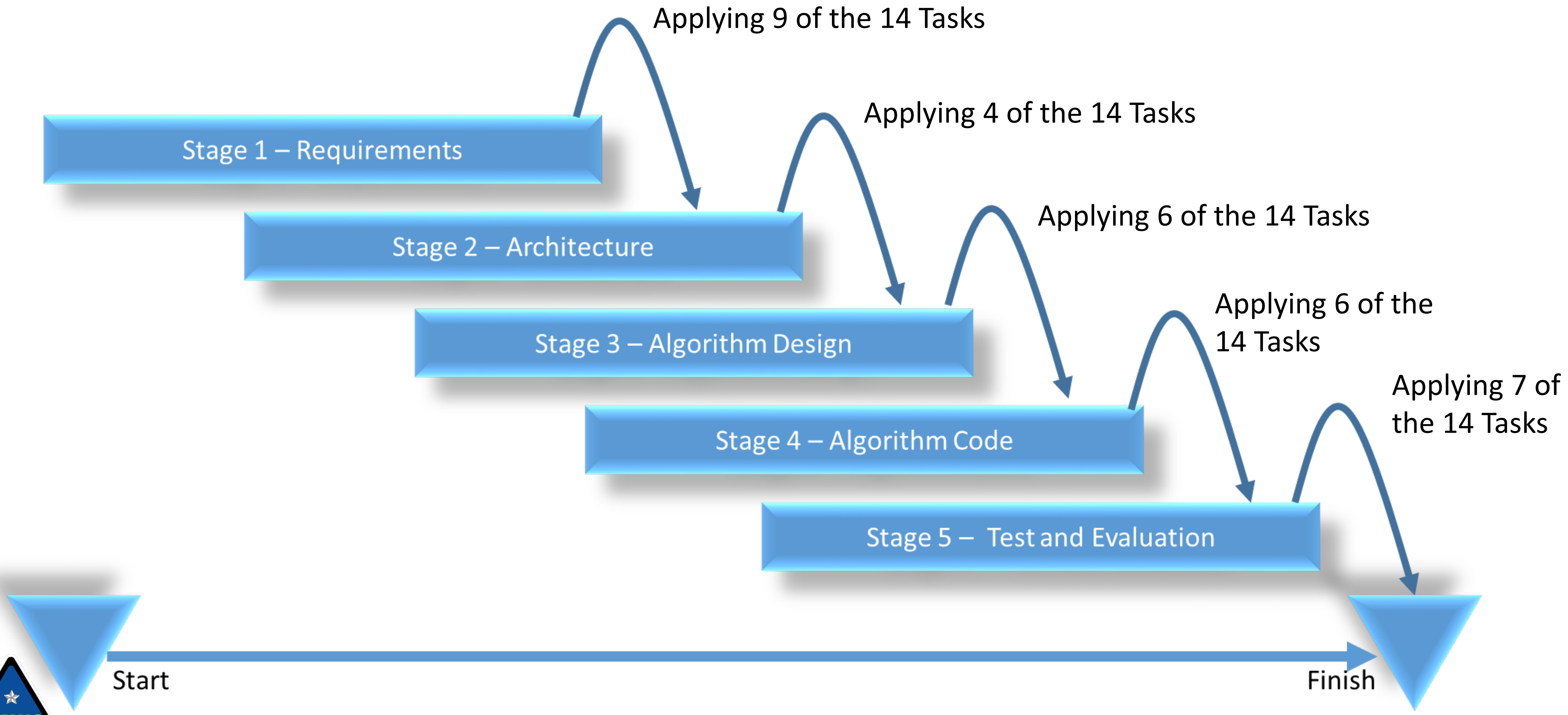
(Panel) Machine Learning: Challenges and Opportunities for the Acquisition Workforce

