INSTITUTE FOR DEFENSE ANALYSES

The Importance of M&S in Operational Testing and the Need for Rigorous Validation

Kelly McGinnity

April 2016

IDA Document NS D-5807
Log: H 16-000555

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

Modeling and simulation (M&S) is often an important element of operational evaluations of effectiveness, suitability, survivability, and lethality. In order to have an adequate understanding of, and confidence in, the results obtained from M&S, statistically rigorous techniques should be applied to the validation process wherever possible. Design of experiments methodologies should be employed to determine what live and simulation data are needed to support rigorous validation, and formal statistical tests should be used to compare live and simulated data. This briefing discusses the importance of M&S in operational testing through a few examples, provides an overview of existing DOT&E guidance on M&S validation, and outlines several statistically rigorous techniques for validation.

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-5807

The Importance of M&S in Operational Testing and the Need for Rigorous Validation

Kelly McGinnity

Executive Summary

Modeling and simulation (M&S) is often an important element of operational evaluations of effectiveness, suitability, survivability, and lethality. In order to have an adequate understanding of, and confidence in, the results obtained from M&S, statistically rigorous techniques should be applied to the validation process wherever possible. Design of experiments methodologies should be employed to determine what live and simulation data are needed to support rigorous validation, and formal statistical tests should be used to compare live and simulated data.

This briefing discusses the importance of M&S in operational testing through a few examples, provides an overview of the existing Director, Operational Test and Evaluation (DOT&E) guidance on M&S validation, and outlines several statistically rigorous techniques for validation. All data and graphical representations in the brief are notional, and the methodologies presented are certainly exhaustive. No specific solution is endorsed; rather, the briefing aims to highlight the type of statistical thinking that should be applied before accrediting M&S capabilities for use in OT.

Additionally, the authors recognize that a statistical comparison of the model output to live data is only one part of a larger validation plan. Both quantitative and qualitative evaluations are necessary to understand the strengths and weaknesses of the model across the operational envelope. This briefing is an initial effort to clarify the intent of DOT&E's recent push for increased rigor in the validation and accreditation process. Further research on best practices for statistically validating M&S is ongoing.

The Importance of M&S in Operational Testing and the Need for Rigorous Validation

Kelly McGinnity

Institute for Defense Analyses

April 29, 2016

IDA

Outline

- **Modeling and Simulation in OT&E**
  - Examples
  - Terminology
- **Guidance on M&S**
- **Statistical Tools for VV&A of M&S**
- **Common Myths and Pitfalls**

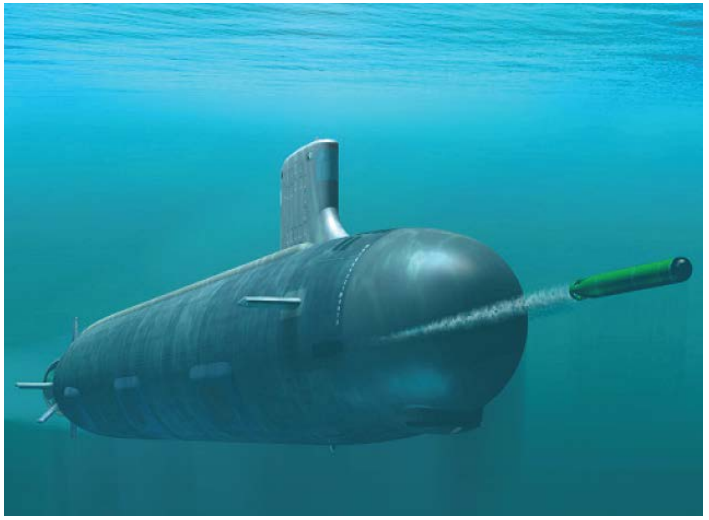
M&S in Operational Testing

- **Expansion of the operational space from what can be done live**
  - High threat density (air and ground)
- **Frame the operational space**
  - Large number of factors contribute to performance outcomes
- **Improve understanding of operational space**
  - Limited live data available
- **Ensure coverage of rare threats/occurrences**
- **End-to-end mission evaluation**
- **Translation of test outcomes to operational impact**

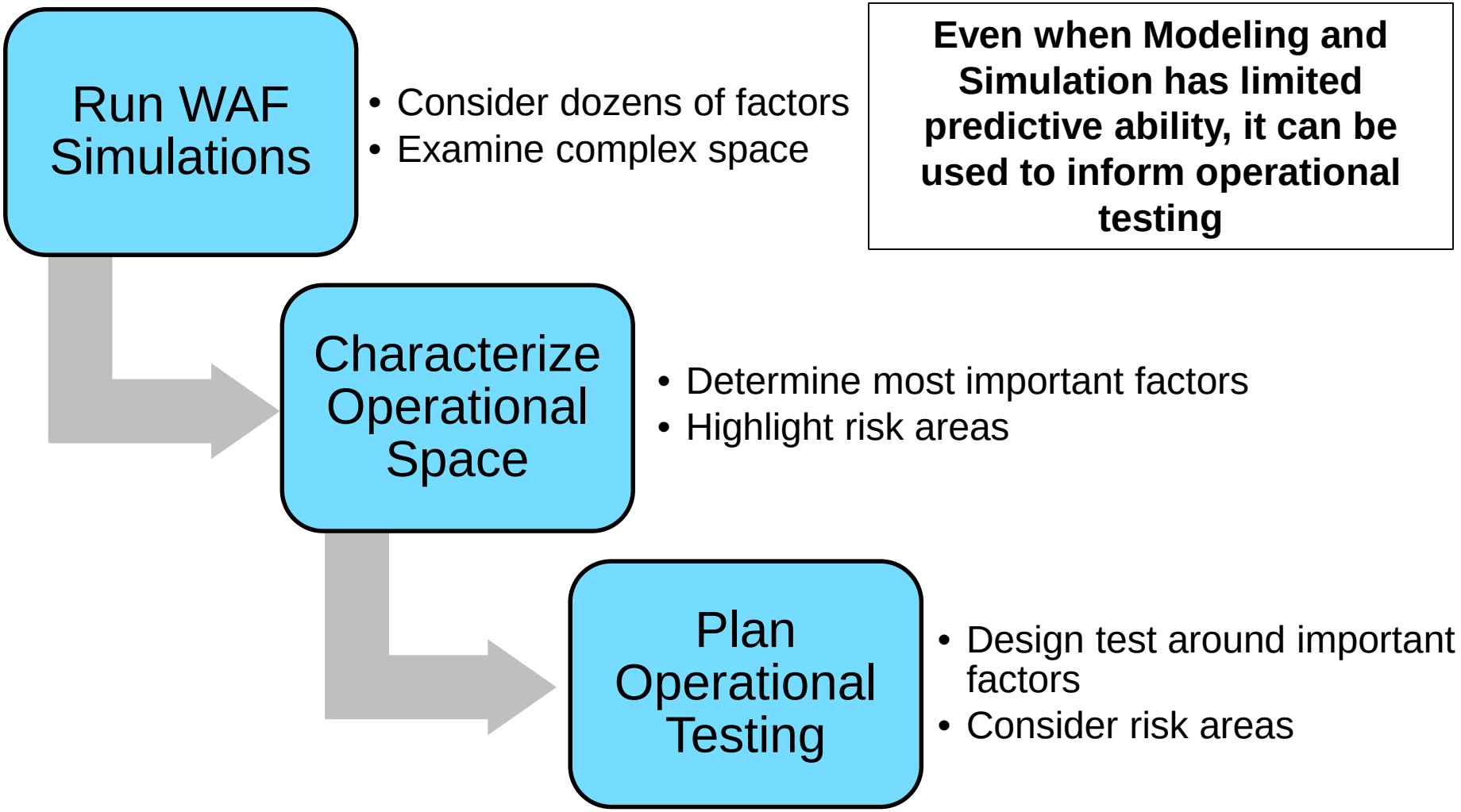
Always strive to do as much testing in the actual operational environment (open air, at sea, etc.) as possible

Example 1:
Weapons Analysis Facility (WAF)

- **Hardware-in-the-loop simulation capability for lightweight and heavyweight torpedoes**
- **Creates simulated acoustic environment**
  - Sonar propagation
  - Ocean features
  - Submarine targets
- **Interfaces with torpedo guidance and control sections**
- **Why we need M&S?**
  - Complex operational space where performance is a function of many environmental and scenario factors
  - In-water torpedo shots are costly
  - Serves primarily as a test-bed for new software
- **Limitations**
  - Computer processing prohibits full reproduction of full ocean conditions which have limited prediction accuracy



Weapons Analysis Facility (WAF) (cont.)



