Presented at the 2010 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com

Simulation-based Experimentation: How to Scope and Estimate Development Projects

Karen Mourikas SD&E - Experimentation

Denise Nelson BR&T – Affordability

ISPA/SCEA Workshop

June 2010

How to Estimate Costs of Simulation-based Experimentation

- Overall Problem Statement
  - Need to estimate the cost of simulation-based experimentation
    - Accurately
    - Consistent approach
    - Easy to use
    - Defendable estimates
    - Value-added
- Solution
  - Develop a standard process and toolkit to estimate the cost and schedule of simulation-based experimentation
    - Using standard cost estimating tools
    - Incorporating standard processes
    - Based on historical data

Objective of Presentation

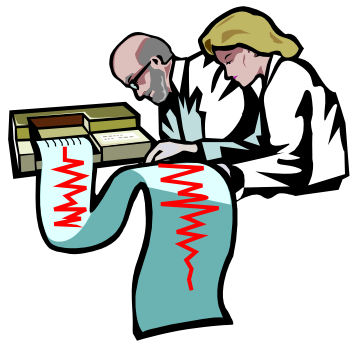
- To provide overview of Simulation-based Experimentation
- To introduce “CostX for Experimentation” Process
- To review Statistical Analysis options used to address the problem

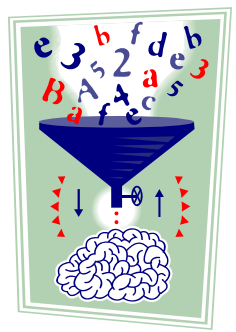
- Outline
  - Problem to be addressed
  - Simulation-based Experimentation
    - In a Nutshell
  - Experimentation Cost Estimating
    - Process and Data
  - Statistical Analysis
    - Options

Experimentation

- Definition of Experimentation *
  - The process of conducting a test under controlled conditions to
    - examine the validity of a hypothesis, or
    - determine the efficacy of something previously untried.

- Experimentation in general
  - Consists of gathering and examining data
  - Explores and Answers Questions with Analyses and Observations
  - Provide Insights to Questions

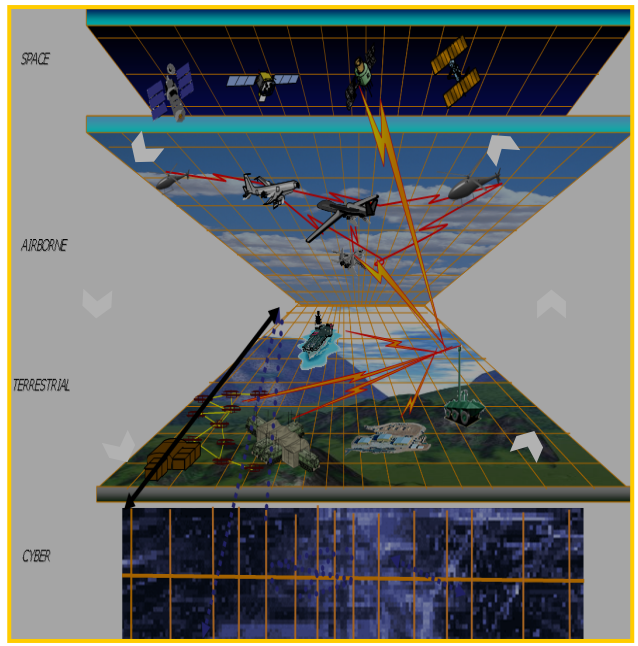


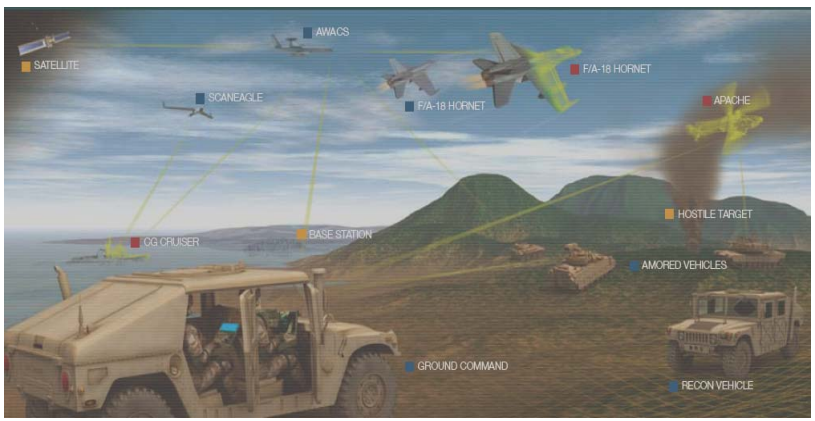


* www.thefreedictionary.com

Simulation-based Experimentation

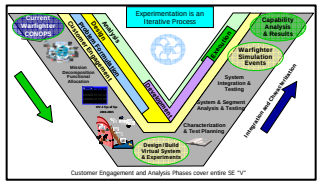
- Simulation-based Experimentation
  - Conducting experiments with the use of virtual and constructive
    - Models
    - Simulations
    - Visualizations
  - Live-Virtual-Constructive (LVC) Components
    - Live : Real people operating real systems
    - Virtual : Real people operating simulated systems
    - Constructive : Automated systems simulating the actions of people and systems