Presented by Bruce Nagy

bruce.m.nagy.civ@us.navy.mil

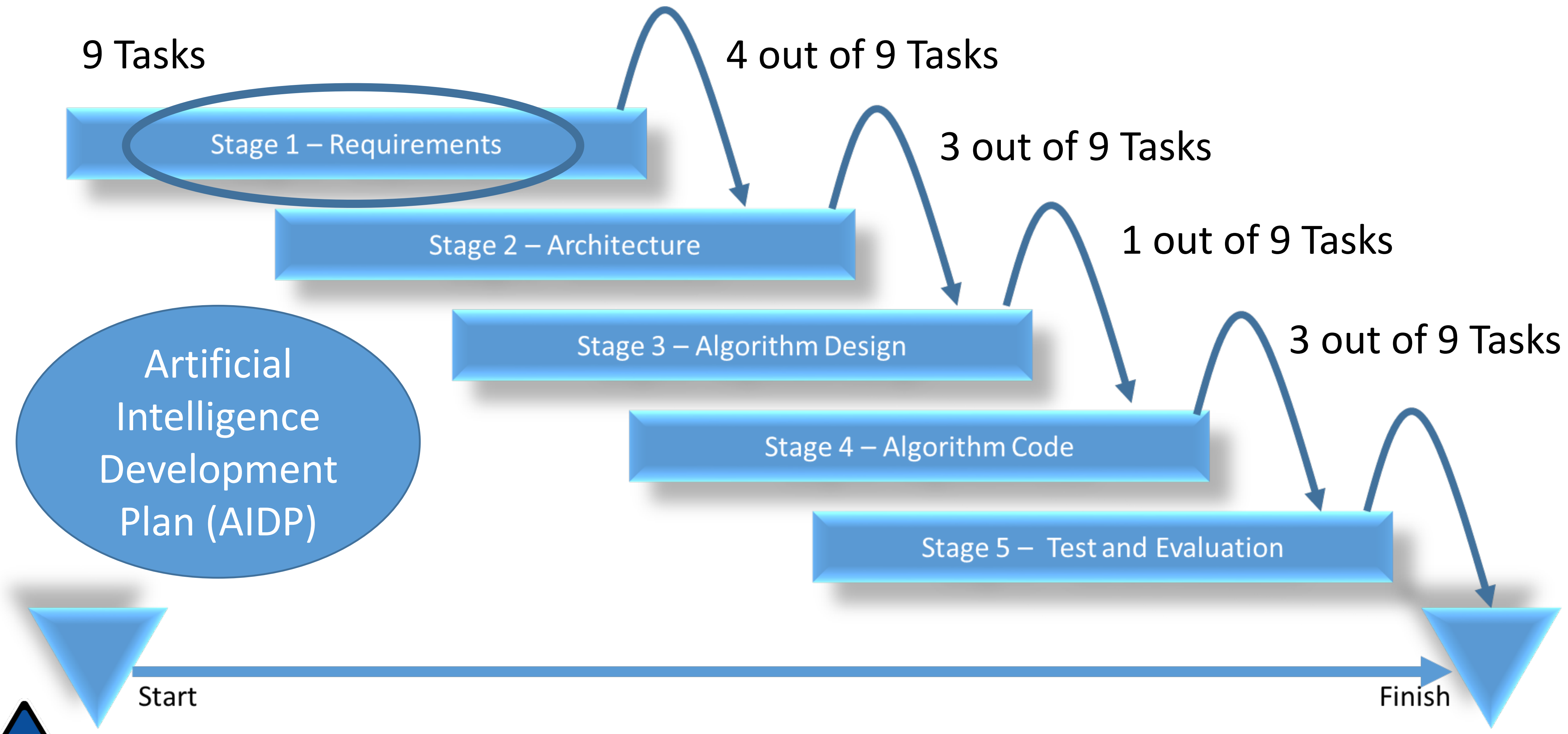
NAVAIR, NAWCWD China Lake

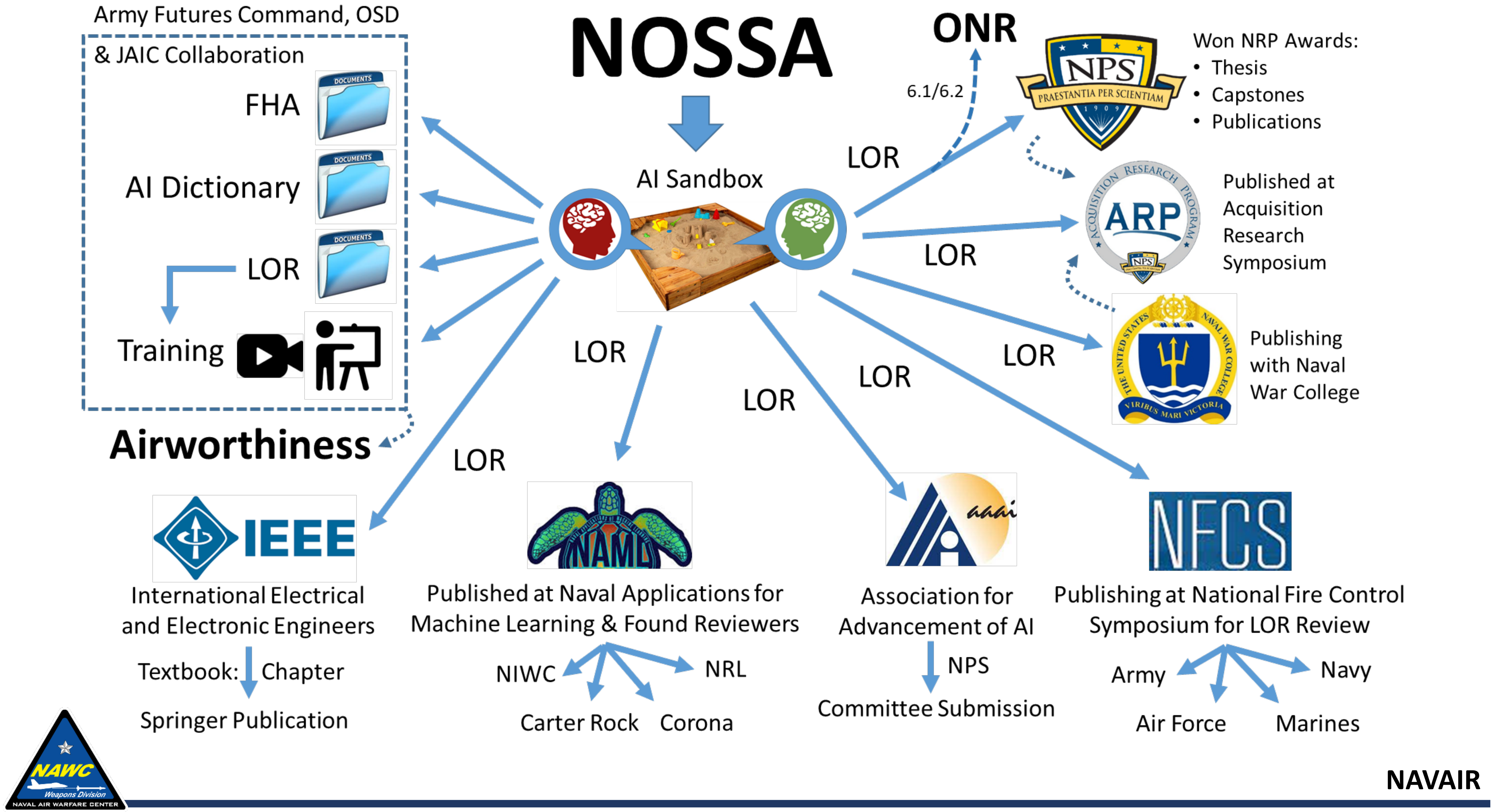
Background: AI Level of Rigor (LOR) Fourteen Tasks across Five Stages