Example 2:
Probability of Raid Annihilation (PRA) Test Bed

- **Question to be addressed:**
  - Self-defense requirements for Navy combatants include a Probability of Raid Annihilation (PRA) requirement
  - To satisfy the PRA requirement, the ship can defeat an incoming raid of anti-ship cruise missiles (ASCM) with any combination of missiles, countermeasures, or signature reduction
- **Why we need M&S:**
  - Safety constraints limit testing
  - No single venue where missiles, countermeasures and signature reduction operate together in OT

PRA Test Bed (cont.)

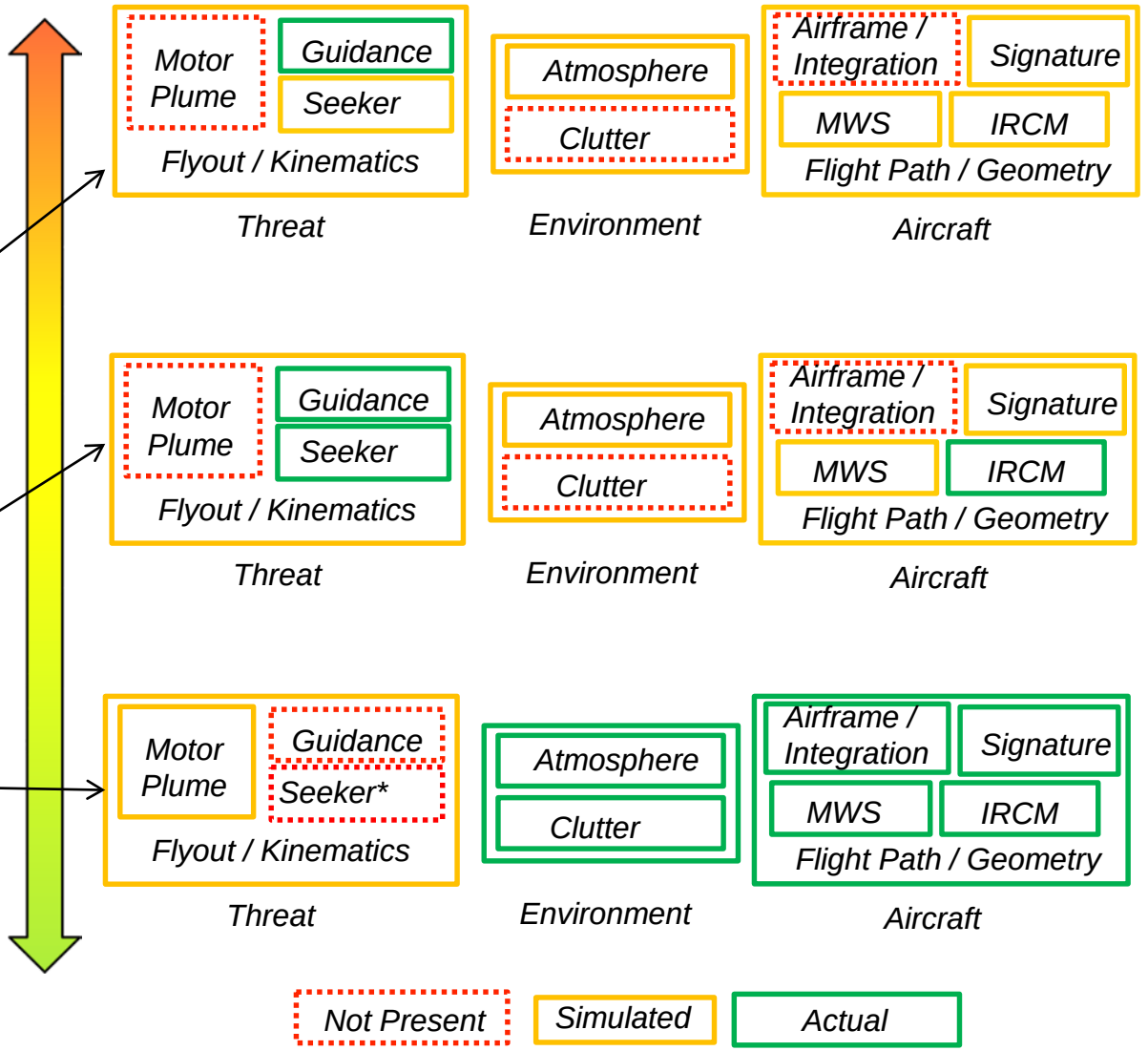
- **PRA is a federation of models that is fully digital**
  - Many system models are tactical code run on desktop computers
  - High-fidelity models of sensors include propagation and environmental effects
  - High-fidelity six-degree-of-freedom missile models
- **Small amount of "live" data from the Self Defense Test Ship provides limited understanding of PRA**
- **Architecture will be useful for a variety of ship classes**
  - LPD 17 was the first successful implementation – provided more information on PRA
  - LHA 6, DDG 1000, Littoral Combat Ship, CVN 78 will be examined

Example 3:
Common Infrared Counter Measures (CIRCM)

- **System Overview:**
  - Multiband infrared (IR) pointer/tracker/laser jammer for small/medium rotorcraft and small fixed wing aircraft
- **Why we need M&S:**
  - Shooting live missiles at aircraft is not feasible
- **M&S Solution**
  - Simulate end-to-end missile engagements by combining results from multiple test facilities using identical initial conditions
  - Allows the full sequence from detecting a threat to using a countermeasure to be assessed

Common Infrared Counter Measures (cont.)

- **Integrated Threat Warning Lab**
  - Assess flight path/geometry
- **Threat Signal Processing in the Loop (T-SPIL)**
  - Actual Threat Tracking
- **Guided Weapons Evaluation Facility (GWEF)**
  - Inclusion of actual seekers and countermeasures supports wider operational space
- **Open Air Range, Missile Plum Simulators**
- **Free-Flight Missile Test**
  - Missiles are not threat representative



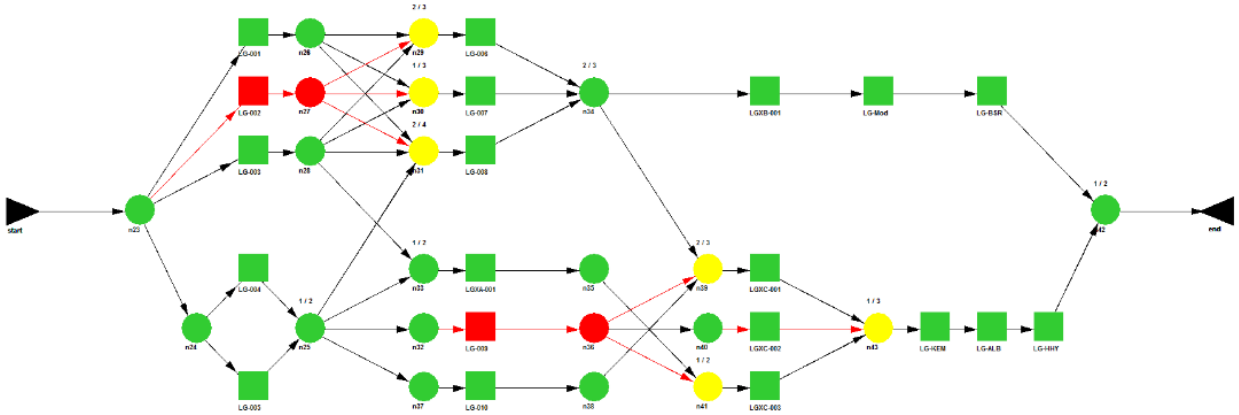
Acronyms this slide: Infrared (IR) Countermeasures (IRCM); Missile Warning System (MWS);

