Statistical Thinking in DoD Test & Evaluation: F-35 Case Study

Dr. Laura Freeman

IDA

Improving Operational Testing: A case study from my past 8 years

Goal of Operational Test: Evaluate Operational Effectiveness, Suitability, and Survivability

Operational Environment

Representative Users

"Real" Threats

Conducting Missions

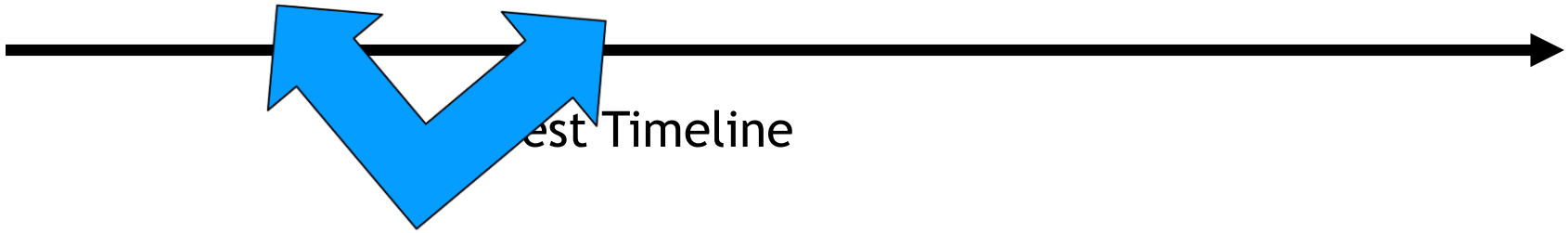
IDA

DoD Test Paradigm In Terms of Your New Corolla

Contractor Testing

Developmental Testing

Operational Testing



Tend to be requirements driven

Requirements documents are often missing important mission considerations

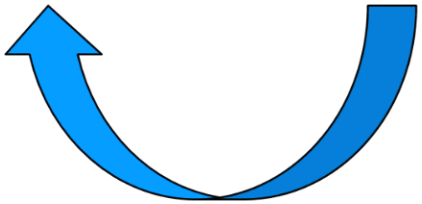
DoD Test Paradigm In Terms of Your New Corolla

Contractor Testing

Developmental Testing

Operational Testing

Test Timeline



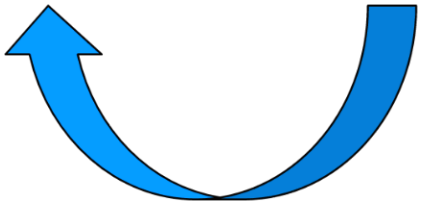
DoD Test Paradigm In Terms of Your New Corolla

Contractor Testing

Developmental Testing

Operational Testing

Test Timeline



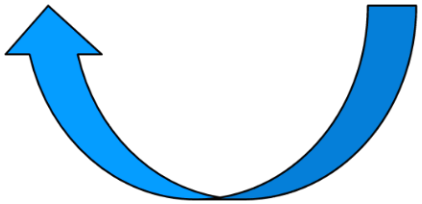
DoD Test Paradigm In Terms of Your New Corolla

