AIR ACADEMY ASSOCIATES

Simplify, Perfect, Innovate

Predictive Modeling for BIG (and small) Data

19th ITEA Test Instrumentation Workshop
14th Directed Energy Test & Evaluation Workshop
May 14, 2015
Las Vegas, NV

15-PREDMOD-5A

Mark J. Kiemele, Ph.D.
President and Co-Founder
Air Academy Associates

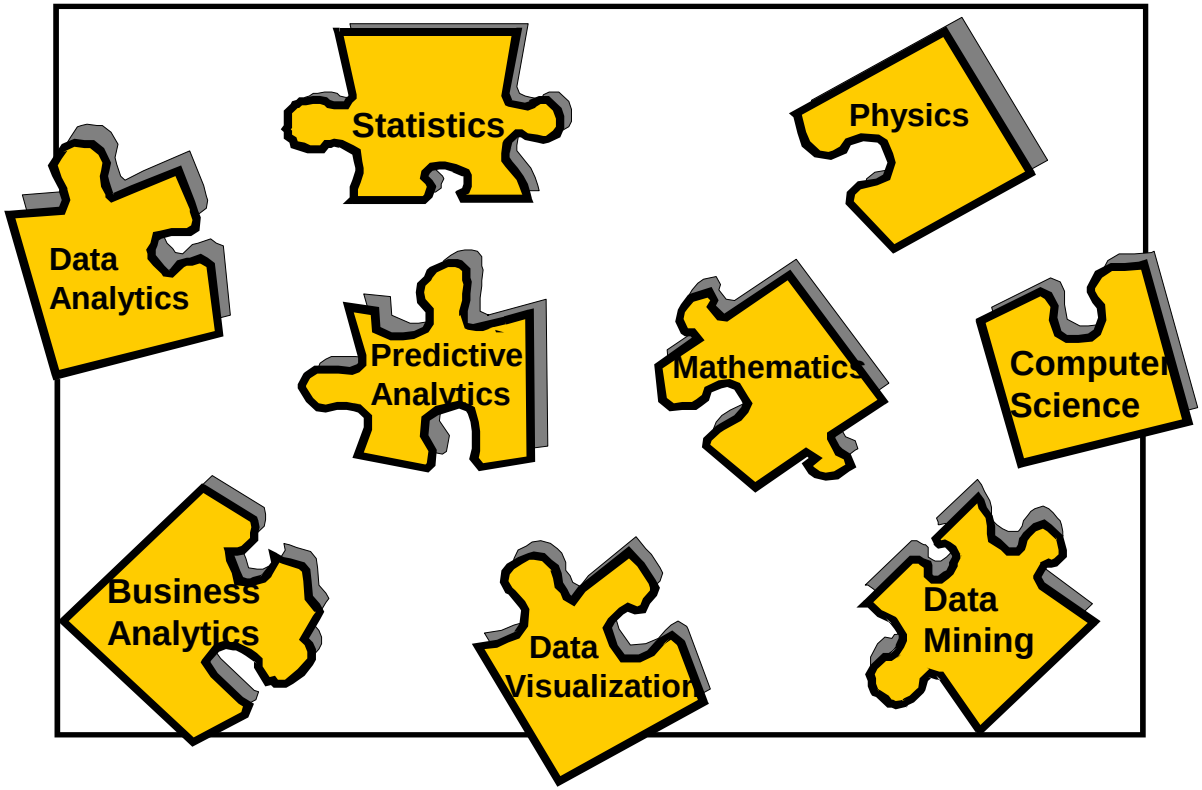
Office: 719-531-0777
Cell: 719-337-0357
mkiemele@airacad.com
www.airacad.com

Outline

- **Big Data and the Advent of Data Science**
- **Descriptive, Predictive, and Prescriptive Analytics**
- **Predictive Modeling Using Regression**
  - Using historical data
  - Using data collected from DOE
- **Best Practices for "operationalizing" predictive modeling**
- **Cautions in the Big Data world**