Example 4: Operational Availability

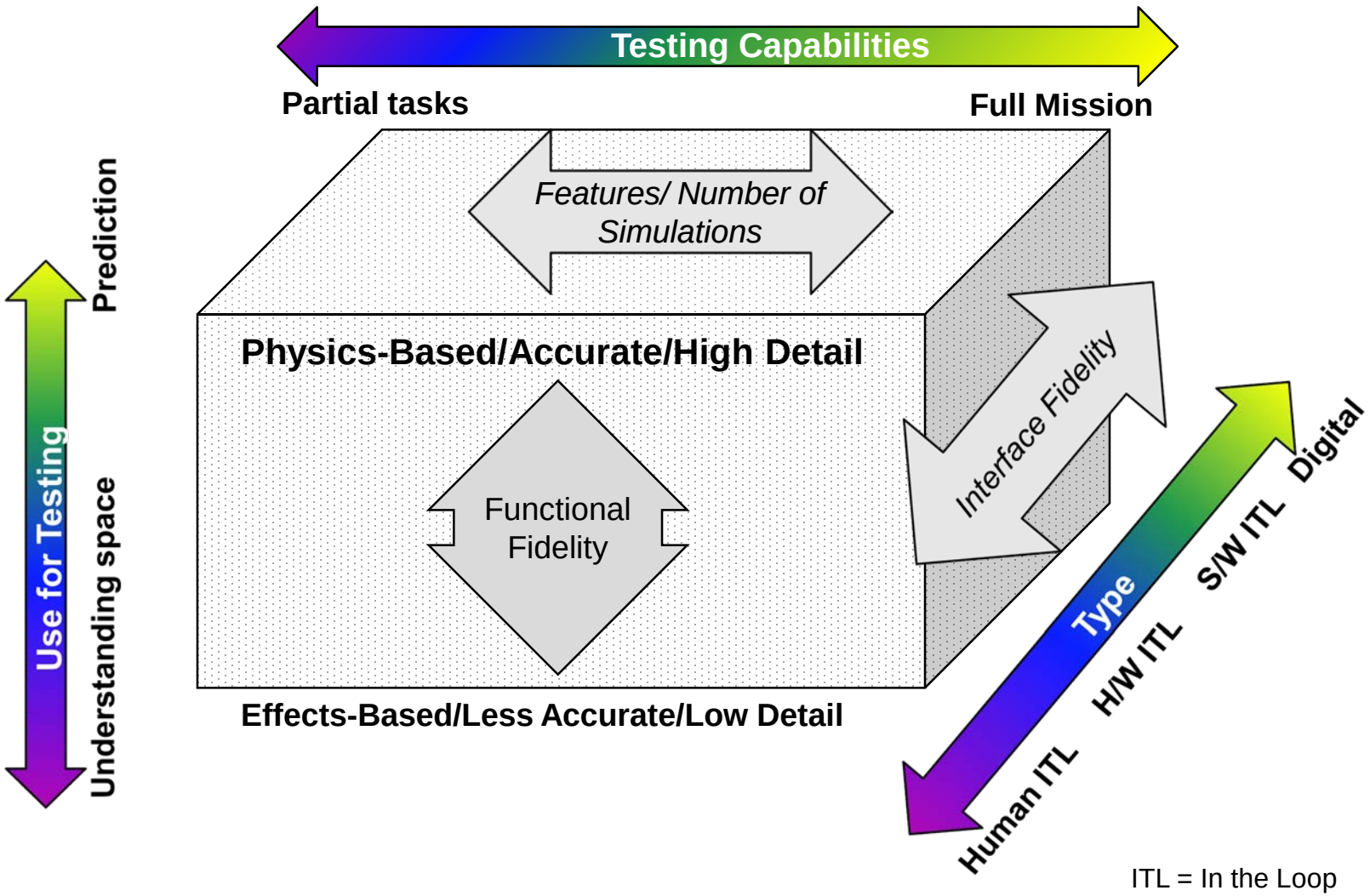
- **For complex systems, the Services use discrete event simulations to model Operational Availability (A_O)**
  - e.g., Raptor, LCOM

- **These digital simulations are based on:**
  1. Reliability block diagrams
  2. Expected component reliability
  3. Expected maintainability

- **Why we need M&S:**
  - Operational Availability cannot be assessed across all mission types during live testing
  - Models help assess the sensitivity of operational availability to changing conditions



Modeling Fidelity and the M&S Space



ITL = In the Loop

Outline

- **Modeling and Simulation in OT&E**
  - Examples
  - Terminology
- **Guidance on M&S**
- **Statistical Tools for VV&A of M&S**
- **Common Myths and Pitfalls**

Current Law and DoD Guidance on M&S

- **US Code: Title X**
  - States that DOT&E's operational assessment may not be based exclusively on M&S
- **DoDI 5000.02 (Operation of the Defense Acquisition System)**
  - Requires OTA accreditation and DOT&E approval to use M&S in support of an operational evaluation.
- **DoDI 5000.61 (DoD M&S VV&A)**
  - Assigns DOT&E responsibility for policies, procedures, and guidance on VV&A for DoD models, simulations, and associated data used for OT&E and LFT&E.

DOT&E TEMP Guidebook: M&S Adequacy

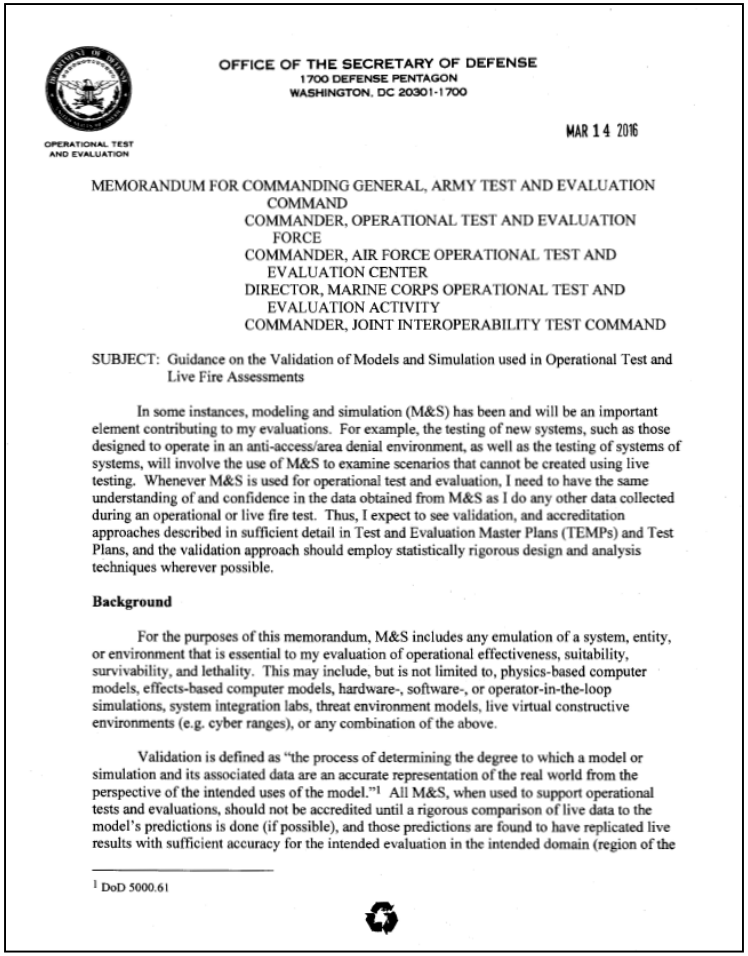
- **M&S capabilities and the approach for assessing credibility of the M&S should be described in the TEMP**

- **Consider the following questions in assessing M&S adequacy for T&E:**
  - What are the strengths and weaknesses of the M&S capability for T&E?
  - What major assumptions will be made in developing the M&S capability, and how would faulty or inaccurate assumptions impact the expected outcome and benefits of M&S use?
  - What are the source(s) and the currency of the data and information used for M&S development and validation, and are these adequate?
  - What field test data are – or will be – available to support validation and accreditation?
  - Under what conditions will the M&S need to be validated for the purpose of accreditation?
  - Has an existing capability gone through a verification, validation, and accreditation process?

"…Design of Experiments techniques should be leveraged to ensure that test data…clearly define the performance envelope…and corresponding statistical analysis techniques should be employed to analyze the data…"

DOT&E Guidance on Validation of M&S

- ***Guidance on the Validation of Models and Simulation Used in Operational Test and Live Fire Assessments,* dated 14 March 2016**
- **TEMPs and Test Plans must describe the validation and accreditation process in sufficient detail to understand the process**
- **Rigorous statistical design and analysis techniques should be used wherever possible**
  - Apply design of experiments principles when planning data collection for the M&S and the live test (if applicable)
  - Employ formal statistical analysis techniques to compare live and M&S data



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

OPERATIONAL TEST AND EVALUATION

MAR 14 2016

MEMORANDUM FOR COMMANDING GENERAL, 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

SUBJECT: Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments

In some instances, modeling and simulation (M&S) has been and will be an important element contributing to my evaluations. For example, the testing of new systems, such as those designed to operate in an anti-access/area denial environment, as well as the testing of systems of systems, will involve the use of M&S to examine scenarios that cannot be created using live testing. Whenever M&S is used for operational test and evaluation, I need to have the same understanding of and confidence in the data obtained from M&S as I do any other data collected during an operational or live fire test. Thus, I expect to see validation, and accreditation approaches described in sufficient detail in Test and Evaluation Master Plans (TEMPs) and Test Plans, and the validation approach should employ statistically rigorous design and analysis techniques wherever possible.