NAVAIR

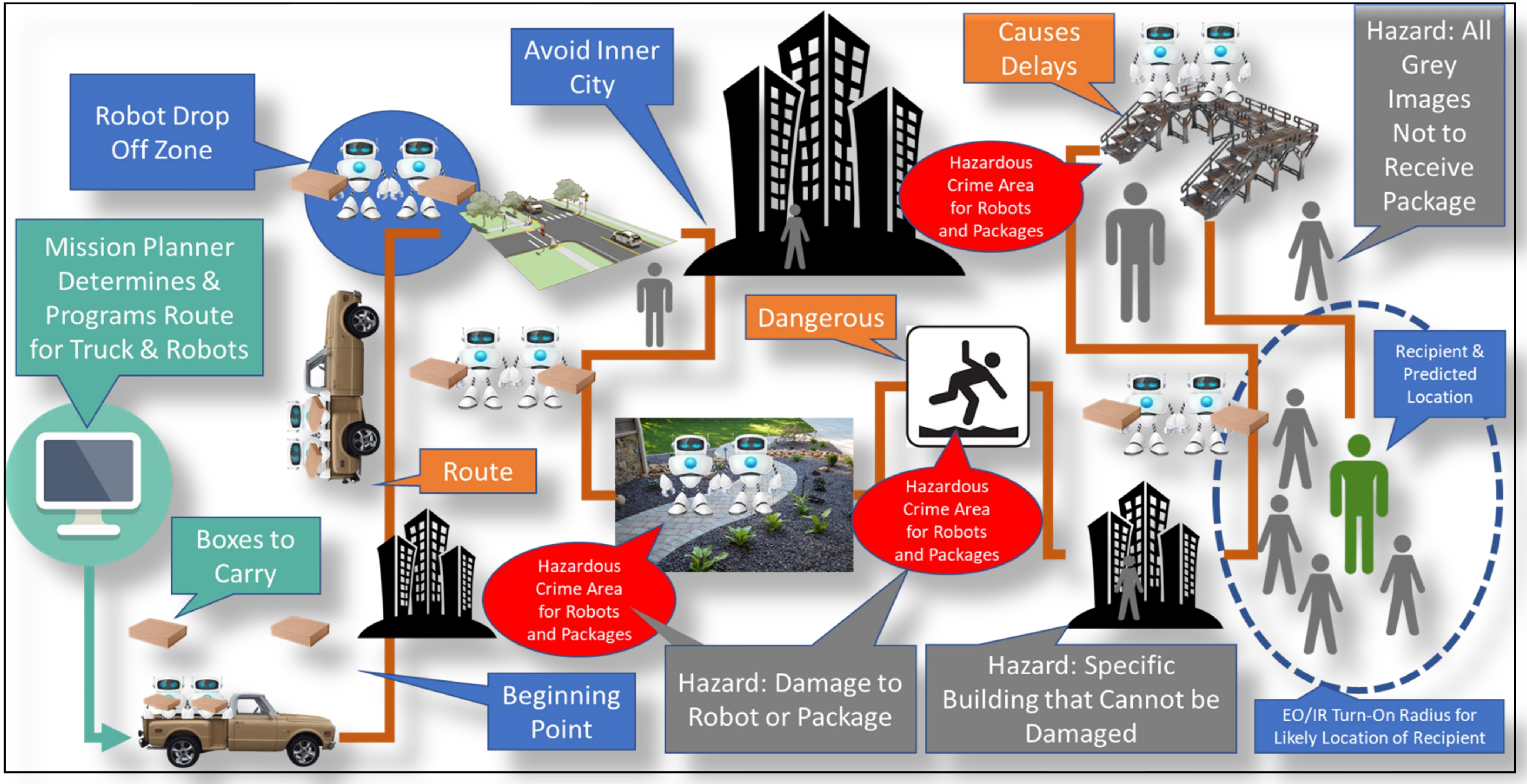
CDRL Item: First Nine Tasks of LOR Listed in AIDP affecting Five Stages