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

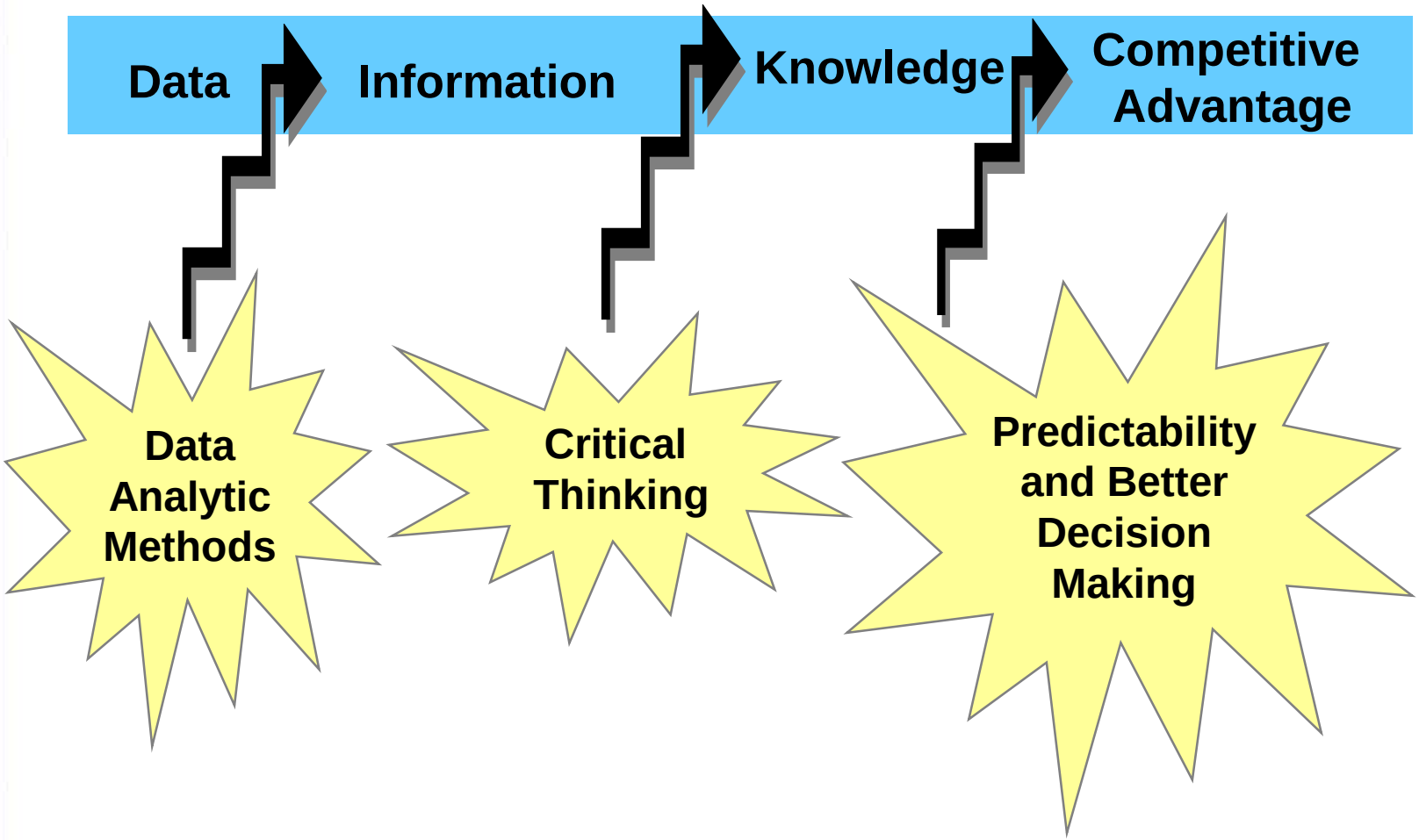
The Big Data Puzzle



"A Set of Disjointed Pieces?" or "Do the Pieces Fit Together?"

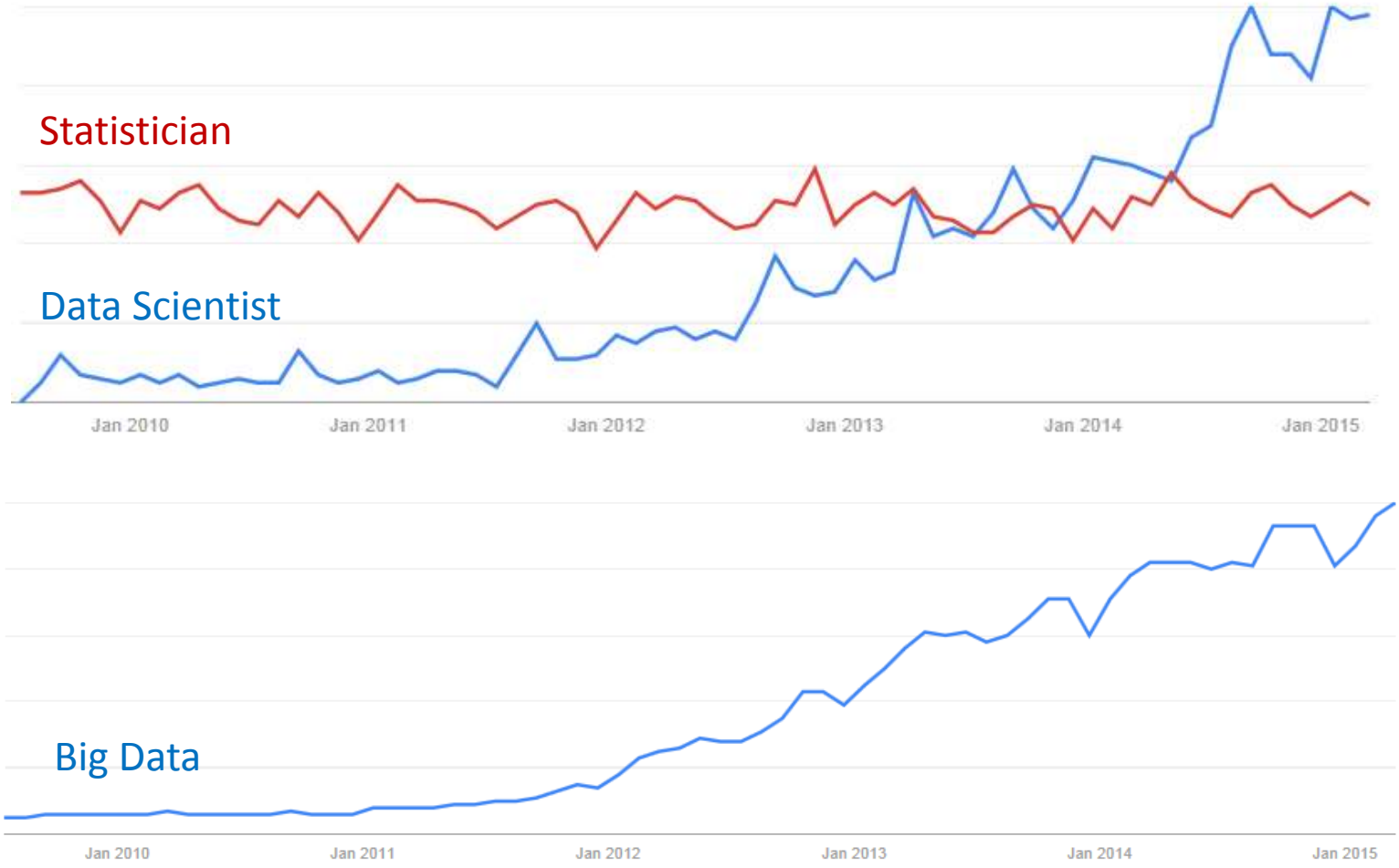
AIR ACADEMY ASSOCIATES
Simplify, Perfect,, Innovate

Data Science:
Transforming Data into Competitive Advantage



AIR ACADEMY ASSOCIATES
Simplify, Perfect,, Innovate

Google Trends



AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

The Big "Data Science" Trend

- "People currently create as much data every two days as was previously created in all of history up to 2003."

  Google CEO Eric Schmidt

- "In the last few years, there has been an explosion in the amount of data that's available. Data is everywhere: sensors, transactions, government, social media, even your body – almost everything has been instrumented. The problem isn't finding data, it's figuring out what to do with it."