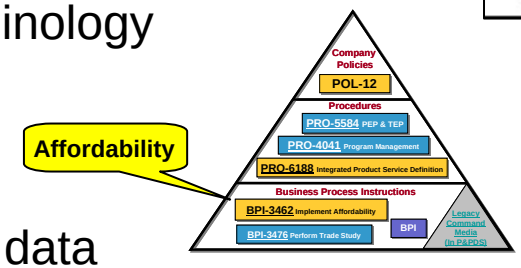




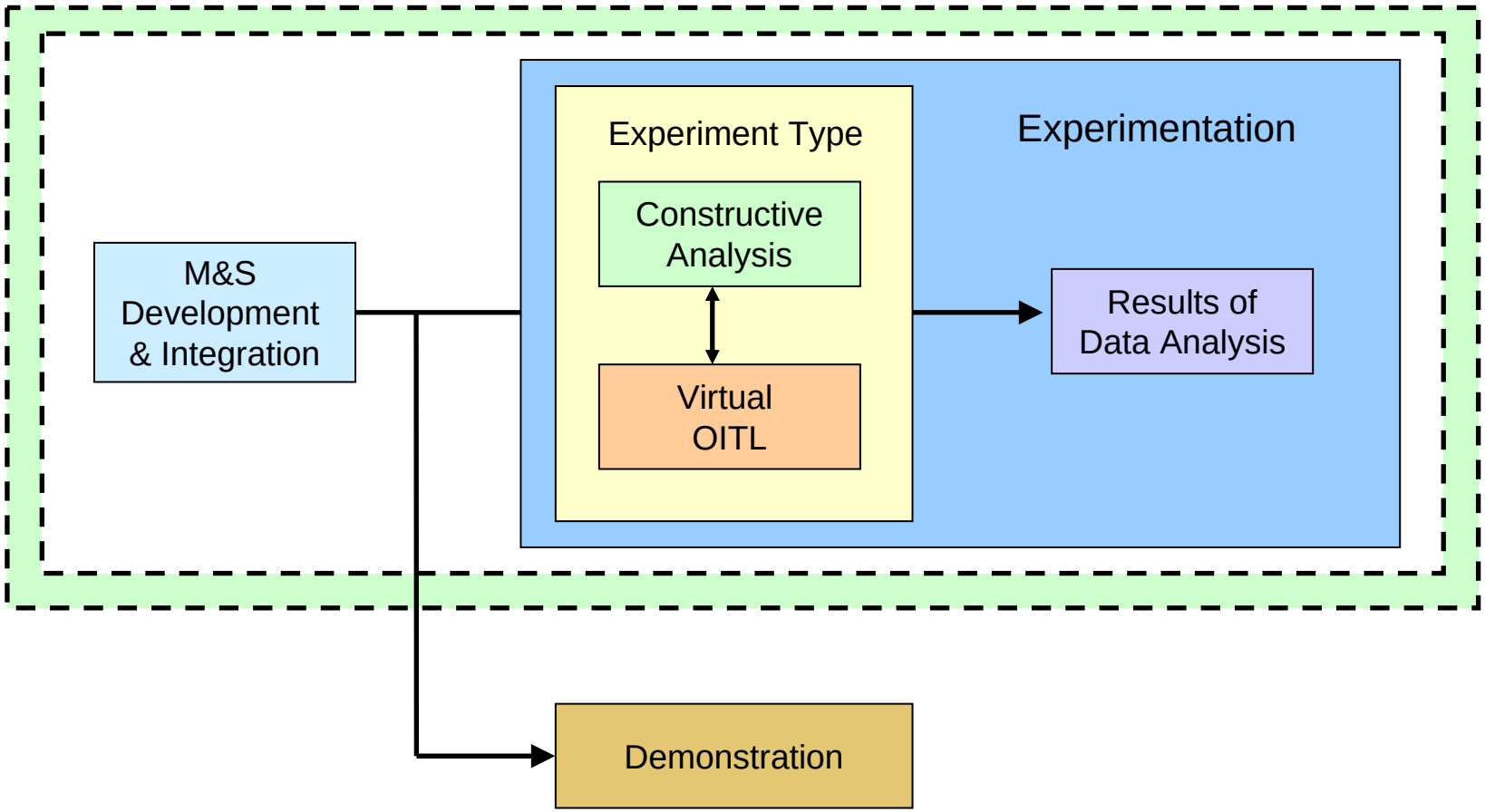
Proposed Solution

- Develop a cost estimating toolkit that is
  - Easy to use
    - Excel-like interface / Web-based GUI (future development)
    - Limited inputs required
    - Quick estimates and standard reports produced
  - Incorporates Standard Processes and Terminology
    - Boeing and Industry
    - Experimentation and Cost Estimating
  - Based on comprehensive, historical Boeing data
    - 70 completed experimentation projects: cost and scope data
    - Multiple sites, organizations, project type and scope
  - Value-added to the users
    - Not just *another* task to do





Focus of Experimentation Data Collection



Project Data

- 70 Completed Projects
  - Recent data (2006-2009)
  - 31 in work – to be collected
- Various project types
  - M&S development & integration
  - Constructive Analysis
  - Virtual OITL simulations
- Leverage reuse
  - Simulations
  - Environment
  - Team Management
  - Data / Analysis
- Scope & Complexity
  - Short to long range duration
  - Simple to Complex simulations and integration

	Proj Info	Complexity			Execution		Cost/Sched			
Project ID	POC	Type	Num Sys	VCL	Num MOEs	Num Sites	Mths	Cost ($K)		POP
a1	John Smith	Modeling	5	C	0	1	4	$	90	4Q 2008
a2	Pocahontas	Analysis	12	C	4	2	10	$	450	mar 07-dec 07
a3										
a4										
a5										
a6										

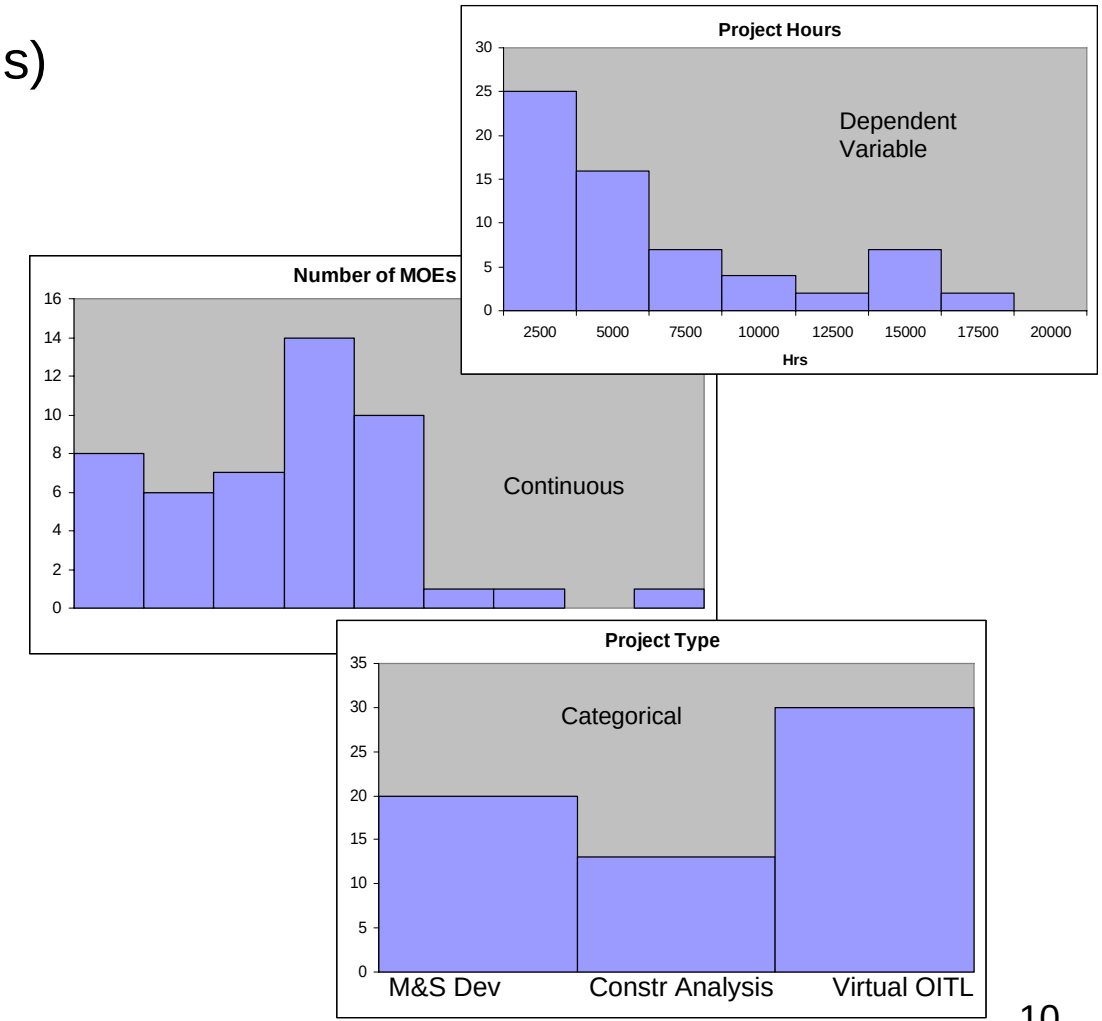
	Reusability							Team	
Project ID	Modeling	Tool Developmt	Tool Integration	Environment	Programatic	Data Collection	Analysis Results	Individual Experience	Team Dynamics
a1	B	B	E+	A	A	B	A	E	E
a2	C	A	B-	B	B	C	B-	B+	B
a3	E	B	A	A	E	E	C	C	D
a4	C	C	A	E	D	A	B	D	E
a5	B	B	A	B	A	B	A	A	B
a6	E	B	E	B	D	E	D	B	B

Statistical Analysis Options

- Database constantly changing
  - Baselined database (configuration control)
  - To be updated 2 times per year
  - Approximately 30 new data points expected per year
- Statistical Analysis
  - Issue with data
    - Lots of qualitative (categorical) data
    - Relatively small dataset
  - Options
    - Random Forests
    - Linear / Multivariate Regression
    - Adjustments to data
  - Goal
    - Determine which variables are most important
    - Develop model with good prediction capability

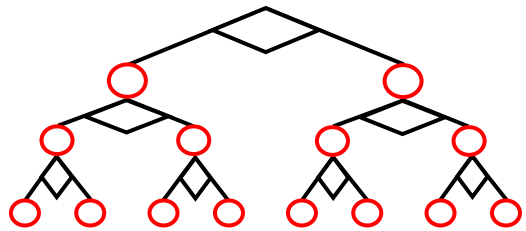
Statistical Analysis: Descriptive Statistics

- 70 data points (observations)
  - 2 Responses
    - Hours (Cost)
    - Duration (Schedule)*
  - 18 Potential Predictors
    - 4 Continuous
    - 4 Categorical Not-ordered
      - 1 with 2 levels (y/n)
      - 2 with 3 levels
      - 1 with 4 levels
    - 10 Ordinal
      - With 5 levels



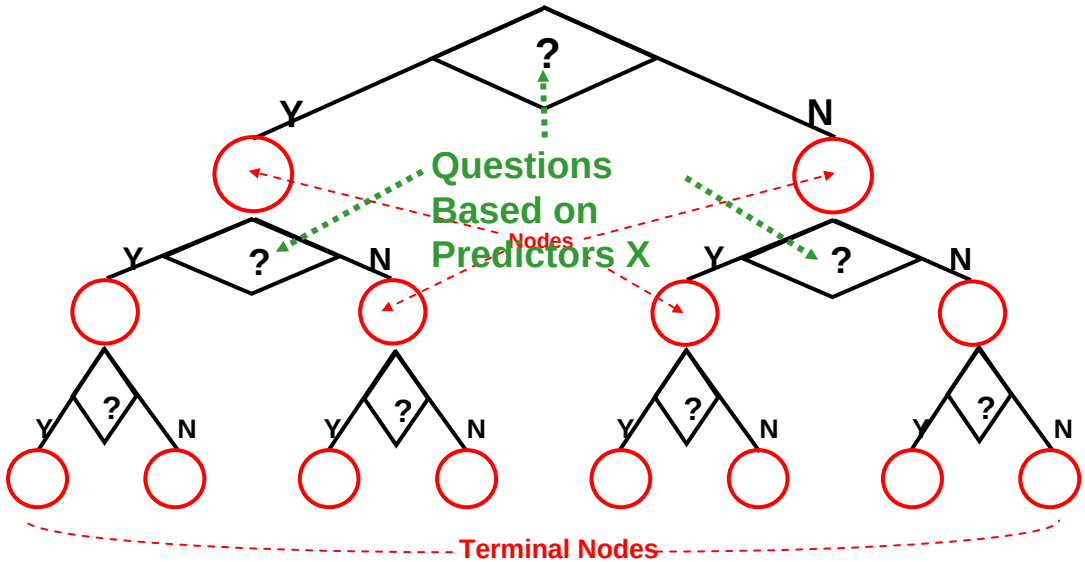
Statistical Analysis: Random Forests

- Random Forests generate multiple decision trees randomly constructed
- A Single Decision Tree is interpretable
  - But not a great predictor