Contractor Testing

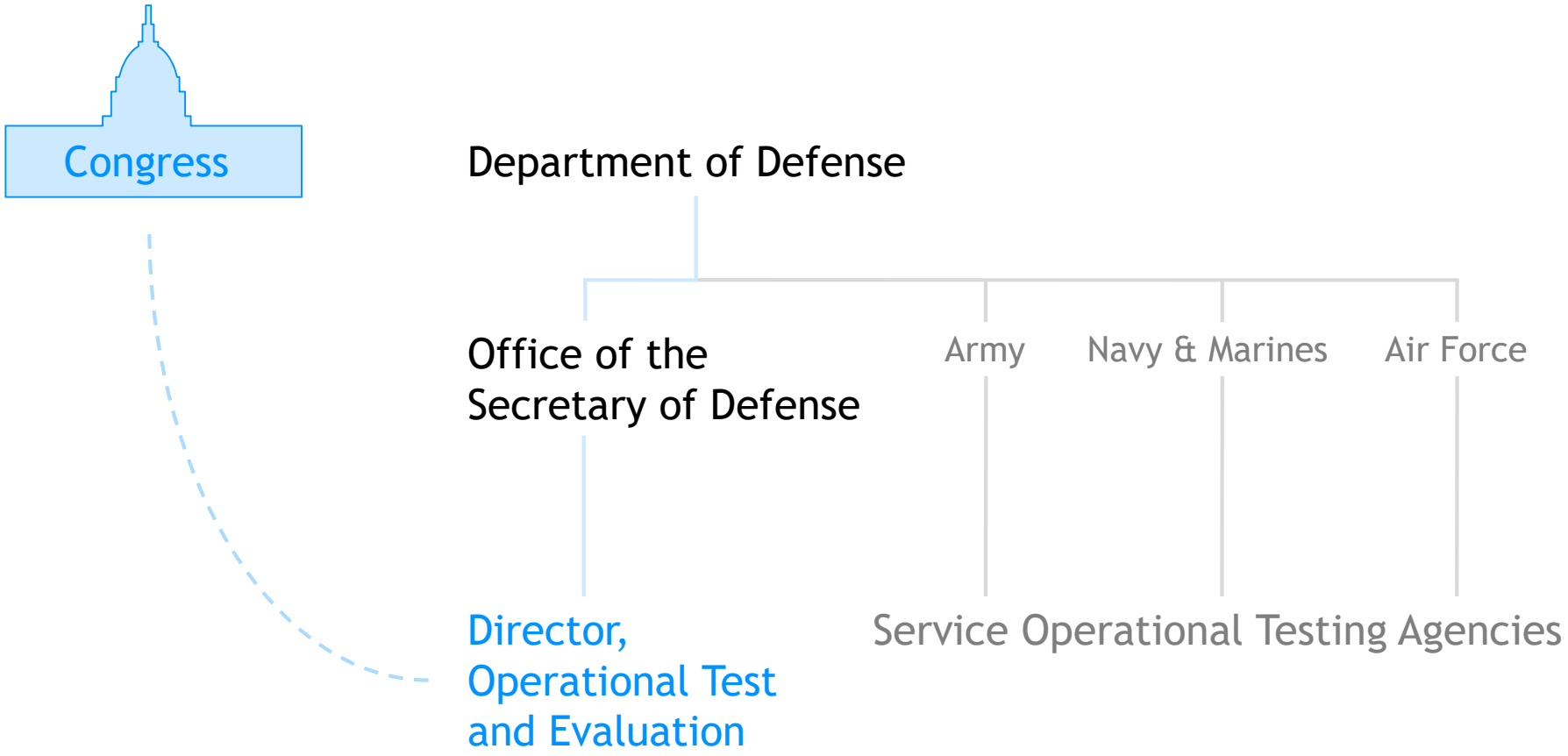
Developmental Testing

Operational Testing

Test Timeline

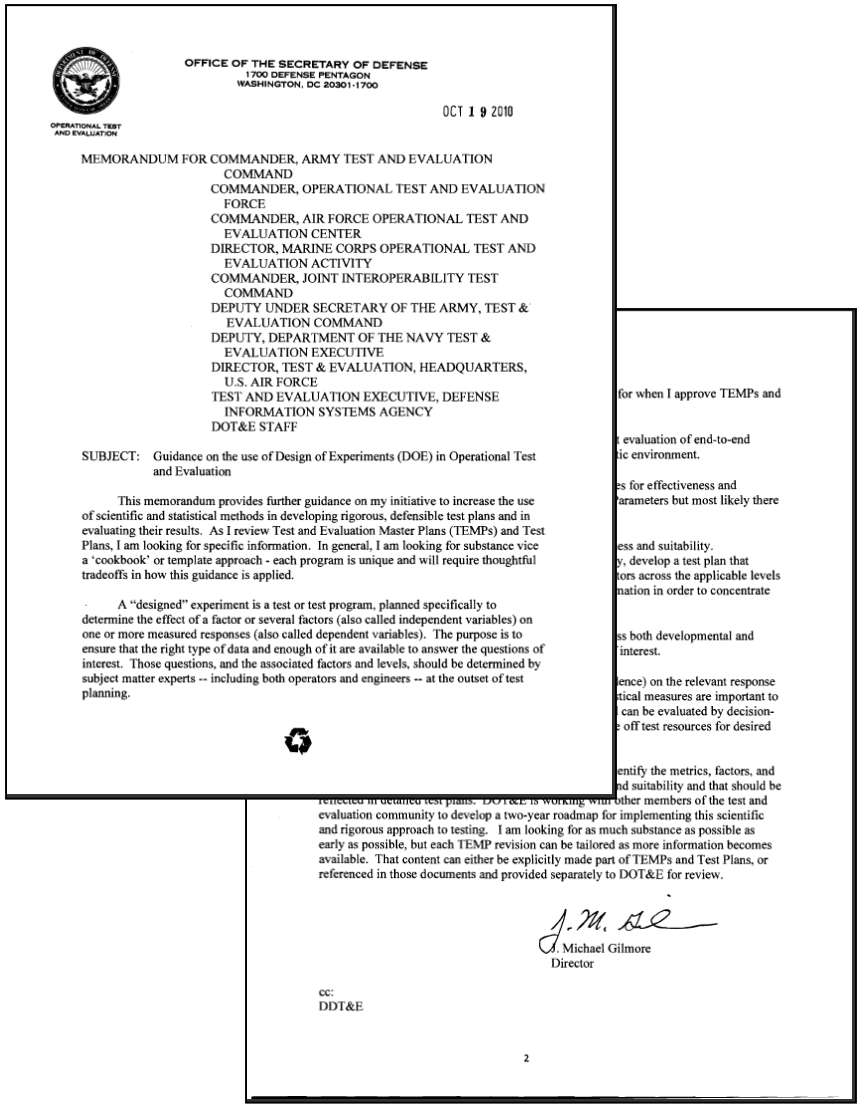


Congress established DOT&E separate from the Services' operational testing agencies



IDA

DOT&E Sets Policy and Guidance for Conducting Operational Testing



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST AND EVALUATION

OCT 19 2010

MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND
DEPUTY UNDER SECRETARY OF THE ARMY, TEST & EVALUATION COMMAND
DEPUTY, DEPARTMENT OF THE NAVY TEST & EVALUATION EXECUTIVE
DIRECTOR, TEST & EVALUATION, HEADQUARTERS, U.S. AIR FORCE
TEST AND EVALUATION EXECUTIVE, DEFENSE INFORMATION SYSTEMS AGENCY
DOT&E STAFF

SUBJECT: Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation

This memorandum provides further guidance on my initiative to increase the use of scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results. As I review Test and Evaluation Master Plans (TEMPs) and Test Plans, I am looking for specific information. In general, I am looking for substance vice a 'cookbook' or template approach - each program is unique and will require thoughtful tradeoffs in how this guidance is applied.

A "designed" experiment is a test or test program, planned specifically to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). The purpose is to ensure that the right type of data and enough of it are available to answer the questions of interest. Those questions, and the associated factors and levels, should be determined by subject matter experts -- including both operators and engineers -- at the outset of test planning.

for when I approve TEMPs and

evaluation of end-to-end
ic environment.

s for effectiveness and
arameters but most likely there

ess and suitability.
y, develop a test plan that
tors across the applicable levels
nation in order to concentrate

ss both developmental and
interest.

ence) on the relevant response
tical measures are important to
can be evaluated by decision-
off test resources for desired

entify the metrics, factors, and
nd suitability and that should be
reflected in detailed test plans. DOT&E is working with other members of the test and evaluation community to develop a two-year roadmap for implementing this scientific and rigorous approach to testing. I am looking for as much substance as possible as early as possible, but each TEMP revision can be tailored as more information becomes available. That content can either be explicitly made part of TEMPs and Test Plans, or referenced in those documents and provided separately to DOT&E for review.

J. Michael Gilmore
Director

cc:
DDT&E

2

- **The goal of the experiment**. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- Quantitative mission-oriented **response variables** for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will be others.)
- **Factors** that affect those measures of effectiveness and suitability. Systematically, in a rigorous and structured way, develop a test plan that provides good breadth of coverage of those factors across the applicable levels of the factors, taking into account known information in order to concentrate on the factors of most interest.
- **A method for strategically varying factors** across both developmental and operational testing with respect to responses of interest.
- **Statistical measures of merit (power and confidence)** on the relevant response variables for which it makes sense. These statistical measures are important to understanding "how much testing is enough?" and can be evaluated by decision makers on a quantitative basis so they can trade off test resources for desired confidence in results.

IDA

Kotter's Process for Leading Change

1. Establish a sense of urgency
2. Form a powerful coalition
3. Create a vision
4. Communicate the vision
5. Empower others to act
6. Create short term wins
7. Consolidate improvements and produce more change
8. Institutionalize new approaches

Project Campions

Rigor and Objectivity in T&E: An Update

J. Michael Gilmore, Ph.D.
Director, Operational Test and Evaluation,
Office of the Secretary of Defense, Washington, D.C.

The Director of Operational Test and Evaluation (OT&E) began four Test and Evaluation (T&E) initiatives after his confirmation by Congress in fall 2009. Underlying his four initiatives were the need for rigorous and objective T&E. Since his original initiatives the Director has advocated for the use of statistically designed experiments as a methodology for increasing the rigor of test planning resulting in efficient tests yielding statistically defensible results. Additionally, he continues to emphasize the need for reliable systems and reliability growth plans and accordingly defensible reliability growth models in T&E.

I began my term as the Director of Operational Test & Evaluation (DOT&E) with four initiatives to increase scientific rigor in T&E. I published those initiatives in the June 2010, *ITEA Journal*, and I am happy to use this opportunity to provide an update. During the past year, I have seen several success stories as well as areas for improvement. I would like to commend ITEA for the theme of this journal, "The Rigor of the Scientific Method." And I appreciate the many articles others have authored on applying rigorous and objective scientific approaches to their specific test challenges.

J. Michael Gilmore, Ph.D.

associated with the test results. Finally, DOE provides the tester with methods for developing and analyzing sequences of tests. Before testing, DOE enables decision makers to clearly see the tradeoffs between test resources and risk. During testing, DOE enables testers to use early results to strengthen and refine subsequent tests. After testing, DOE gives decision makers a framework for understanding and weighing the importance of the results.

In October 2010, I outlined the specific elements of DOE that I am looking for when I review TEMPs and test plans.

Project Campions

Rigor and Objectivity in T&E: An Update

J. Michael Gilmore, Ph.D.
Director, Operational Test and Evaluation,
Office of the Secretary of Defense, Washington, D.C.

The Director of Operational Test and Evaluation (OT&E) began four Test and Evaluation (T&E) initiatives after his confirmation by Congress in fall 2009. Underlying his four initiatives were the need for rigorous and objective T&E. Since his original initiatives the Director has advocated for the use of statistically designed experiments as a methodology for increasing the rigor of test planning resulting in efficient tests yielding statistically defensible results. Additionally, he continues to emphasize the need for reliable systems and reliability growth plans and accordingly defensible reliability growth models in T&E.

