

LEAN SIX SIGMA IMPROVE PHASE

"If you torture the data long enough, it will confess." - Ronald H. Coase



COURSE CONTENT

Coverage:

- ANALYZE PHASE TOOLS
 - Regression techniques - Simple linear regression
 - OFAT experimentas
 - LEAN concepts

Introduction to Predictive Analysis



Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future.

With in the list of Predictive Analytics tools we will go in detail and understand about Regression analysis

Fundamental question...

Can we predict the Outcome Metric (Dependant Metric (or) Response Metric) by analyzing the impact or relationship that one or many Predictor Metric (Independent Metric (or) covariates) have on this outcome?

Answer to this question is **YES**

This takes us to one of the fundamental equations of six sigma

$$Y = f(x_1, x_2, x_3, x_4, \dots, x_n)$$

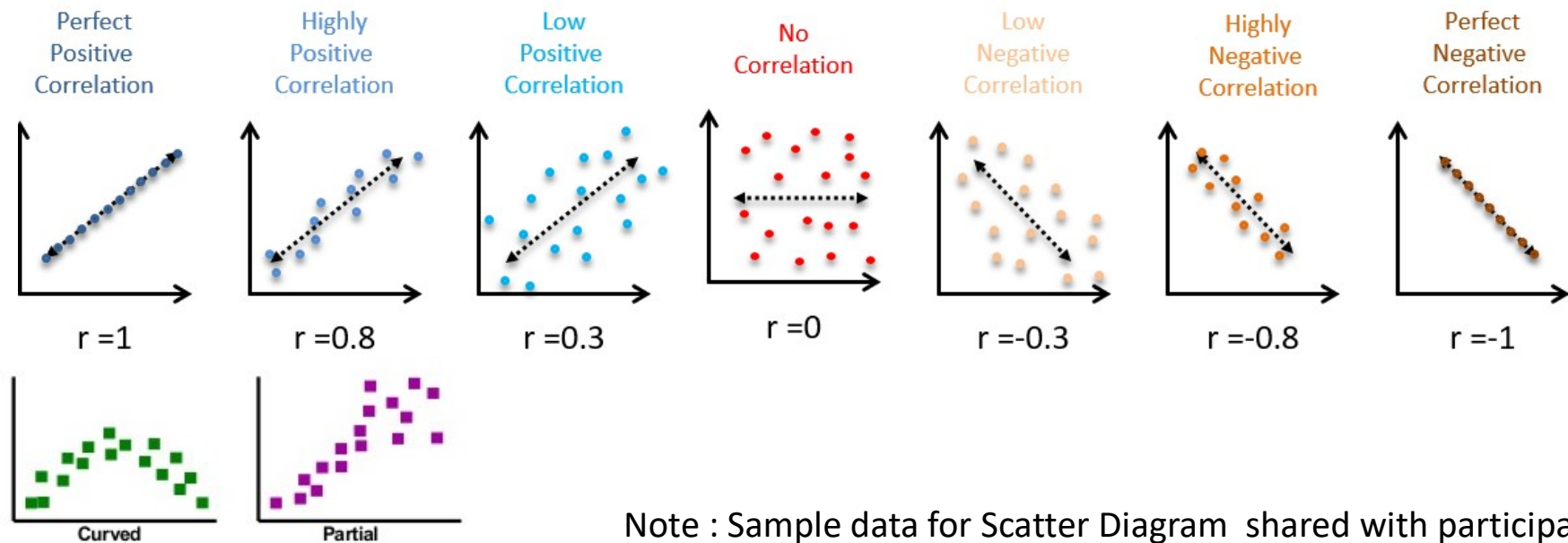
What do we understand from this equation...

We can predict what will be the value of if we can control Xs at a given level

Correlation

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. The strength of correlation is statistically represented “correlation coefficient” and graphically represented by “Scatter Diagram”

Scatter Plots & Correlation Examples



Note : Sample data for Scatter Diagram shared with participants

Correlation

Causation

Co-Linearity

Multi Co-Linearity

Covariance

Confounding variable

Are these terms intimidating?

They are not as difficult concepts as they sound

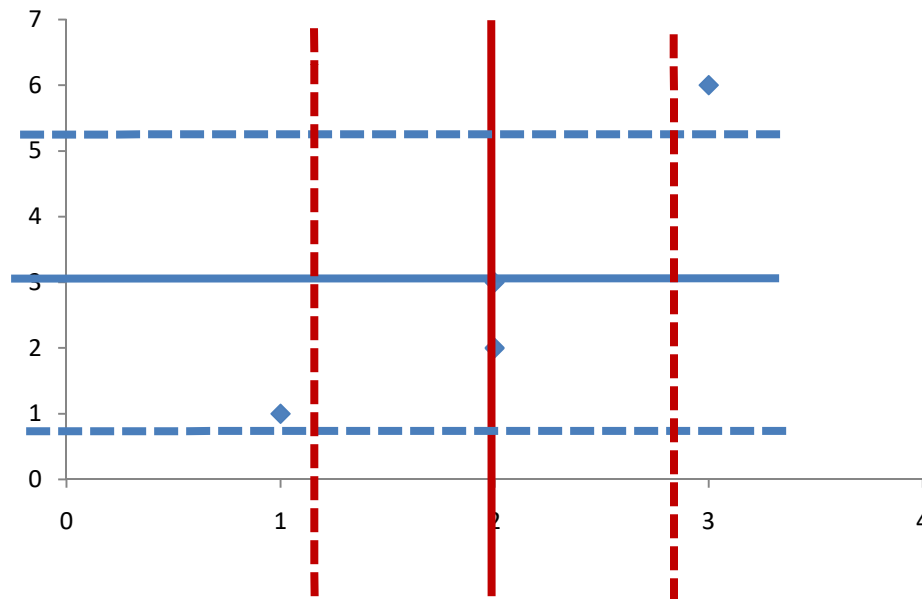
Let's demystify these concepts

Correlation coefficient

There are several different measures for the degree of correlation in data, depending on the kind of data: principally whether the data is a measurement, ordinal, or categorical. Pearson product-moment correlation coefficient is the best-known and most commonly used type of correlation coefficient.



Refer correlation
Tab in excel file

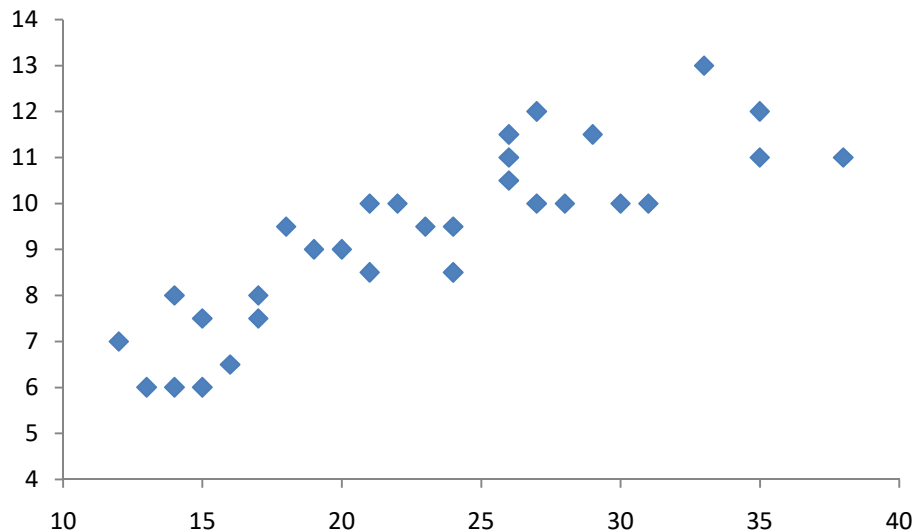


X	Y
1	1
2	2
2	3
3	6

Mean	2	3
Stdev	0.82	2.16

“correlation need not necessarily mean causation”

It's often very tempting to look at statistical information, spot correlation, and then assume causation. It's a mistake that gets made often. It is important to understand what a correlation is and what a causation is. correlation between two events or variables simply indicates that a relationship exists, whereas causation is more specific and says that one event actually causes the other.



Ref correlation but
no causation Tab
in excel file

Note : This data set have a very strong positive correlation but there is no causation

collinearity and Multi - collinearity

For us to understand collinearity and Multi – collinearity lets first understand with is simple linear regression and multiple linear regression

In simple linear regression a single independent variable is used to predict the value of a dependent variable. In multiple linear regression two or more independent variables are used to predict the value of a dependent variable

Collinearity and Multi - Collinearity occurs when we do multiple regression analysis. Collinearity occurs when two predictor variables (e.g., x_1 and x_2) in a multiple regression have a correlation. Multicollinearity occurs when more than two predictor variables (e.g., x_1 , x_2 and x_3) are inter-correlated