Background

For the purposes of this memorandum, M&S includes any emulation of a system, entity, or environment that is essential to my evaluation of operational effectiveness, suitability, survivability, and lethality. This may include, but is not limited to, physics-based computer models, effects-based computer models, hardware-, software-, or operator-in-the-loop simulations, system integration labs, threat environment models, live virtual constructive environments (e.g. cyber ranges), or any combination of the above.

Validation is defined as "the process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model."[1] All M&S, when used to support operational tests and evaluations, should not be accredited until a rigorous comparison of live data to the model's predictions is done (if possible), and those predictions are found to have replicated live results with sufficient accuracy for the intended evaluation in the intended domain (region of the

[1] DoD 5000.61

DOT&E Guidance on Validation of M&S (cont.)

- **Extrapolation outside the domain in which an M&S was validated is dangerous**

- **If inadequate data are available, either:**
  - The model should not be used,
  - Effort should be made to collect the necessary data, or
  - The validation report and any results based on the M&S should be caveated with a clear explanation of which areas are not sufficiently validated

"All M&S, when used to support operational tests and evaluations, should not be accredited until a rigorous comparison of live data to the model's predictions is done, and those predictions are found to have replicated live results with sufficient accuracy for the intended evaluation in the intended domain…"

Outline

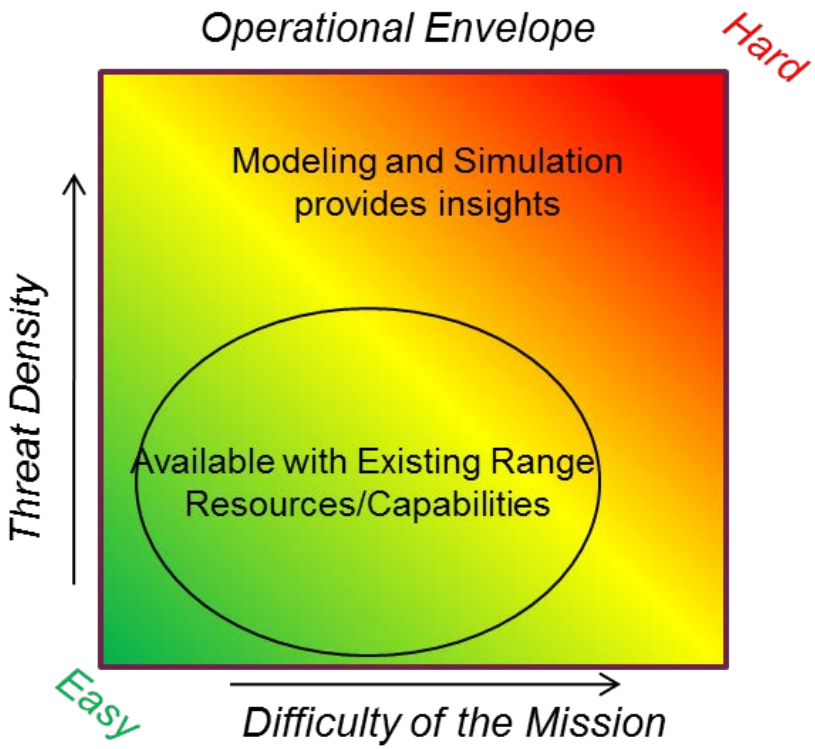
- **Modeling and Simulation in OT&E**
  - Examples
  - Terminology
- **Guidance on M&S**
- **Statistical Tools for VV&A of M&S**
- **Common Myths and Pitfalls**

Design of Experiments (DOE) for M&S

- **DOE provides a framework for selecting:**
  - Which simulation runs?
  - Which live runs?
  - How to validate?

Key Validation Questions

1. *What are the changes in outcomes as we move across test conditions? Do they match live testing? [Factor Effects]*
2. *What is the variability within a fixed condition? Is it representative of live testing? [Run-to-run variation]*
3. *What defines "matching live testing"? What is close enough? [Bias and Variance]*
4. *How do we control statistical error rates? [Type I and Type II errors]*



Validation Methods Overview

- **Typically a combination of validation techniques will be used**
  - Comparison to other models
  - Event validity (does the simulation go through all necessary steps?)
  - Face validity (evaluation by subject matter experts)
  - Comparison to historical data
  - Extreme condition comparisons
  - Internal validity

- **Methods that should be used more frequently**
  - Sensitivity analysis – changes to inputs produce reasonable changes to outputs
  - Predictive validation – can the model predict live test outcomes

When M&S will be used in DOT&E's assessment of Operational Effectiveness, Suitability, Survivability, or Lethality, the data collected for validation purposes needs to support sensitivity analysis and predictive validation

Outline of Statistical Methodologies

- **One sample tests w/ single "roll-up" score**
- **Series of partial roll-ups across conditions**
- **Tests that account for conditions via statistical models**

Least rigorous

Most rigorous

Prefer methodologies that look at how M&S validity varies across the envelope versus roll-up assessments

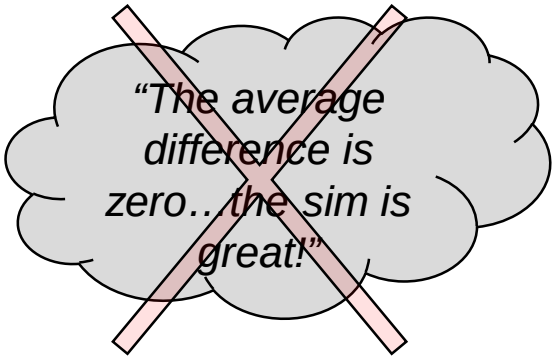
Matched Pairs

One-Sample / Single Score

- **Point-by-point comparison**
- **Single roll-up summary**
- **Limitations**
  - Not statistically rigorous
  - Doesn't control for conditions
    - » One number summary masks possible differences between conditions
  - Not robust to permutations
    - » If the live data and sim data contain the same numbers, but in different orders (as in this example), the metric will say it the live matches the sim when in reality it could be very different
- **<u>Avoid using this method</u>**

Live	Sim	Difference
15	9	6
12	15	-3
6	12	-6
9	6	3
⋮	⋮	⋮

Sum = 0



Simple Hypothesis Tests

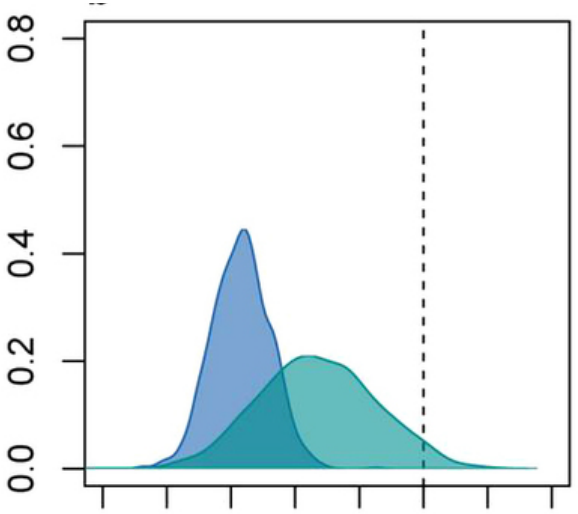
One-Sample / Single Score

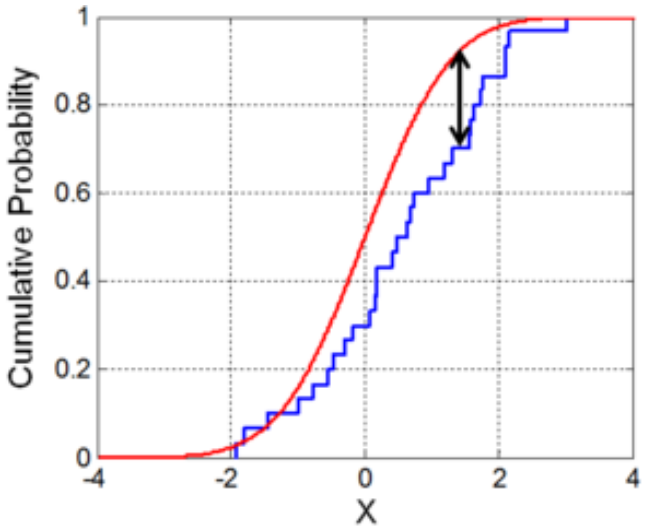
- **Single score "roll-up" comparison**
  - Compare a single metric (e.g. the mean) from the live data to that from the M&S data
  - Only valid under certain selections of data where the conditions are the same, e.g. one location only
- **Specific techniques: t-test, test for two proportions, variance test, nonparametric tests such as Wilcoxon rank-sum**
- **Limitations**
  - Sampling restraints
  - Assumptions require replication of live data
  - Doesn't test for factor effects
  - Not robust to permutations
- **<u>Avoid using these methods except in very specific cases</u>**