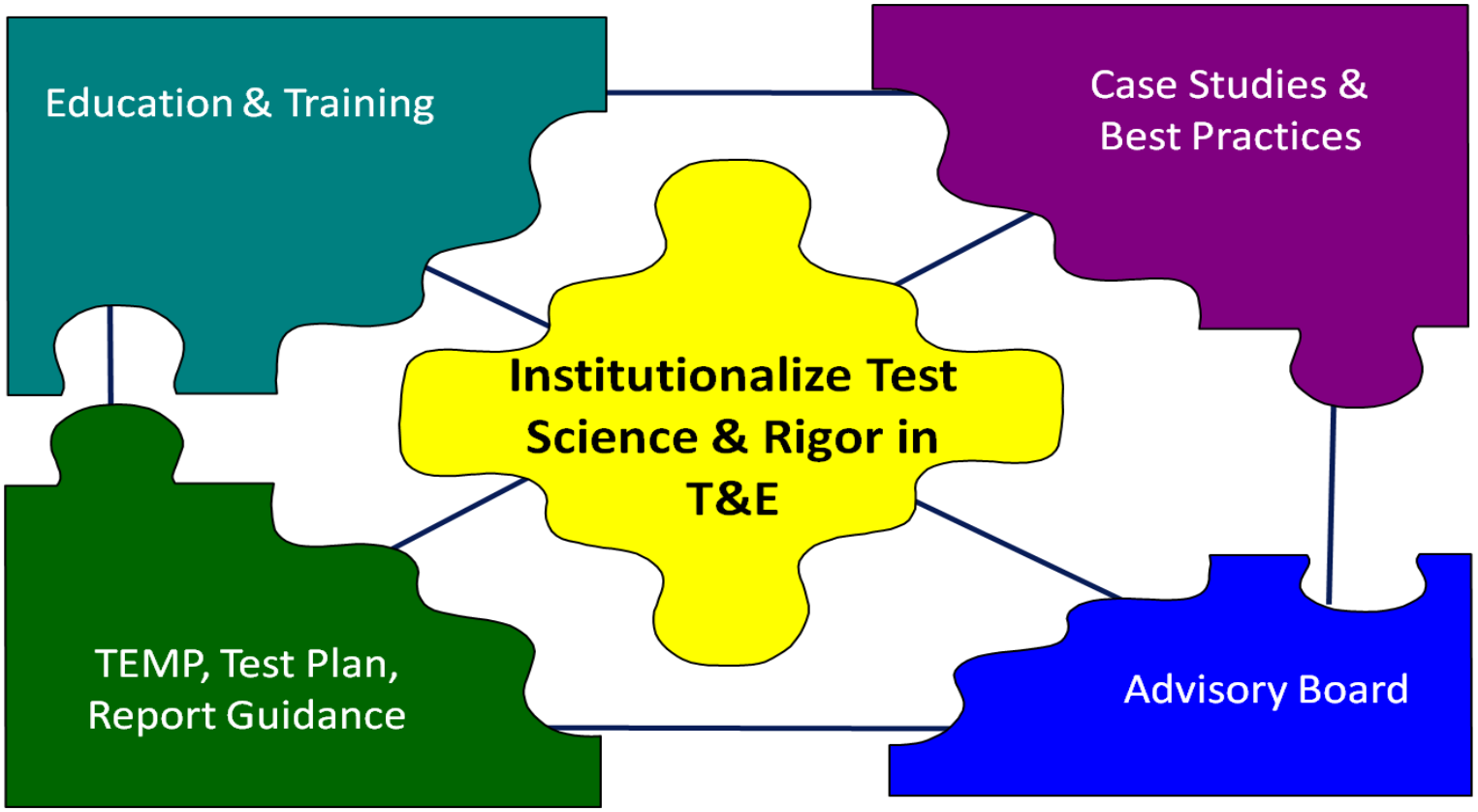
I began my term as the Director of Operational Test & Evaluation (DOT&E) with four initiatives to increase scientific rigor in T&E. I published those initiatives in the June 2010, *ITEA Journal*, and I am happy to use this opportunity to provide an update. During the past year, I have seen several success stories as well as areas for improvement. I would like to commend ITEA for the theme of this journal, "The Rigor of the Scientific Method." And I appreciate the many articles others have authored on applying rigorous and objective scientific approaches to their specific test challenges.

J. Michael Gilmore, Ph.D.

associated with the test results. Finally, DOE provides the tester with methods for developing and analyzing sequences of tests. Before testing, DOE enables decision makers to clearly see the tradeoffs between test resources and risk. During testing, DOE enables testers to use early results to strengthen and refine subsequent tests. After testing, DOE gives decision makers a framework for understanding and weighing the importance of the results.

In October 2010, I outlined the specific elements of DOE that I am looking for when I review TEMPs and test plans.

Strategic Plan



Design of Experiments for Test Planning
F-35 Case Study

The F-35 Program is Complex even by DoD Standards

Conventional	Short takeoff/vertical landing	Carrier variant
		

And Required to Accomplish Many Diverse Missions

Conventional	Short takeoff/vertical landing	Carrier variant
		

Mission Areas	Air Threat	Ground Threat
Air-Surface Strike		
Destruction/Suppression of Enemy Air Defenses		
Defensive counter air		
Offensive counter air		
Close air support		
Search and rescue		

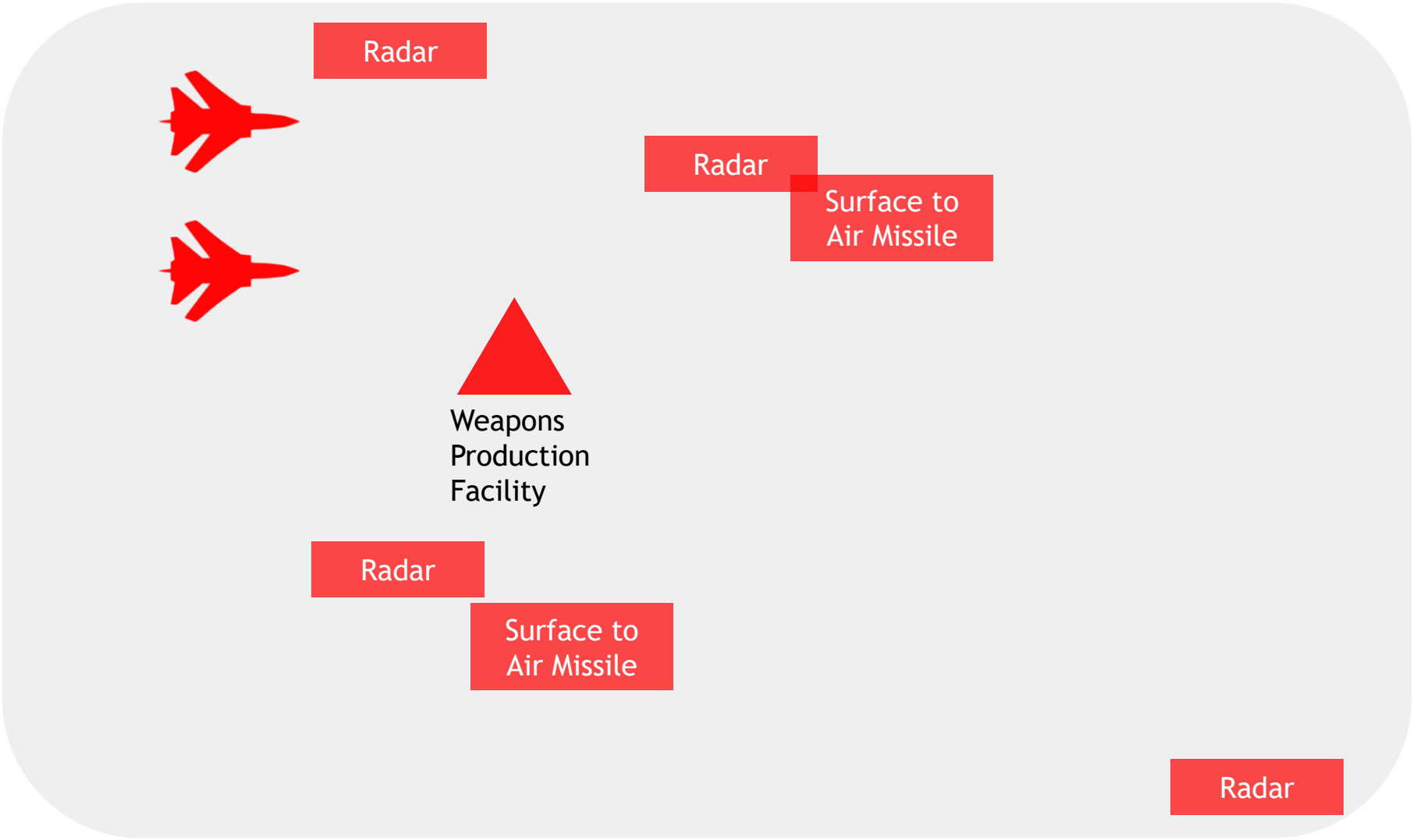
Problem Identification

How do you evaluate the F-35's ability to accomplish a diverse set of operational missions with limited test resources?