  Mike Loukides, O'Reilly Research

Air Academy Associates
Simplify, Perfect, Innovate

The Big “Data Science” Trend

- “For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have come to feel that my central interest is in data analysis which is intrinsically an empirical science.”

  John Tukey, Bell Labs

- “The key word in “Data Science” is not Data, it is Science.”

  Jeff Leek, “Simply Statistics”

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

The Big "Data Science" Trend

- "I think data scientist is a sexed-up term for a statistician." Nate Silver, "The Signal and the Noise"

- "What differentiates data science from statistics is that data science is a holistic approach. We're increasingly finding data in the wild, and data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others." Tim O'Reilly, O'Reilly Research

- "The half-life of a buzzword is about 7 years, but 'data science' could stick around longer than that, just like another incongruous joining of two words, 'computer science,' did. Gil Press, Forbes

Types of Analytics*

- **Descriptive**
  - Reports on the past, usually in summary or graphical form
- **Predictive**
  - Uses models based on past (historical) data to predict future outcomes
- **Prescriptive**
  - Uses embedded, continuously updated models in key processes to specify immediate optimal behaviors and actions

* ***Analytics 3.0*** by Thomas H. Davenport, Harvard Business Review, Dec., 2013

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Predictive Modeling Techniques

- **Regression Analysis**
- **Decision Trees**
- **Discriminant Analysis**
- **Neural Networks**
- **Cluster Analysis**

Air Academy Associates
Simplify, Perfect, Innovate

Historical Data Analysis

- **Using historical data analysis is a very efficient way to use data that may already be available.**
- **Can be used in any scenario.**
- **Can be used to develop a mathematical (prediction) model of a process without conducting a designed experiment.**
- **The major drawback to historical data is that there is more noise in it than is typically found in data obtained from a designed experiment.**

Air Academy Associates
Simplify, Perfect, Innovate

Historical (Haphazardly Collected) Data from a Plating Process

	A	B	C	D	E	F
1	Time	Bath Temp	% Nickel	Vendor	Phos	Y1
2	5	15	9	1	29	49.5
3	6	13	11	1	33	47.5
4	7	14	13	1	28	47.4
5	6	17	20	2	51	55.6
6	4	18	16	2	53	28.2
7	7	13	15	2	55	44.8
8	6	30	8	2	28	21.4
9	5	32	11	2	30	16.5
10	3	34	13	2	35	8.2
11	5	33	19	1	55	18.4
12	5	31	17	1	51	17.9
13	4	30	18	1	54	19.4
14	14	18	9	2	34	65.4
15	11	15	12	2	27	79.4
16	10	17	11	2	34	72.6
17	13	14	20	1	51	74.3
18	14	19	16	1	54	70.1
19	15	17	19	1	56	78.4
20	11	31	8	1	27	71.2
21	12	33	11	1	34	68.7
22	14	37	13	2	31	93.1
23	11	32	12	1	28	66.4
24	12	30	17	2	56	74.6
25	14	29	18	2	52	70.6
26	13	28	20	2	57	68.4
27	9	26	16	1	43	50.4
28	10	23	13	2	46	78.6
29	8	24	14	1	40	49.7
30	8	24	15	2	43	53.8
31	11	28	12	1	46	73.1
32	5	23	13	1	43	33.9

	A	B	C	D	E	F
1	Time	Bath Temp	% Nickel	Vendor	Phos	Y1
33	7	35	15	1	40	38.5
34	9	14	16	1	46	59.1
35	8	24	19	1	57	50.2
36	7	23	11	1	32	51.3
37	11	26	16	2	44	66.8
38	5	24	15	2	43	27.6
39	9	31	14	2	37	61.2
40	7	19	12	2	38	54.4
41	10	24	19	2	58	62.4
42	7	27	8	2	34	58.3
43	6	31	20	1	52	38.3
44	14	24	12	2	29	66.9
45	3	35	15	1	42	17.6

The output variable Y1 is Plating Thickness

AIR ACADEMY ASSOCIATES

Simplify, Perfect, Innovate

Final Regression Output

Y-hat Model		Thickness			
Factor	Name	Coeff	P(2 Tail)	Tol	Active
Const		58.686	0.0000		
A	Time	31.332	0.0000	0.9819	X
B	Bath Temp	-3.977	0.0540	0.9523	X
AA		-12.200	0.0059	0.8457	X
AB		10.663	0.0041	0.8430	X

R^2	0.8841
Adj R^2	0.8722
Std Error	7.5236
F	74.3631
Sig F	0.0000
F_{Lof}	11.8710
Sig F_{Lof}	0.0806

Source	SS	df	MS
Regression	16837.3	4	4209.3
Error	2207.6	39	56.6
$Error_{Pure}$	10.0	2	5.0
$Error_{Lof}$	2197.6	37	59.4
Total	19045.0	43	