Distribution Tests

One-Sample / Single Score

- **Compare distribution of live data to distribution of M&S data under same conditions**
- **Specific techniques: Kolmogorov-Smirnov, Anderson-Darling**
- **Strengths**
  - Good filter / big picture look
  - Computationally easy
- **Limitations**
  - Multiple live data points required (need enough to form a distribution)
  - Sampling restraints (M&S and live need to have come from same conditions)
  - Not robust to permutations





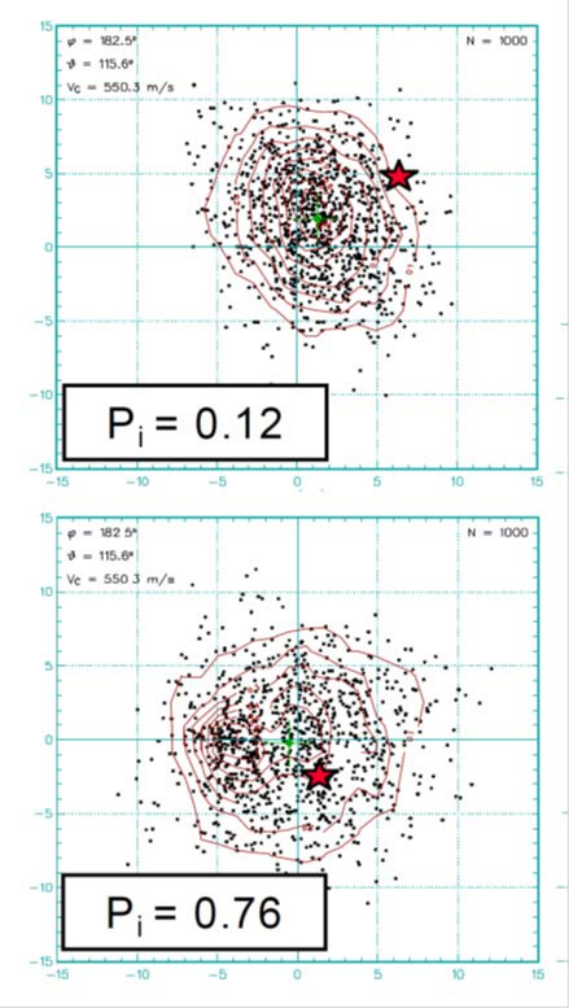
Fisher's Combined Probability Test

One-Sample / Single Score

- **Continuous data, e.g. missile miss distance**
  - 1 live shot per condition
  - Null hypothesis is that the live shot comes from the same distribution as the simulation "cloud"
  - Tail probabilities under each condition combined using a chi-squared test statistic
    - » $X = -2 \Sigma \ln(p)$ follows a chi-square distribution with 2N degrees of freedom

- **Strengths**
  - Intuitive way to handle limited data
  - Preferred to the t-test which ignores the variability of the "cloud"
  - Preferred to goodness-of-fit tests for most alternative hypotheses

- **Limitations**
  - Sensitive to one failed test condition
  - Requires computation
  - Requires adjustment if more than 1 live shot per condition is obtained
  - No formal test of factor effects



Series of Tests

Series of partial roll-ups

- **Separate out data by condition and perform multiple simple hypothesis tests (one in each condition)**
- **Specific techniques: t-test, test for two proportions, variance test, nonparametric tests such as Wilcoxon rank-sum**
- **Strengths**
  - Performance can be evaluated on a condition-by-condition basis
- **Limitations**
  - Not practical; need adequate data in each condition
  - Fails to leverage information across bins
  - No logical “summary” of all tests

Regression Modeling: Parameterizing Live vs. Sim

Statistical Modeling

- **Pool live and M&S data and build a statistical model**
  - Include a term that indicates whether the data point comes from live or M&S (*test type*), as well as interaction terms between *test type* and other factors of interest
  - For example,

    $$\textit{Detection Range} = \beta_0 + \beta_1 TestType + \beta_2 Threat + \beta_3 (TestType * Threat) + \epsilon$$

  - If the *Test Type* effect is statistically significant, then the M&S runs are not providing data that are consistent with the live runs
  - If the interaction term is significant, there many be a problem with the simulation under some conditions but not others

- **Strengths**
  - M&S runs can be formally compared to the live test events, even when there is limited live data
  - Model allows for testing of factor main effects and interactions with the test type

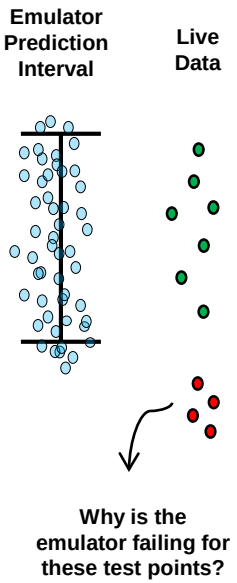
- **Limitations**
  - Assumes a good match between live and sim and requires sufficient evidence to *disprove* the validity of the model (backwards from what we normally do)
  - Need adequate power, otherwise this is a weak test
  - May need to consider higher order interaction terms to avoid rolling up results; requires more data
  - Given limited data, cannot differentiate between problems with bias vs. variance

Appropriate when we have extremely limited data in both the sim and live environments and we need to leverage them simultaneously to understand the space, e.g. PRA Testbed

Emulation and Prediction

Statistical Modeling

- **Build an empirical emulator (e.g. a logistic regression model) from the simulation**
  - As a new set of live data becomes available, compare each point with the prediction interval generated from the emulator under the same conditions
    - » If a live point falls within the prediction interval, that is evidence that the simulation is performing well under those conditions
  - Compare/model the live points that do vs. don't fall within the emulator prediction intervals and test for any systematic patterns
    - » Will help explain where / why the simulation is failing in certain cases
  - Once the live data is classified or "tested", it can then be used to update the simulation and continue to "train" the model



- **Strengths**
  - Applicable to any amount of live data
  - Can test for factor effects, as well as differentiate between problems with bias and variance (in the case of >1 live shot per condition)
  - Live data serves dual purposes of validating and updating the model
  - Emulator can help inform the live test

- **Limitations**
  - Not reasonable in the case of 1 or very few simulation runs per condition
  - Requires adjustment for a non-continuous response variable

Appropriate when there is adequate M&S data to build a model prior to live testing and the factor spaces for M&S and live testing overlap, e.g. WAF