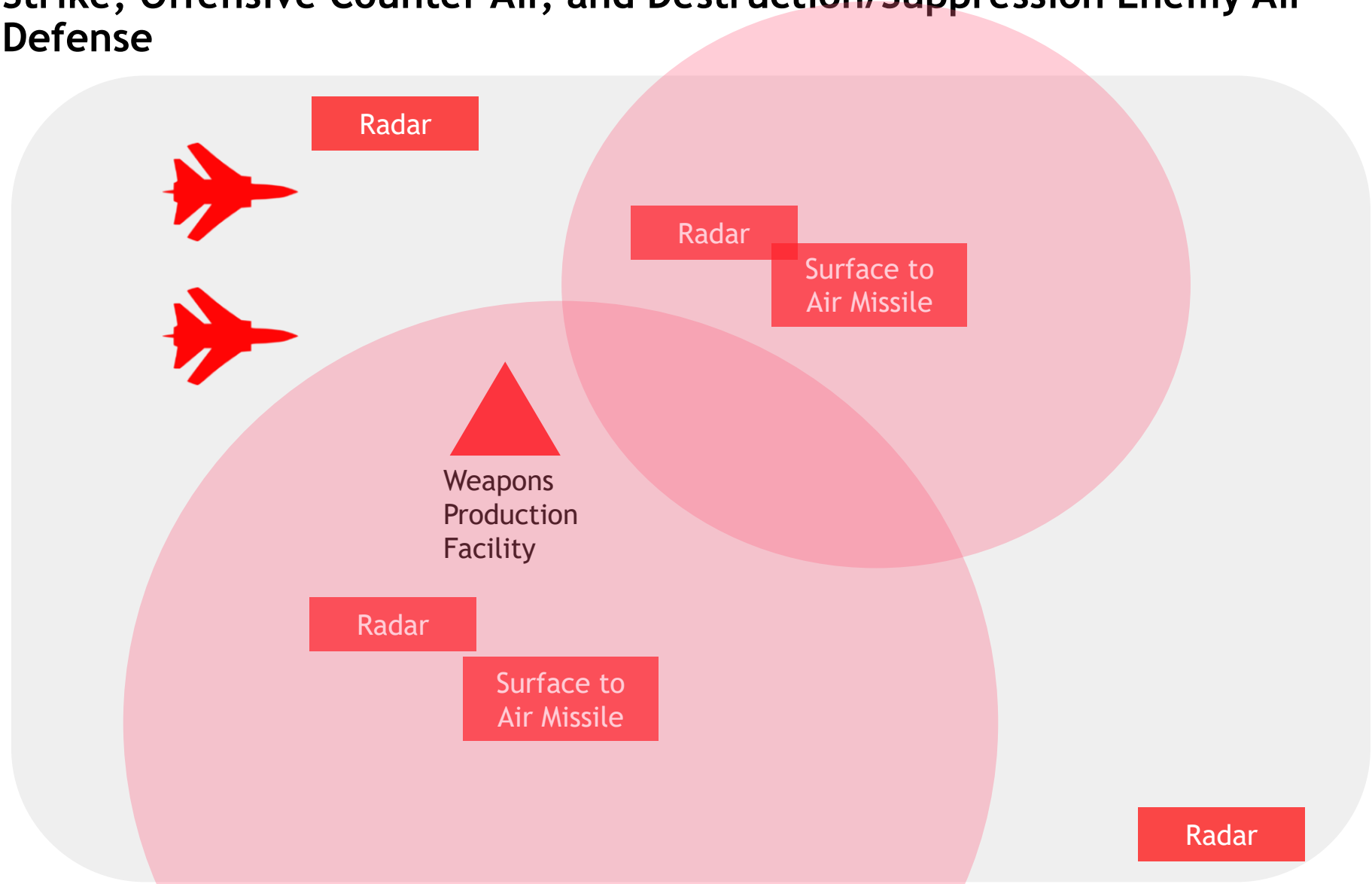
Characterization across operational envelope -
Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense

Weapons
Production
Facility

Characterization across operational envelope -
Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense

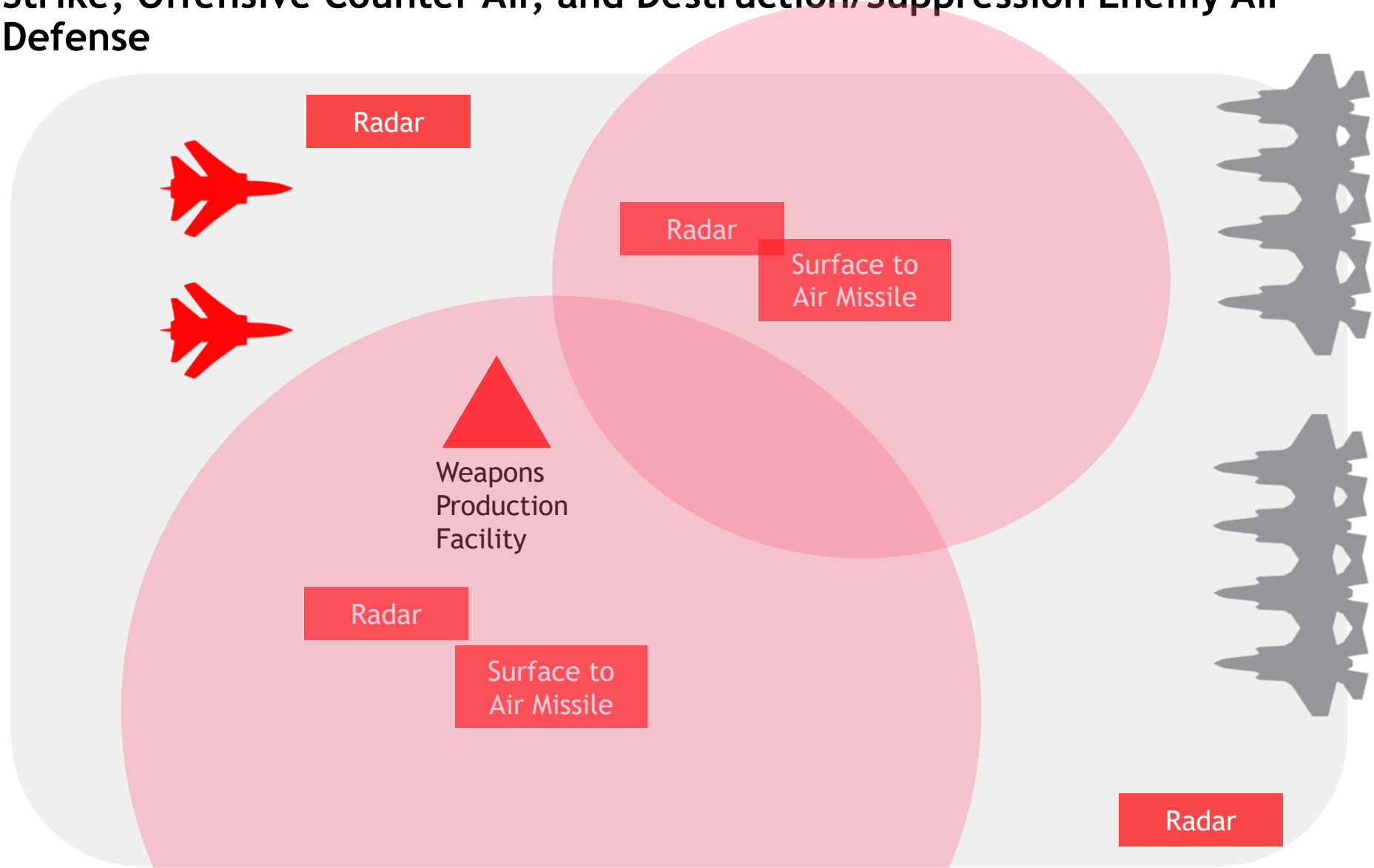


Characterization across operational envelope -
Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense



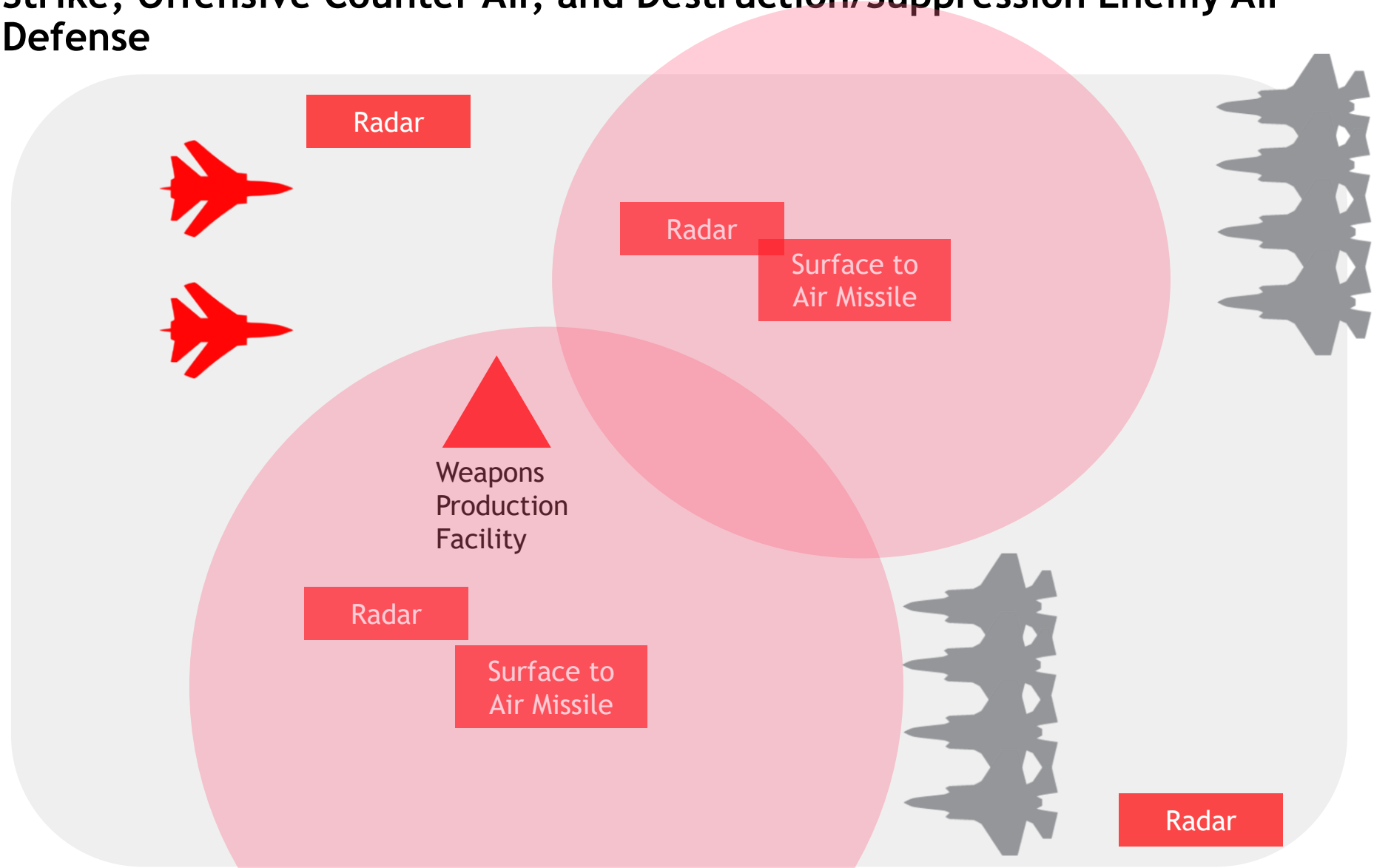
Characterization across operational envelope -

Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense

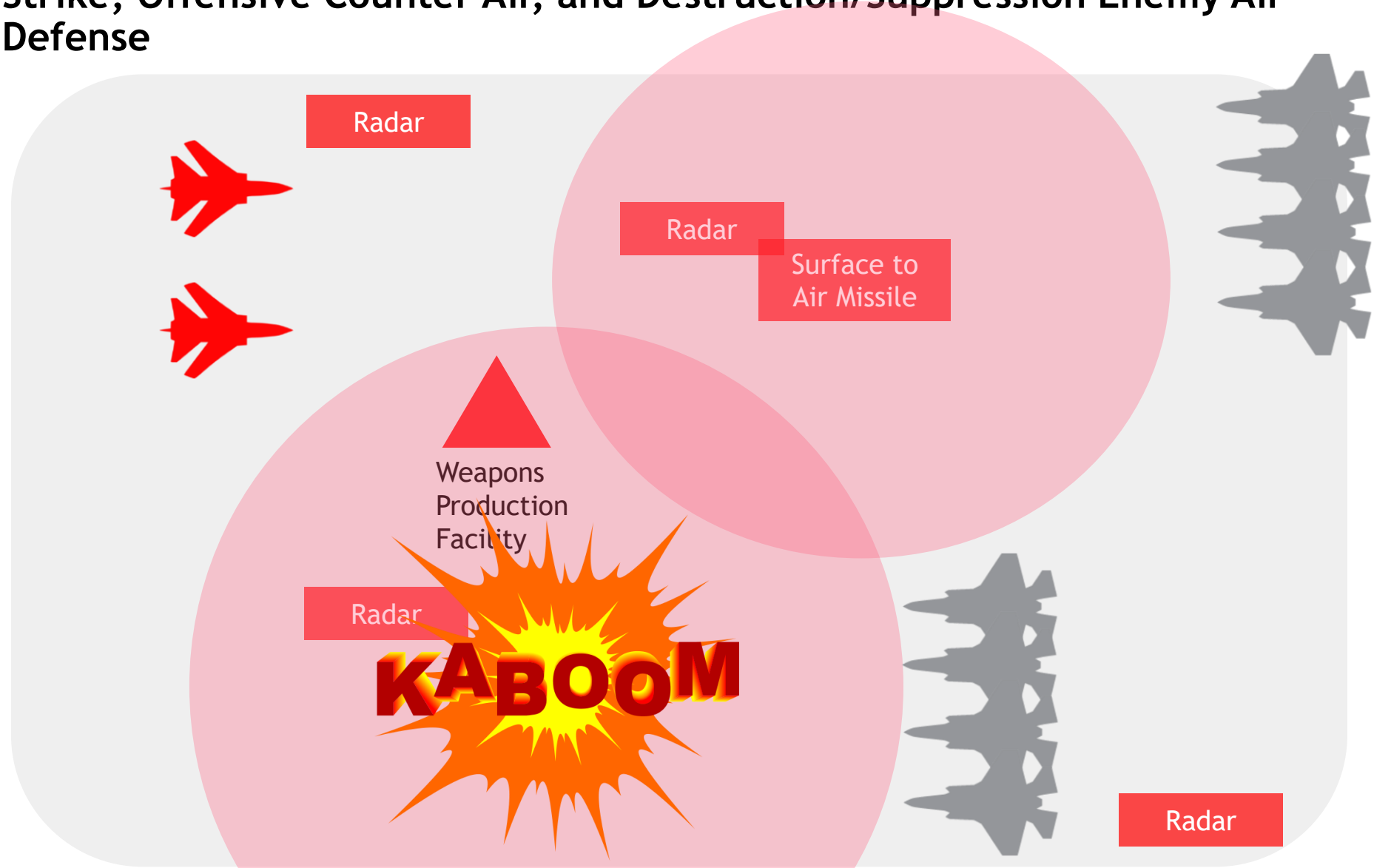


IDA

Characterization across operational envelope -
Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense

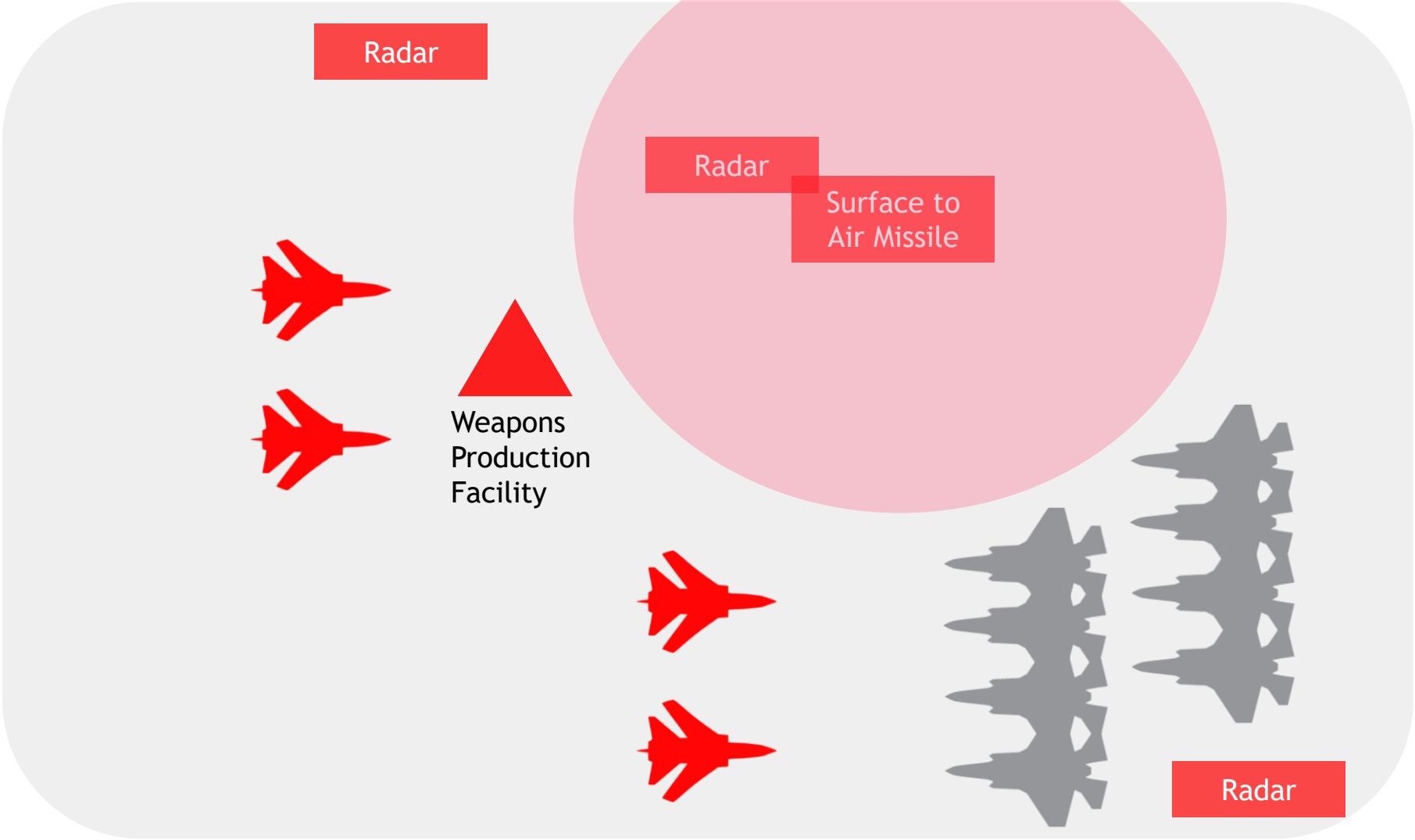


Characterization across operational envelope -
Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense



IDA

Characterization across operational envelope -
Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense



Characterization across operational envelope - Response Variables

Lots of measures to capture:

Mission outcomes

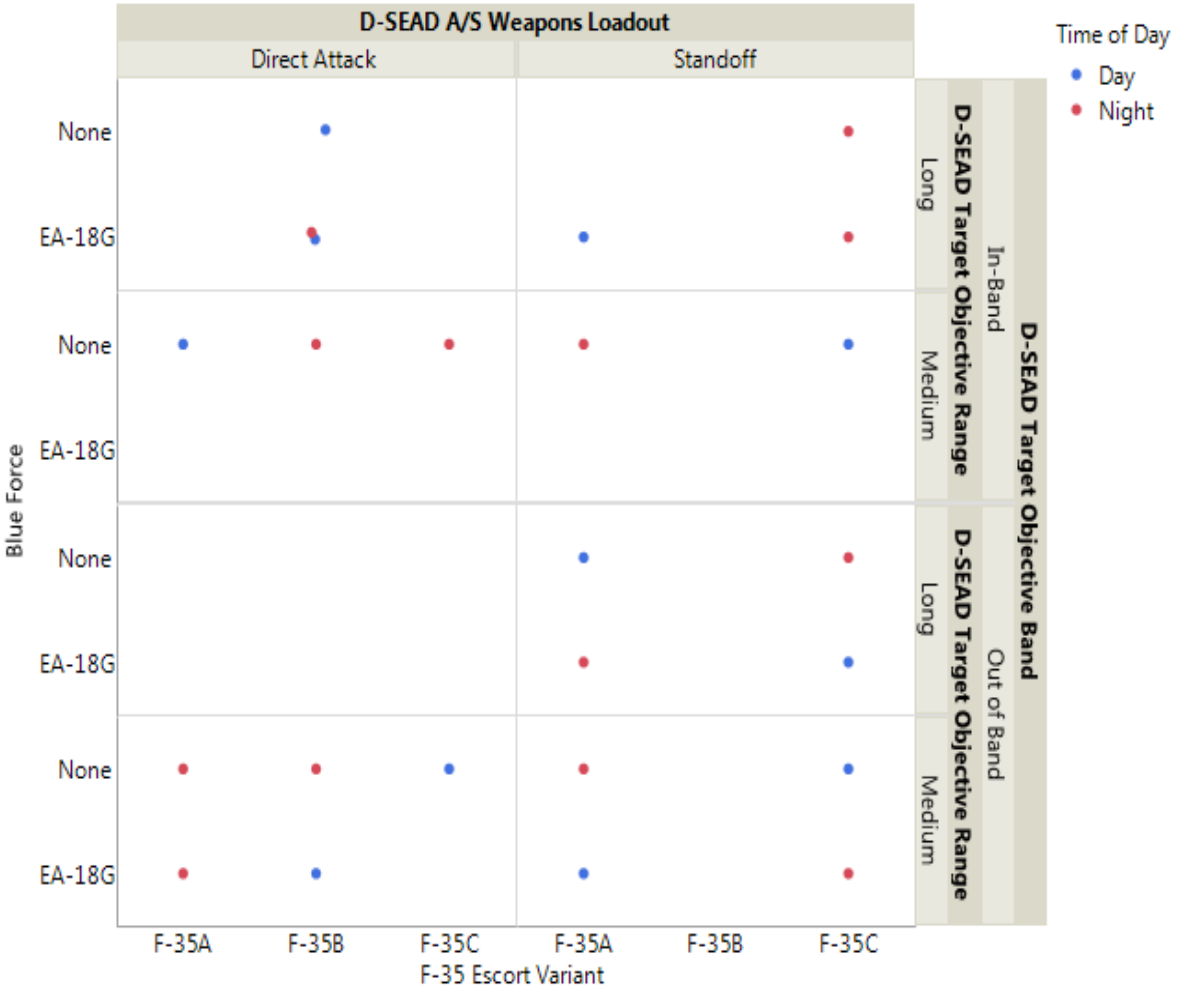
Air to Air Performance

Air to Surface Performance

System sensor capabilities

Targeting Accuracy Striker
Striker First Track Range
Striker First Hostile Declaration Range
Striker First Shot Range
Red Air First Detection Range
Red Air First Shot Range
Striker SAM Track Time
Proportion of Valid Weapon Releases to Number of Valid Weapon Releases Required to Meet Mission Tasking
Proportion of Assigned Air to Surface Targets Removed
Proportion of Striker Kill Removed
Striker to Red Air Exchange Ratio
Geolocation Find Time
Fix Time
DEAD Time
Targeting Accuracy Escort
Escort SAM Track Time
Proportion of Assigned SAM Elements Removed
Proportion of Assigned SAM Elements Engaged
Exchange Ratio
Closest Red Air Range to Strike Package
Blue Striker Encroachment Range
Escort First Track Range
Escort First Hostile Declaration Range
Escort First Shot Range
Red Air First Detection Range
Red Air First Shot Range
Proportion of Escort Blue Strikers that reach their Weapons Release Point
Proportion of Protected Aircraft (Strikers) Not Kill Removed
Proportion of Escort F-35 Kill Removed
Escort to Red Fighter Exchange Ratio