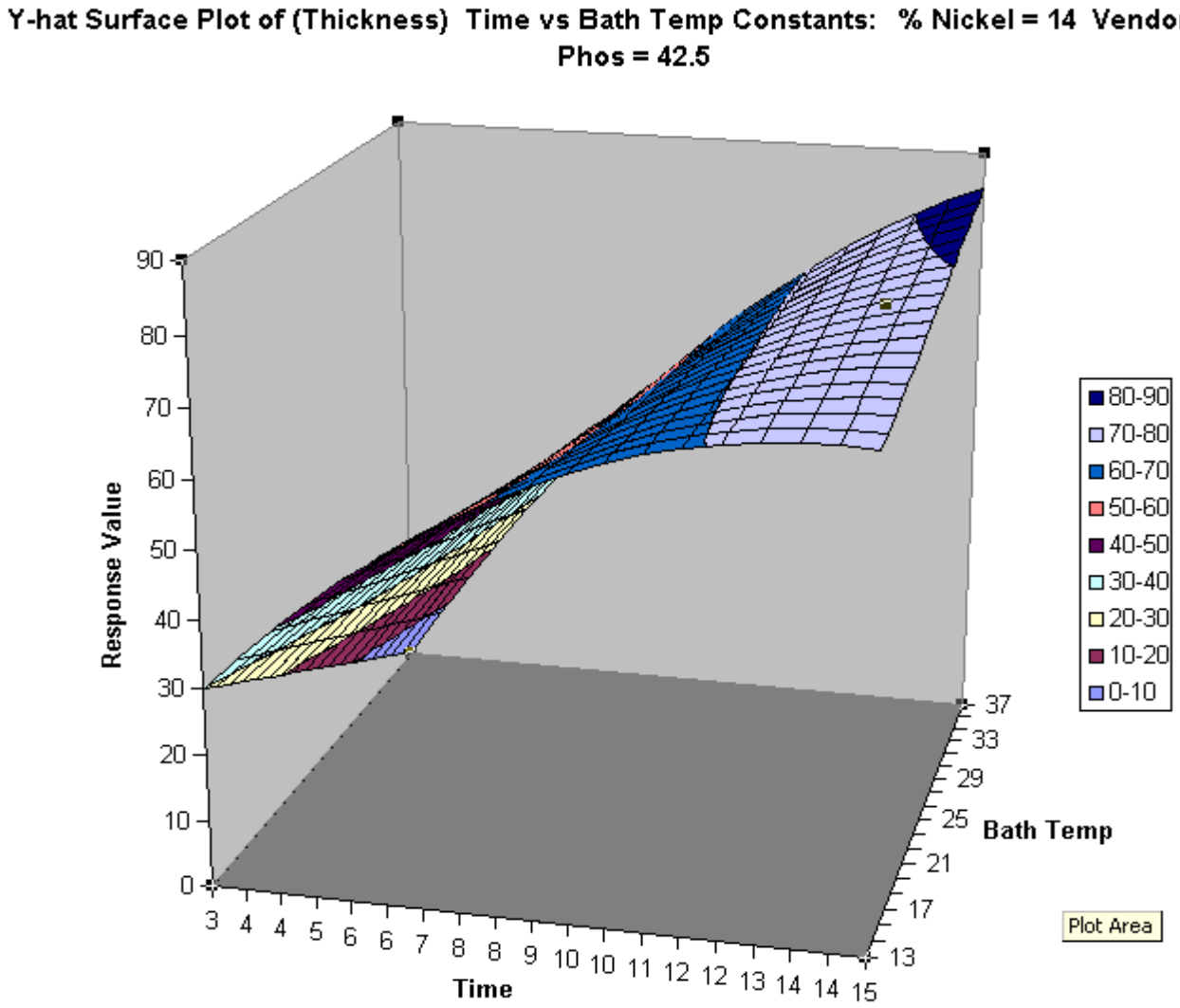
Factor	Name	Low	High	Exper
A	Time	3	15	9
B	Bath Temp	13	37	25
C	% Nickel	8	20	14
D	Vendor	1	2	1.5
E	Phos	27	58	42.5

Multiple Response Prediction				
			99% Confidence Interval	
	Y-hat	S-hat	Lower Bound	Upper Bound
Thickness	58.6863	7.5236	36.115	81.257

Prediction Model: $\hat{Y} = 58.7 + 31.3A - 3.98B - 12.2A^2 + 10.7AB$

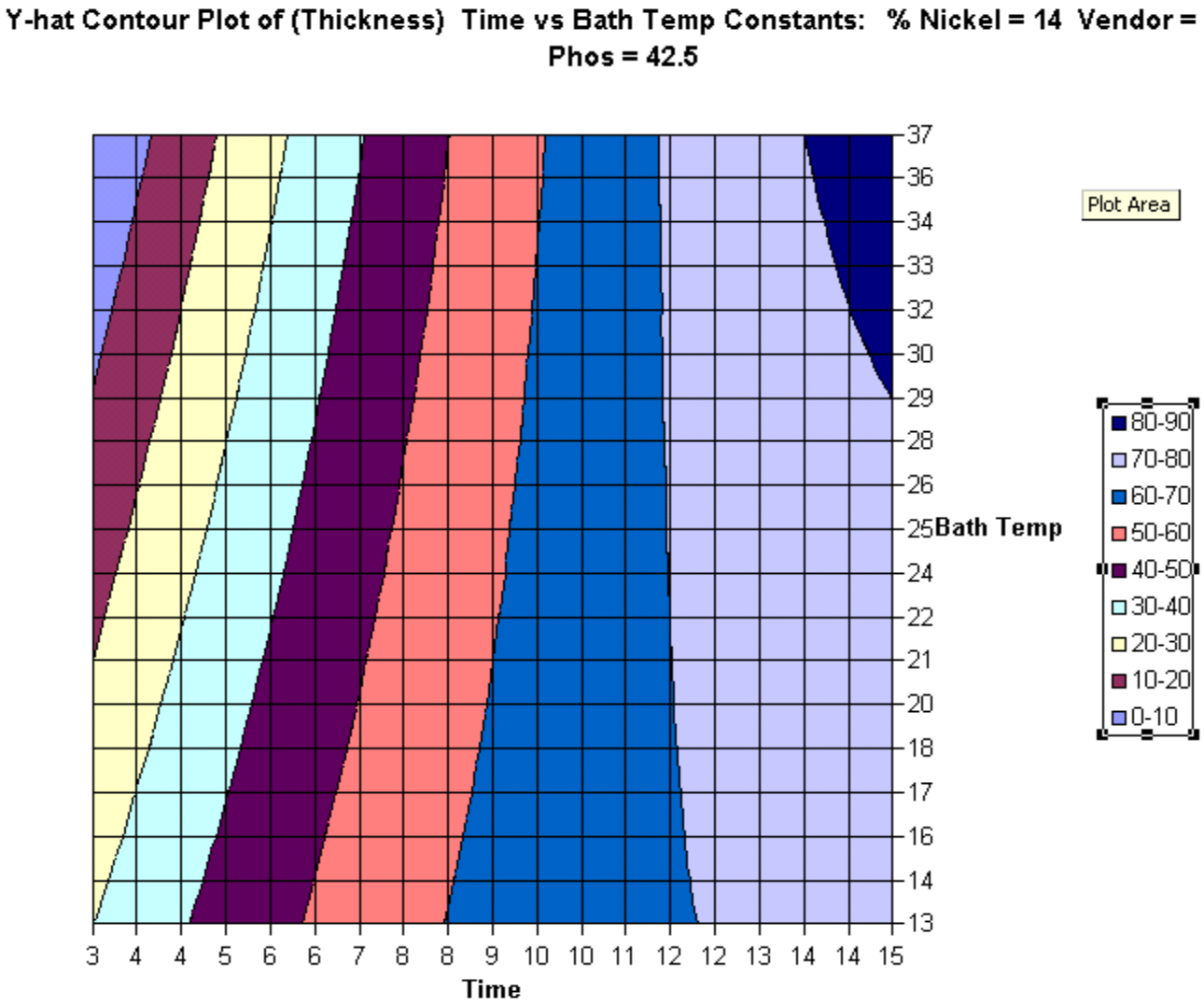
AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Surface Plot