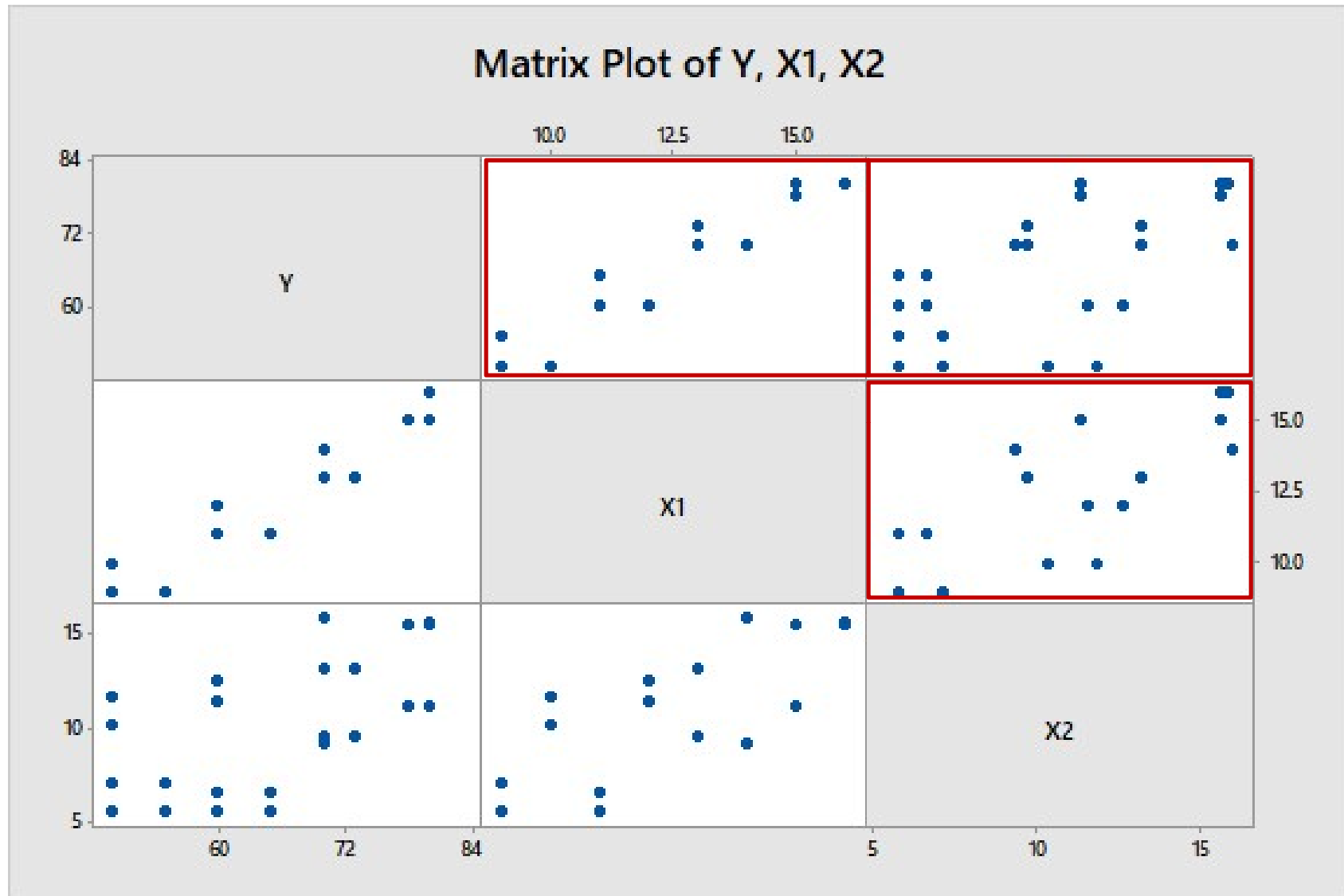
collinearity

LEAN SIX SIGMA

IMPROVE

Matrix Plot of Thickness, FilmDens, Tensile, Airbubbles, PreHeat, ...





Correlation (vs) Covariance

Covariance is a statistical tool that is used to determine the relationship between the movement of two asset prices. When two stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative

In simple words, both the terms measure the relationship and the dependency between two variables. “Covariance” indicates the direction of the linear relationship between variables. “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables

Confounding variable

A confounding variable is an outside influence that changes the effect of a dependent and independent variable

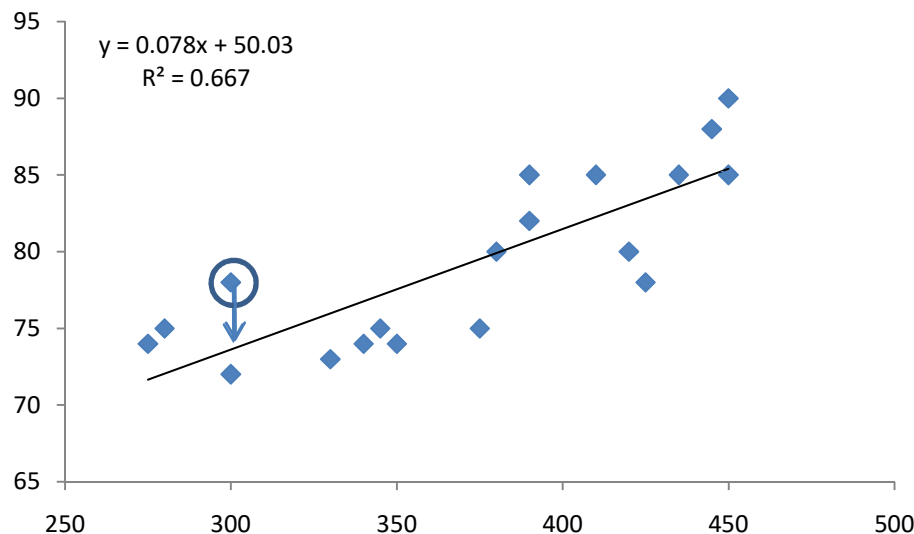
Example 1 : You find that more workers are employed in states with higher minimum wages. Does this mean that higher minimum wages lead to higher employment rates?

Not necessarily. Perhaps states with better job markets are more likely to raise their minimum wages, rather than the other way around. You must consider the prior employment trends in your analysis of the impact of the minimum wage on employment, or you might find a causal relationship where none exists.

Example 2 : You find that babies born to mothers who smoked during their pregnancies weigh significantly less than those born to non-smoking mothers. However, if you do not account for the fact that smokers are more likely to engage in other unhealthy behaviors, such as drinking or eating less healthy foods, then you might overestimate the relationship between smoking and low birth weight.

Simple Linear regression

Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line



The line that get fitted here is called regression line

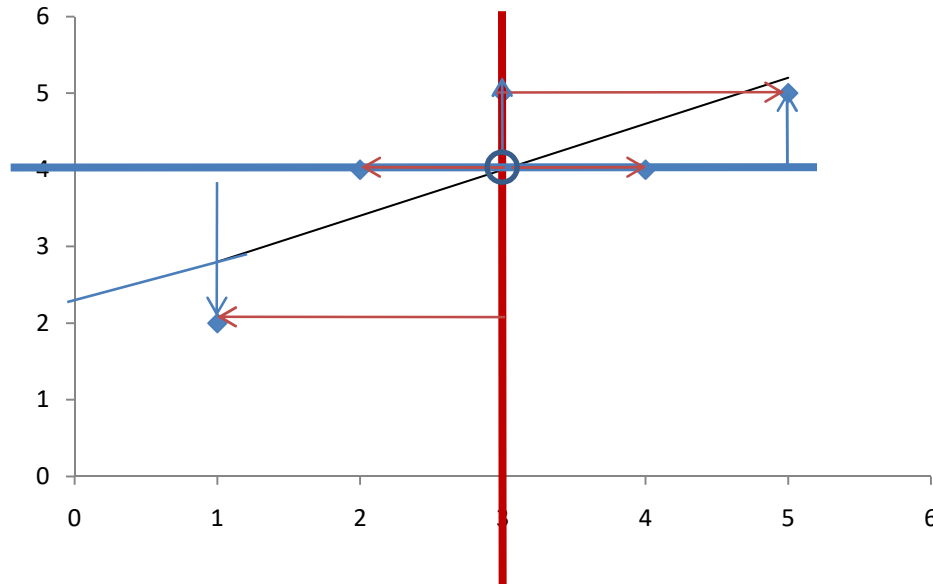
This line is fitted using least square method

Least squared method is nothing but to minimize the error between estimated value and actual value

Regression equation $Y = mX + C$

Lets understand how do we achieve this

Simple Linear regression



X	Y	X-Xbar	Y-Ybar	X-Xbar^2	(X-Xbar)*(Y-Ybar)
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
Average	3	4		10	6

Regression equation $Y=mX+C$

$$m = ((X-Xbar)*(Y-Ybar)) / (x-xbar^2)$$

$$m = 6 / 10$$

$$m = .6$$

$$Y = mX+C$$

$$Y = .6X+c$$

$$4 = .6(3)+c$$

$$C = 4-1.8$$

$$C=2.2$$

Regression equation $Y=mX+C$

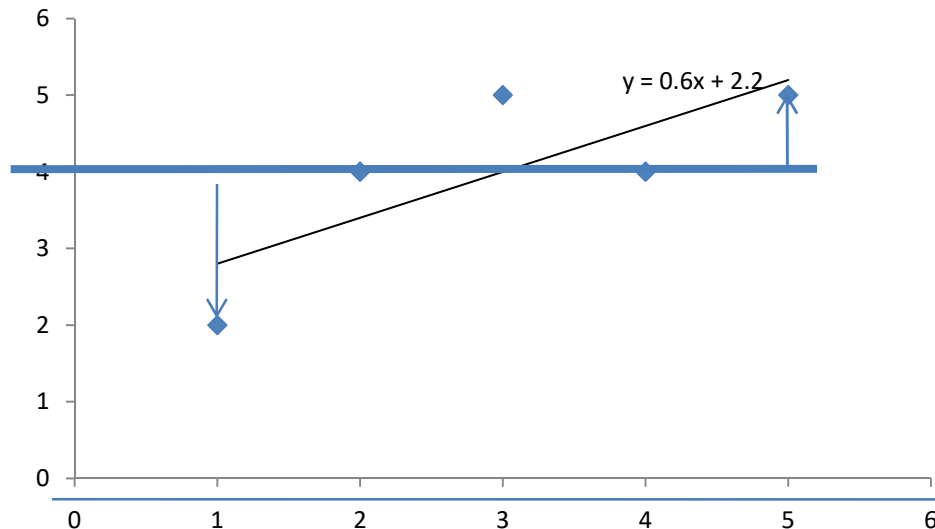
m = Slope of the

c = y intercept when x is 0

What do we mean by $(r)^2$

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs

Simply $(r)^2$ R-squared is a comparison of Distance between observed value from the mean (vs) Distance between expected value from the mean for my dependant variable

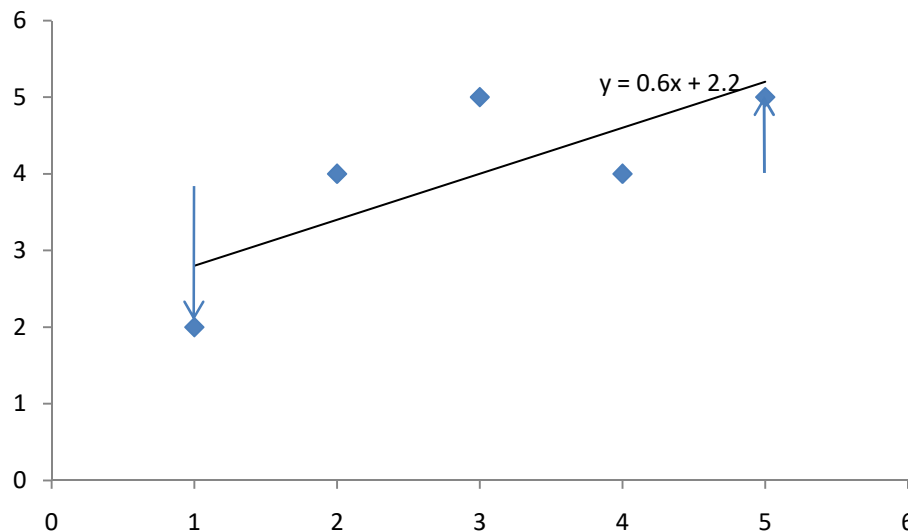


X	Y	Y-Ybar	Y-Ybar ²	EV of Y	EV - Ybar	EV - Ybar ²
1	2	-2	4	2.8	-1.2	1.44
2	4	0	0	3.4	-0.6	0.36
3	5	1	1	4	0	0
4	4	0	0	4.6	0.6	0.36
5	5	1	1	5.2	1.2	1.44
Average	3	4	6			3.6

$$\begin{aligned}
 R\text{-squared} &= \frac{EV - Ybar^2}{Y-Ybar^2} \\
 &= \frac{3.6}{6} \\
 &= .6 \\
 &\text{i.e 60\% strength}
 \end{aligned}$$

What do we mean by standard error

The standard error of the regression (S), also known as the standard error of the estimate, represents the average distance that the observed values fall from the regression line. Conveniently, it tells you how wrong the regression model is on average using the units of the response variable.