Experimental designs determine test adequacy

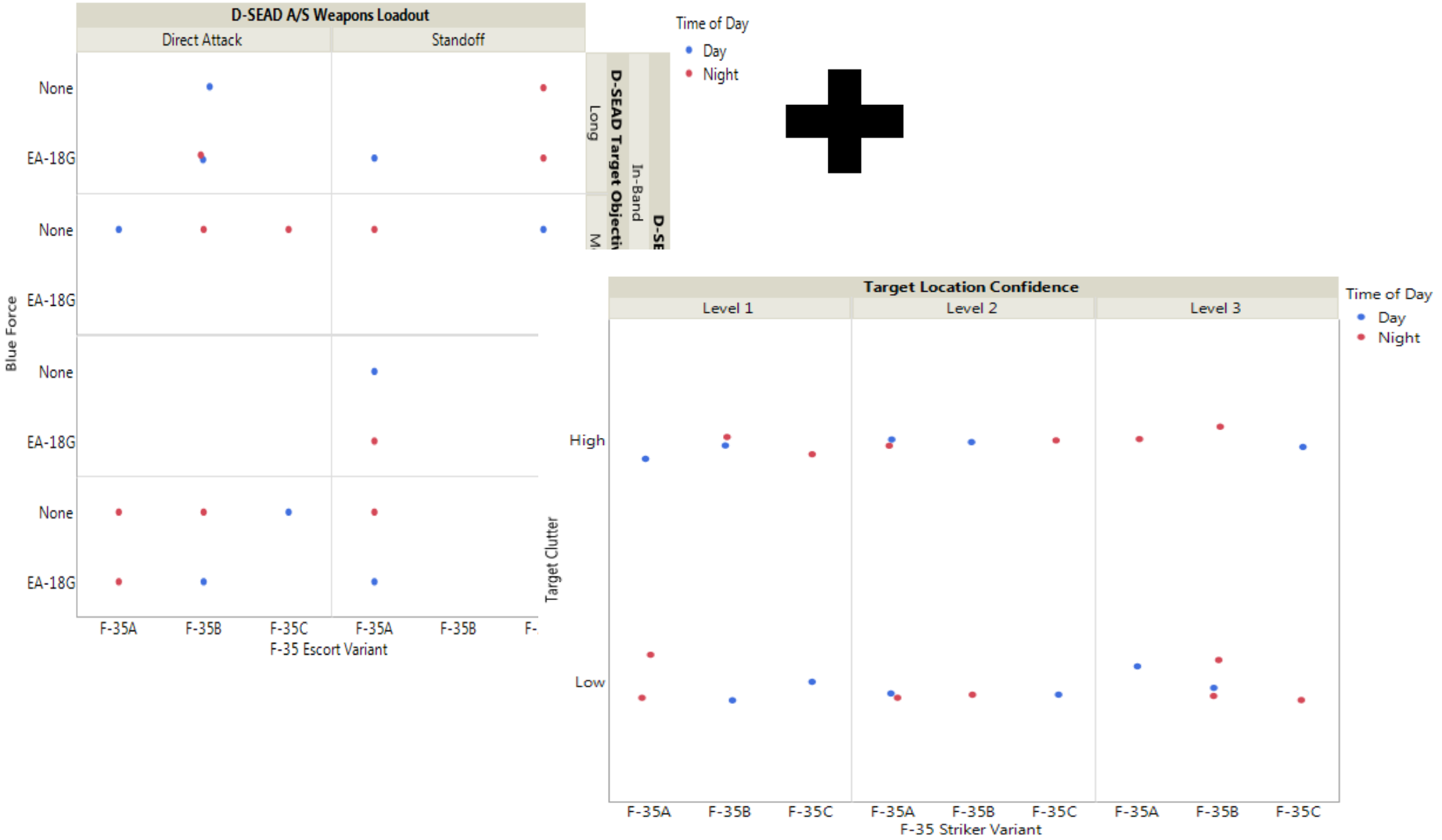


24 Run, D-Optimal

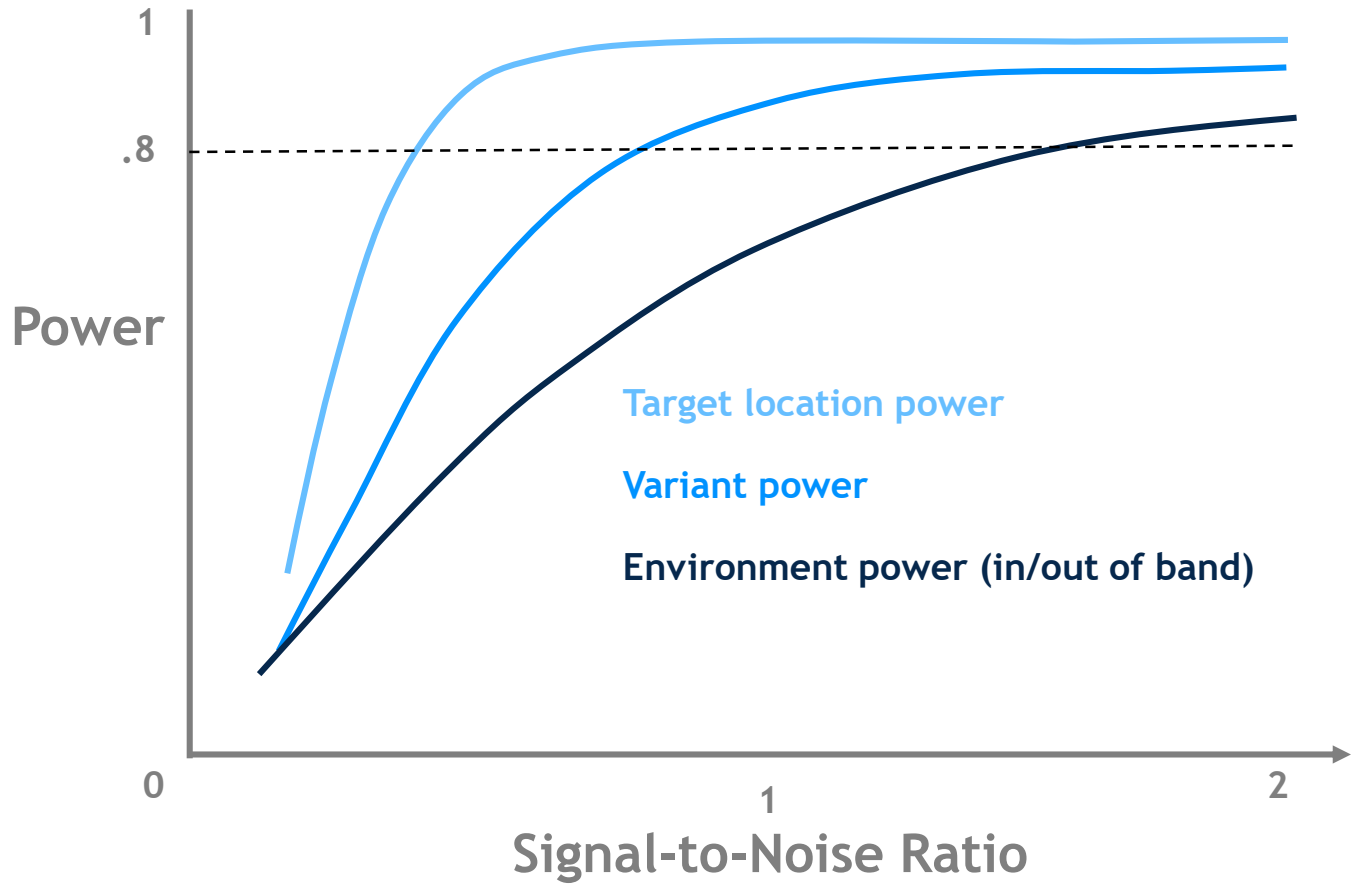
2nd Order Design

Disallowed Combinations

Two mission designs, executed in a 5th generation scenario



Power calculations provided justification for number of trials



We took a scientific approach to all operational testing

Conventional	Short takeoff/vertical landing	Carrier variant
		

Mission Areas	Air Threat	Ground Threat
Air-Surface		
Strike		
Destruction/Suppression of Enemy Air Defenses		
Defensive counter air		
Offensive counter air		
Close air support		
Search and rescue		