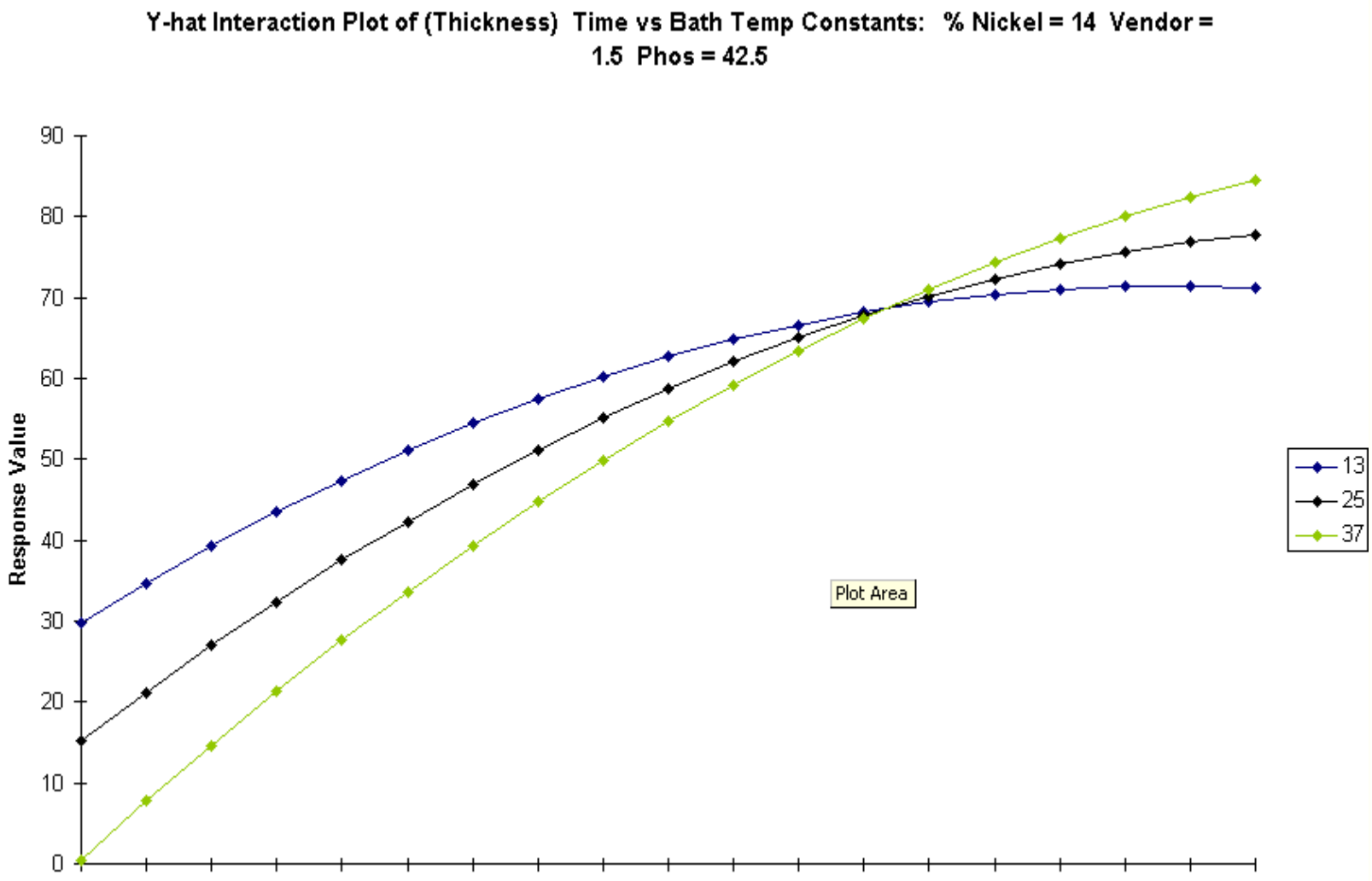
AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Contour Plot



Simplify, Perfect, Innovate

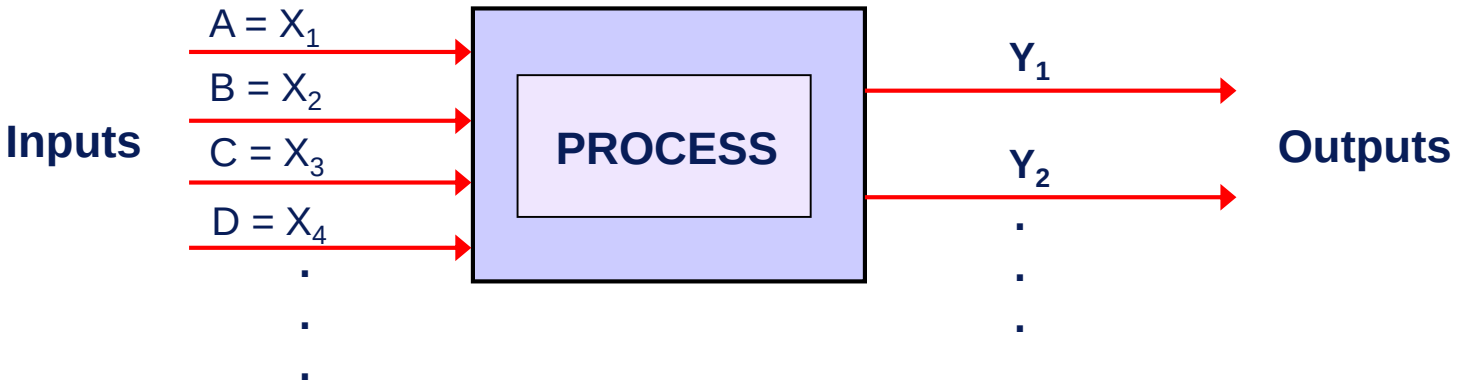
Interaction Plot



Air Academy Associates
Simplify, Perfect, Innovate

What is a Designed Experiment?