X	Y	EV of Y	EV of y - Y	(EV of y - Y) ²
1	2	2.8	0.8	0.64
2	4	3.4	-0.6	0.36
3	5	4	-1	1
4	4	4.6	0.6	0.36
5	5	5.2	0.2	0.04
Average	3	4		2.4

$$\text{Std error} = \text{Sqrt} \left(\frac{\text{sum}((\text{EV of } y - Y)^2)}{n-2} \right)$$

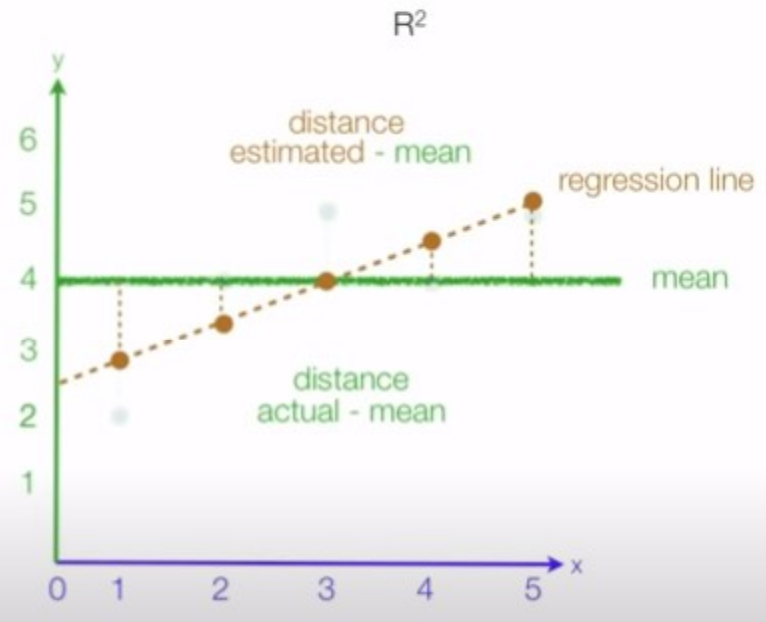
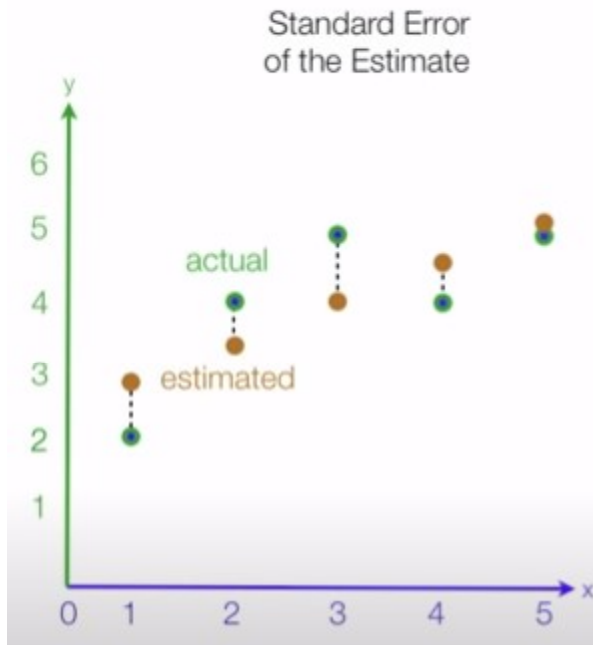
$$\text{Std error} = \text{Sqrt} ((2.4)/(5-2))$$

$$= \text{Sqrt} (2.4/3)$$

$$= \text{Sqrt} (0.8)$$

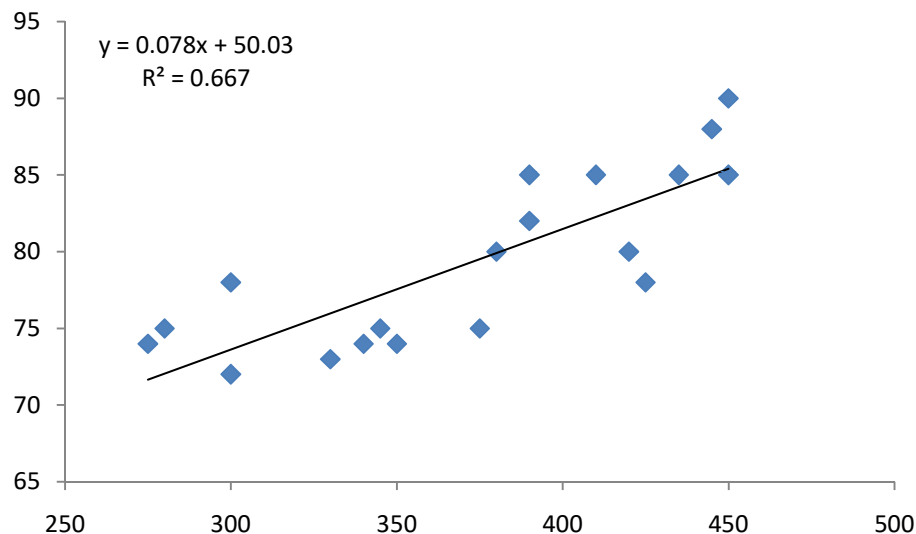
$$= .89$$

R-squared (vs) standard error



Simple Linear regression

Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line



The line that get fitted here is called regression line

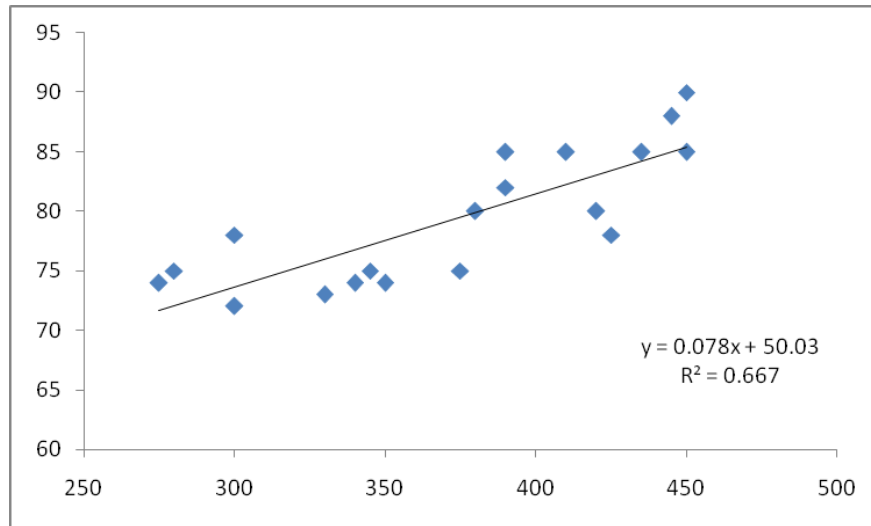
This line is fitted using least square method

Least squared method is nothing but to minimize the error between estimated value and actual value

Regression equation $Y = mX + C$

Lets understand how do we achieve this

Simple Linear regression



SLR

$$\text{\$ Value} = 50.03 + 0.0786 \text{ T (d)}$$

$$100 = 50.03 + 0.0786$$

$$\text{T (d)} = (100 - 50.03) / 0.0786$$

$$\text{T(d)} = 636$$

Regression Analysis: \$ Value Sold Per Day versus Tenure in days

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	376.83	376.83	34.15	0.000
Tenure in days	1	376.83	376.83	34.15	0.000
Error	17	187.59	11.03		
Lack-of-Fit	14	152.59	10.90	0.93	0.607
Pure Error	3	35.00	11.67		
Total	18	564.42			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.32185	66.76%	64.81%	58.42%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	50.03	5.08	9.85	0.000	
Tenure in days	0.0786	0.0135	5.84	0.000	1.00

Regression Equation

$$\text{\$ Value Sold Per Day} = 50.03 + 0.0786 \text{ Tenure in days}$$

Residual Plots for \$ Value Sold Per Day

Multiple Linear regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is an extension of linear regression that uses just one explanatory variable.

Best Subset

Best Subsets compares all possible models using a specified set of predictors, and displays the best-fitting models that contain one predictor, two predictors, and so on. The end result is a number of models and their summary statistics. It is up to you to compare and choose one.



MLR

Best Subsets Regression: Thickness versus FilmDens, Tensile, ...

Response is Thickness

						A i F r P V M D i T b P l A a o w l e u r u i c l e m n b e g r P d l D s b H T P r T T e i l e e r e e i n l e a m e s m m s e s t p s s p e											
Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S												
1	72.1	71.8	70.8	55.5	0.089424												
1	50.2	49.6	47.1	156.4	0.11948											X	
2	77.8	77.2	76.0	31.2	0.080300										X		X
2	75.5	74.9	73.7	41.9	0.084369										X		X
3	80.2	79.4	78.2	22.3	0.076407									X	X		X
3	79.7	78.9	77.3	24.5	0.077336									X	X		X
4	83.8	82.9	80.7	7.5	0.069508									X	X		X
4	81.8	80.8	78.3	16.8	0.073683									X	X		X
5	84.5	83.4	80.6	6.5	0.068554									X	X		X
5	84.3	83.2	81.0	7.3	0.068927									X	X		X
6	85.2	83.9	81.2	5.4	0.067524									X	X		X
6	84.6	83.3	79.8	8.1	0.068838									X	X		X
7	85.3	83.8	81.0	6.7	0.067694									X	X		X
7	85.2	83.7	80.4	7.0	0.067833									X	X		X
8	85.5	83.7	80.3	8.1	0.067841									X	X		X
8	85.3	83.6	80.4	8.6	0.068139									X	X		X
9	85.5	83.5	79.7	10.0	0.068318									X	X		X