- Rather than pick one “Best” model (tree)
  - Fit collection of decision tree models (forest)
  - Leads to greater accuracy but more difficult to interpret
  - Can handle “wide” problems: small n, large p
- “Black Box” Process
  - Not easily interpretable
  - Predicts a cost number
    - Not a parametric equation

Picture of Single Tree



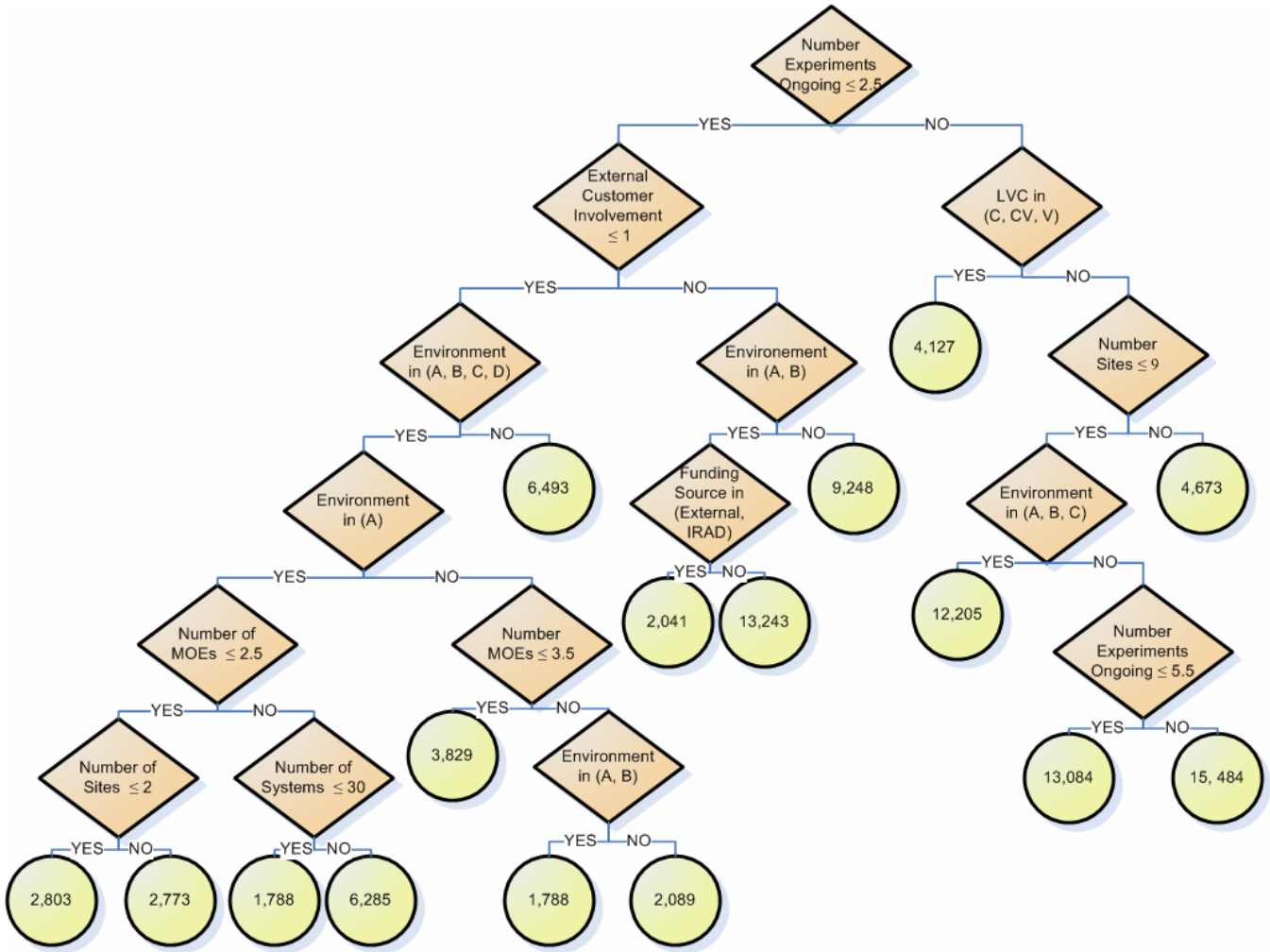
Build Model:

- At each node, split on a predictor variable that divides the data into two parts that most reduce prediction error.
- Each node is predicted by the average Hours.