Purposeful changes of the inputs (factors) in order to observe corresponding changes in the output (response).



Run	X_1 X_2 X_3 X_4	Y_1 Y_2	$\bar{Y}$	S_Y
1				
2				
3				
.				
.				

Air Academy Associates
Simplify, Perfect, Innovate

Famous Quote

**"All experiments (tests) are designed;
some are poorly designed,
some are well designed.**

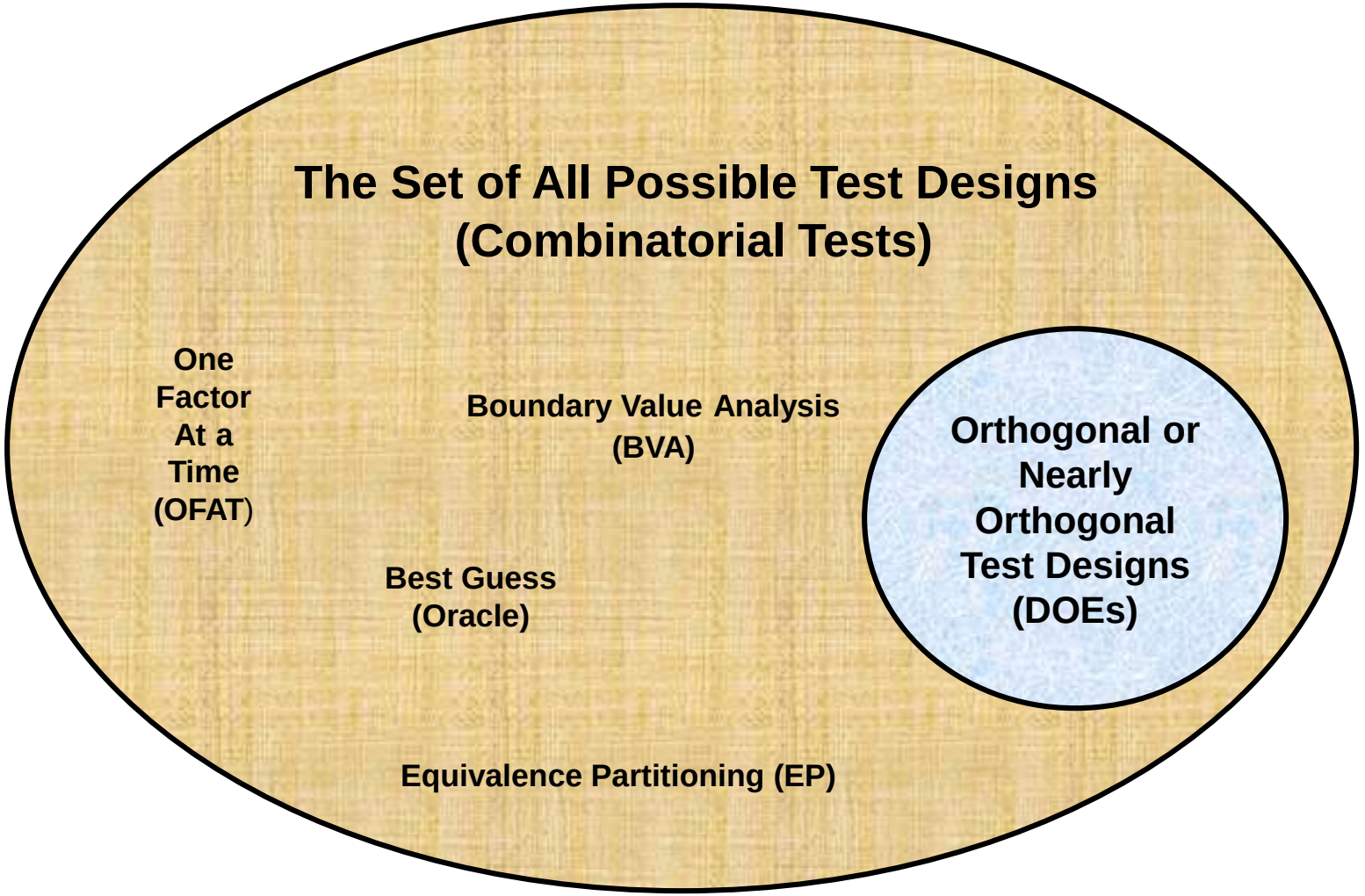
George Box (1919-2013), Professor of Statistics, DOE Guru

A Corollary:

**"Great data doesn't happen by accident.
It has to be designed into the process."**

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Design of Experiments (DOEs): A Subset of All Possible Test Design Methodologies



Air Academy Associates
Simplify, Perfect, Innovate

Design of Experiments (DOE):
Orthogonal or Nearly Orthogonal Designs

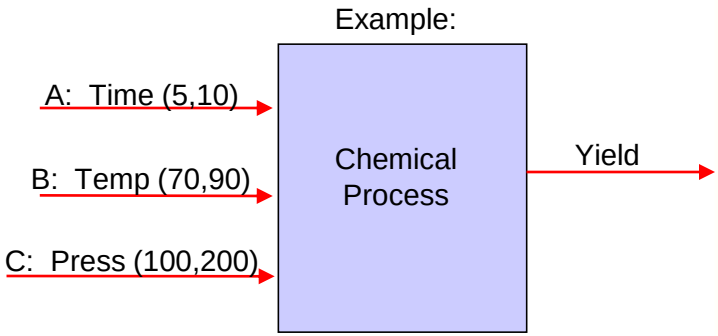
- FULL FACTORIALS (for small numbers of factors)
- FRACTIONAL FACTORIALS
- PLACKETT - BURMAN
- LATIN SQUARES
- HADAMARD MATRICES