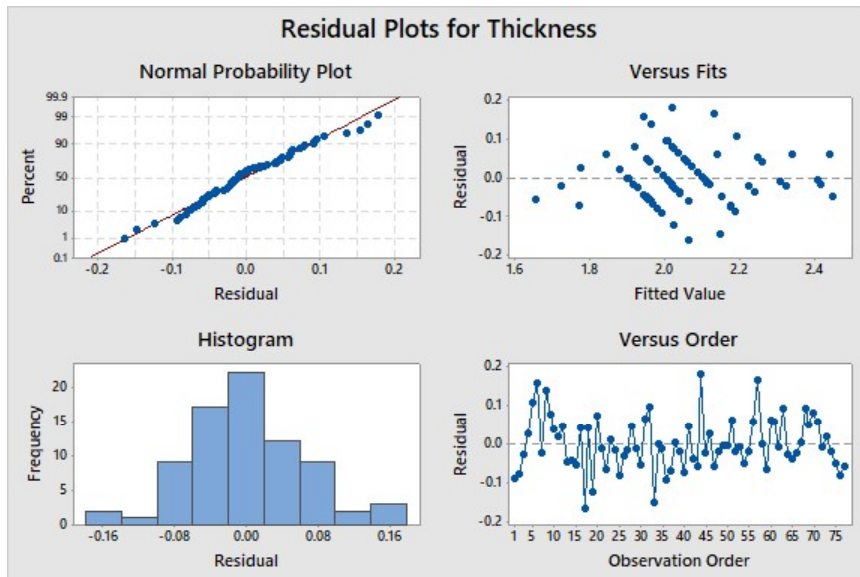
Multiple Linear regression - VIF

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Typically in practice there is a small amount of collinearity among the predictors. As a rule of thumb, under ideal conditions $VIF < 3$ is recommended in some cases $VIF < 5$ is also accepted

Mallows' Cp

Mallows' Cp (compares the precision and bias of the full model to models with a subset of the predictors. Usually, you should look for models where Mallows' Cp is small and close to the number of predictors in the model plus the constant

Multiple Linear regression



MLR

Regression Analysis: Thickness versus FilmDens, Tensile, PreHeat, VacPress

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	1.80383	0.450958	93.34	0.000
FilmDens	1	0.07832	0.078321	16.21	0.000
Tensile	1	0.11628	0.116280	24.07	0.000
PreHeat	1	0.32575	0.325749	67.42	0.000
VacPress	1	0.08874	0.088743	18.37	0.000
Error	72	0.34786	0.004831		
Lack-of-Fit	71	0.34286	0.004829	0.97	0.688
Pure Error	1	0.00500	0.005000		
Total	76	2.15169			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0695078	83.83%	82.94%	80.67%

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	20.96	2.69	7.80	0.000	
FilmDens	-0.1095	0.0272	-4.03	0.000	1.56
Tensile	0.00493	0.00101	4.91	0.000	2.21
PreHeat	-0.05104	0.00622	-8.21	0.000	2.06
VacPress	-0.01855	0.00433	-4.29	0.000	1.63

Regression Equation

Thickness = 20.96 - 0.1095 FilmDens + 0.00493 Tensile - 0.05104 PreHeat - 0.01855 VacPress

Fits and Diagnostics for Unusual Observations

Obs	Thickness	Fit	Resid	Std Resid	R	X
5	2.3000	2.1935	0.1065	2.08	R	X
6	2.1000	1.9448	0.1552	2.33	R	
8	2.1000	1.9638	0.1362	2.06	R	
17	1.9000	2.0646	-0.1646	-2.41	R	
21	2.3000	2.3101	-0.0101	-0.20	X	
33	2.0000	2.1476	-0.1476	-2.15	R	
44	2.2000	2.0218	0.1782	2.64	R	
57	2.3000	2.1343	0.1657	2.47	R	

R Large residual
X Unusual X

Response Optimizer

Response Optimization: Thickness

Parameters

Response	Goal	Lower	Target	Upper	Weight	Importance
Thickness	Target	1.6	2.5	2.75	1	1

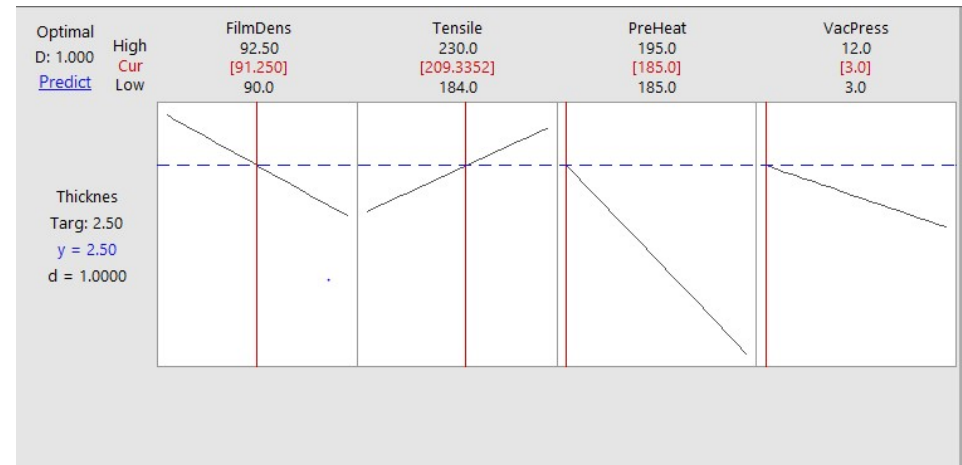
Solution

Solution	FilmDens	Tensile	PreHeat	VacPress	Thickness	Composite
1	91.25	209.335	185	3	2.5	1

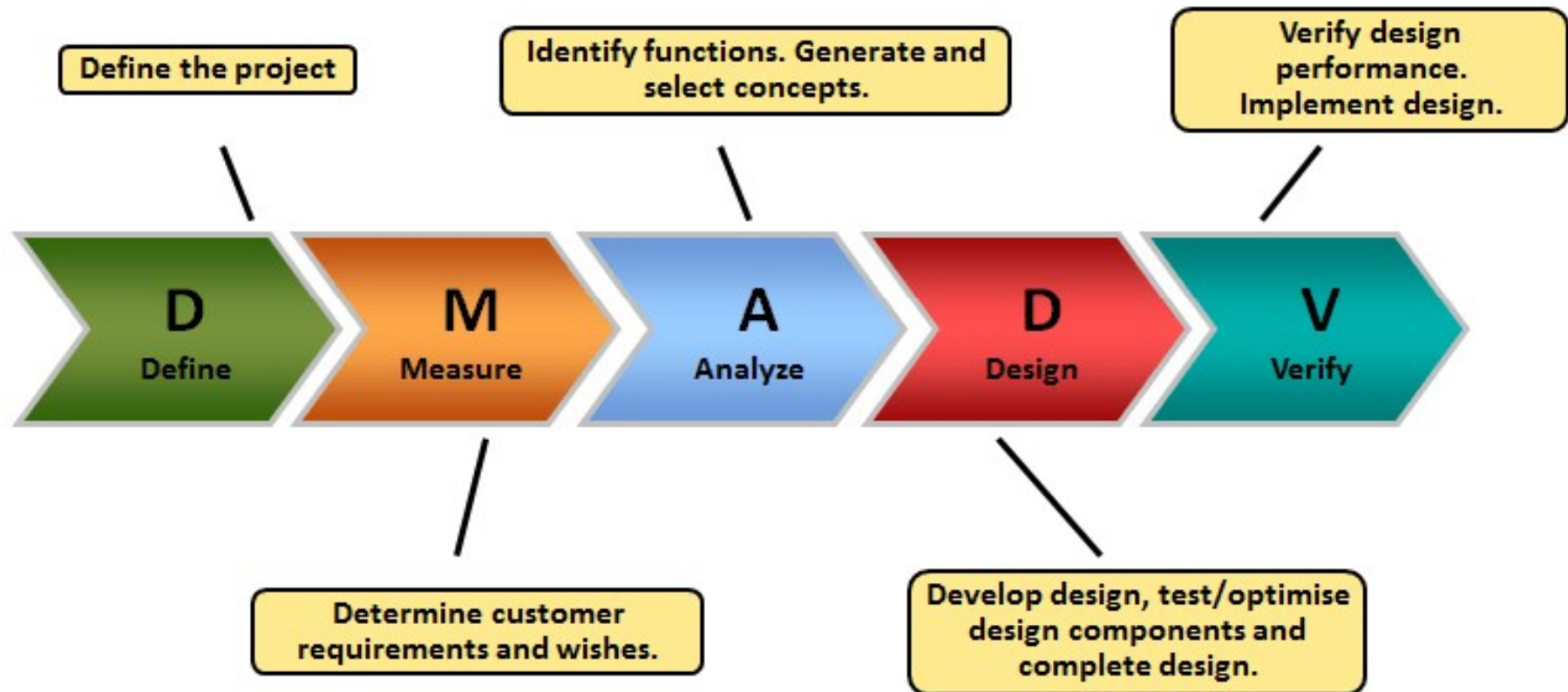
Multiple Response Prediction

Variable	Setting
FilmDens	91.25
Tensile	209.335
PreHeat	185
VacPress	3

Response	Fit	SE Fit	95% CI	95% PI
Thickness	2.5000	0.0308	(2.4387, 2.5613)	(2.3485, 2.6515)