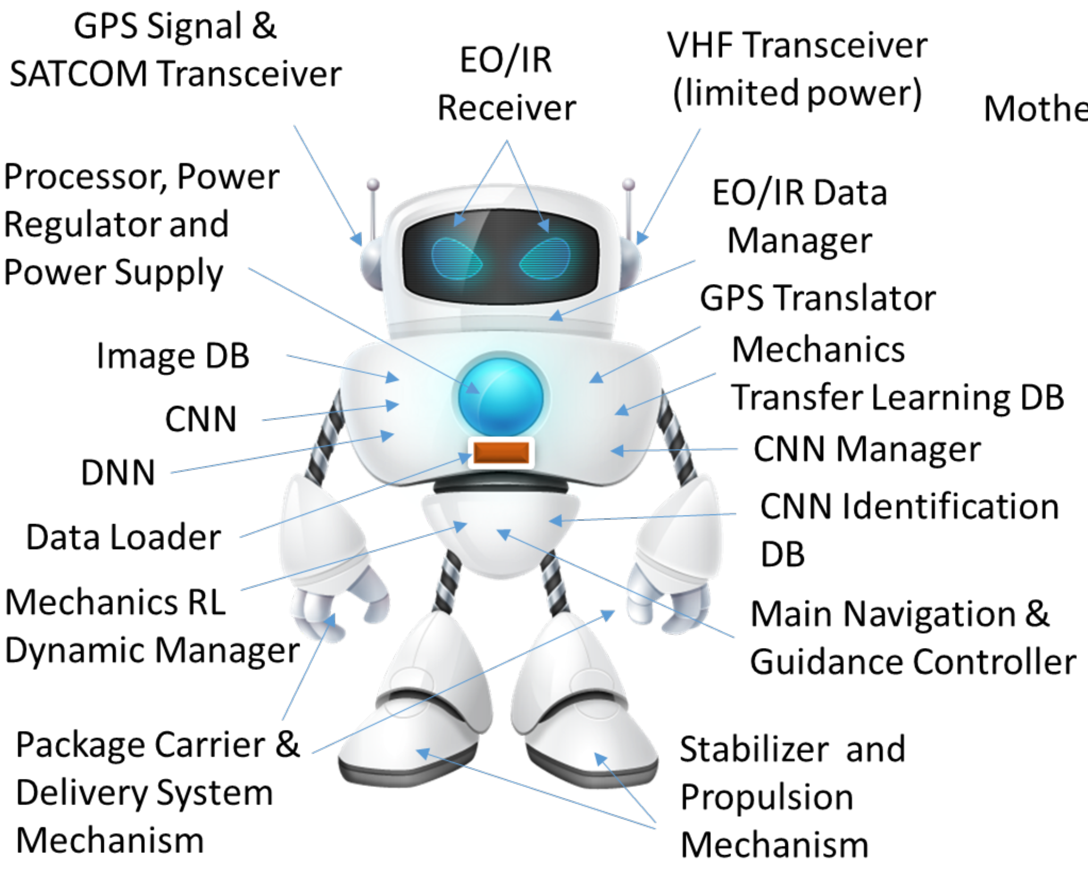
NOSSA
AI Sandbox
Army Futures Command, OSD
& JAIC Collaboration
FHA
AI Dictionary
LOR
Training
DOCUMENTS
DOCUMENTS
DOCUMENTS
Airworthiness
ONR
6.1/6.2
Won NRP Awards:
• Thesis
• Capstones
• Publications
NPS
PRAESTANTIA PER SCIENTIAM
1909
ACQUISITION RESEARCH PROGRAM
ARP
Published at Acquisition Research Symposium
Publishing with Naval War College
THE UNITED STATES NAVAL WAR COLLEGE
VIRIBUS MARI VICTORIA
LOR
LOR
LOR
LOR
LOR
LOR
LOR
IEEE
International Electrical and Electronic Engineers
Textbook: Chapter
Springer Publication
NAML
Published at Naval Applications for Machine Learning & Found Reviewers
NIWC
NRL
Carter Rock
Corona
aaai
Association for Advancement of AI
NPS
Committee Submission
NFCS
Publishing at National Fire Control Symposium for LOR Review
Army
Navy
Air Force
Marines
NAWC
Weapons Division
NAVAL AIR WARFARE CENTER
NAVAIR