Taguchi Designs

- BOX - BEHNKEN DESIGNS
- CENTRAL COMPOSITE DESIGNS

Response Surface Designs

- HIGH THROUGHPUT TESTING (ALL PAIRS)
- NEARLY ORTHOGONAL LATIN HYPERCUBE DESIGNS



SIMPLE DEFINITION OF A TWO-LEVEL ORTHOGONAL DESIGN

Run	Actual Settings			Coded Matrix			Interactions	
	(5, 10) A: Time	(70, 90) B: Temp	(100, 200) C: Press	(A) Time	(B) Temp	(C) Press	(AB) Uncoded	(AB) Coded
1	5	70	100	-1	-1	-1	350	+1
2	5	70	200	-1	-1	+1	350	+1
3	5	90	100	-1	+1	-1	450	-1
4	5	90	200	-1	+1	+1	450	-1
5	10	70	100	+1	-1	-1	700	-1
6	10	70	200	+1	-1	+1	700	-1
7	10	90	100	+1	+1	-1	900	+1
8	10	90	200	+1	+1	+1	900	+1

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

The Power of Orthogonality in a Test Design
(Vertical and Horizontal Balance)

A Full Factorial Design for 3 Factors A, B, and C, Each at 2 levels:

Run	A	B	C	AB	AC	BC	ABC
1	-	-	-	+	+	+	-
2	-	-	+	+	-	-	+
3	-	+	-	-	+	-	+
4	-	+	+	-	-	+	-
5	+	-	-	-	-	+	+
6	+	-	+	-	+	-	-
7	+	+	-	+	-	-	-
8	+	+	+	+	+	+	+

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Summarizing Design of Experiments (DOE)

- An optimal data collection methodology
- “Interrogates” the process
- Used to identify important relationships between input and output factors
- Identifies important interactions between process variables
- Can be used to optimize a process
- Changes “I think” to “I know” (with a certain level of confidence)

Air Academy Associates
Simplify, Perfect, Innovate

Three Major Reasons for Using a DOE

- **Screening** (Data Cleansing)
  - For testing many factors in order to **separate** the critical factors from the trivial many.
- **Modeling** (Prediction and Optimization)
  - For building **functions** that can be used to predict outcomes, assess risk, and optimize performance. These **include** the ability to evaluate **interaction and higher order effects**.
- **Performance Verification and Validation**
  - For **confirming** that a system performs in accordance with its specifications/requirements.

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Motivation for DOE from Dr. Gilmore (DOT&E)

(from his 26 June 2013 memo on Flawed Applications of DOE)

1. One of the most important goals of operational testing is to **characterize** a system's effectiveness over the operational envelope.

2. I advocate the use of DOE to ensure that test programs are able to **determine the effect of factors on** a comprehensive set of operational mission-focused and **quantitative response variables**.

3. Future test plans must state clearly that data are being collected to measure a particular response variable (possibly more than one) in order to **characterize** the system's performance by examining the effects of multiple factors … and clearly delineating what statistical **model** (e.g., main effects and interactions) is motivating … the variation of the test.

4. Confounding factors must be avoided.

5. Another pitfall to avoid is relying on binary metrics as the primary response variable.

Value Delivery: Reducing Time to Market for New Technologies in the Design Phase

INPUT

OUTPUT

Pitch <) (0, 15, 30)

Roll <) (0, 15, 30)

W1F <) (-15, 0, 15)

W2F <) (-15, 0, 15)

W3F <) (-15, 0, 15)

Modeling Flight Characteristics of New 3-Wing Aircraft

Six Aero-Characteristics

- **Total # of Combinations = 3^5 = 243**
- **Central Composite Design: n = 30**

Patent Holder: Dr. Bert Silich

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Predictive Models for Aircraft Performance

C_L = .233 + .008$(P)^2$ + .255(P) + .012(R) - .043(WD1) - .117(WD2) + .185(WD3) + .010(P)(WD3) - .042(R)(WD1) + .035(R)(WD2) + .016(R)(WD3) + .010(P)(R) - .003(WD1)(WD2) - .006(WD1)(WD3)

C_D = .058 + .016$(P)^2$ + .028(P) - .004(WD1) - .013(WD2) + .013(WD3) + .002(P)(R) - .004(P)(WD1) - .009(P)(WD2) + .016(P)(WD3) - .004(R)(WD1) + .003(R)(WD2) + .020$(WD1)^2$ + .017$(WD2)^2$ + .021$(WD3)^2$

C_Y = -.006(P) - .006(R) + .169(WD1) - .121(WD2) - .063(WD3) - .004(P)(R) + .008(P)(WD1) - .006(P)(WD2) - .008(P)(WD3) - .012(R)(WD1) - .029(R)(WD2) + .048(R)(WD3) - .008$(WD1)^2$

C_M = .023 - .008$(P)^2$ + .004(P) - .007(R) + .024(WD1) + .066(WD2) - .099(WD3) - .006(P)(R) + .002(P)(WD2) - .005(P)(WD3) + .023(R)(WD1) - .019(R)(WD2) - .007(R)(WD3) + .007$(WD1)^2$ - .008$(WD2)^2$ + .002(WD1)(WD2) + .002(WD1)(WD3)