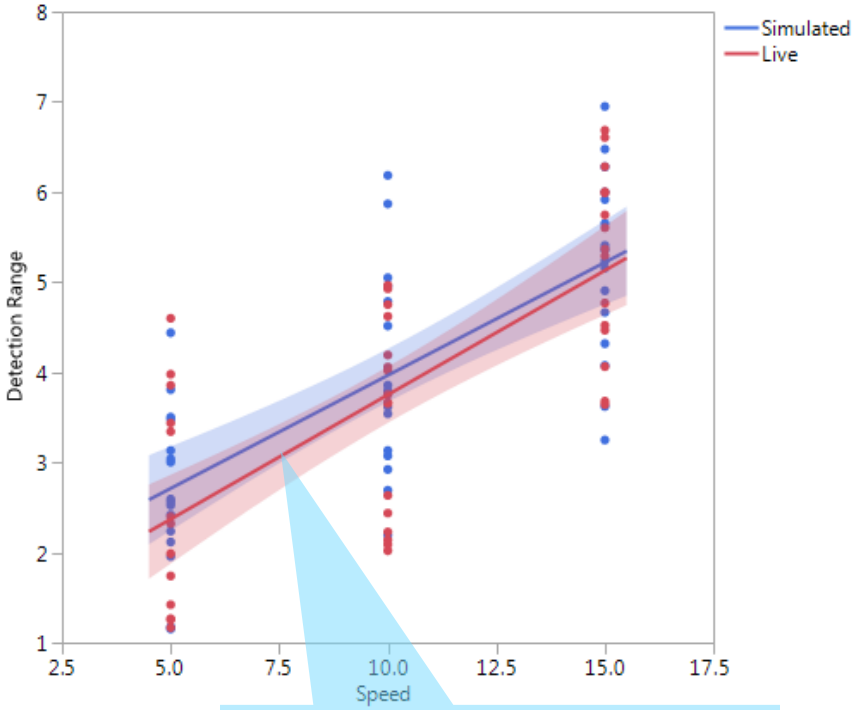
Other Modeling Techniques

Statistical Modeling

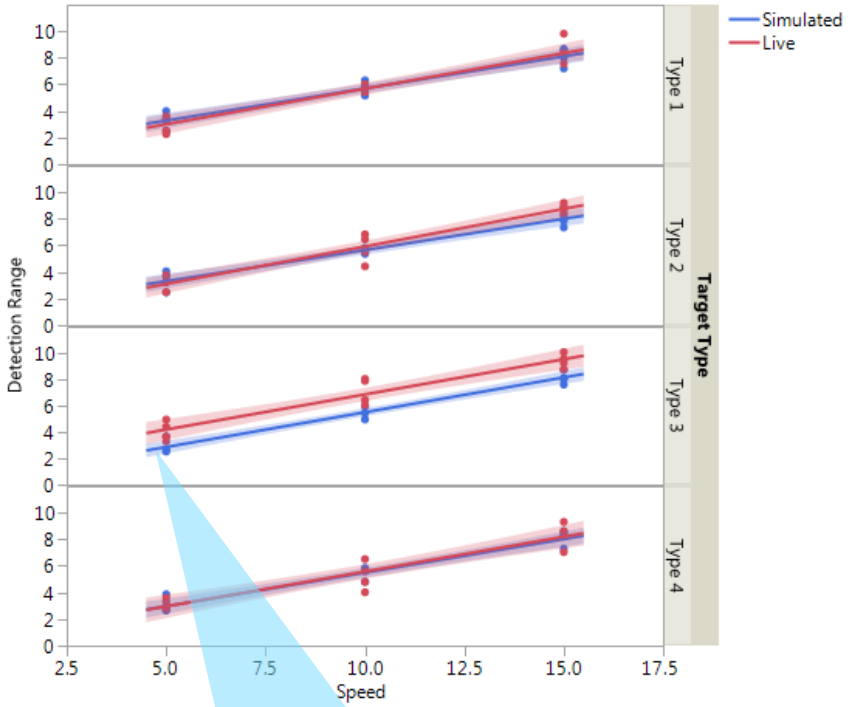
- **Bayesian modeling**
  - Use computer model outputs and expert opinion to improve estimation and prediction of a physical process
- **Hierarchical linear models**
  - Remove the variation due to covariates first, then test live vs. sim
- **Parameter calibration / Gaussian process models**
  - Use physical data to calibrate the computer experimental data and estimate unknown parameters
- **Limitations**
  - Complex methodologies limit DoD application
  - Current M&S designs do not support Gaussian Stochastic Process models
  - Focus is on improving prediction, we simply need to validate and state limitations

Benefits of Statistical Modeling

- **Models allow us to see where in the operational space M&S outcomes statistically match live outcomes**



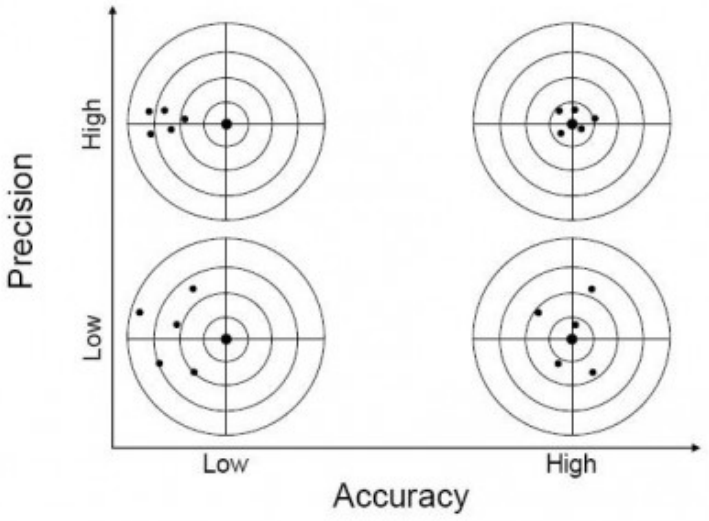
Regression Models can formally test for differences in slope/outcome