Based on Sandbox Development…

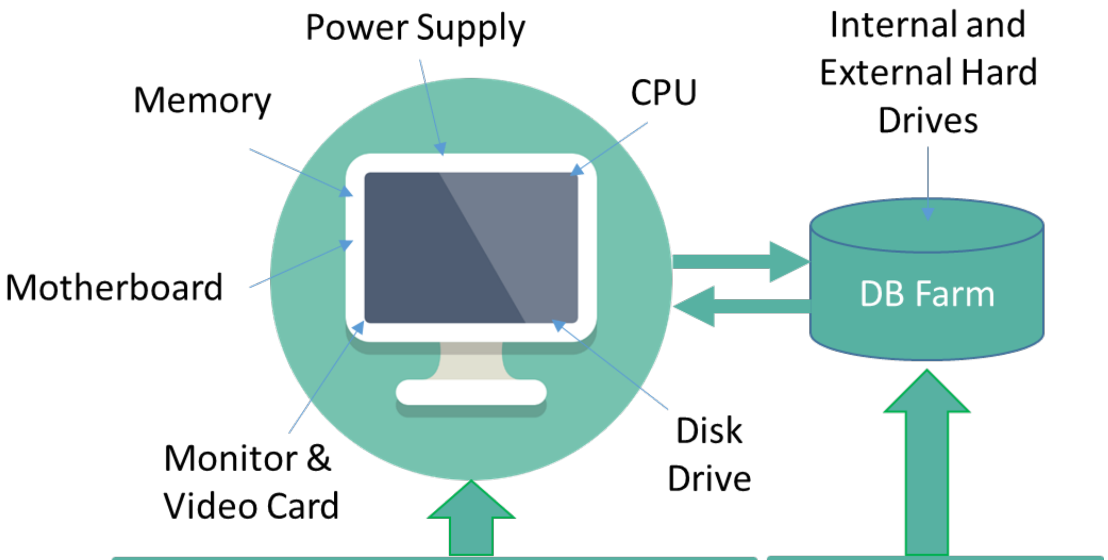


Operational View

Robot and Planner Systems



Note: Robot includes other passive sensors not listed but used to support navigation and guidance.



Loaded Software Containing the Following Code:

- Route Segments Algorithm
- Mapping Segments Algorithm
- Legs ID Algorithm
- Mapping Characteristics Algorithm
- Candidate Selection Algorithm
- Naïve Bayes for Classification Algorithm
- Random Forest Algorithm
- Naïve Bayes for Statistics Determination Algorithm
- Minimax Algorithm
- Group Legs Algorithm
- Temporal Greedy Search Algorithm
- GUI Objects
- GUI Object Manager
- Data Base Manager
- Statistical Graphics Package
- Data Loader Manager
- Data Collection & Processing Algorithm

Loaded DBs/Data Farm:

- Robot & Truck Availability & Performance Characteristics DB
- Metamodels (includes input characteristics to output probabilities) by Generic Segment and Leg DB
- Mission Segments, Constraints and Coordinates Template DB
- Upload Package DB
- Weather & Police Intel per Geo Region DB
- Geographic Route, Legs and related Obstacles DB

Nr	LOR Task	Stages Involved	Requirement (Stage 1)	Architecture (Stage 2)	Algorithm Design (Stage 3)	Algorithm Code (Stage 4)	Test and Evaluation (Stage 5)
1	AI Function Type Definition	1	x				
2	Justifying AI Function Needed	1	x				
3	Justifying AI Function Needs to be Autonomous or Semi-Autonomous	1	x				
4	Best Practice discussion about AI/ML Development	3	x	x			x
5	Training Set Alignment Test (TSAT)	3	x	x		x	
6	Sources to Attribute Ratios for 1, 2 or 3 (nth) (StAR-n) Order Matrix	3	x	x		x	
7	Best Practice discussion on data set generation	3	x	x		x	
8	k-Fold Variation	2	x				x
9	Missing and Sparse Data (MSD) Class Requirements and Actual Results Tables	3	x		x		x
10	Confusion Matrix Creation	2			x		x
11	ROC AUC or PR AUC Analysis	3			x	x	x
12	Bias, Variance and the Sweet Spot	1			x		
13	Subsystem Hazard Analysis Format	3			x	x	x
14	Best Practice discussion about AI/ML Algorithm Development	3			x	x	x

Artificial Intelligence Development Plan (AIDP)

Confidence Level	Meaning	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	LOR Tasking
Level 1	Highly Confident	X	X	X	X	X	All
Level 2	Most Confident	X	X	X		X	All
Level 3	Confident	X	X			X	1 to 9, 11
Level 4	Somewhat Confident	X				X	1 to 9, 11
Level 5	Minimally Confident	X					1 to 4

Level 1 -- *Highly Confident (All Development Stages, All Tasks)*

- Stage 1: Recommend Using Tasks 1 to 9
- Stage 2: Recommend Using Tasks 4 to 7
- Stage 3: Recommend Using Tasks 9 to 14
- Stage 4: Recommend Using Tasks 5 to 7, 11, 13, 14
- Stage 5: Recommend Using Tasks 4, 8 to 11, 13, 14

Level 2 -- *Most Confident (Stages 1, 2, 3, 5, All Tasks)*

- Stage 1: Recommend Using Tasks 1 to 9
- Stage 2: Recommend Using Tasks 4 to 7
- Stage 3: Recommend Using Tasks 9 to 14
- Stage 5: Recommend Using Tasks 4, 8 to 11, 13, 14

Level 3 -- *Confident (Stages 1, 2, 5,* Tasks *1 to 9, 11)*

- Stage 1: Recommend Using Tasks 1 to 9
- Stage 2: Recommend Using Tasks 4 to 7
- Stage 5: Recommend Using Tasks 4, 8 to 11

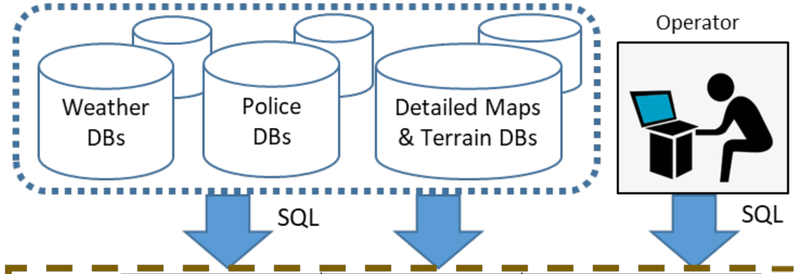
Level 4 -- *Somewhat Confident (Stages 1, 5,* Rigor *1 to 9, 11)*