DMADV

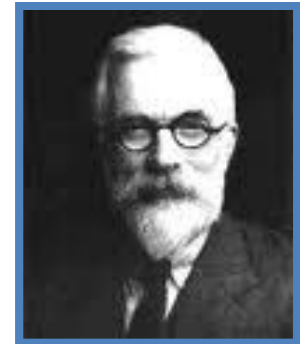


DESIGN techniques

- DOE
- Design Thinking
- 6 Thinking Hats

History of DOE

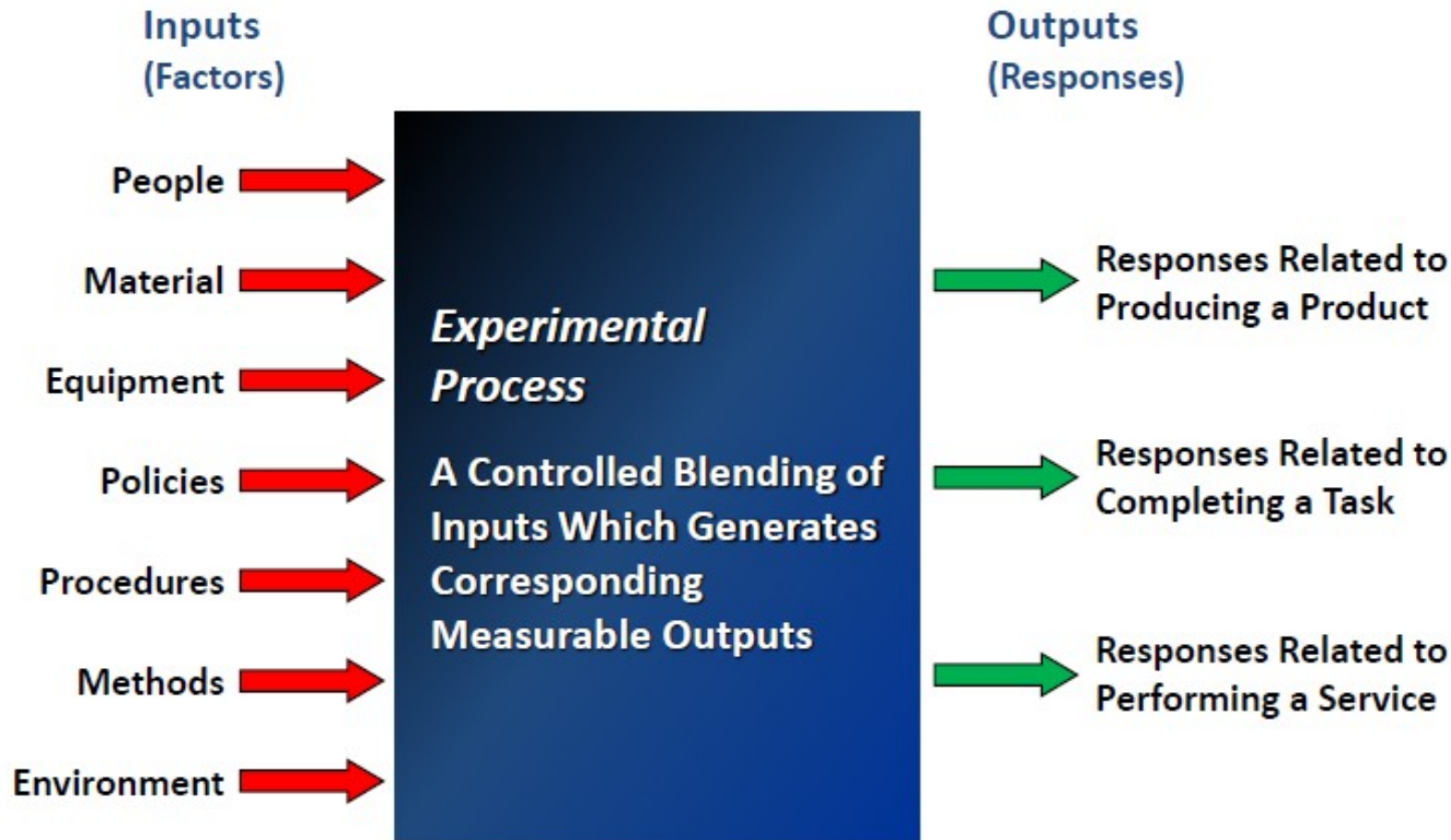
- Sir Ronald A. Fisher, the renowned mathematician and geneticist, first developed this method in 1920s.
- Agri origin : The Rothamsted experimental station in England was conducting a series of experiments to measure the effects of different fertilizers on various crops.



Design of Experiments?

- Identification of critical factors to improve performance
- Identification of unimportant factors to reduce costs
- Reduction in cycle time
- Reduction of scrap/rework
- Scientific method for setting tolerances

What is an Experimental Design?



Why Use DOE?

- The structured methodology provides a directed approach
- The designed experiment gives a mathematical model relating the variables and responses.
- The model is easily optimized.
- The statistical significance of the results is known, so there is much greater confidence in the results.
- Can determine how multiple input variables interact to affect results.

Methods of Experimentation

- Experimentation has been used for a long time. Some experiments have been good, some not so good.
- Our early experiments can be grouped into the following general categories:
 - Trial and Error
 - One-Factor-At-a-Time (OFAT)
 - Full Factorial
 - Fractional Factorial
 - Others

Planning for Designs

- Define the problem.
- Define the objective and budget.
- Develop an experimental plan that will provide meaningful information.
- Make sure the process and measurement systems are in control.

Screening Designs

- The purpose of a screening design is to look at a relatively large group of factors to determine which factors have a significant effect on the response variable of interest.
- There are different opinions about the number of factors that can be looked at with a screening design.
- Based on input from the process experts, all of the factors and possible interactions need to be considered.

Refining Designs

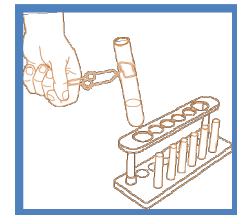
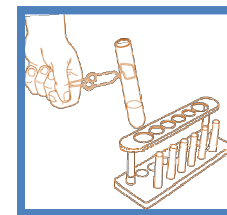
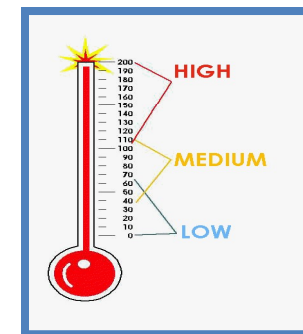
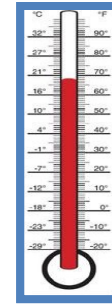
- Refining designs are the same as screening designs, except that only a few critical factors from the screening design are involved.
- The purpose is to more thoroughly characterize the relationship between the critical factors and the response(s).
- A design large enough to give us sufficient information about interactions is used.
- Refining designs also often allow for testing of non-linear effects.

Optimizing Designs

- These designs allow us to create very accurate mathematical models for a few critical factors.
- We can model curved relationships as well.
- Mathematical models allow us to predict the response we will obtain at various factor level settings.

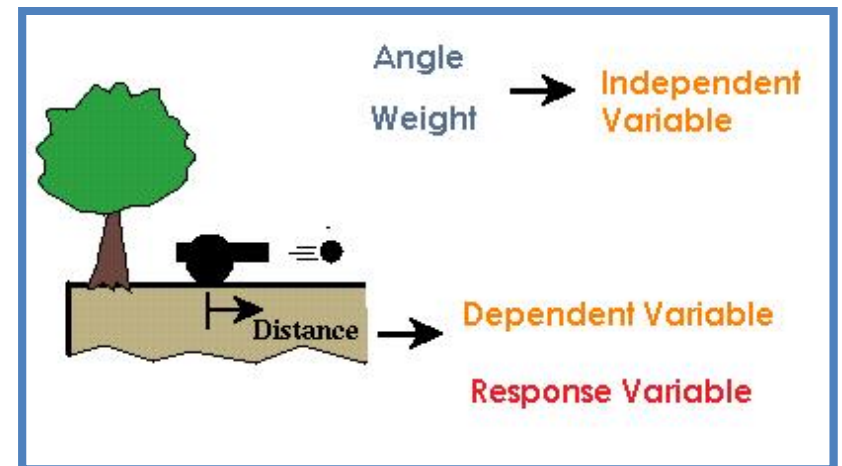
Basic Terms

- **Factor**- Variable un/controlled by the experimenter
 - E.g.- Furnace temperature, Follow up
- **Levels**- Specific values of factor.
 - E.g.- Three different temperature, Daily/Weekly/Monthly follow up
- **Replication**- Re-run experiment to reduce measurement error. It is not Repetition



Basic Terms

- **Dependent Variables-** Variables dependent on another variable.
 - E.g. Force on the paper flight
- **Independent Variables-** Variables presumed to affect a dependent variable. They are most commonly the input in an experiment.
 - E.g. Right hand or Left hand
- **Response-** Variables that are outcome of the experiment.
 - E.g.-Paper flight distance

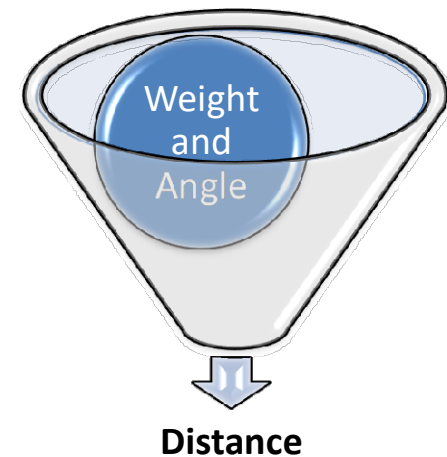


Experimental Design Types

- **Randomized experiments** are appropriate when only one factor is analyzed.
- **Factorials** are appropriate when investigating several factors at two or more levels are interaction is necessary.

Experimental Design Types

- **Blocked factorials** reduce the number of runs and use blocking to run the experiments in subsets.
- **Fractional factorials** reduce the combinations of factors and levels required to be run while resulting in a close estimate to the full factorial.



	Factor - 1	Factor -2	Factor - 3
Trial 1			
Trial 3			
Trial 4			
Trial 7			

DOE Design Principles

- **Randomization** is a method of designing and organizing the experiment by randomly assigning the test run.
- **Blocking** is a method of organizing the experiment by grouping the experiments in batches.
- **Interaction** occurs when one or more factors have an influence on the behavior of another factor

DOE Steps



DOE – Food for Thought

- What is the measurable objective
- What it will cost to conduct a DOE
- How to randomize sample
- How long it will take to conduct a DOE
- How the data are going to be analyzed
- Communication to customers and similar process owners

Full Factorial DOE

- Full factorials examine every possible combination of factors at the levels tested. The full factorial design is an experimental strategy that allows us to answer most questions completely.
- Full factorials enable us to:
 - Determine main effects of the factors manipulated on response variables
 - Determine effects of factor interactions on response variables
 - Estimate levels to set factors for best results