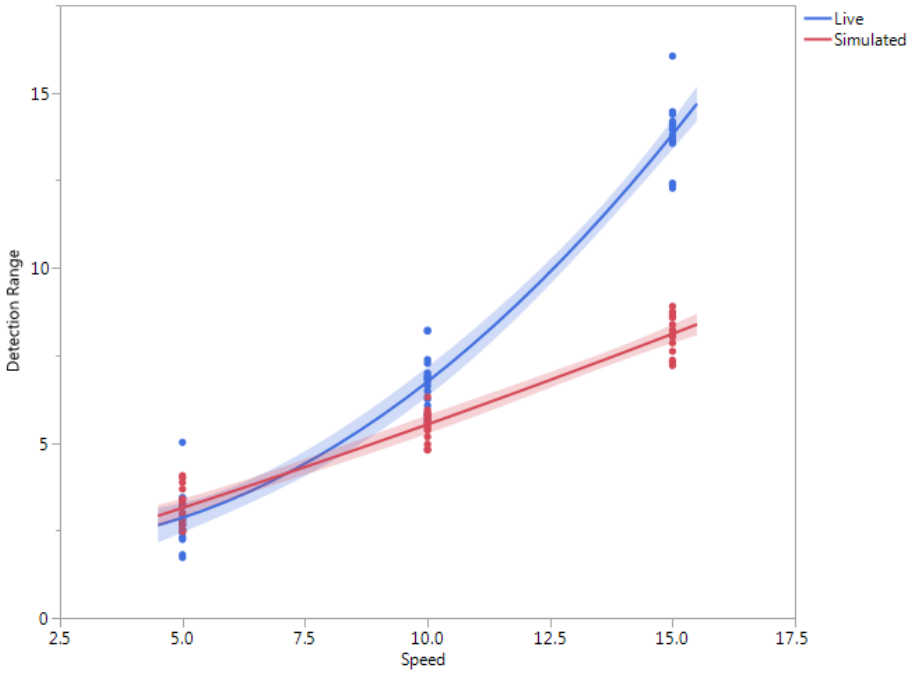


Same data with factors included in model shows differences

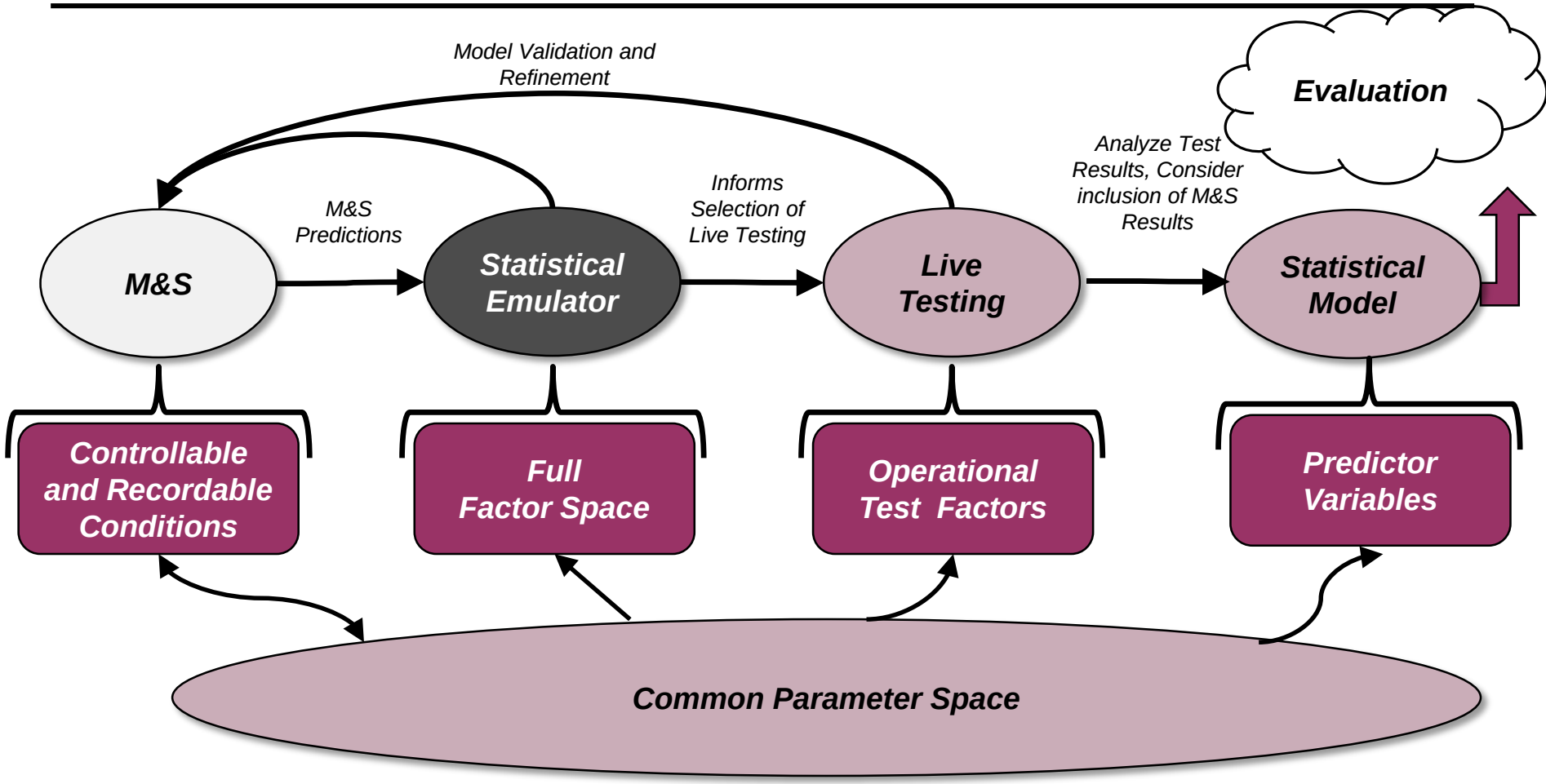
Benefits of Statistical Modeling (cont.)

- **Allows us to understand what parts of the space the modeling and simulation is valid for and what parts have large differences**
- **If sufficient data is available we can test for shifts in means and variances to determine if model is accurate and precise**





Framework for M&S Use in T&E



Identify the common set of variables that spans the operational space

Outline

- **Modeling and Simulation in OT&E**
  - Examples
  - Terminology
- **Guidance on M&S**
- **Statistical Tools for VV&A of M&S**
- **Common Myths and Pitfalls**

Common Myths

- **Myth: This model was accredited for developmental testing, therefore, it is valid for operational testing.**
- **Reality: The objectives of operational testing are different from developmental testing. Therefore, models need to be re-evaluated and compared to operational test data to understand their applicability to operational test assessments.**
- **Myth: We can replace operational test trials with modeling and simulation trials if there is an accredited end-to-end mission M&S**
- **Reality: We can never fully reproduce the operational space in a digital space. M&S can be used to help scope the operational space, expand beyond what is possible in live testing, and interpolate with in the operational space. However, it cannot replace individual points one-for-one.**

Common Pitfalls

- **Averaging validation results across conditions rather than discussing where the M&S is valid and where it isn't**
- **Faulty assumptions in developing or using M&S such as assuming independence between events that actually have some type of dependency or relationship**
- **Using M&S results outside their validation domain which are uncharacterized and include unknown uncertainties**
- **Improper use of data for M&S development or validation such as relying solely on heart-of-the-envelope performance data or using specification values instead of actual performance data when the latter is available**

Conclusions

- **M&S is increasingly becoming a key element of DOT&E's evaluations of effectiveness, suitability, survivability, and lethality**
- **DOT&E has issued a guidance memo emphasizing the importance of rigorous approaches to validation**
  - Statistical design and analysis techniques should be employed wherever possible
  - There is no one-size-fits-all solution
  - We are working to develop additional materials to aid the community

IDA

BACKUP

References

- Sargent, Robert G. "Verification and validation of simulation models." *Proceedings of the 35th conference on Winter simulation*. IEEE Computer Society Press, 2003.
- Oberkampf, William L., and Timothy G. Trucano. "Verification and validation in computational fluid dynamics." *Progress in Aerospace Sciences* 38.3 (2002): 209-272.
- Rao, Lei, Larry Owen, and David Goldsman. "Development and application of a validation framework for traffic simulation models." *Proceedings of the 30th conference on Winter simulation*. IEEE Computer Society Press, 1998.
- Kleijnen, Jack PC, and Robert G. Sargent. "A methodology for fitting and validating metamodels in simulation." *European Journal of Operational Research* 120.1 (2000): 14-29.
- Kleijnen, Jack PC, and David Deflandre. "Validation of regression metamodels in simulation: Bootstrap approach." *European Journal of Operational Research*170.1 (2006): 120-131.
- Rivolo, A. Rex, Fries, Arthur, Comfort, Gary C. "Validation of Missile Fly-out Simulations", IDA Paper p-3697, 2004.
- Thomas, Dean and Dickinson, R. "Validating the PRA Testbed Using a Statistically Rigorous Approach." IDA Document NS D-5445, 2015.
- Law, Averill M. *Simulation modeling and analysis*. Vol. 5. New York: McGraw-Hill, 2013.
- Rolph, John E., Duane L. Steffey, and Michael L. Cohen, eds. *Statistics, Testing, and Defense Acquisition:: New Approaches and Methodological Improvements*. National Academies Press, 1998.
- Easterling, Robert G., and James O. Berger. "Statistical foundations for the validation of computer models." *Computer Model Verification and Validation in the 21st Century Workshop*, Johns Hopkins University. 2002.

VV&A

- **All M&S used in T&E must be accredited by the intended user (PM or OTA). DOT&E determines if a model has been adequately VV&A'd to use in Operational Testing.**
- **"Verification is the process of determining if the M&S accurately represents the developer's conceptual description and specifications and meets the needs stated in the requirements document."**
- **"Validation is the process of determining the extent to which the M&S adequately represents the real-world from the perspectives of its intended use."**
- **"Accreditation is the official determination that the M&S is acceptable for its intended purpose."**

"A model should be developed for a specific purpose (or application) and its validity determined with respect to that purpose" (Sargent 2003)

Terminology

- **Classes of Modeling and Simulation**
  - **Digital models**: represent functions using programming (software) code in a manner that mimics real-world equipment, events, processes, etc.
  - **Software-in-the-loop**: employ one or more elements of operational software (computer programming code)
  - **Hardware-in-the-loop**: employ one or more pieces of operational equipment (to include computer hardware) within the simulation
  - **Human-in-the-loop**: employ one or more human operators in direct control of a key support function (e.g., decision making)
  - **Simulation Federation**: A system of interacting models and/or simulations