- Stage 1: Recommend Using Tasks 1 to 9
- Stage 5: Recommend Using Tasks 4, 8 to 11

Level 5 -- *Minimally Confident (Stages 1, 5,* Tasks *1 o 4)*

- Stage 1: Recommend Using Tasks 1 to 4

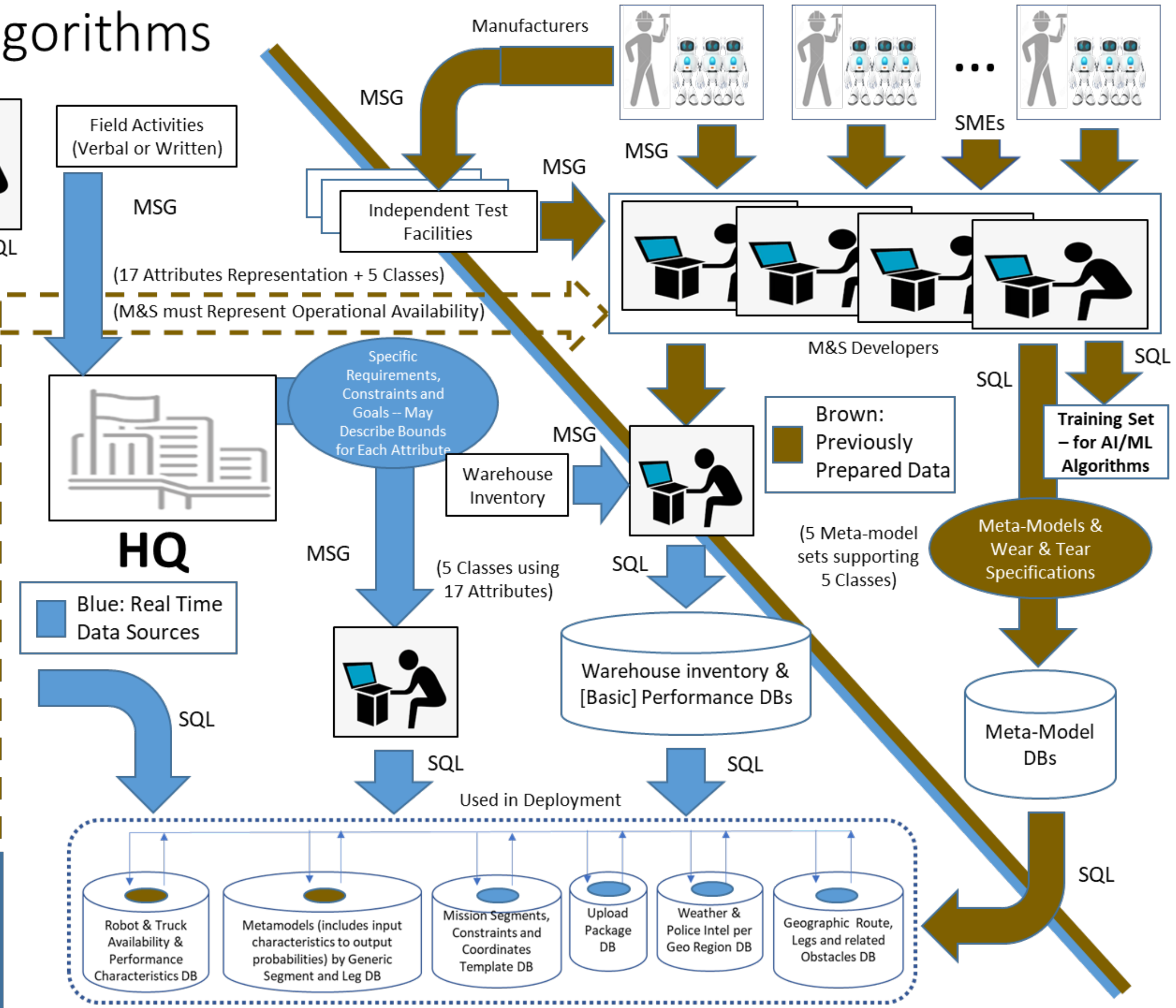
Data Sources Used by ML Algorithms



Inputs & Attributes	Used to Understand Evironment Weather & Police Intel per Geo Region DB	Used to Understand Route Geographic Route, Legs and related Obstacles DB	Used by Meta-Model (Variables to Affect Robot and Route Selecton) 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



Nr	LOR Task	Stages Involved	Requirement (Stage 1)	Architecture (Stage 2)	Algorithm Design (Stage 3)	Algorithm Code (Stage 4)	Test and Evaluation (Stage 5)
1	AI Function Type Definition	1	x				
2	Justifying AI Function Needed	1	x				
3	Justifying AI Function Needs to be Autonomous or Semi-Autonomous	1	x				

LOR Task 1: Conduct AI Type Function Definition grading of Proposed Functions that may include an AI Type algorithm in order to document decision to follow the AI rigor.