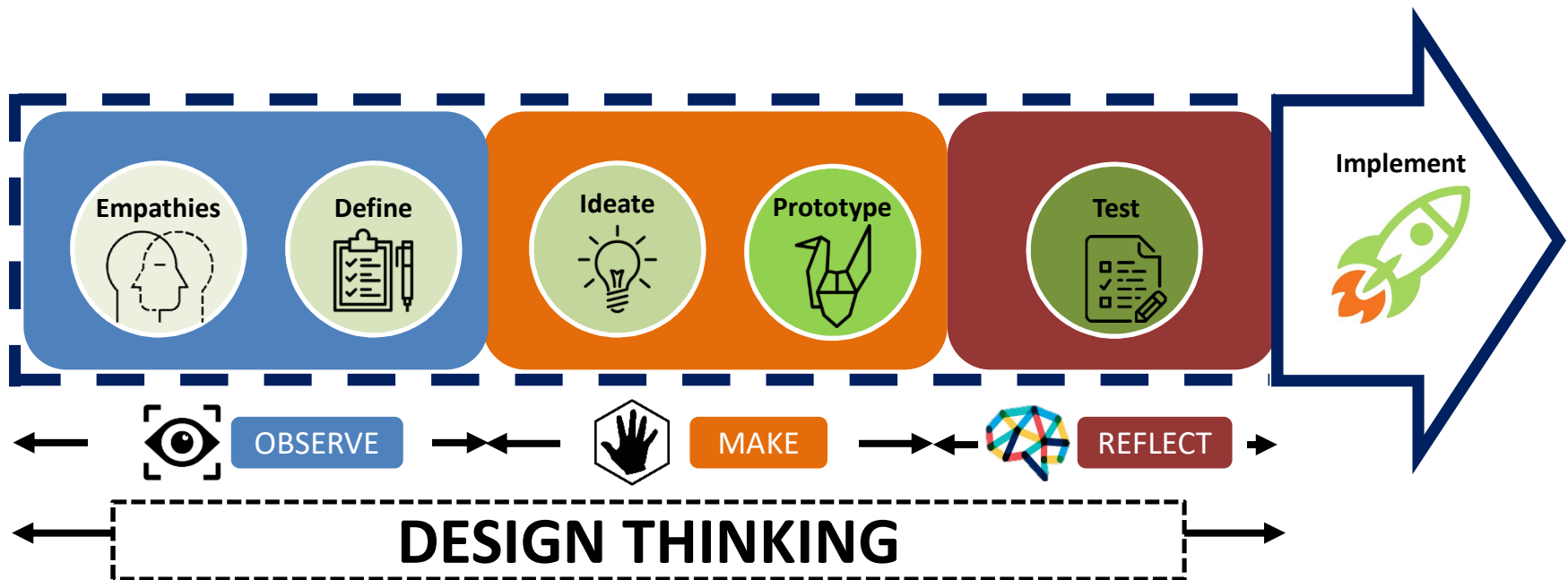
Full Factorial

- Analyzing Full Factorial with Replicates
- Displaying Plots
 - Factorial plots for an experiment with three factors
 - Contour and surface plots
- Response Optimization
 - Overlaid contour plot
 - Response optimization

DOE References

- *Statistics for Experimenters, G.E.P. Box, William G. Hunter, and J. Stuart Hunter, Wiley, 1978.*
- *Off-line Quality Control, Parameter Design and the Taguchi Method, R.N. Kacker, Journal of Quality Technology, 17, 1985.*
- *Design and Analysis of Experiments, Douglas Montgomery, Wiley, 1997.*
- Schmidt and Launsby, *Understanding Industrial Designed Experiments, Air Academy Press and Associates, 1997*
- Thomas B. Barker, *Quality by Experimental Design, Marcel Dekker 1994, ISBN 0-8247-8910-5*

DESIGN Thinking



Note : The five stages are not always sequential — they do not have to follow any specific order and they can often occur in parallel and be repeated iteratively. As such, the stages should be understood as different modes that contribute to a project, rather than sequential steps.

DESIGN Thinking

What is Design Thinking and Why Is It So Popular?

“In its simplest applications, design thinking is a process of creating and testing innovative ideas to improve a product/service or solve existing problems. Design thinking provides people with a “generic” sequence of events to come up with solutions where User, Technology and Business views are considered”

DESIGN Thinking



Empathy

Empathize : The ability to share someone else's feelings or experiences by imagining what it would be like to be in that person's situation



Define

Define : In Define phase of Design Thinking synthesise your observations about your users from the first stage, the Empathize stage. A great definition of your problem statement will guide you and your team's work and kick start the ideation process (third stage) in the right direction.

DESIGN Thinking



Ideate

Ideate : “Ideation is the mode of the design process in which you concentrate on idea generation. Mentally it represents a process of “going wide” in terms of concepts and outcomes. Ideation provides both the fuel and also the source material for building prototypes and getting innovative solutions into the hands of your users.”

– d.school, An Introduction to Design Thinking PROCESS GUIDE



Prototype

Prototype : This method involves producing an early, inexpensive, and scaled down version of the product in order to reveal any problems with the current design. Prototyping offers designers the opportunity to bring their ideas to life, test the practicability of the current design, and to potentially investigate how a sample of users think and feel about a product.

“They slow us down to speed us up. By taking the time to prototype our ideas, we avoid costly mistakes such as becoming too complex too early and sticking with a weak idea for too long.” – Tim Brown

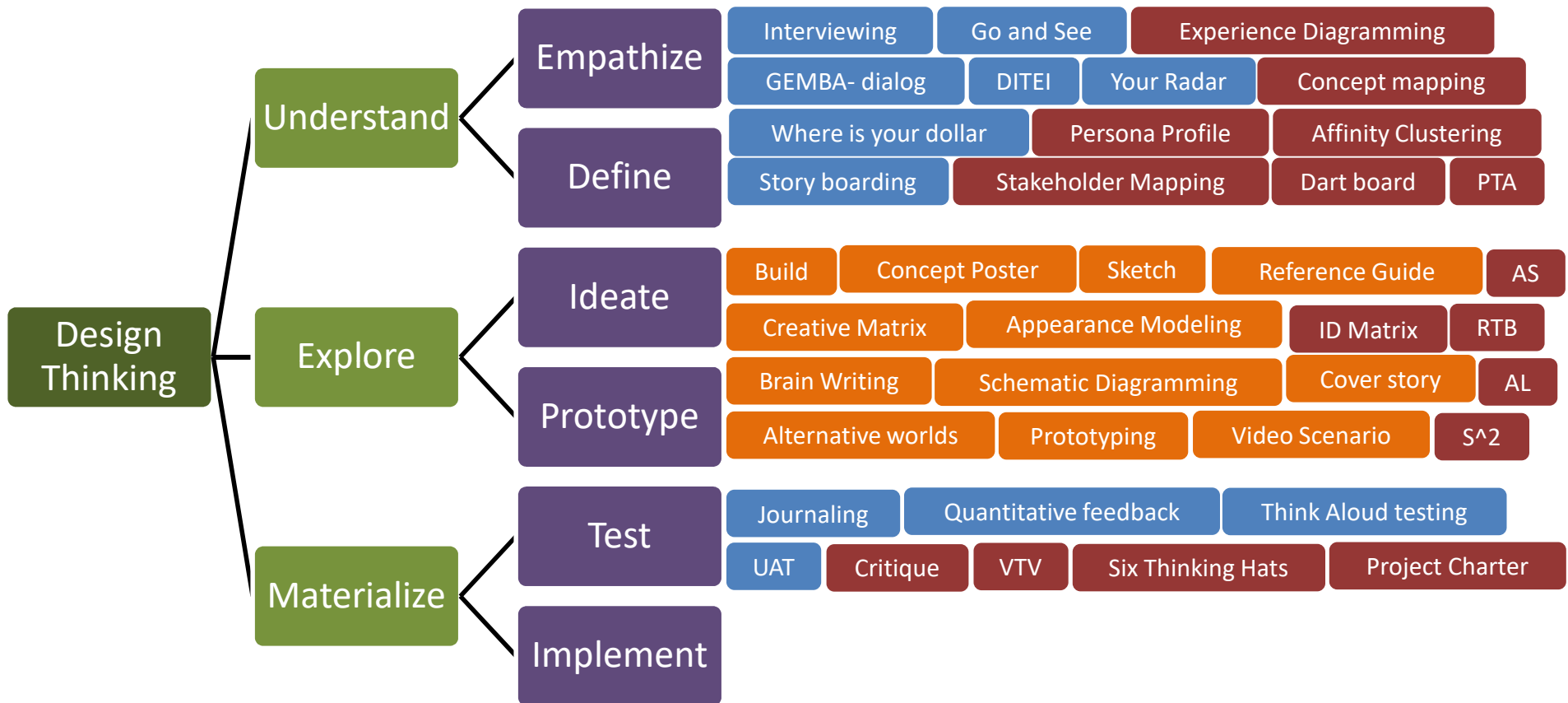
DESIGN Thinking



Test

Test : Testing can be undertaken throughout the progress of a Design Thinking project, although it is most commonly undertaken concurrently with the Prototyping stage. Testing, in Design Thinking, involves generating user feedback as related to the prototypes you have developed, as well as gaining a deeper understanding of your users. When undertaken correctly, the Testing stage of the project can often feed into most stages of the Design Thinking process

DESIGN Thinking



OBSERVE

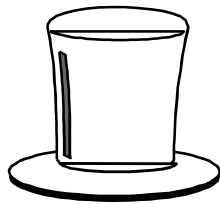


MAKE



REFLECT

6 Thinking Hats



The SIX thinking Hats



6 Thinking Hats - History

Edward de bono in his famous book "The Use of Lateral Thinking" (published in 1967) introduced a term "Lateral Thinking"

According to De Bono, lateral thinking is deliberately distancing oneself from the standard perception of creativity as "vertical" logic, the classic method for problem solving.

Subsequently in his other books like

Lateral Thinking for Management (1971)

Children Solve Problems (1972)

Beyond Yes and No (1973)

Teaching Thinking (1976)

He explains how can we coach ourself to become a lateral thinker

Finally in his master piece "six thinking hat" (1985) he chalks out a method for groups to plan thinking processes in a detailed and cohesive way, and in doing so to think together more effectively.

6 Thinking Hats - History



The WHITE hat

The WHITE Hat is the information hat.

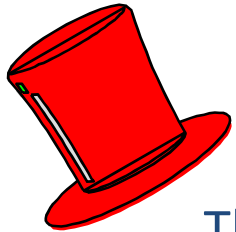
It is neutral and objective.

It calls for the facts, and just the facts.

It imitates the computer.

It identifies information that is missing.

6 Thinking Hats - History



The RED Hat

The RED Hat is the emotional hat.

It is the hat of feelings, fears, likes, dislikes.

It is the hat that legitimizes emotions and feelings.

It is the opposite of neutral.

You don't need to give reasons with this hat.

6 Thinking Hats - History



The BLACK Hat

The BLACK Hat is a caution hat.

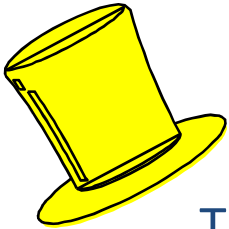
It acts as devil's advocate.

It points out dangers and problems.

It must be logical.

It is the risk analysis hat.

6 Thinking Hats - History



The YELLOW Hat

The YELLOW Hat is the positive hat.

It symbolizes sunshine, brightness and optimism.

It permits visions and dreams.

It probes and explores for value & benefit.

It creates concrete proposals and suggestions.

6 Thinking Hats - History



The GREEN Hat

The GREEN Hat is the creative hat.

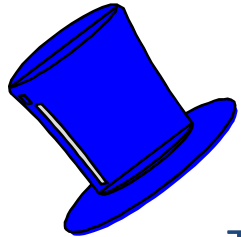
It does not need to be logical.

It calls for creative thinking.

It moves from one idea to reach other ideas.

It generates new concepts and new perceptions.