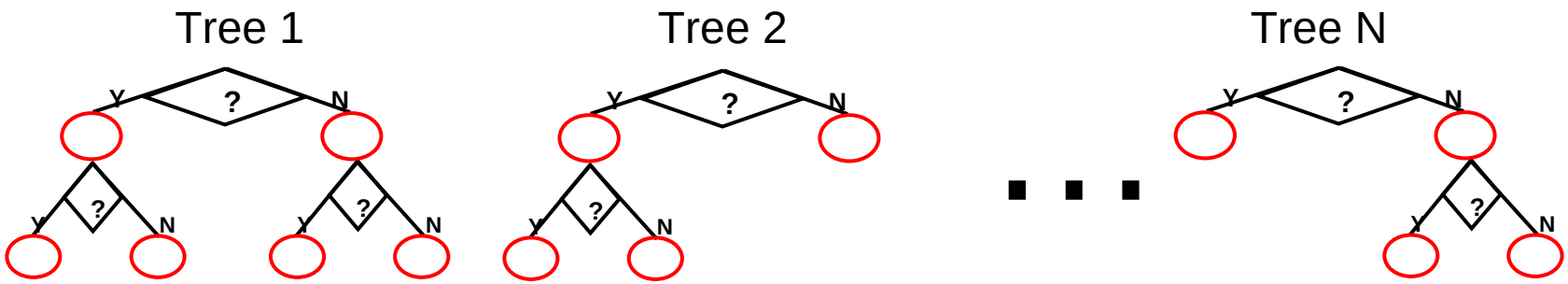
Predict:

- Drop a supplier down the tree to a terminal node
- Predicted Hours = average Hours at each terminal node

Example of Random Forest Generated Tree



Random Forest



Build Model:

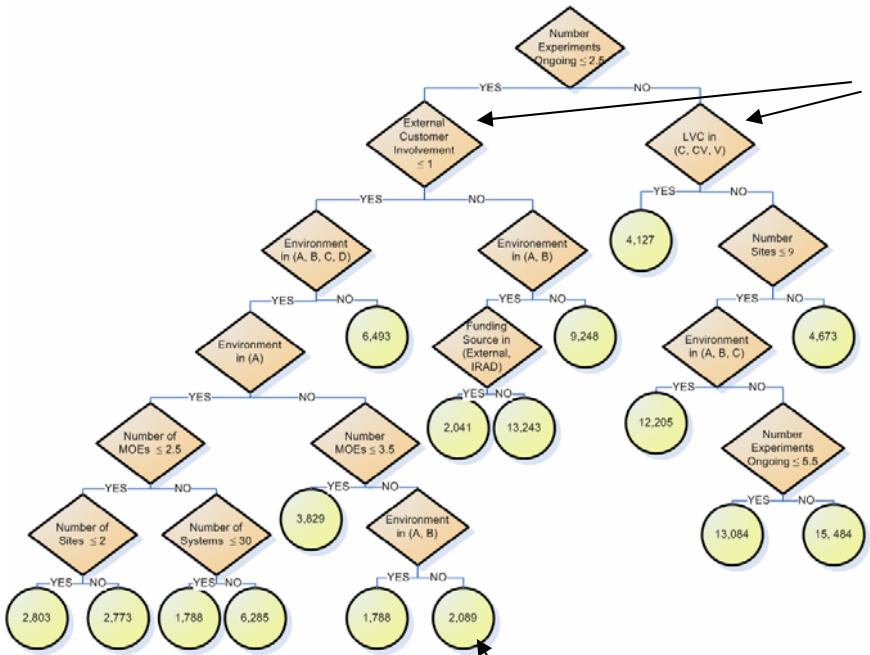
•Create diverse collection of trees (forest) by sampling rows and columns of the data

Predict:

• Drop a supplier down each tree to a terminal node

• Predicted Hours = Average Hours in each terminal node averaged over all the trees

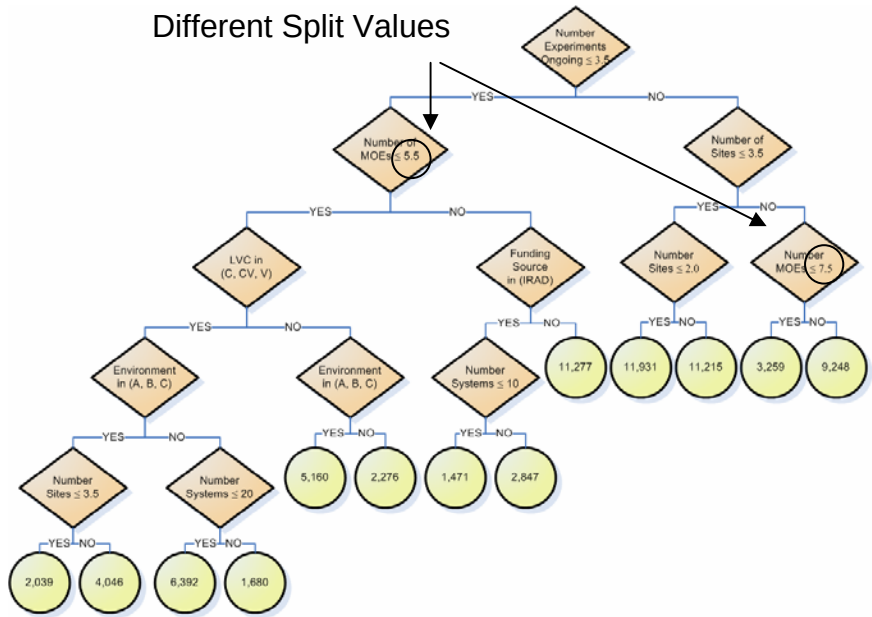
Examples of Trees



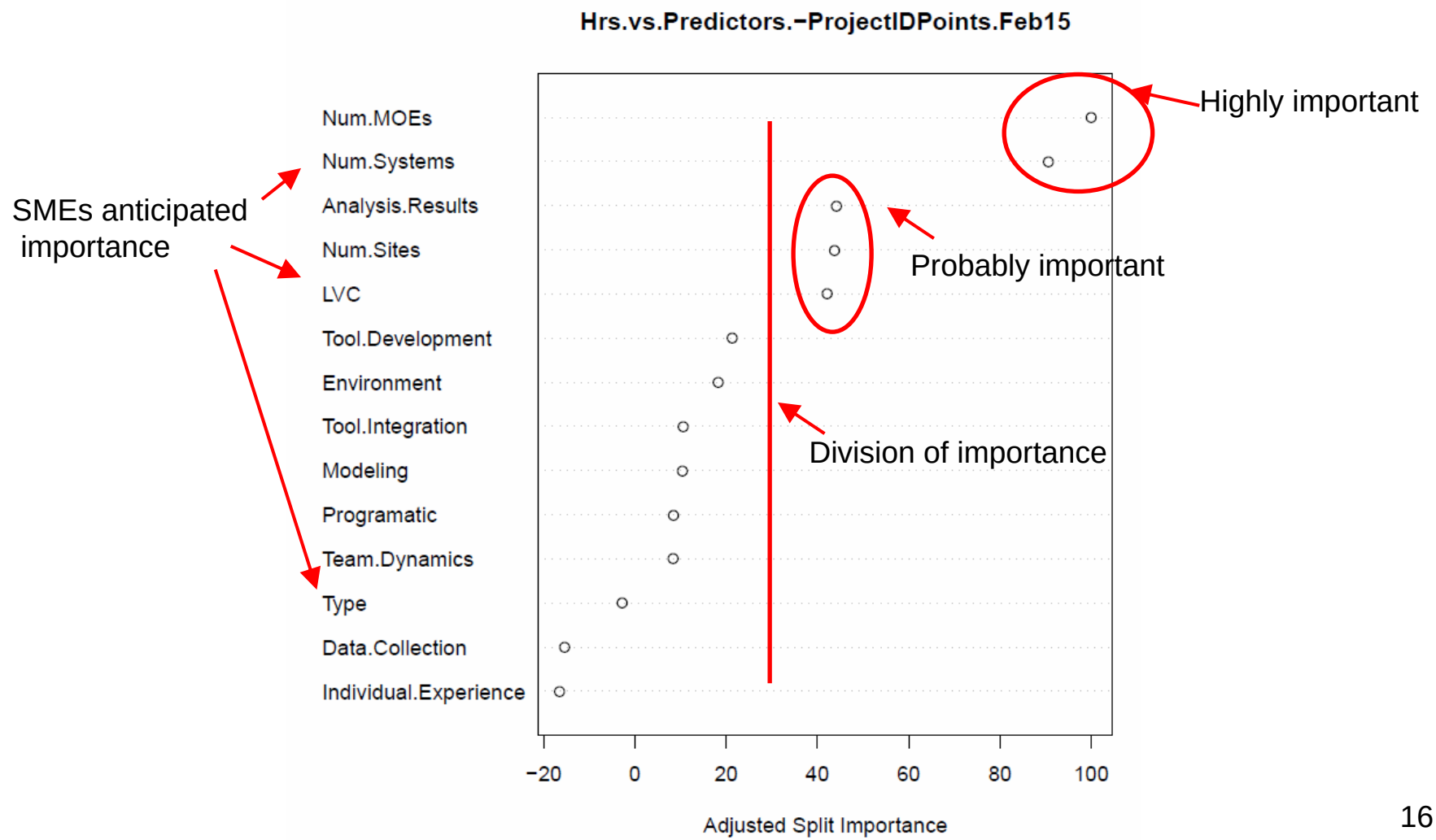
Different Predictor Variables

Average of hours for this path

Different Split Values

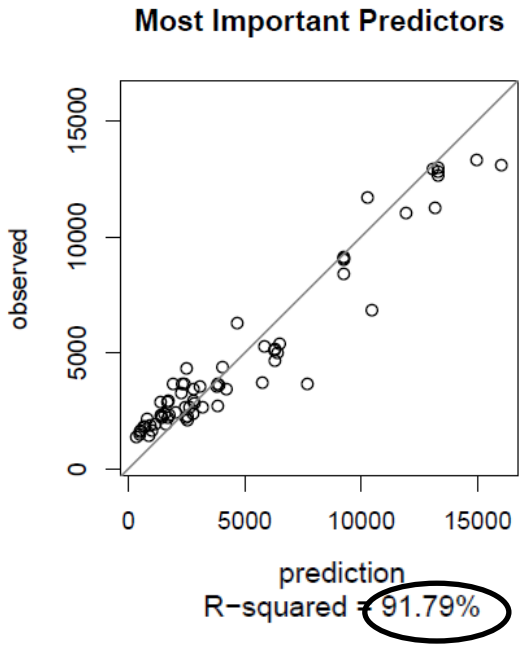


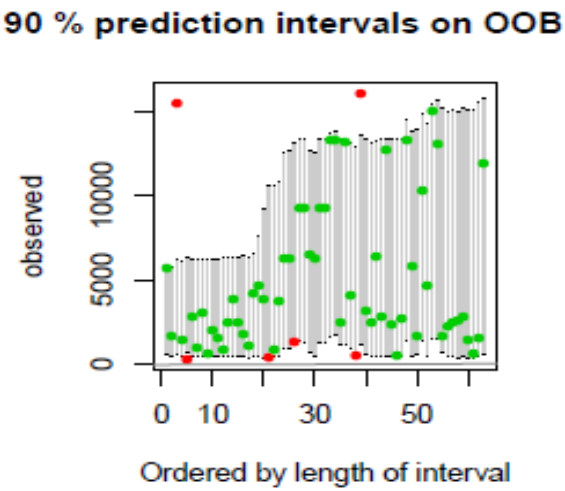
Variable Importance from Random Forest



Results of Random Forest Analysis

- Accuracy / Goodness of Fit
  - Removed some variables based on Variable Importance
    - Improved R^2
  - $R^2 \sim 92\%$
  - Wide Prediction Intervals
    - Need more data to develop smaller prediction intervals





Statistical Analysis: Regression

- Assumption: model is NOT linear
  - Expect interaction and/or quadratic terms
- Easily understandable relationship between input variables & responses
- Lots of categorical data
  - Numerical data easier to analyze
  - Converted Categorical data
    - Based on engineering judgement
  - Is this statistically valid?
    - Goal: Develop model with good prediction capability
    - Hence reasonable approach to take
- Accuracy / Goodness of Fit
  - R^2 (Variability explained by model) ~70%

Analysis Options Evaluation

Random Forests

- Pros
  - Can be used with categorical data
  - Can be used with small data sets
  - Statistical Results R2 ~ .92
  - Prediction accuracy: very good
  - Data Mining community big supporter
- Cons
  - Black Box
  - Results in a number, not a CER
  - Prediction Intervals need lots of data (more than we have)

Regression

- Pros
  - Familiar Process
  - Results in a CER
  - Statistical Results : R2 ~ .70
  - Prediction Accuracy: good
- Cons
  - Small dataset => limited DOF
  - Categorical to numeric data conversion
  - Prediction intervals easier to obtain (with assumptions)
  - Validity of approach

Next Steps for 2010

- Data Collection/Analysis
  - Collect new data
    - Update database
  - Analyze additional data
    - Update cost estimating relationships
  - Analyze pre and post project scope
    - Adjust estimate based on leads' knowledge

Contributor Information:

Karen Mourikas: Phantom Works/SD&E	714-580-9609	karen.mourikas@boeing.com
Denise Nelson: BR&T/Affordability	714-389-8501	denise.j.nelson@boeing.com
Nancy Gove: BR&T/Cyber Analytics	425-373-2698	nancy.gove@boeing.com
Jim Schimert : BR&T/Cyber Analytics	425-373-2728	james.schimert@boeing.com
Julio Peixoto: EOT/Applied Statistics	425-373-2720	julio.l.peixoto@boeing.com