- Criteria 1 – Algorithm contain Data Approximations
- Criteria 2 – Algorithm contain Data Samples

LOR Task 2: Discuss and document a justification for the Proposed AI/ML Algorithm's development application vs. a Traditional Code development to explain why an AI/ML algorithm is a "better" fit to the function requirement.

LOR Task 3: Discuss and document a justification for the Need of an AI/ML algorithm's level of functional autonomy to explain why an AI/ML algorithm requires this level of autonomy to support the function requirement.

Nr	LOR Task	Stages Involved	Requirement (Stage 1)	Architecture (Stage 2)	Algorithm Design (Stage 3)	Algorithm Code (Stage 4)	Test and Evaluation (Stage 5)
4	Best Practice discussion about AI/ML Development	3	x	x			x

LOR Task 4: Review of Best Practices to document various discussions that need to be translated into needed requirements.

- What ML Training Data modality type are you representing in your deployed system and your data generation process?
- Does the synthetic or live data represent all the training data needed to train the algorithm to identify each label/class within the needed success rate?
- Does each Class have an appropriate number of attributes, or values, that can be learned by the algorithm for the Class/number being determined?
- How do we know that the synthetic or live data creating the training data is aligned with the mission parameters?
- How are we ensuring that the algorithm being deployed, after using training data, provides the correct answer when data input issues occur?
- Can other control entities (such as a human operator) be inserted into the loop to reduce the autonomy?
- What are the ratio requirements of *Sparse* and *Missing data* occurrences to normal operations when creating training data from synthetic or live data?
- Will the architecture, design and code support sparse and missing data management, or more specifically, will it filter or use a selection of less significant attributes to do the calculations?
- What processes are being defined, to support Data Management curation, to ensure that the ML algorithm provides accurate data input?
- How well does the particular ML algorithm support increased battle complexity and how does that affect sparse and missing data issues?

Nr	LOR Task	Stages Involved	Requirement (Stage 1)	Architecture (Stage 2)	Algorithm Design (Stage 3)	Algorithm Code (Stage 4)	Test and Evaluation (Stage 5)
5	Training Set Alignment Test (TSAT)	3	x	x		x	

LOR Task 5: For each ML class, define requirements that rank the importance of attributes, i.e., creating a priority list, within each instance that the AI algorithm will be trained to recognize. This ranking represents a baseline to determine if a quality training set is being used.

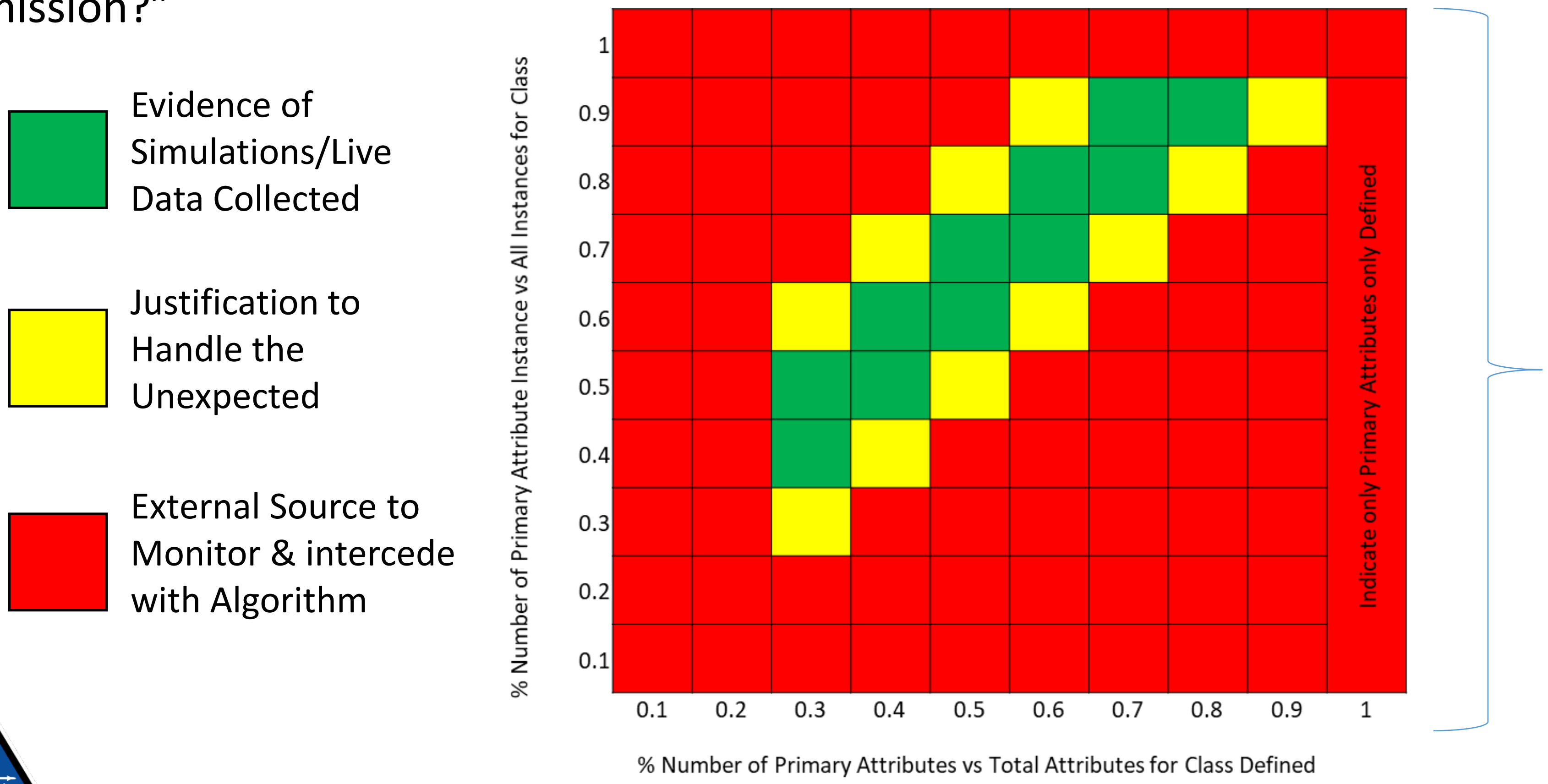
DOE Signficance Ranking for LT
6
8
7
10
9
1
3
1
2
5
4
1
1
6
1
1
1