6 Thinking Hats - History



The BLUE Hat

The BLUE Hat is the process control hat.

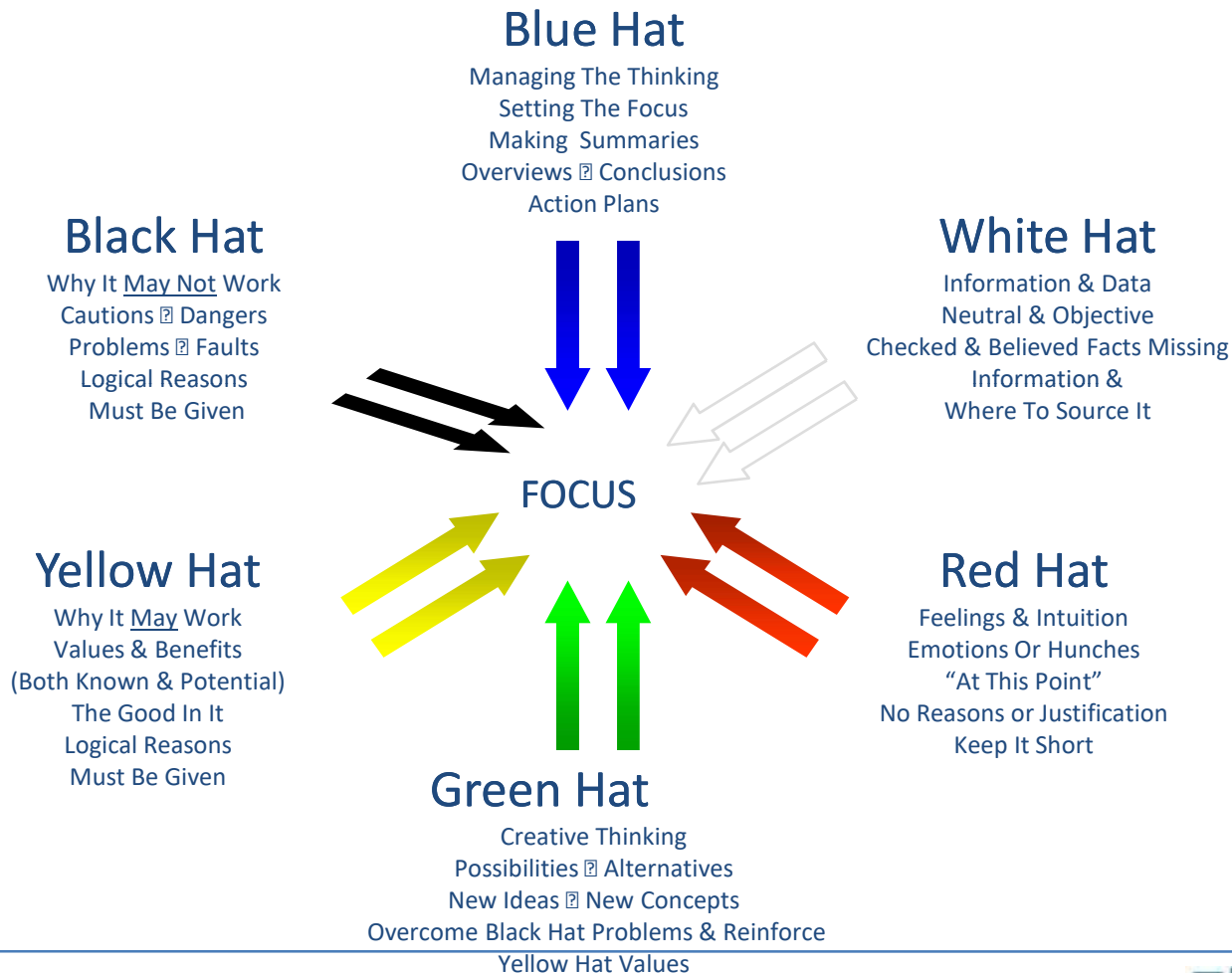
It conducts and organizes the thinking.

It calls for the use of other hats.

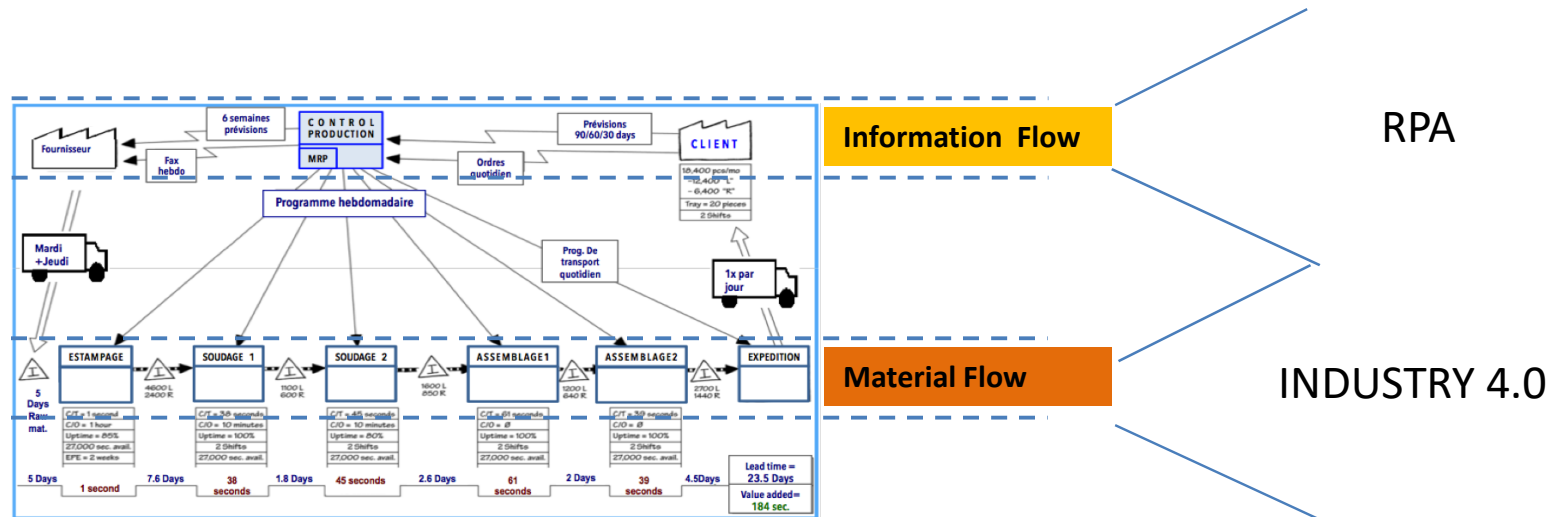
It sets the focus by defining the problem and shaping the questions.

It ensures we keep to our time.