IDA

Impact so far

Congressional review of Close Air Support Testing

Still to come

Test Execution and Analysis

Execution Considerations

- Challenges with aircraft availability
- Confounding variables

Analysis Considerations

- Demand for quick answers
- Big Data, Little Information

Statistical Engineering Shortcomings

Initial focus was on tools

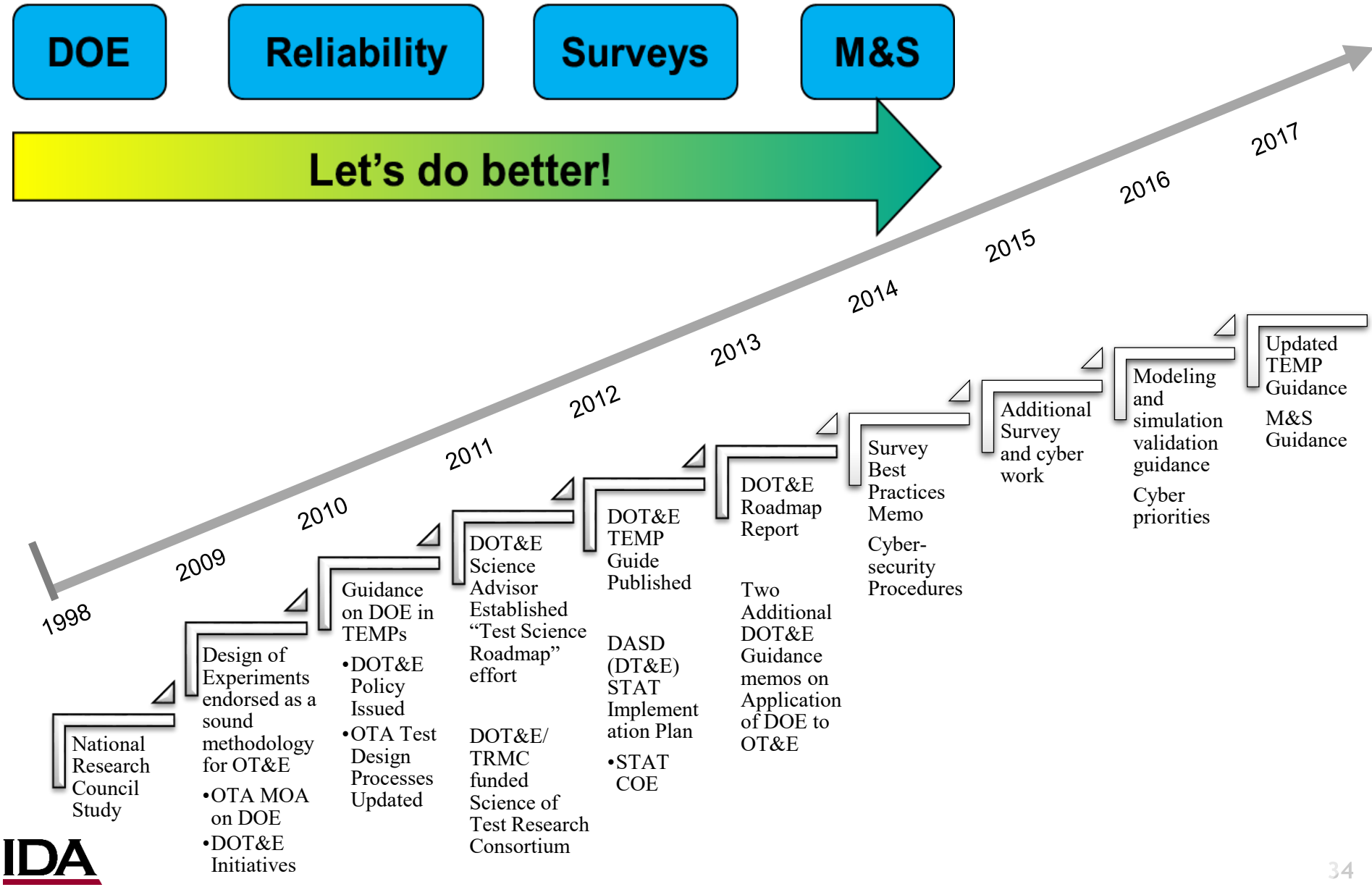
Processes are still highly dependent on individuals involved

Adherence to statistical rules

Leadership changes & final solution not fully deployed

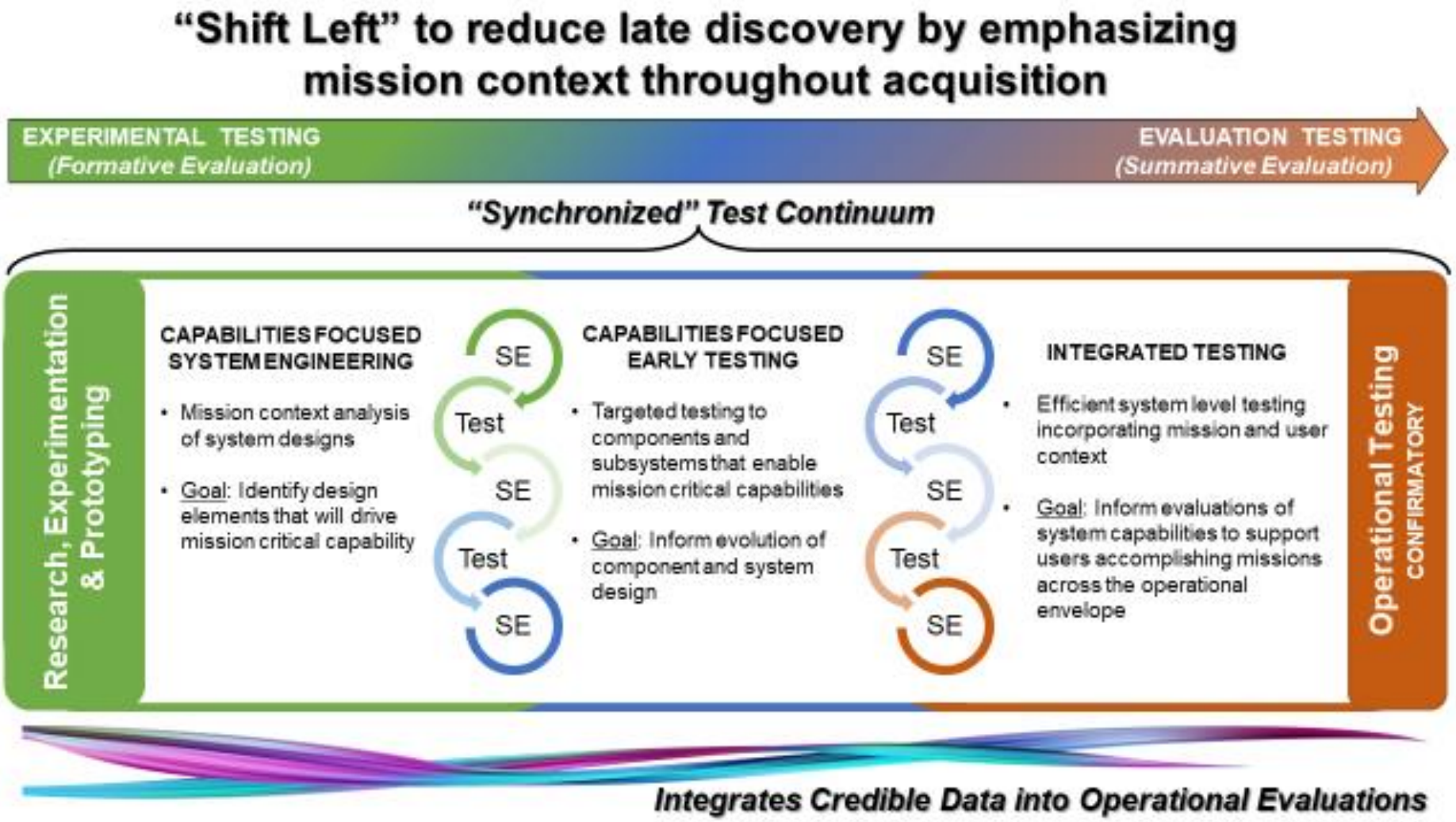
Failing to see the big picture

We continue to increase the statistical defensibility of DoD Test and Evaluation



IDA

Needed a larger focus for statistical engineering efforts



Thank you!

Innovation Adoption

I consistently meet brilliant, creative, entrepreneurial people in DoD with novel and implementable ideas, but they are fighting against entrenched processes and regulations that – in some cases – haven't been modified in decades. Incentives are often misplaced. Decision-making seems surprisingly diffuse for an organization known for its hierarchical structure and decisive leaders. Some of these *intrapreneurs* find workarounds to inflexible systems or receive temporary shelter under a like-minded commander; far more do not. Even the most senior leaders described responsibilities being so intricately nested across the organization that a sense of true ownership proved elusive to them. Early on, I reached a fundamental conclusion that has been borne out over time: DoD does not have an innovation problem; it has an innovation *adoption* problem.

Dr. Eric Schmidt,

Testimony to House Armed Services Committee

April 17, 2018

Laura's conjecture

Statistician's are uniquely equipped to lead & implement change, especially in data-centric fields!