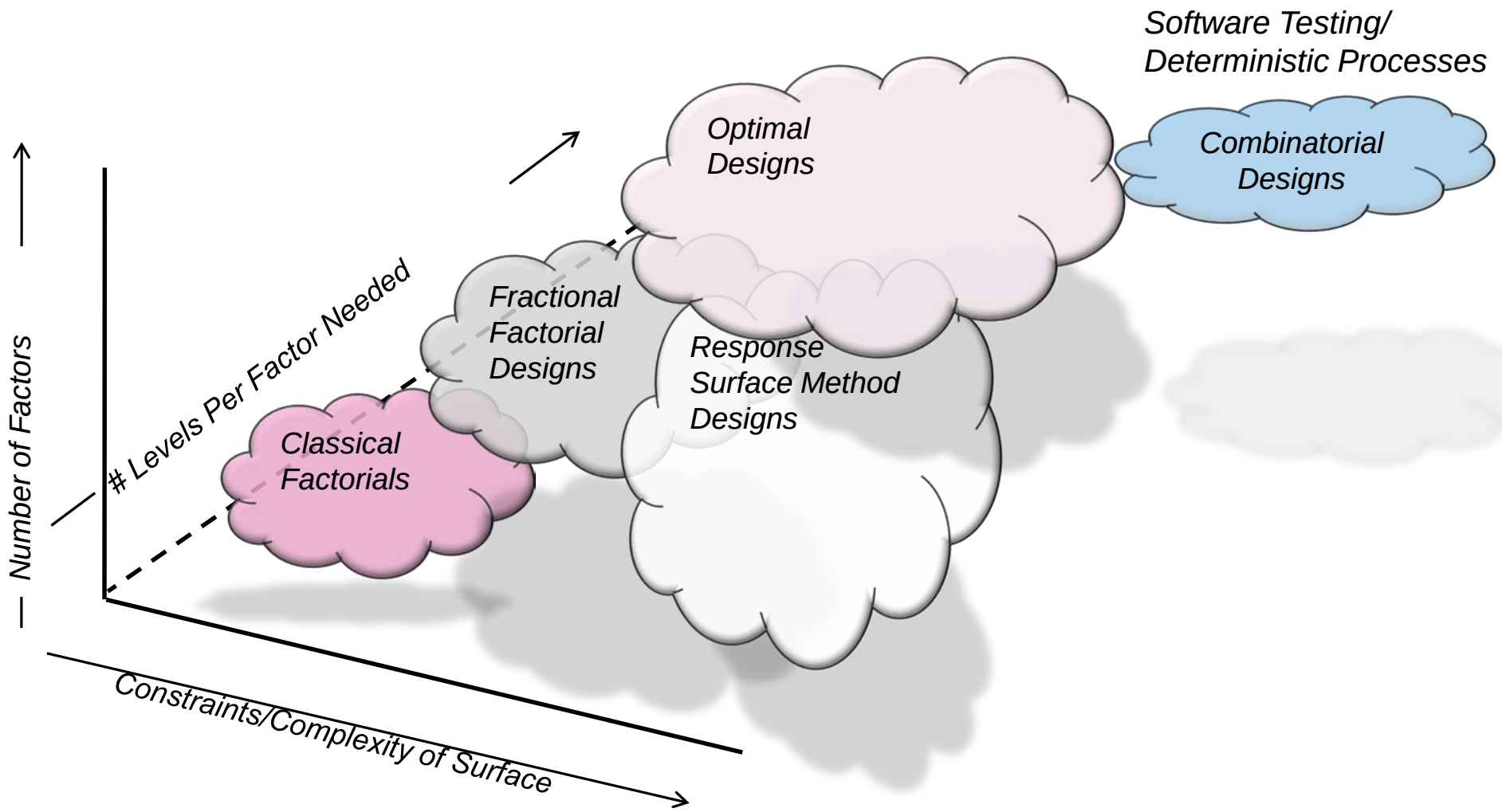
- **Modeling and Simulation can be classified as:**
  - Effects Based
  - Model Based

Challenges

- **Are the approaches different for different types of models?**
  - Software in the loop
  - Hardware in the loop
  - Purely deterministic models
  - Purely stochastic models (Monte Carlo models based on empirical data)
- **What measures do we use to compare?**
  - Absolute measures from the live data and M&S
  - Differences between live and M&S
- **How do we account for lots of replications on the M&S and very limited live data?**

Case studies will help us address these challenges

Types of Designs – Overview



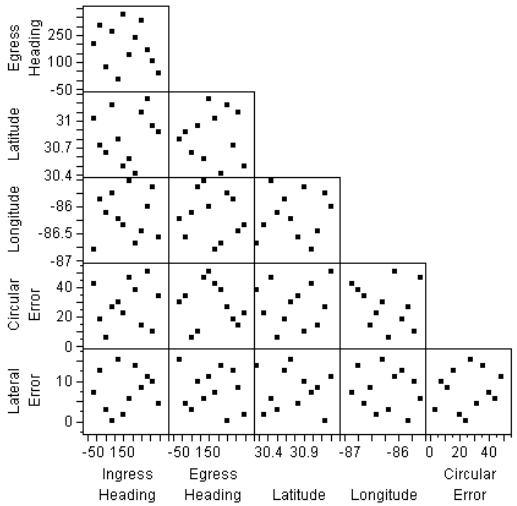
How do we choose the best design?

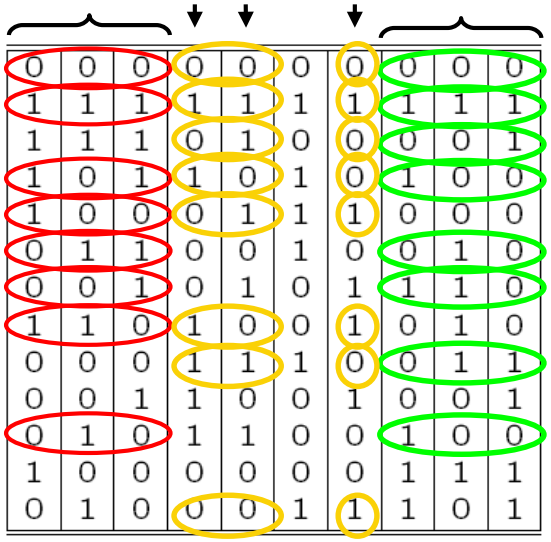
- **Most appropriate design choice depends on:**
  - The purpose of the M&S / goal of the validation analysis
  - The type of simulation (deterministic vs. stochastic)
  - The nature of the data (categorical vs. discrete)
  - The model terms desired to be estimated (e.g. what the “emulator” should look like)

- **Various selection criteria for design evaluation:**
  - High statistical power for important effects
  - Robustness to missing data
  - Low correlation between factors
  - Maximize the number of estimable main effects, two factor interactions and other higher order terms (depending on the goal of the test)
  - Minimize correlation between two-factor interactions and main effects

DOE for Deterministic M&S

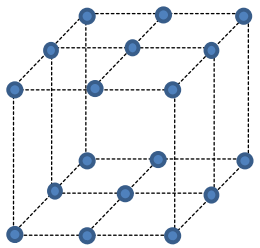
- **Space Filling Designs**
  - An efficient way to search or cover large continuous input spaces
  - Algorithms spread out test points using tailored optimality criteria
  - Analyzed via Gaussian process models
- **Factor Covering Arrays**
  - Type of combinatorial design; used to find problems
  - An efficient way to test when the space is large and made up of combinations of selections (categorical / binary input)
- **Computer simulation experiments**
  - Many recent methods in academic literature
  - Parameter calibration using Gaussian Stochastic Process Models
  - Bayesian techniques



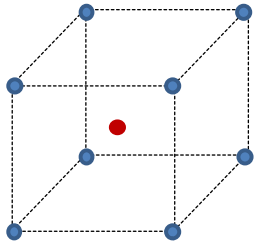


DOE for Stochastic M&S

- **Classical Factorial Designs**
  - Full coverage
  - Highest fidelity
  - All model terms estimable
- **Screening Designs (e.g. Fractional Fact.)**
  - Good for testing many factors at once
  - Lower fidelity
  - Some aliasing / inestimable terms
- **Response Surface Designs**
  - Best for a characterizing a few continuous factors
  - Allows testing for curvature
- **Optimal Designs**
  - Most efficient and flexible
  - Allows for constrained spaces, disallowed combinations, etc.



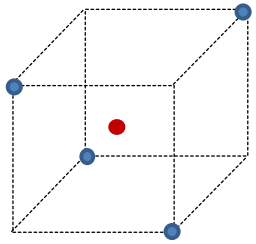
General Factorial
3x3x2 design



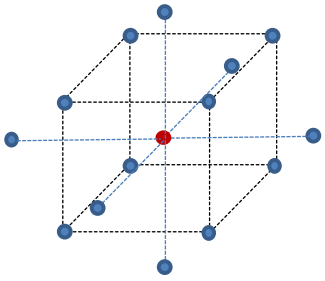
2-level Factorial
2^3 design

single point

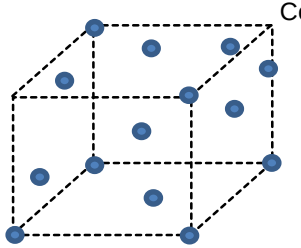
replicate



Fractional Factorial
2^{3-1} design



Response Surface
Central Composite design



Optimal Design
IV-optimal