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	

Training Set Alignment Test (TSAT)

Nr	LOR Task	Stages Involved	Requirement (Stage 1)	Architecture (Stage 2)	Algorithm Design (Stage 3)	Algorithm Code (Stage 4)	Test and Evaluation (Stage 5)
6	Sources to Attribute Ratios for 1, 2 or 3 (nth) (StAR-n) Order Matrix	3	x	x		x	

LOR Task 6: Once attributes are ranked in terms of priority, the next question should be, "Does the ranking indicate a grouping of attributes based on the importance and availability of data during a mission?"



Training Set Alignment Test (TSAT)

Nr	LOR Task	Stages Involved	Requirement (Stage 1)	Architecture (Stage 2)	Algorithm Design (Stage 3)	Algorithm Code (Stage 4)	Test and Evaluation (Stage 5)
7	Best Practice discussion on data set generation	3	x	x		x	

LOR Task 7: Review of Best Practice to document discussions on data set generation or data set collection (from "live" data) to identify needed requirements.

- How do you know if the quality and quantity of Training Data is sufficient?
- How do you assess the operational limits described by the training data? (Consider the "You don't know what you don't know" issue.)
- Did the training set include enough noise/clutter for each class (in this case, less significant attributes determined by SMEs for a particular meta-model class) to ensure that the function works properly when deployed? Are there sparse data and/or mission data issues? How is the bias of the training set and variance of the test results determined?
- For simulation generation of the training data:
    - How would you ensure synthetic or live data configurations work, i.e., is the training data covering the real-world experiences? (Optimizing bias [how well it fits the training set] and variance [how well it predicts using the test set], including considerations of overfitting/under-fitting).
    - What quality of synthetic or live training data, i.e., attribute composition on each instance, and how many of these various compositions are really enough to train an algorithm?

Nr	LOR Task	Stages Involved	Requirement (Stage 1)	Architecture (Stage 2)	Algorithm Design (Stage 3)	Algorithm Code (Stage 4)	Test and Evaluation (Stage 5)
8	k-Fold Variation	2	X				X
9	Missing and Sparse Data (MSD) Class Requirements and Actual Results Tables	3	X		X		X

LOR Task 8: Has a process been identified to ensure that randomly selected T&E data is available for testing from the curated training data before any developer uses it? If not, why not?

CM + BLOCKCHAIN

LOR Task 9: Will a Missing and Sparse Data (MSD) Class Requirements Table, consisting of four sections, be used?

- **Section 1:** Create a table or list by class the expected training data quantities/numbers based on ML Training Data.
- **Section 2:** Create a table or list that describes, within the training set, an expected percentage of how often primary attributes occur in an instance/sample compared to the total number of instances being used for training.
- **Section 3:** Create a table or list that describes the expected success rate when combining attributes from various priority groups of the algorithm (e.g., as a percentage).
- **Section 4:** Create a table or list that provides an expected majority or minority class analysis of how balanced (equal quantities) the classes are with each other.

Nr	LOR Task	Stages Involved	Requirement (Stage 1)	Architecture (Stage 2)	Algorithm Design (Stage 3)	Algorithm Code (Stage 4)	Test and Evaluation (Stage 5)
1	AI Function Type Definition	1	X				
2	Justifying AI Function Needed	1	X				
3	Justifying AI Function Needs to be Autonomous or Semi-Autonomous	1	X				
4	Best Practice discussion about AI/ML Development	3	X	X			X
5	Training Set Alignment Test (TSAT)	3	X	X		X	
6	Sources to Attribute Ratios for 1, 2 or 3 (nth) (StAR-n) Order Matrix	3	X	X		X	
7	Best Practice discussion on data set generation	3	X	X		X	
8	k-Fold Variation	2	X				X
9	Missing and Sparse Data (MSD) Class Requirements and Actual Results Tables	3	X		X		X
10	Confusion Matrix Creation	2			X		X
11	ROC AUC or PR AUC Analysis	3			X	X	X
12	Bias, Variance and the Sweet Spot	1			X		
13	Subsystem Hazard Analysis Format	3			X	X	X
14	Best Practice discussion about AI/ML Algorithm Development	3			X	X	X

Artificial Intelligence Development Plan (AIDP)