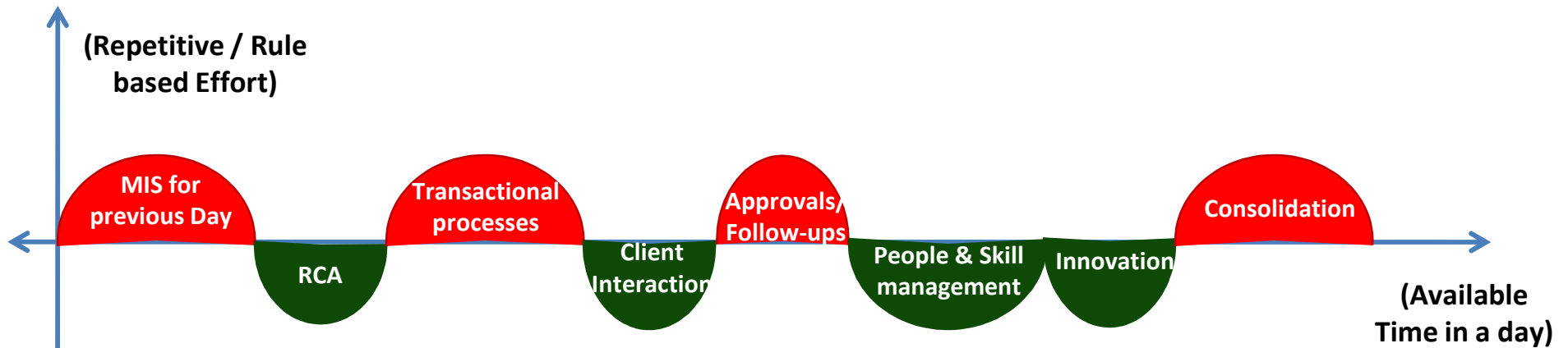
6 Thinking Hats - History



RPA & INDUSTRY 4.0

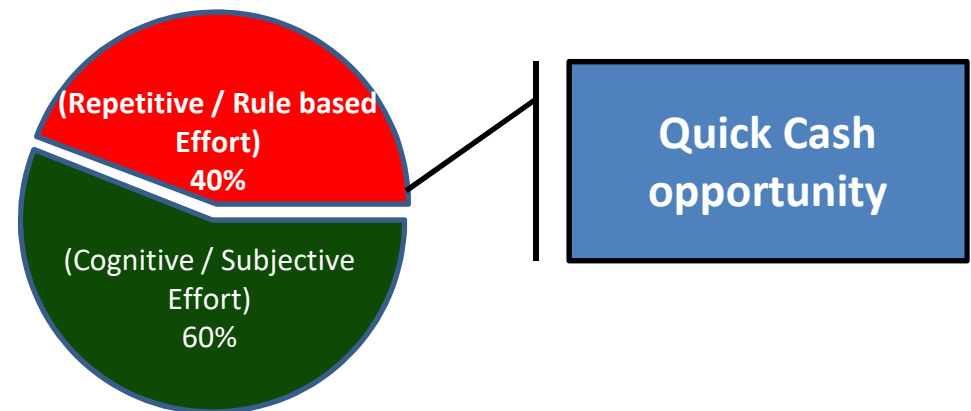


RPA



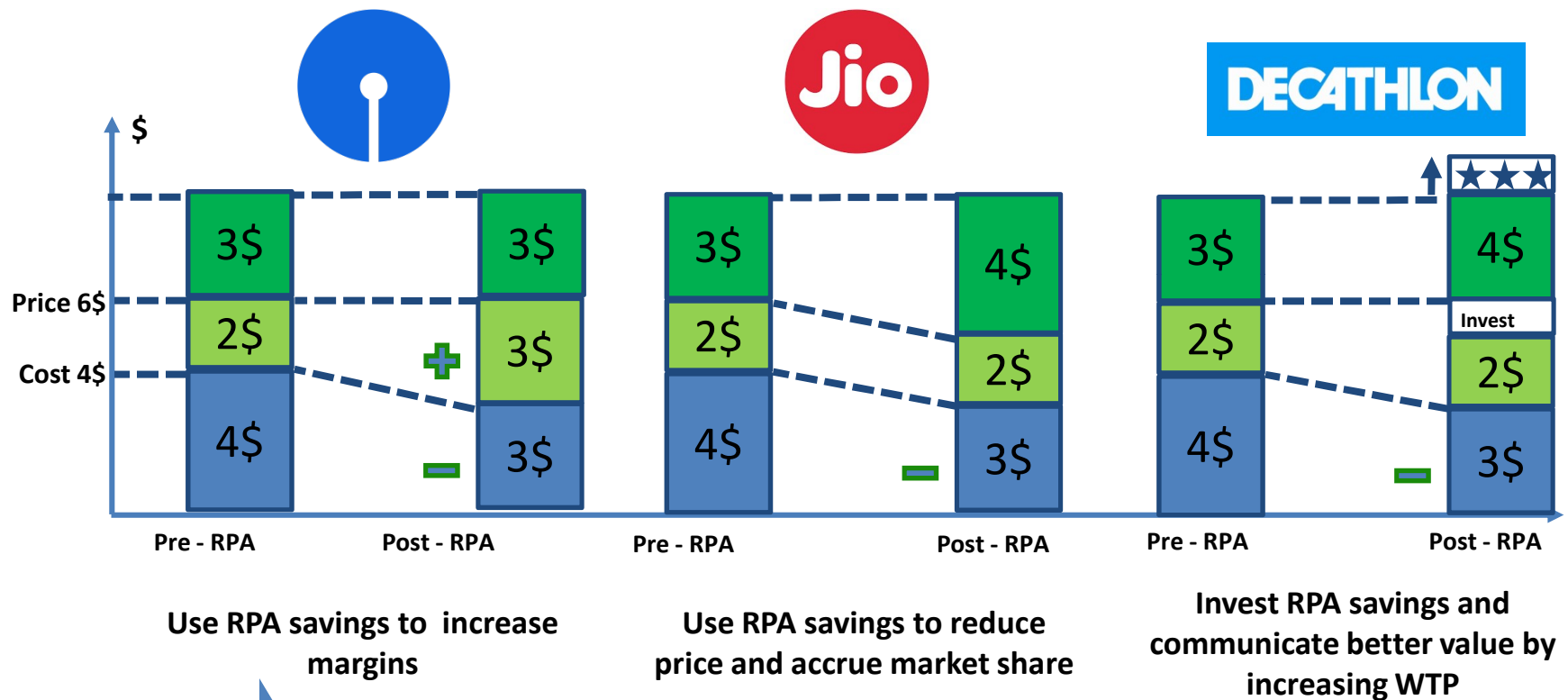
My favorite question no : 1

In our typical workday how much of time is spent in repetitive / Rule based activities ?



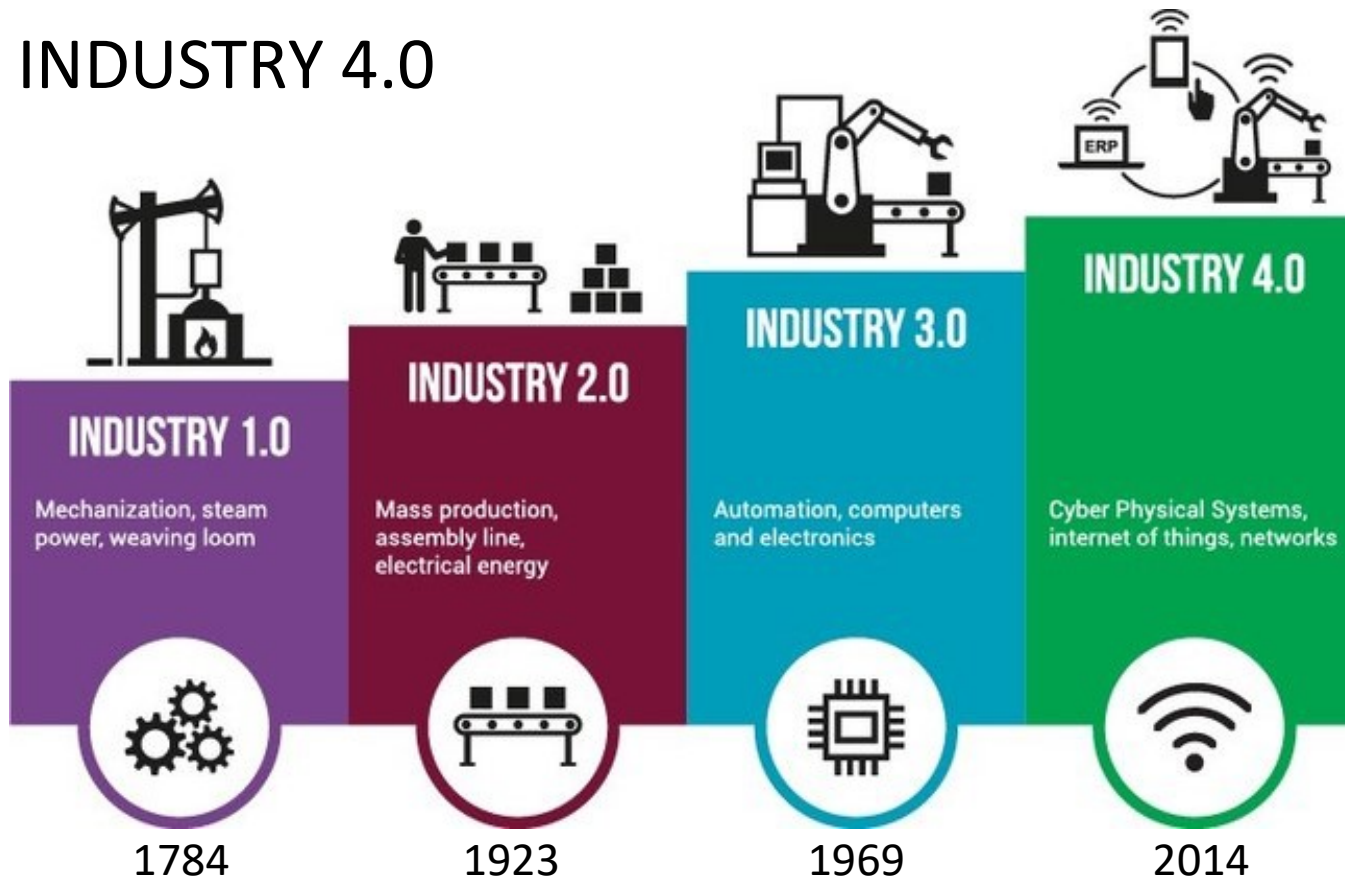
RPA

“RPA will give significant benefits” is now agreed beyond doubt, but how organizations use these benefits is what differentiates market leaders from others



RPA LEADING TO PRICING STRATEGY

INDUSTRY 4.0



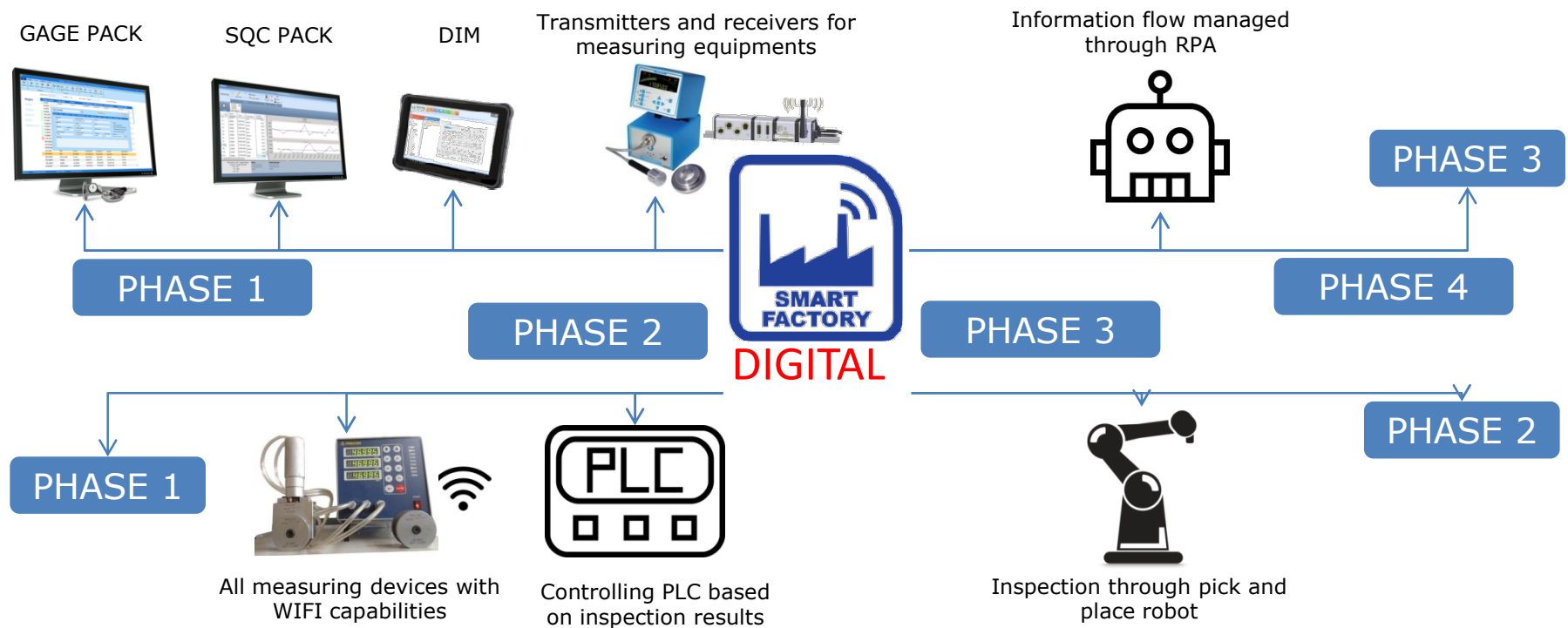
The fourth industrial revolution is upon us. Although the first three industrial revolutions of mechanization, mass production, and computerization have come to define the world we live in today, the fourth industrial revolution will usher in a new age of innovation and transformation. This is characterized by the advent of cyber-physical systems, arising from the convergence of the digital and physical worlds.

These industrial revolutions will have a profound impact on companies and economies around the world. In the future, Industry 4.0 will create a world where processes are increasingly digitalized and integrated; where devices, machines, and systems can autonomously optimize processes and manage operations; and where humans and machines work together to create smart facilities that are efficient, flexible, and adaptive.

LEAN SIX SIGMA

IMPROVE

INDUSTRY 4.0





Lean is a way of thinking- not a list
of things to do

— *Shigeo Shingo* —