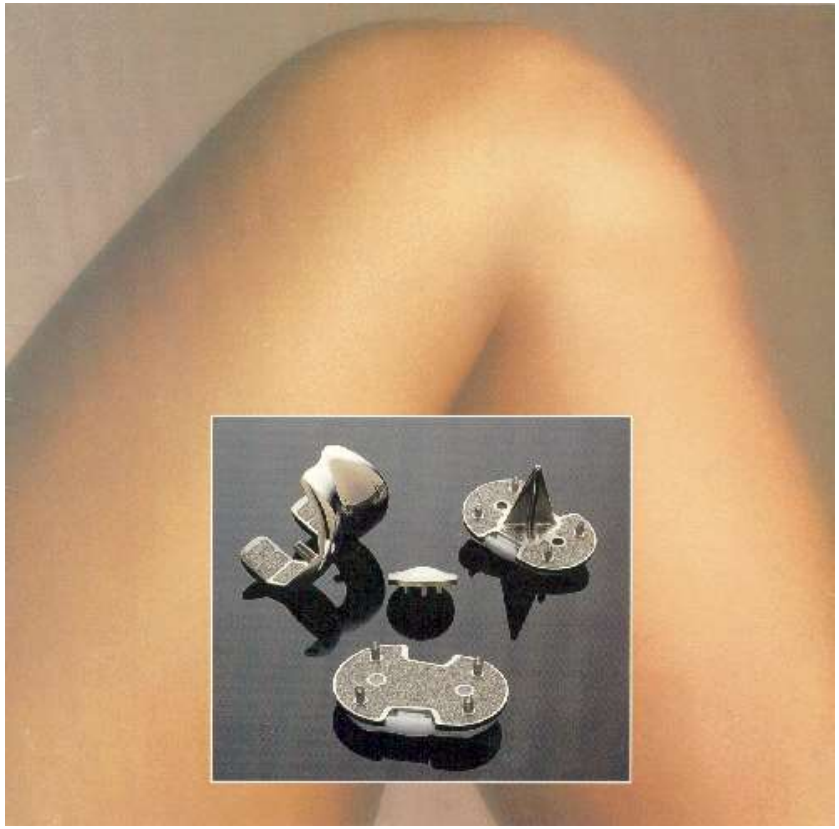
C_{YM}= .001(P) + .001(R) - .050(WD1) + .029(WD2) + .012(WD3) + .001(P)(R) - .005(P)(WD1) - .004(P)(WD2) - .004(P)(WD3) + .003(R)(WD1) + .008(R)(WD2) - .013(R)(WD3) + .004$(WD1)^2$ + .003$(WD2)^2$ - .005$(WD3)^2$

C_e = .003(P) + .035(WD1) + .048(WD2) + .051(WD3) - .003(R)(WD3) + .003(P)(R) - .005(P)(WD1) + .005(P)(WD2) + .006(P)(WD3) + .002(R)(WD1)

AIR ACADEMY ASSOCIATES

Simplify, Perfect, Innovate

Fusing Titanium and Cobalt-Chrome



Courtesy Rai Chowdhary

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

DOE "Market Research" Example

Suppose that, in the auto industry, we would like to investigate the following automobile attributes (i.e., factors), along with accompanying levels of those attributes:

A: Brand of Auto:	**-1 = foreign**		**+1 = domestic**
B: Auto Color:	**-1 = light**	**0 = bright**	**+1 = dark**
C: Body Style:	**-1 = 2-door**	**0 = 4-door**	**+1 = sliding door/hatchback**
D: Drive Mechanism:	**-1 = rear wheel**	**0 = front wheel**	**+1 = 4-wheel**
E: Engine Size:	**-1 = 4-cylinder**	**0 = 6-cylinder**	**+1 = 8-cylinder**
F: Interior Size:	**-1 ≤ 2 people**	**0 = 3-5 people**	**+1 ≥ 6 people**
G: Gas Mileage:	**-1 ≤ 20 mpg**	**0 = 20-30 mpg**	**+1 ≥ 30 mpg**
H: Price:	**-1 ≤ $20K**	**0 = $20-$40K**	**+1 ≥ $40K**

In addition, suppose the respondents chosen to provide their preferences to product profiles are taken based on the following demographic:

J: Age:	**-1 ≤ 25 years old**	**+1 ≥ 35 years old**
K: Income:	**-1 ≤ $30K**	**+1 ≥ $40K**
L: Education:	**-1 < BS**	**+1 ≥ BS**

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

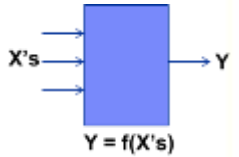
Google on DOE
(quotes* from Daryl Pregibon, Google Engineer)

"From a user's perspective, a query was submitted and results appear. From Google's perspective, the user has provided an opportunity to test something. What can we test? Well, there is so much to test that we have an Experiment Council that vets experiment proposals and quickly approves those that pass muster."

" We <u>evangelize</u> experimentation to the extent that we even provide a mechanism for advertisers to run their own experiments."

*** Taken From: *Statistics @ Google* in Amstat News, May 2011**

The Value of Predictive Models



- Simple and compact way of understanding relationships between performance measures or response variables (Y's) and the factors (X's) that influence them.
- Allows us to
  - Predict the response variable (y), with associated risk levels, before any change in the product or process is made.
  - Assess the product/process capability in the presence of uncontrolled variation or noise using Monte Carlo Simulation (DFSS tool: Expected Value Analysis).
  - Understand the impact of the factors (sensitivity analysis)
  - Optimize performance easily using DFSS tools such as parameter design and tolerance allocation.
- Greatly enhance one's knowledge of a product or process.
- In general, they are the gateway to systematic innovation.
- Provide a meaningful metric for the maturity in DFSS for any organization.

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Modeling The Drivers of Turnover*
(this predictive model is based on historical data)

*Adapted from Harvard Business Review article on Boston Fleet Bank, April 2004, pp 116-125

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Best Practices for "Operationalizing" Predictive Modeling

(i.e., changing the culture to one of habitually using it)

1. Coaching on projects is an absolute must.

2. A Keep-It-Simple-Statistically (KISS) approach with easy-to-comprehend materials and easy-to-use software.

3. Gaining and propagating quick-hitting successes.

4. Getting leadership on board and continuously re-invigorating them is necessary.

5. Developing a culture of continuously generating prediction models for the purpose of optimization, prediction, and risk assessment.

AIR ACADEMY ASSOCIATES

Simplify, Perfect, Innovate

Key Take-Aways

- Big Data is here to stay.
- We must still be cautious of big data. Why?
  - Correlation is not necessarily causality.
  - How data is collected makes a big difference on how easy or how hard it is to infer relationships between variables
  - If Data Science concerns itself only about data analysis techniques and does not include the science of properly collecting and gathering data, its "science" will suffer.
- Regression analysis provides a straightforward approach to building predictive models, even if the data was not obtained using a DOE.
- Learning about and making predictive modeling an integral part of an organization does NOT have to be difficult. Following the 5 Best Practices for "operationalizing" predictive modeling in an organization can make it happen.

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate