



***Parametric Cost Estimating
Training Track For The***

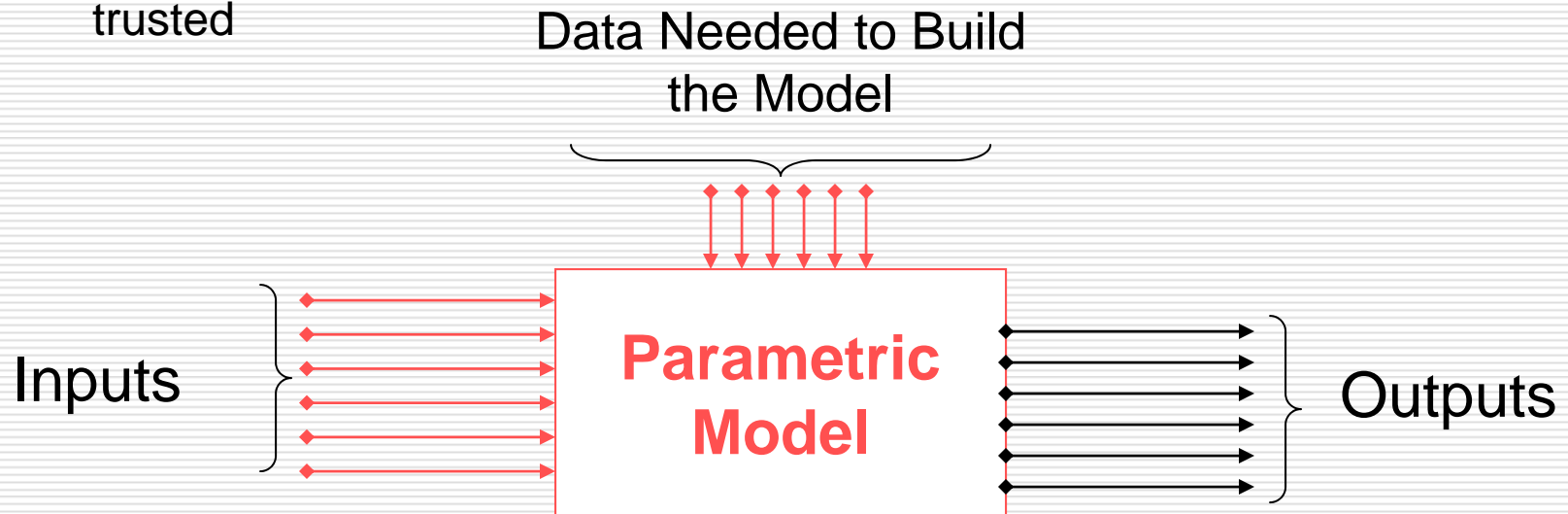
***Parametric Estimating Handbook
4th Edition
Chapter 2 Data Collection & Analysis***

2009



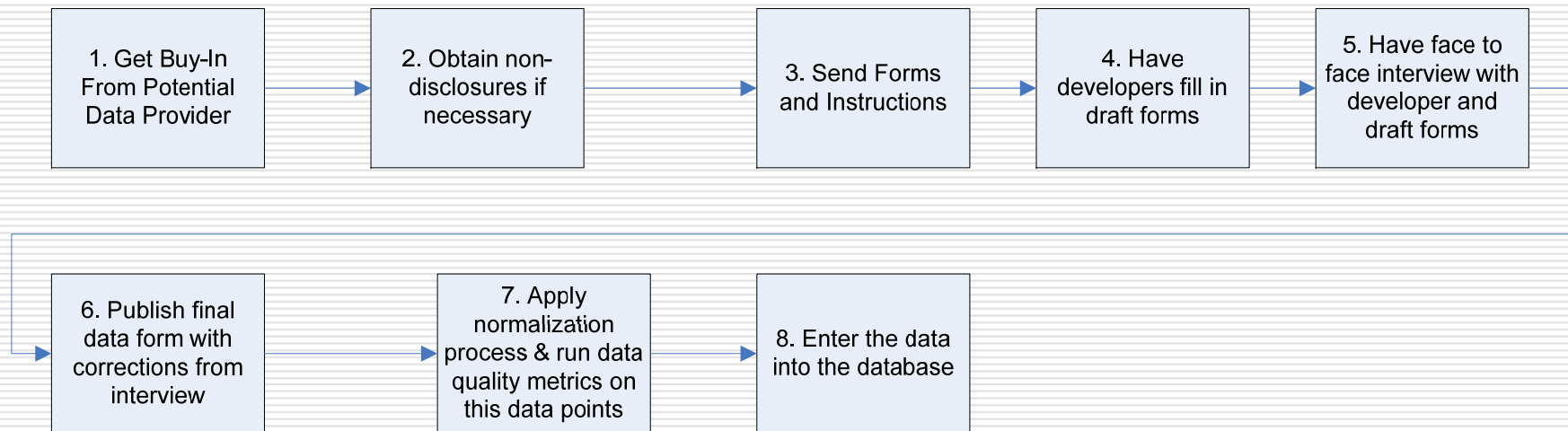
2.0 Why We Need Data Collection & Analysis, Overview

- ❑ Parametric estimating techniques, including cost estimating relationships (CERs) and complex models, require credible data before they can be used effectively.
- ❑ Calibration of existing models requires data
- ❑ Credible data are needed in two ways:
 1. **Build the model**--if the data are not credible, the model cannot be trusted
 2. **Input to the model**--If the data are not credible, the output cannot be trusted





Data Collection Process





2.1 Data Types

- Not all parametric tools require all kinds of data--what is needed depends on the nature and purpose of the tool
- Sophisticated models may require some or most of the following:
 - Costs--labor, material, other
 - Duration--length of time to do the work
 - Size (hardware)--weight, count, area, length, volume, etc.
 - Material (hardware)
 - Manufacturing process (hardware)
 - Size (software)--SLOC, function points, object classes, etc.
 - Process (software, production)
 - Intended use of the product and quality requirements
 - Team experience and ability, and tools available to the team
 - Volatility of the requirements...and more!



Formats Typically Required for Data Collection

- ❑ Unless data are collected in some structured manner, they are generally useless for either model building or model input
- ❑ Work breakdown structure (WBS)--it is typically product oriented
- ❑ Another is activity based costing (ABC)
 - Focuses on being inclusive of all activities
- ❑ Many other formats are possible:
 - Project phase (time and activity oriented)
 - Chart of accounts (accounting oriented--not popular with project teams, but this is the way the books are kept)
 - Precision manufacturing costing (emphasis on equivalent “rental” of factory work stations and giving loss responsibility to marketing, rather than manufacturing, if the machines are kept busy)
 - Throughput accounting
 - **Most projects use a mixture of formats!**



“In God we trust,
all others bring data.”

- W. Edwards Deming



Best Practices

Regardless of the method of formatting, a contractor's data collection practices should be consistent with the processes used in estimating, budgeting and executing the projects on which the data were collected. If this is not the case, the data collection practices should contain procedures for mapping the costs used in the database to the specific model elements.



Data Sources

□ Primary

- Primary data are obtained from the original source
- Considered the best in quality, and ultimately the most reliable

□ Secondary

- Data derived (possibly "sanitized") from primary data
- Not obtained directly from the source
- May be of lower overall quality and usefulness.

Source	Source Type
Basic Accounting Records	Primary
Cost Reports	Either (Primary or Secondary)
Historical Databases	Either (Primary or Secondary)
Functional Specialist	Either (Primary or Secondary)
Technical Databases	Either (Primary or Secondary)
Other Information Systems	Either (Primary or Secondary)
Contracts	Secondary
Cost Proposals	Secondary

Primary data is the best and most reliable



(1) WBS-1

- As an example:
 - Defined in MIL-HDBK-881 (for guidance only)
- Definition: A product-oriented family tree composed of hardware, software, services, data, and facilities
- A WBS displays and defines the product, or products, to be developed and/or produced
 - Relates elements of work to be accomplished to each other and to the end product
 - WBS can be expressed down to any level of interest
 - Generally the top three levels are as far as any program or contract need go unless the items are high cost or high risk



(1) WBS-2

- ❑ Mil-HDBK 881 Level 1 is the entire materiel item--e.g., a radar system
- ❑ Level 2 elements are the major elements of the item--e.g., a power supply for a radar system
- ❑ Level 3 elements are elements subordinate to level 2 elements--e.g., a rectifier in a power supply for a radar
- ❑ What is not a WBS element, according to MIL-HDBK-881:
 - Functional efforts such as design engineering and requirements analysis
 - Material resources such as aluminum or composites
 - Accounting classifications such as direct cost
 - Program phases, such as design, development, production
 - Rework, refurbishment, and retesting
 - Recurring and non-recurring
 - Warranty, could cost, cost savings efforts
 - Organizational structure



Data Collection “Gotchas”

The Case of the Made Up Data

- ❑ Collecting data independently from the same developer for two purposes
 1. For a database
 2. For an estimate
- ❑ The numbers were completely different

- ❑ What do you do?



Best Practices

A common problem in data collection is that different projects use different WBSs. Getting data to be “apples to apples” can sometimes be a daunting task. Organizing data into a common structure for similar products can be very helpful in model building.

Yes, a WBS is product oriented. But the best data are collected if every WBS element breaks out all of the “not a WBS” excluded elements on the preceding chart!



(2) ABC-1

- Activity based costing is an attempt to get at what products “really cost”
- Consider a factory that builds two products, A and B
 - Both have the same direct cost, but A is mostly material and B is mostly labor
 - The company applies a 100% burden to labor
 - The distortion shown below is removed by looking at the true activity costs--example: A requires 5 times as many purchase orders as B per unit; B requires twice as much use of tools as A

Conventional Accounting					
Product	Material	Touch Labor	Direct	Overhead	Total
A	\$10	\$2	\$12	\$2	\$14
B	\$2	\$10	\$12	\$10	\$22

ABC Accounting					
Product	Material	Touch Labor	Purchasing	Tools	Total
A	\$10	\$2	\$5	\$4	\$21
B	\$2	\$10	\$1	\$8	\$21



Traditional vs.. Activity View

Source (J. Kammerer)

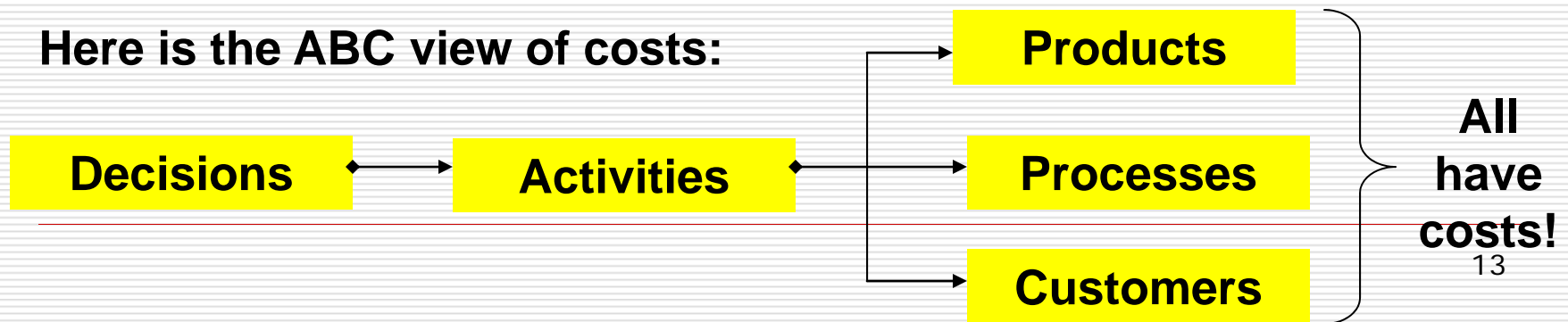
Traditional View

<input type="checkbox"/>	Military personnel	\$175m
<input type="checkbox"/>	Civilian personnel	\$ 25m
<input type="checkbox"/>	Equipment	\$ 10m
<input type="checkbox"/>	Supplies	\$ 10m
<input type="checkbox"/>	AVPOL	\$ 50m
<input type="checkbox"/>	DLRs	\$ 45m
<input type="checkbox"/>	TDY	\$ 2m
<input type="checkbox"/>	Total	\$317m

Activity View

<input type="checkbox"/>	Fly Missions	\$300m
<input type="checkbox"/>	Repair Aircraft	\$10m
<input type="checkbox"/>	Maintain Grounds	\$ 3.0m
<input type="checkbox"/>	Operate Library	\$ 0.5m
<input type="checkbox"/>	Operate CDC	\$ 2.0m
<input type="checkbox"/>	Process Payroll	\$. 25m
<input type="checkbox"/>	Provide Phone Svc	\$1.25m
<input type="checkbox"/>	Total	\$317m

Here is the ABC view of costs:





(3) Other Formats-1

- Chart of accounts--the view of the accounting department
 - Be aware and beware: the accounting department's customers are senior management, stockholders, investors, and the IRS, not project teams and cost analysts
 - Be sure you understand the accounting rules behind any numbers they give you
- Order of assembly--the view of the manufacturing department
 - What is a wing?
 - WBS view (possibly): structure and skin
 - Manufacturing view (possibly): structure, wiring, plumbing, flap actuators, fuel cell, navigation lights, spoiler actuators, skin
- Throughput Accounting (Goldratt)
- Many others



Best Practices

Whatever formatting method(s) you choose, be sure you know what is included in your cost data!



C) Data Collection Practices

- Subjects to be discussed:
 - Consistent with current practices
 - Automation
 - Accounting procedures
 - Earned value reports
 - Proposals
 - Technical data



(1) Consistent with Current Practices

Regardless of the method of formatting, a contractor's data collection practices should be consistent with the processes used in estimating, budgeting and executing the projects on which the data were collected. If this is not the case, the data collection practices should contain procedures for mapping the costs used in the database to the specific model elements.



Data Collection "Gotchas" The Case of the undefined data

- Small company was selling a software database
- Included the size for over 10,000 programs
- When queried, they didn't know the definitions of size, nor were the sizes consistent

- How much you pay for such a database?



(2) Automation

- Automation affects data collection and analysis in at least two ways:
 - Ability to store and rapidly process more and more data
 - Affects the amount and quality of data available for parametric analysis
 - Tends to improve the quality of parametric models as time goes on and data automation improves
 - Improved productivity of project teams due to better tools
 - This affects the comparability of historical data to current estimates because of improvements in automation
 - Treat older data (five+ years old) specially to ensure improvements in efficiency and technology are dealt with



(3) Accounting Procedures-1

- Accounting procedures have as their primary customers senior executives, board members, stock and bond holders, lenders, and the IRS
 - The rules for accounting procedures are based on Generally Accepted Accounting Principles (GAAP), a huge body of principles from many sources
 - They represent the result of political maneuverings and compromises between users, practitioners, and government interests
 - They are not a stable platform for parametric analysis
 - Even experts don't always agree on the interpretation of GAAP
- The main reason ABC and related methods have evolved has been dissatisfaction with GAAP style accounting among product teams
 - **Except for valuation of inventory at the end of an accounting period, GAAP is essentially unconcerned with product costs**
 - The analyst must continually be aware and beware of the differences



Best Practices

Think ABC and beware of the GAAP!



(4) Earned Value Reports-1

- ❑ Earned value management (EVM) systems are mandatory on major defense acquisition programs, including large subcontracts—
- ❑ Cost Performance Reports are the main source of data for parametric analysis
- ❑ There are several forms
 - WBS element
 - Organizational categories
 - Baseline
 - Staffing
 - Explanations and problem analysis
- ❑ Definition of contract baseline:

Negotiated
cost

+

Definitized
changes

+

Authorized
undefinitized

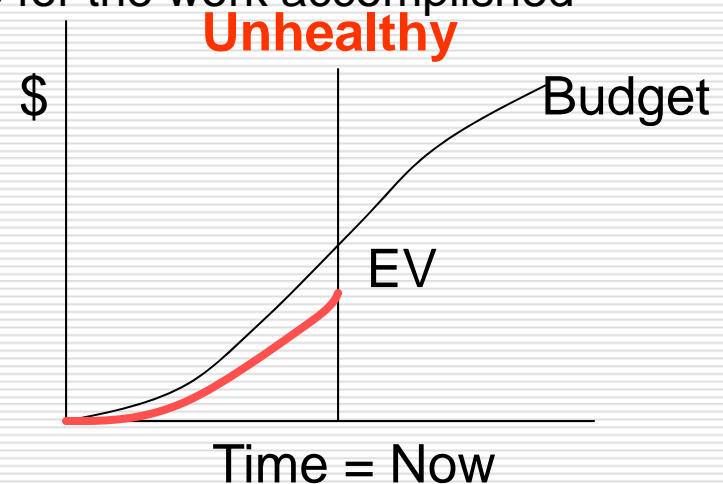
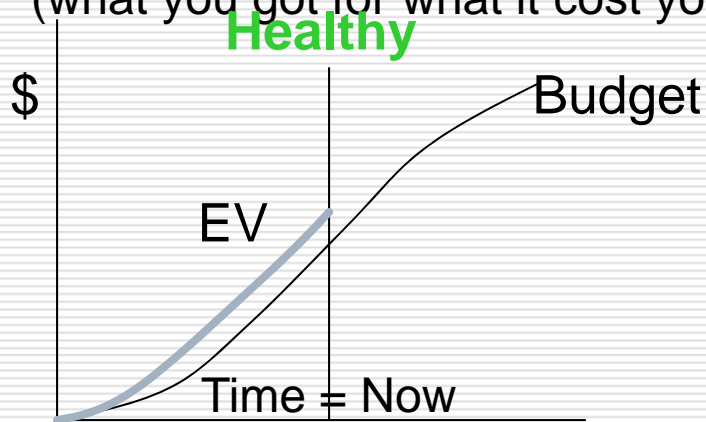
=

Contract
baseline



(4) Earned Value Reports-2

- The main concern of EVM is what has been accomplished in a given time and budget, versus what was planned for the same time and budget
 - A project is generally deemed healthy if what has been accomplished is what was planned, or more
 - A project is deemed unhealthy if accomplishment lags expectations
- Definition: Earned value = budgeted value for the work accomplished (what you got for what it cost you)





Data Collection Gotchas

The Case of The Unearned Value

- ❑ Earned value analysis showed a certain productivity and cost to produce
- ❑ Earned value showed the task was 75% complete
- ❑ What are the risks of using the earned value before something is completed as “data” for modeling, calibrating or predicting future systems?



(5) Proposals

- Proposals can be a useful source of cost and schedule information under certain conditions
- Conditions:
 - The work has not been won, or is still in its early stages (else actual results are preferred)
 - The proposal was carefully done with a serious intent to perform the work at a fair price (not a courtesy bid or an attempt to buy in)
 - The proposal was not widely at variance with the winning bid
 - The subject matter was similar to situations where the data are to be used for estimating



(b) Cost Drivers-1

- ❑ Cost driver: a parameter that tends to increase or decrease cost or is highly correlated with cost
- ❑ To qualify as a cost driver, a parameter must be a significant predictor of cost at least in the statistical sense



(b) Cost Drivers-2

□ Typical hardware cost drivers

- Weight
- Thrust
- Energy
- Power
- Count
- Area
- Box volume
- Complexity (various measures)
- Operating environment
- Reliability
- Maintainability
- Inheritance
- Team capability

□ Typical software cost drivers

- Source lines of code
- Function point count
- Object class count
- Number of screens
- Language
- Application type
- Complexity (various measures)
- Build environment
- Target environment
- Reliability
- Maintainability
- Degree of reuse or COTS
- Team capability



Sources of Data

Basic accounting records – primary

- Plant labor collection systems
- Best source of cost data
 - Local cost experience
- SAP
 - Native format not designed for DoD type of business model
 - Changes the way cost is collected and reported
 - Data mining more difficult



1) Management Information Systems

MIS-1

- MIS: Any computer oriented system that stores large quantities of data and data manipulation tools for authorized users. Frequently includes:
 - Data warehousing
 - Enterprise resource planning
 - Strategic planning
 - Workflow planning
 - Business intelligence
 - Supply chain management
 - Electronic commerce and procurement
 - Customer relations management
 - Voice and data communications
 - Information security
 - Parametrics practitioners often associated with engineering, data regarded as “property” of finance or accounting departments

- Another problem is understanding the true nature of the data—recall ABC vs.. GAAP



2) MRP / ERP -1

- MRP systems simplify manufacturing planning of dependent items
 - A dependent demand item is one whose demand depends on the demand for one or more parent items
 - MRP systems also accommodate planning for independent items, which require human scheduling input
 - MRP combines several planning functions, such as bill of materials explosion, netting, and back scheduling
 - Extended to distribution requirements planning and manufacturing resource planning (MRP II)
- The chief benefit of MRP/MRP II is its ability to generate valid schedules and keep them that way
- MRP reports are a rich source of data on what has actually happened in manufacturing, and what is planned for the future



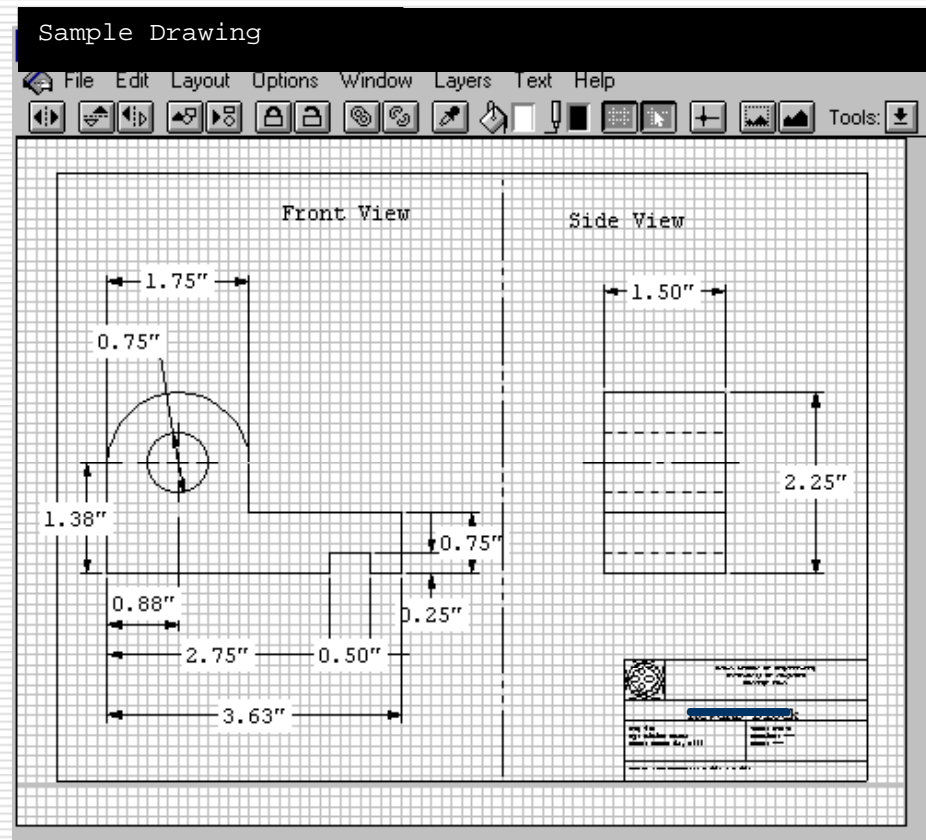
3) Engineering Drawings-1

- An engineering drawing is the visual expression of what the designer thinks the product should look like, and how it should be made
 - It is the factory's guide to building the item
- Normally, a drawing specifies form, fit, function, finish, and materials; related documents may specify other properties
 - Other properties may also be specified, such as test and acceptance requirements, interfaces, and make vs.. buy
 - Other analytical documents referencing a drawing may be said to be included as “drawings”, e.g., weight statements, loads and stress analyses, power consumption analyses, etc.
- Usual engineering practice is to define a product in terms of a “drawing assembly breakdown”
 - This is the engineers' logical view of how the product is built up from basic components—it may differ somewhat from the WBS or the manufacturing order of assembly



3) Engineering Drawings-2

- Many parametric models focus strongly on the information to be found in engineering drawings
 - This is no accident—the physical nature of the product is always strongly related to the effort and therefore the cost to create it
- Access to drawings today is much easier than in years past because most drawings are created and stored electronically





3) Engineering Drawings-3

- A valuable adjunct to drawings is the Bill of Materials
 - Typically a small database that identifies the source of every item on the drawing
 - It gives part numbers, quantities, and vendor names
- Another valuable part of drawings is the Notes
 - Notes typically provide information about unusual manufacturing processes or special testing that must be done



4) Specifications-1

- A specification is a word description of the required features and characteristics of a specific product

- Specifications are not generally a rich source of parametric data
 - But they do “bound” the product in ways that may be useful to the analyst
 - Examples:
 - “The widget shall not weigh more than x pounds.”
 - “The gizmo shall be gold plated.”
 - “The whatizit must have folding protuberances in accordance with ANSI xyz.”
 - “The software shall read instructions from a script file and present them to the user on request.”



5) Interviews-1

- An interview is a vehicle for obtaining parametric data from subject matter experts
- Some steps for a successful interview:
 - Identify the goal of the interview; determine what information is needed
 - Decide on the structure and approach of the interview
 - Decide on the number and sequence of questions
 - Draft questions and critique them
 - Decide how you will summarize and report the interview
 - Choose a recording method
 - Make arrangements for the interview—be sure you have the right subject matter expert



5) Interviews-2

- An interview can be unstructured if you are not sure what data are available
 - An unstructured interview might be followed up later with a structured interview
- General (open) questions are best if you want a wide range of responses, but ask specific (closed) questions if you want particular data
 - Be aware and beware: Closed questions asked too early may mislead the respondent into thinking you don't need certain information that could in fact be valuable
- It is better to err in the direction of asking too many questions rather than too few
- Be sure your questions are understandable to your respondent—tell your respondent how the data will be used



Best Practices

Be clear on what you are going to do with the data as you plan the interview, else you may not ask the right questions.



Interview “Gotchas”

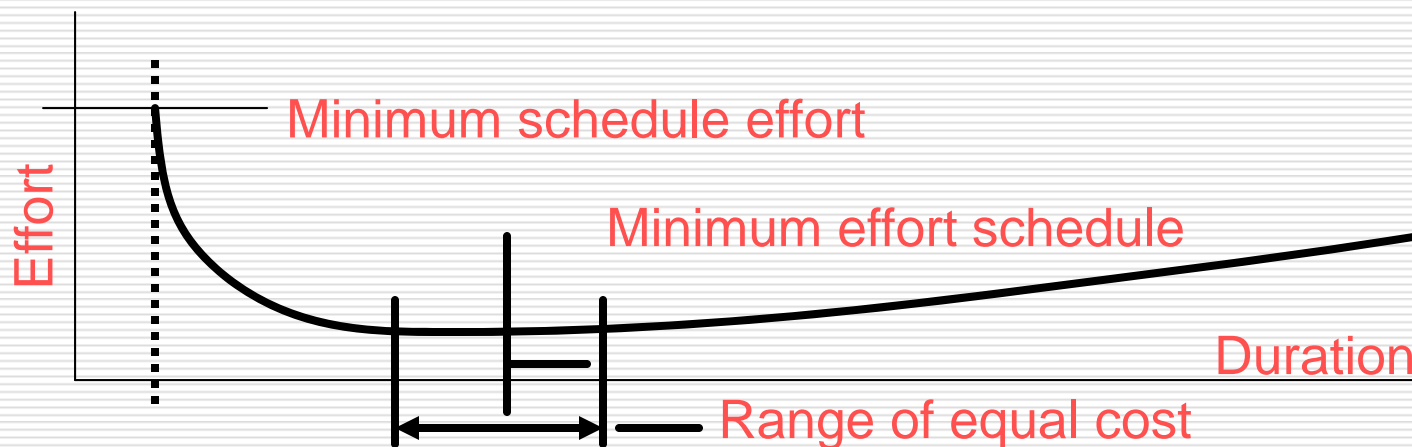
The case of the bad memory

- ❑ Interviewing engineering personnel they generally remember costs but...
- ❑ They don't remember what is included on excluded in the cost data
- ❑ They don't remember subcontractor loading nor other key information
- ❑ They often remember what they thought something should cost, not what it did cost.
- ❑ What can you do for the bad memory?



(a) Schedule-1 (Programmatic Data)

- Projects typically have a:
 - Minimum time schedule, virtually impossible to decrease
 - Minimum effort schedule, such that to shorten or lengthen it will increase the effort
 - Range of approximately equal cost around the minimum effort schedule
- Decreasing the schedule is much more expensive than increasing a similar amount
- Decreasing the schedule tends to introduce more defects; increasing it tends to decrease defects





(a) Schedule-2

- Implications for data collection and analysis
 - If possible, it should be learned whether a historical project was conducted under minimum time, normal or relaxed conditions--the cost and schedule differences can be dramatic
 - A rushed development has been observed to both flatten the improvement curve for early units (while "bugs" are removed from both the product and the production processes), and to increase the first unit cost



(b) Quantities-1 (Programmatic Data)

- Production quantity is a huge driver of effort, cost, and schedule.
- At small quantities (say, under 10), the relationship shown tends to prevail

$$\text{Total Cost} = \text{Unit Cost} \times \text{Quantity}$$

- At higher quantities, various economies of scale may come into play:
 - Quantity cost breaks from suppliers
 - The ability to use better and more efficient, but more expensive, equipment, facilities and processes
 - Improvement curves, related to learning by workers, more efficient procurement practices, increased yield, reduced scrap, better processes, etc.



(c) Production Rates-1

- Effects of production rate on costs are very complex
- A production line has a natural pace that tends to increase gradually over time, up to a limit
- Tools for increasing the pace include:
 - Incentives for the workers
 - Giving the workers better tools
 - Adding more production lines
 - Simplifying the product
- A sometimes asked question is, what happens if the rate is cut substantially, to accommodate lower funding in the future?
 - This is a difficult question for a parametric model, because it happens infrequently, and every instance seems to be unique
 - It generally requires an industrial engineering study



(c) Production Rates-2

- A model sometimes used to account for small changes in rate is:

$$\text{Total Cost} = T_1 * \text{Qty}^{B_1} * \text{Rate}^{B_2}$$

- This a modified improvement curve , with rate as an additional parameter
- Some manufacturers will have difficulty obtaining enough data to support modeling this curve
- Implications for data collection and analysis
 - This is a difficult and complex subject
 - Data are likely to be scarce



(d) Improvement Curves-1

- ❑ A phenomenon first noted in the production of aircraft in WW II was that as production quantity increased, labor hours per aircraft decreased, rapidly at first, less rapidly as time went on
- ❑ It was first believed that this was due to “learning” by the workers—they simply got better at what they were doing as they repeated it
- ❑ Today, learning is an accepted component of the phenomenon, but other factors are also seen as contributing, such as better tools and methods, and better management
- ❑ Because more than just learning by the workers seems to be involved, the term “improvement” has come to be preferred over “learning”



(d) Improvement Curves-2

- Two models of the improvement effect have achieved wide acceptance:
 - **Unit model:**
 - $U_n = U_1 n^B$
 - This relates the hours for the nth unit to the hours for the first unit
 - B is called the natural slope of the learning curve—it is generally a negative number less than unity
 - **Cumulative average model:**
 - $A_n = A_1 n^B$
 - This relates the average hours for units one through n to the hours for the first unit
 - Again, B is called the natural slope, and it is generally a negative number less than unity
- Both models rely on the so-called “power law” relationship characterized by $y = ax^b$



Underlying Power Law

The “power law” formula is the basis of both models

$$y = ax^b$$

b is called the natural slope—it represents the rate of learning [always a negative number except for (rare) “forgetting”]

The two models differ in their interpretation of y (next chart)

a always represents the theoretical labor hours required to build the first unit produced (a positive number)

x always represents the number (count) of an item in the production sequence (unit #1, #2, #3, etc.)



Lore of the Slope-1 (Source Stump)

- Fit learning curves to historical data when available
 - This is usually the best source, but not always
 - Guidelines for use when historical data are not available:
 - Operations that are fully automated tend to have slopes of 100%, or a value very close to that (no learning can happen).
 - Operations that are entirely manual tend to have slopes in the vicinity of 70% (maximum learning can happen). (cont.)
-



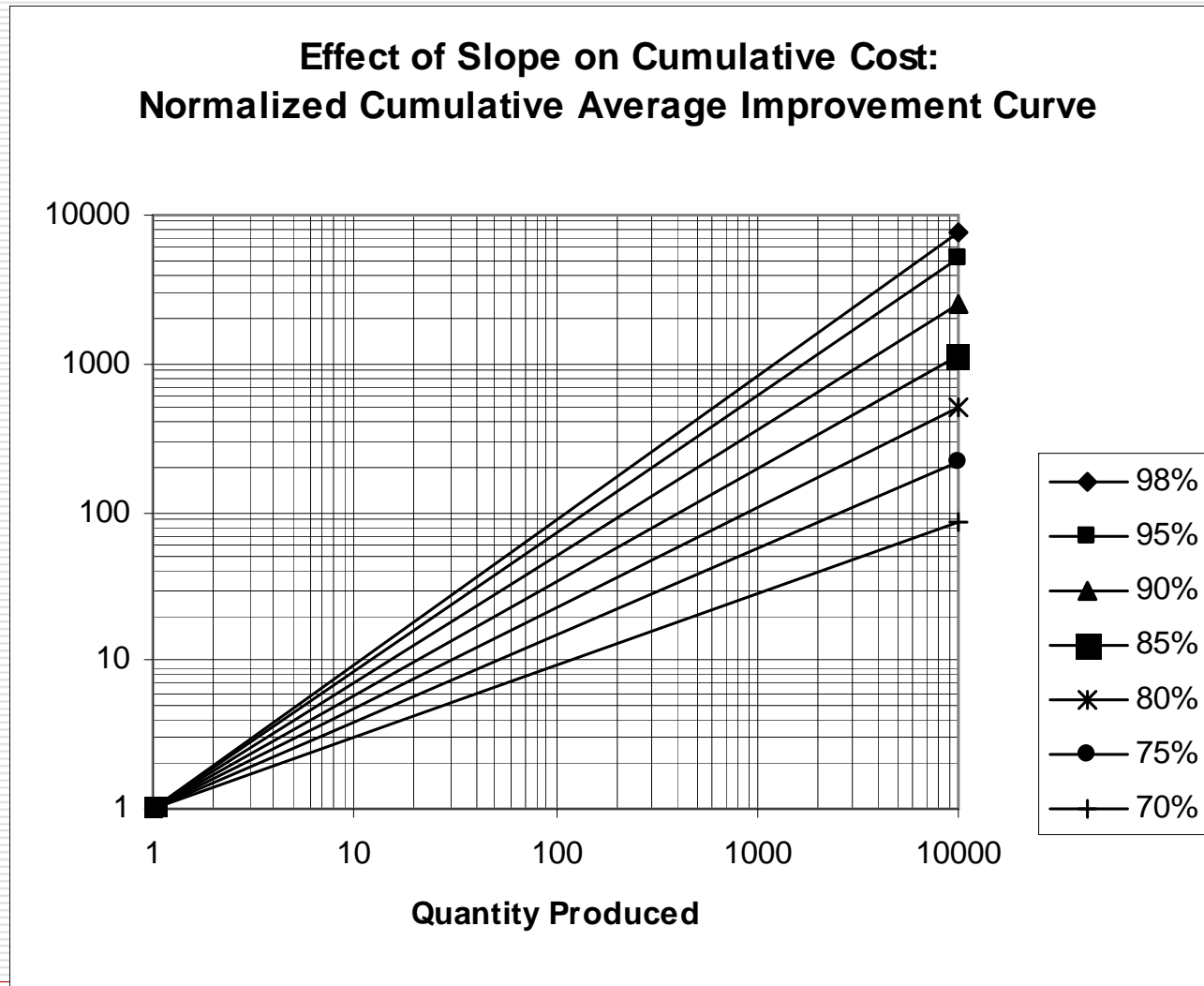
Lore of the Slope-2 (Source Stump)

□ Guidelines (cont.)

- If an operation is 75% manual and 25% automated, slopes in the vicinity of 80% are common.
 - If it is 50% manual and 50% automated, expect about 85%.
 - If it is 25% manual and 75% automated, expect about 90%.
 - The average slope for the aircraft industry is about 85%. But there are departments in a typical aircraft factory that may depart substantially from that value.
-
- Shipbuilding slopes tend to run between 80 and 85%.



(d) Improvement Curves-5



Note 100% = No Learning or Improvement



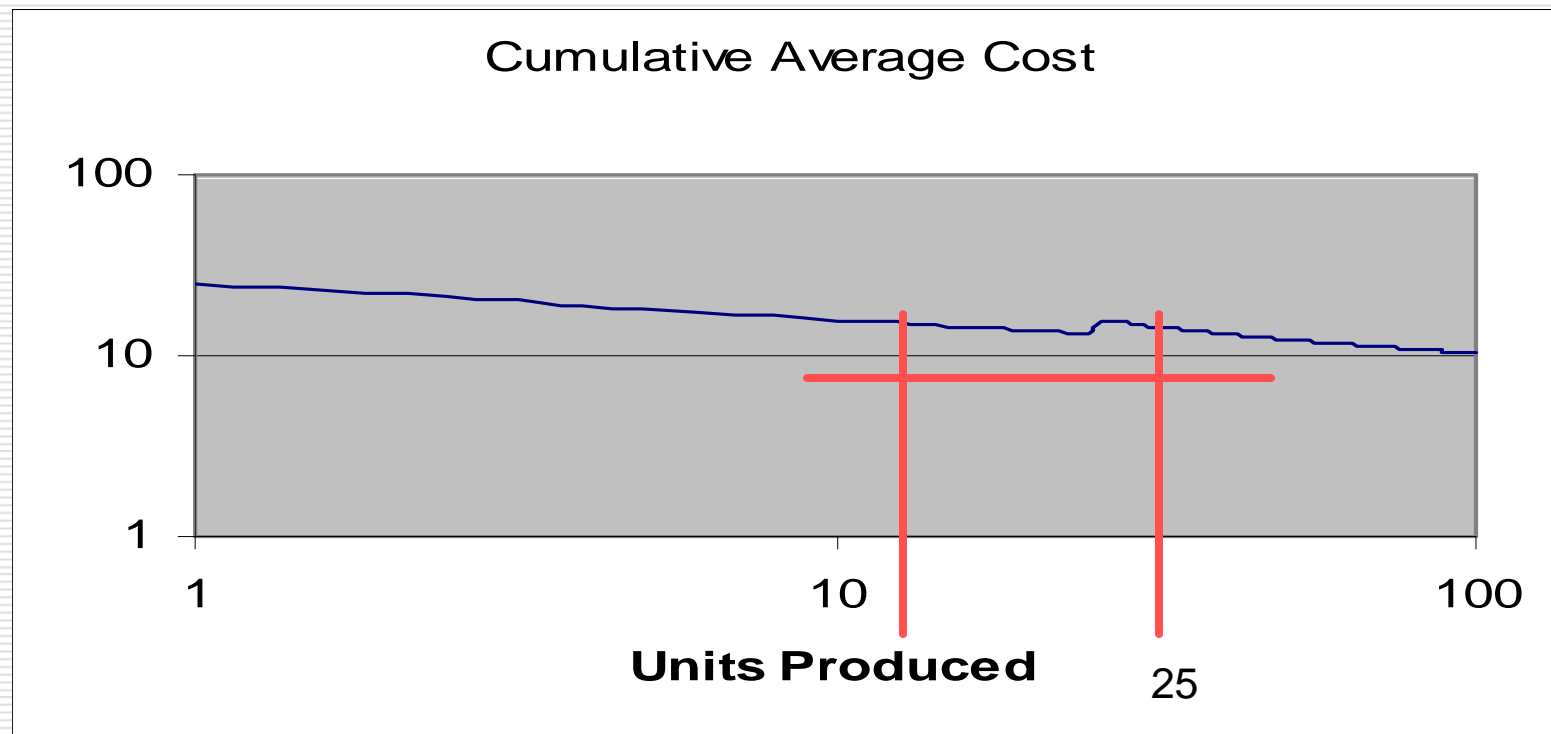
(e) Production Breaks-1

- A break in production is any significant temporary discontinuance of production for any reason
- Common reasons are:
 - Design changes
 - Introduction of major production methods improvements
 - Replacement of worn tooling
 - Funding limitations
 - Strikes
- A break in production can cause a “loss of learning”
- The prevailing wisdom is that:
 - The longer the break, the greater the loss
 - Several factors are at play in determining the amount of loss
- The theory that seems to have the greatest acceptance is due to George Andelohr (“What Production Breaks Cost,” *Industrial Engineering*, September 1969.)



(e) Production Breaks-2

- Applying Andelohr's theory: break in production at unit 25 with loss of learning back to unit 10



Implications for estimating and risk
An often overlooked factor in both estimating
and risk analysis is the possibility of breaks



(f) Design Changes

- A common reason for a break in production is a design change
- This complicates the production cost data analysis picture considerably
 - The entire design is seldom changed—the unchanged part may or may not proceed without a break in production
 - If it proceeds without a break, it should follow the existing improvement curve, all else being equal
 - If there is a break, Andelohr's method can be used
 - The part that is changed typically starts on a new improvement curve
- The non-recurring cost of design changes is usually modeled using the same parametric methods used for the initial design—but some companies create special models for design changes



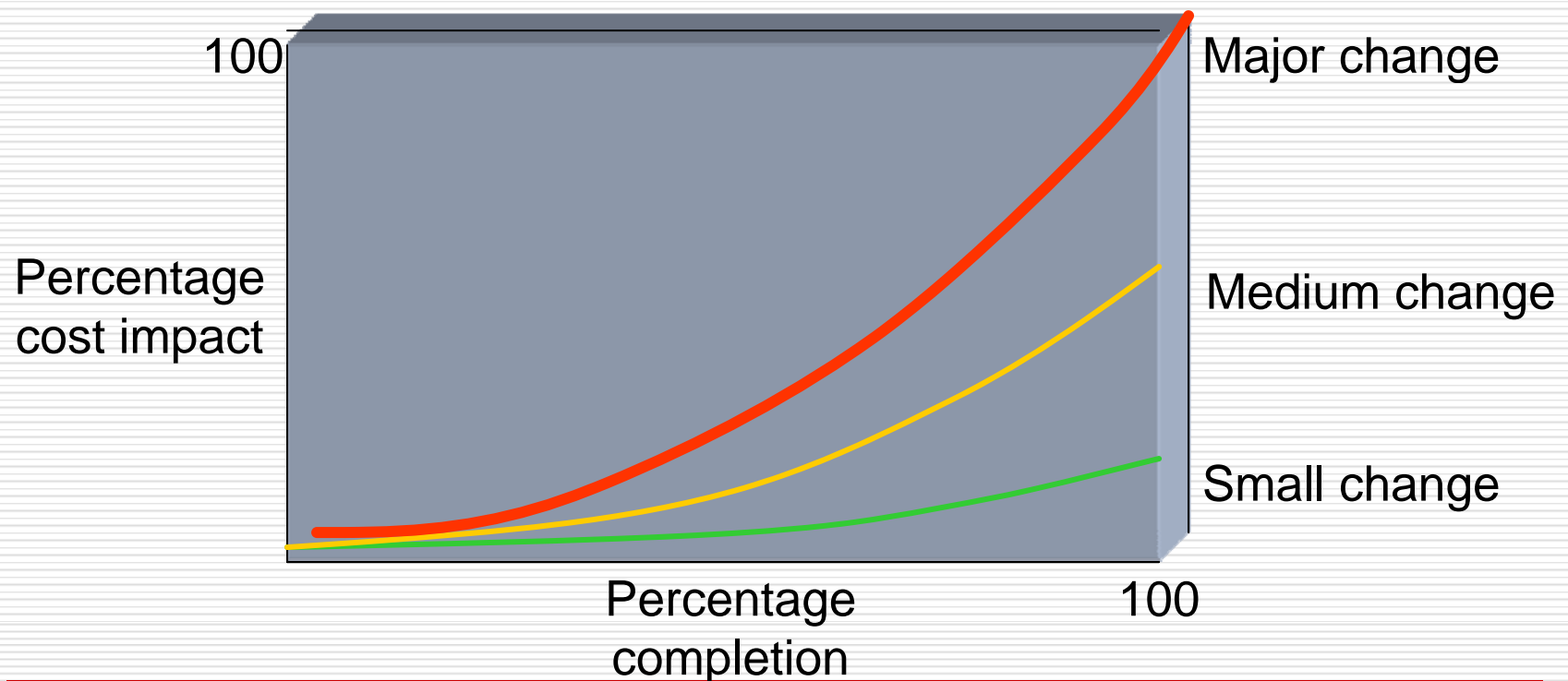
(g) Funding Changes

- The most common funding change is a reduction or even a total cancellation
 - Parametric analysis is usually not appropriate for cancellations, because they involve complex and unique issues of termination costs and indemnification that have not been modeled
 - Parametric analysis may have a role in a reduction or “stretch-out,” if it is not too extreme
- A funding change frequently involves a change in production quantity
 - May also involve design changes to reduce costs
 - May involve breaks in production



(h) Requirement Changes-1

- ❑ A requirement change when little work has been done is of little consequence
- ❑ A requirement change when a lot of work has been done can be a major problem, and will likely have a major cost impact
- ❑ Notionally, the plots below describe the situation





(i) Other Factors

- Here are some other programmatic factors that can influence costs, and that should be considered, if possible, in data collection and analysis
 - Experience and capability of the project team
 - Inherent difficulty of the work
 - Maturity of technology
 - Unusual technology integrations
 - Working conditions
 - Available tools and facilities
 - Ease of communications
 - Inheritance from previous projects

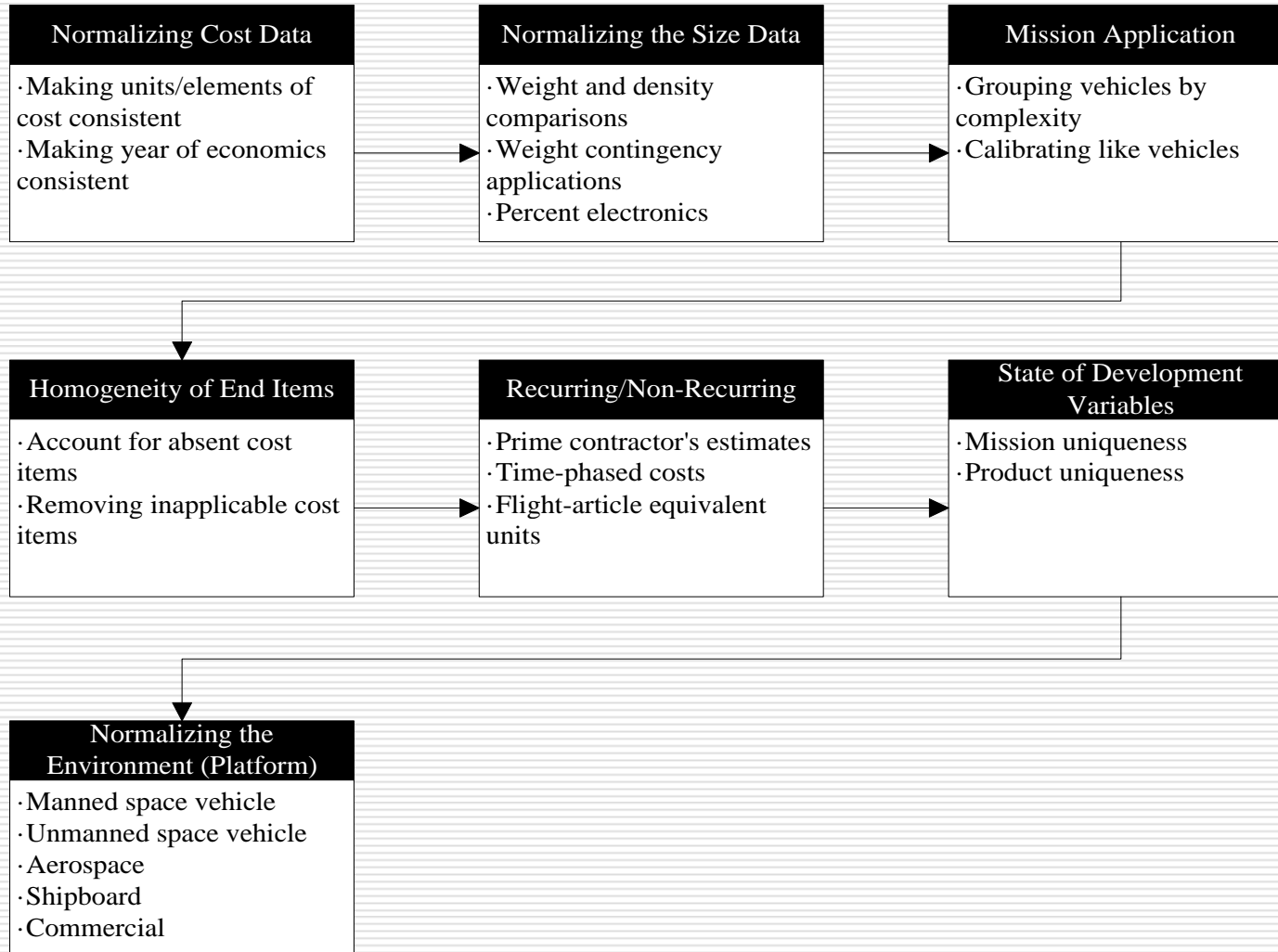


Best Practices

Consider all the dimensions of your data. Costs have many drivers. Get to know them.



Data Normalization





Normalization

Adjustments for homogeneity

- Data comes in all “sizes and shapes”
- Development of CERs and calibration of models require that format and scale of data used be internally consistent
 - Inches versus centimeters
 - Full dollars versus “K” dollars versus Euros
 - Labor phases and activities
 - Handling of overhead
 - ... and the list goes on



Normalization

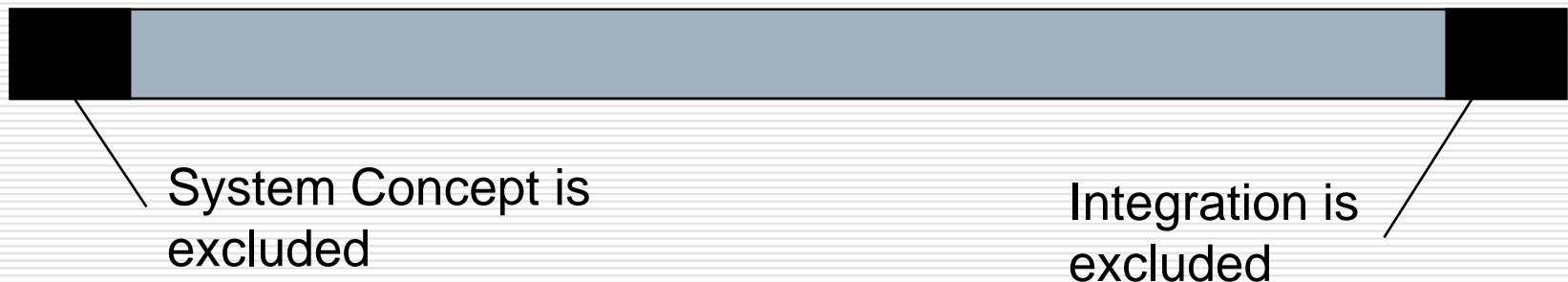
Adjustments for homogeneity

- Time phased data can be particular difficult to handle
 - As spent dollars are influenced by inflation
 - Requires knowledge of time period involved
 - May require converting to “fixed dollar” units
 - Process data
 - Need to know specific sub-phase of the process
 - May require allocation to sub-phase



What To Look For In An Actual - Phases

Phase : All *activities* may not be included.



Phases

System Concept

System Req Design

System Req Analysis

Preliminary Design

Detailed Design

Coding / Unit Testing

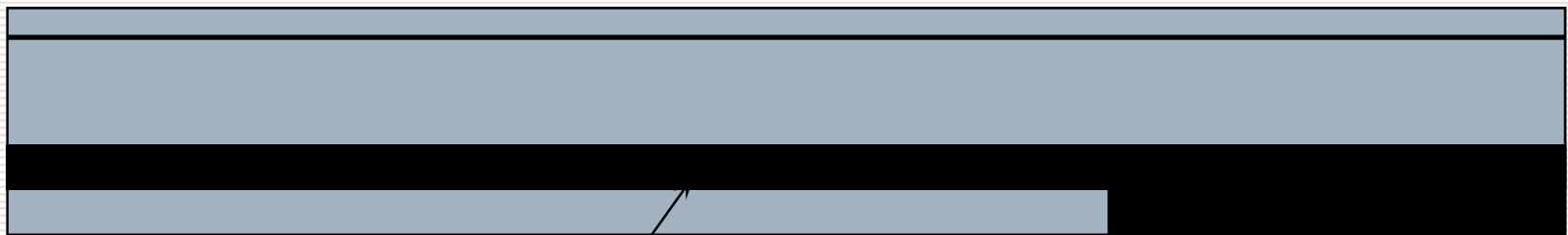
Software Test

System Integration / OT&E



What To Look For In An Actual - Labor

Labor : All *categories* may not be included.



Configuration Mgt.
is excluded

Quality Assurance is
excluded

Labor Categories

Management

Design

Data Preparation

Configuration Mgt

System Engineering

Coding

Test

Quality Assurance



What Else To Look For In An Actual

Forensics

Schedule. Was it “stop and start”? Were there schedule constraints?

Resources. Were there hard-hitting resource constraints?

Volatility. Did requirements undergo extraordinary evolution?

Manager’s Objectives. Was it to complete the project in *minimum time* or *at least cost*?

Effort. Are effort figures actually derived from cost figures?

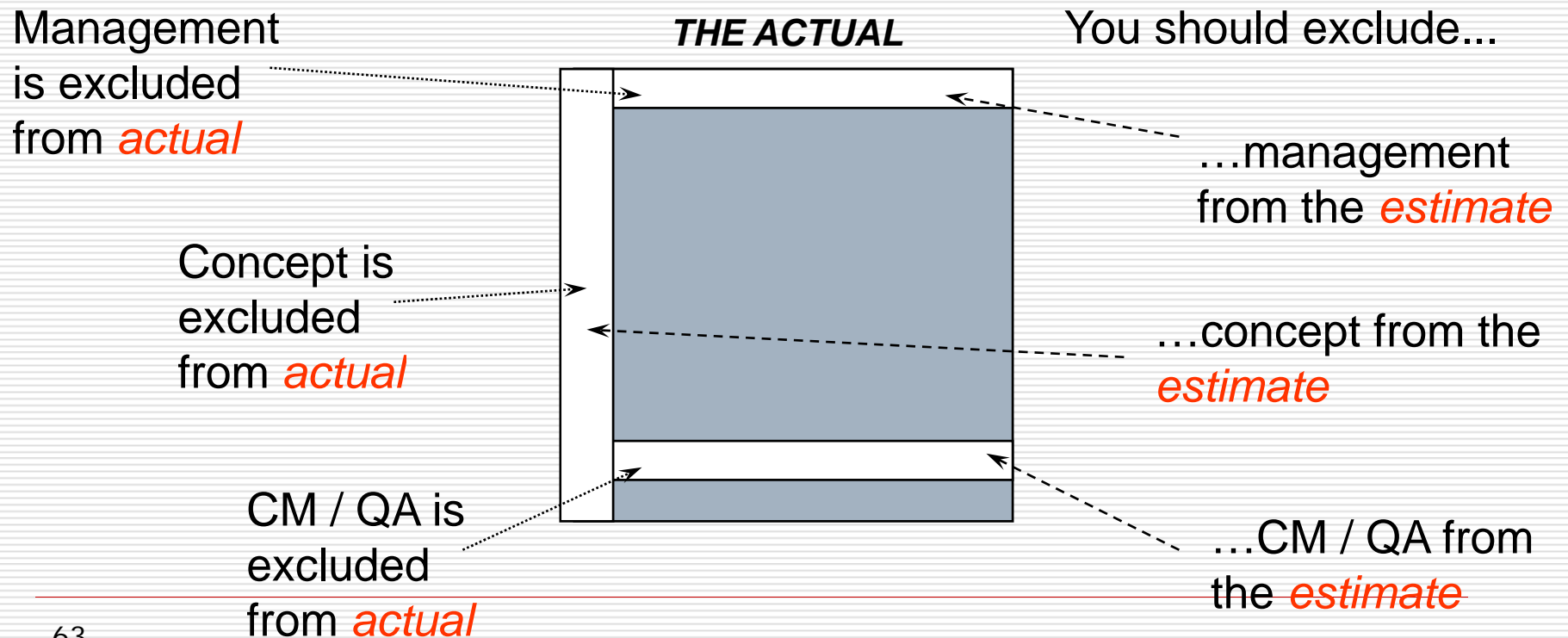
When creating the estimate --- adjustments for *extraordinary* conditions may be possible *within the software estimating model*.



Normalizing Between An Estimate and Actual

What's missing from the actual?

*Models often break out **activities** and **labor**...
...allowing you to exclude these from the estimate.*





Stratifying Data

Calibration factors that are more **specific** can be more **accurate**
But patterns may not be **evident** in samples that are **too small**.

Categories Concerning

Development Organization (team, contractor, government...)

Application Type (command / control, signal, MIS...)

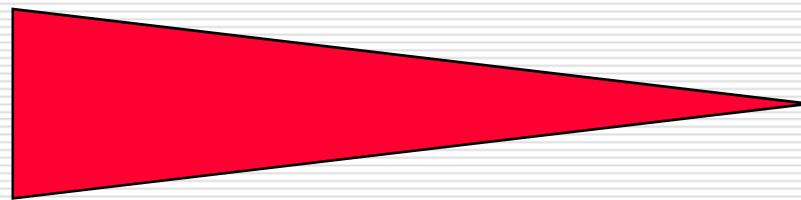
Platform (space / ground, PC / client-server...)

Estimating hierarchy (unit, program, project)

Chronology (recent, ancient)

BEWARE!
With more
subcategories...

Size of sample



...samples get
smaller -- and
statistical validity
decreases.

Categories →



Normalization Examples

Technology

- Technology Identification
 - VLSIC VHSIC SIC
 - Advanced Materials (Composites)
 - Use of Integrated Workbench
 - Advanced Designs (V/STOL)

- Technology Techniques
 - Use of Time as a proxy variable
 - Dummy Variables
 - Stratify the data set



Evaluation Issues

- Sufficient data available?
- Established methodology to obtain relevant data on completed projects?
- Cost, technical, & programmatic data collected in consistent format?
- Data collection consistent with contractor's estimating practices?
- Was the source data adjusted?
- Did adjustments affect the logic, reasonableness, and defensibility of the data?
- Anomalies identified?



Normalization Example: Normalizing the ISBGS Software Benchmark

Lee Fischman & Karen McRitchie
Galorath Incorporated



ISBSG Data Set Background

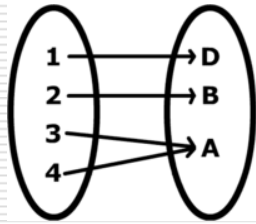
- ❑ The International Software Benchmarking Standards Group collects data for software development and enhancement projects
- ❑ They more recently added a repository for maintenance projects
- ❑ ISBSG offers data sets for purchase which are based on actual software projects
- ❑ The subject of this study was ISBSG release 9 which contains 3023 observations



see www.isbsg.org for more information



Normalization



Mapping ISBSG
to SEM categories



Acceptable
as-is

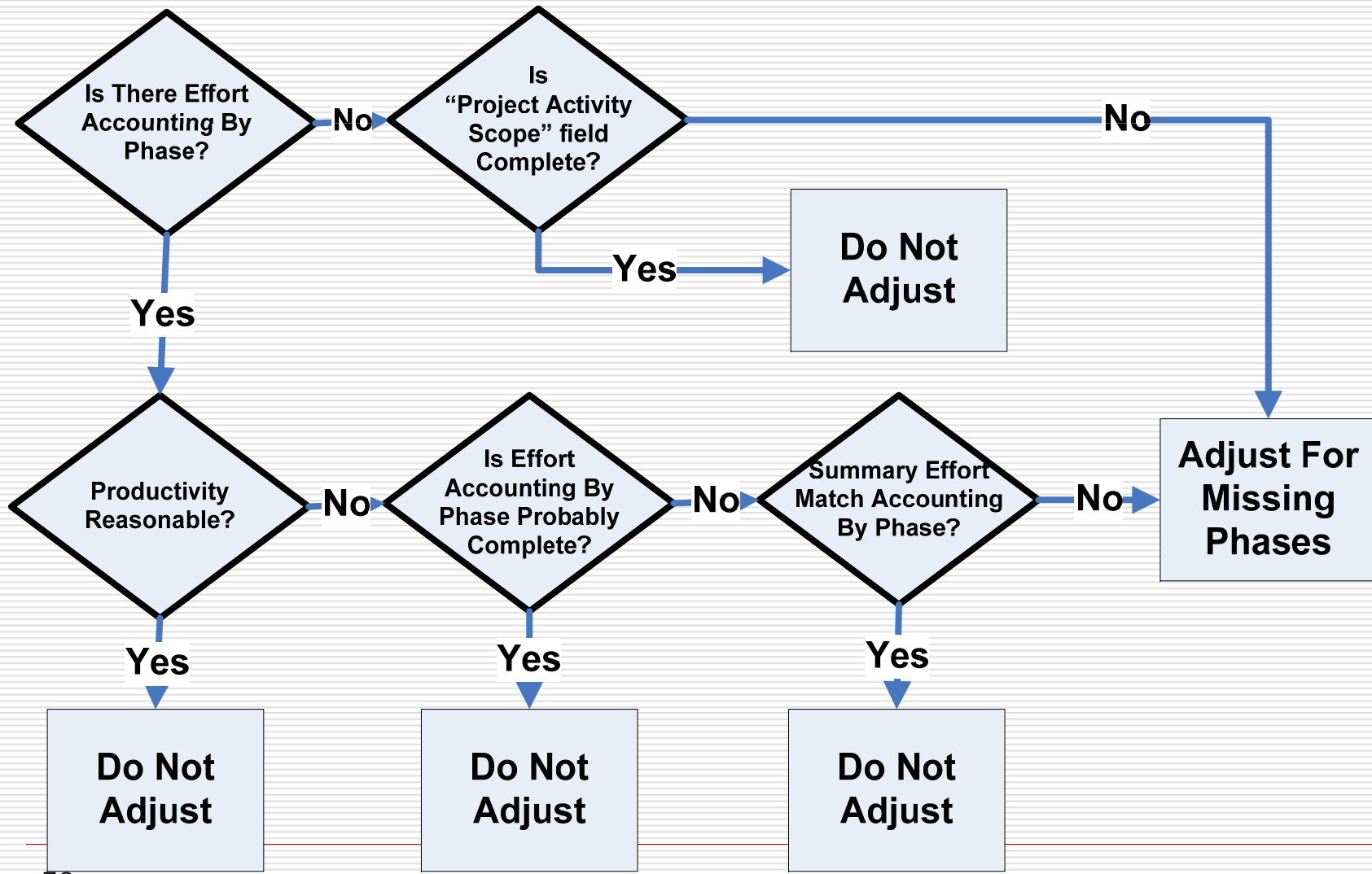


Focus of
our work

- Knowledge bases
 - Mapping ISBSG qualitative descriptors into model attribute names
 - platform, application, method, standard
 - Repeatable process developed
- Size normalization
 - Our analysis utilizes unadjusted function points
 - ISBSG provides a well defined measure of size, requiring no special normalization or analysis
- Effort normalization
 - Model assumes effort for certain phases
 - Effort needs to be normalized so that all projects are comparable



Phase Normalization Decision Process





Phase Reporting In ISBSG v9

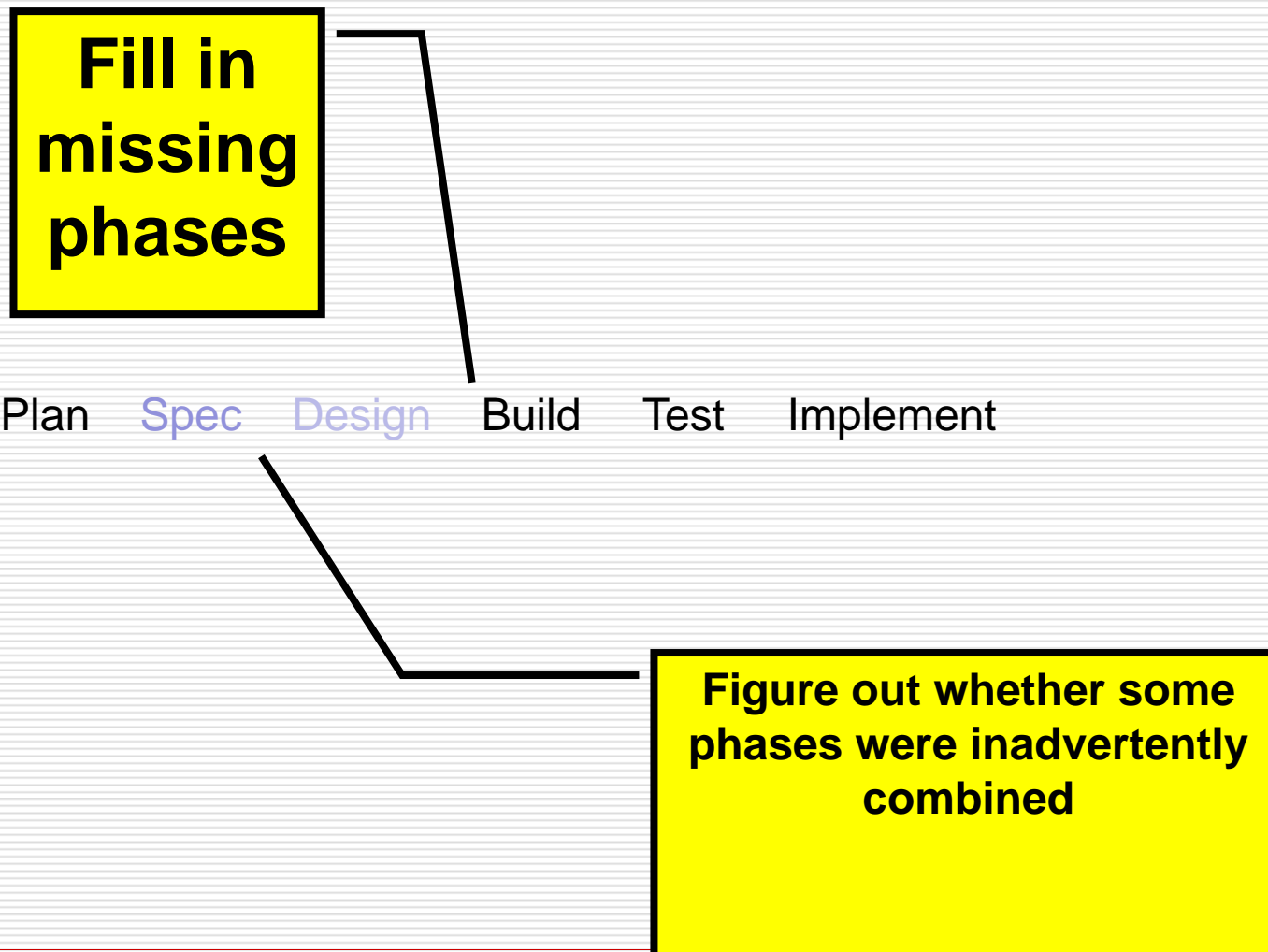
Respondent may have provided list of phases included

Respondent may have provided a detailed effort accounting by phase

Project Activity Scope	Effort Plan	Effort Specify	Effort Design	Effort Build	Effort Test	Effort Implement
Planning;Specification;Build;Test;Implement; Planning;Specification;Build;Test;	100	400		1000	200	150
<i>x</i>		281		359	156	
Specification;Build;Test;Implement; Planning;Specification;Build;Test;	62	17172		136	230	
Planning;Specification;Build;Test;	3	459		2	1	
Specification;Build;Test;		75	<i>x</i>	248	136	
Planning;Specification;Build;Test;Implement;	320	480		800	960	
Planning;Specification;Build; Planning;Build;Test;	427	967		905	358	
Planning;Build;Test;Implement; Specification;Build;Test;	1248	0		2504	8768	12520
Specification;Build;Test;Implement;	<i>x</i>	153		1961	510	255
	90	720		1043	1682	23
Planning;Specification;Build;Test;		640		811	936	
7 Planning;Specification;Build;		102		4		
	39	164		315	65	65

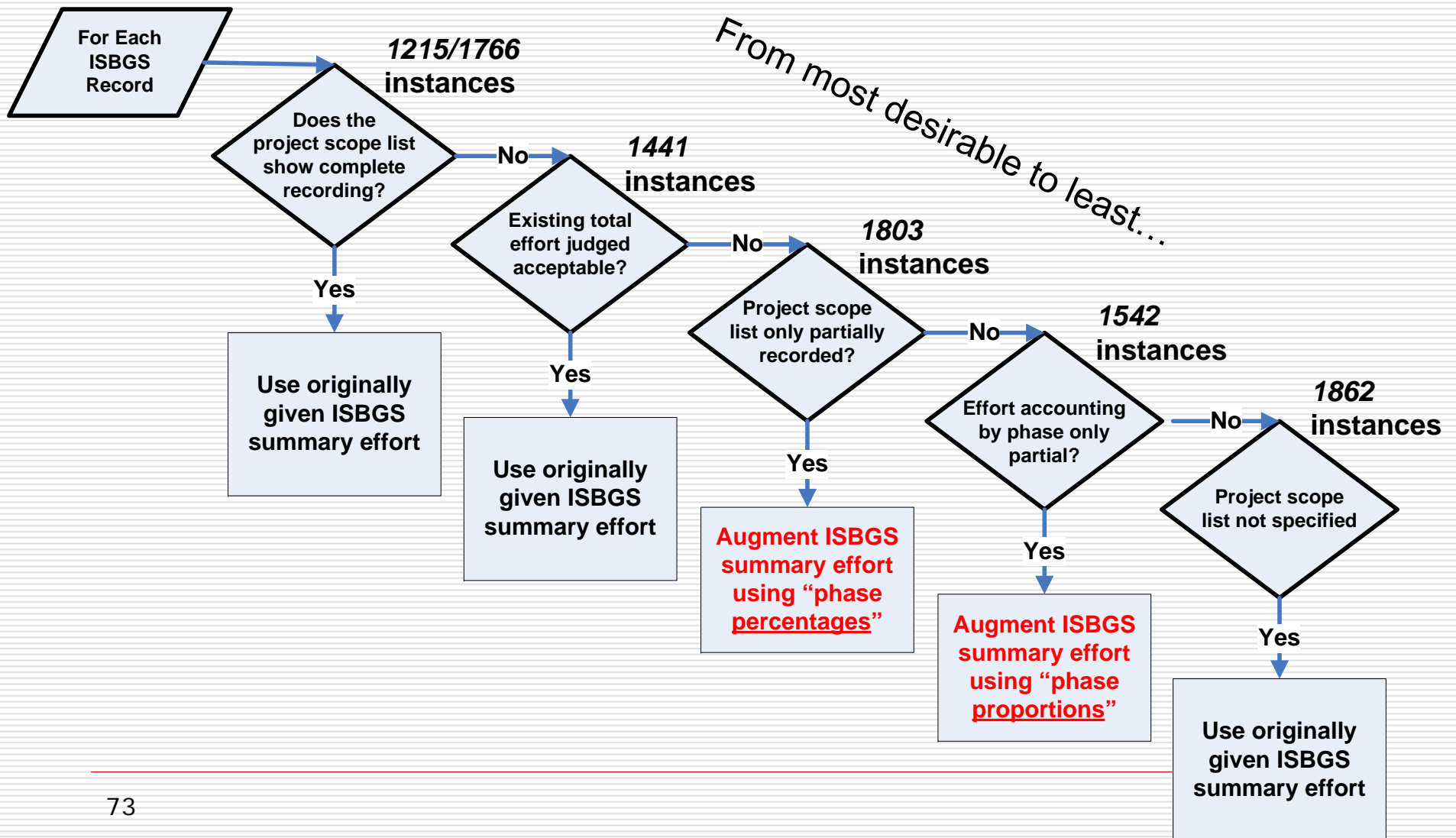


“Project Activity Scope” Normalization Sub-Goal





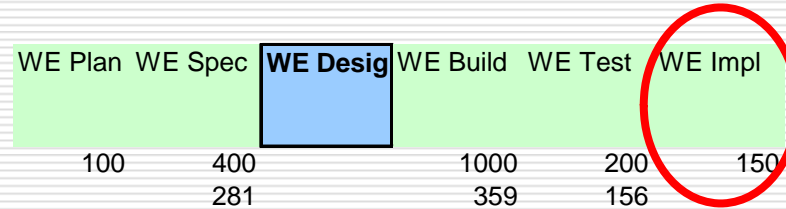
The Normalization Measure Taken Depended On The Info Available





Implementation Effort Removal Approaches

Best case: Explicitly given.



Otherwise: Use an estimated value.

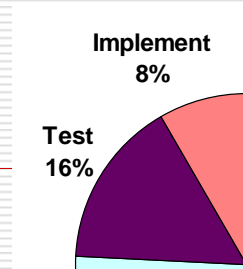
Removal of Implementation phase effort is consistent with the normalization method used:

- Proportions between phases – Implementation removed based on its historically proportionate share, given the balance of other phases.

Matrix summarizing combinations to left
Numbers below are A / B

A	WE Plan	WE Spec	WE Design	WE Build	WE Test	WE Impl
WE Plan	0.354	0.391	0.185	0.400	0.884	0.884
WE Spec		0.659	0.598	1.157	3.133	3.133
WE Design			0.271	0.932	2.688	2.688
WE Build				2.000	6.182	6.182
WE Test					2.613	2.613
WE Impl						2.613

- Percentages of each phase - relative to percentage given for Implementation





Lessons Learned

- Plan your project carefully and adhere to the plan
- Collect data in accordance with a plan
- Develop detailed mapping scheme for data provided in an alternative format
- Automate for simplicity and effectiveness
- Ensure database management hardware and software meets users needs
- Reserve adequate funding for documentation
- Document “as-you-go”



Data Collection Lessons Learned 2

- Clear Definitions are essential: But people will not always follow the definitions even when they are provided

- The most effective data collection process is to send people the forms first, then make a personal visit and interview to capture the data
 - People will not always read the instructions

- Proprietary restrictions will be used as an excuse When people don't want to provide data.

- Prime contractors may not want to offer their subcontractor's data



Data Collection Lessons Learned 3

- ❑ Don't let a subcontractor lock up your database:
- ❑ Collect completed project actuals first:
- ❑ Qualify the data quality:



Backup Slides



Normalization

Consistency

- ❑ Never assume that you know what is in the details
- ❑ Example: You are asked to build a System Engineering (SE) CER
 - ❑ Three program are identified as basis for the CER
 - ❑ All three programs are involved with similar products and scope
 - ❑ But, the System Engineering hours from one of the programs is significantly larger than the other two

“The Devil’s in the Detail!”



Normalization

Consistency

- Digging into the details discloses that for the program in question
 - SE provided normal core competencies, plus
 - A special white paper study
- White paper hours are separated from the core competencies hour to normalize the hours to the same scope base as the other two programs
- Systems En scope of work and resulting hours are now consistent across all three programs



Normalization Examples

- Operating environment
- Consistency in counting lines of code, function points, or objects
- Consistent Scope
- Anomalies
 - Historical cost data should be adjusted for anomalies when it is not reasonable to expect these unusual costs to be present in the new projects.
 - Adjustments and judgments used in preparing the historical data for analysis should be fully documented



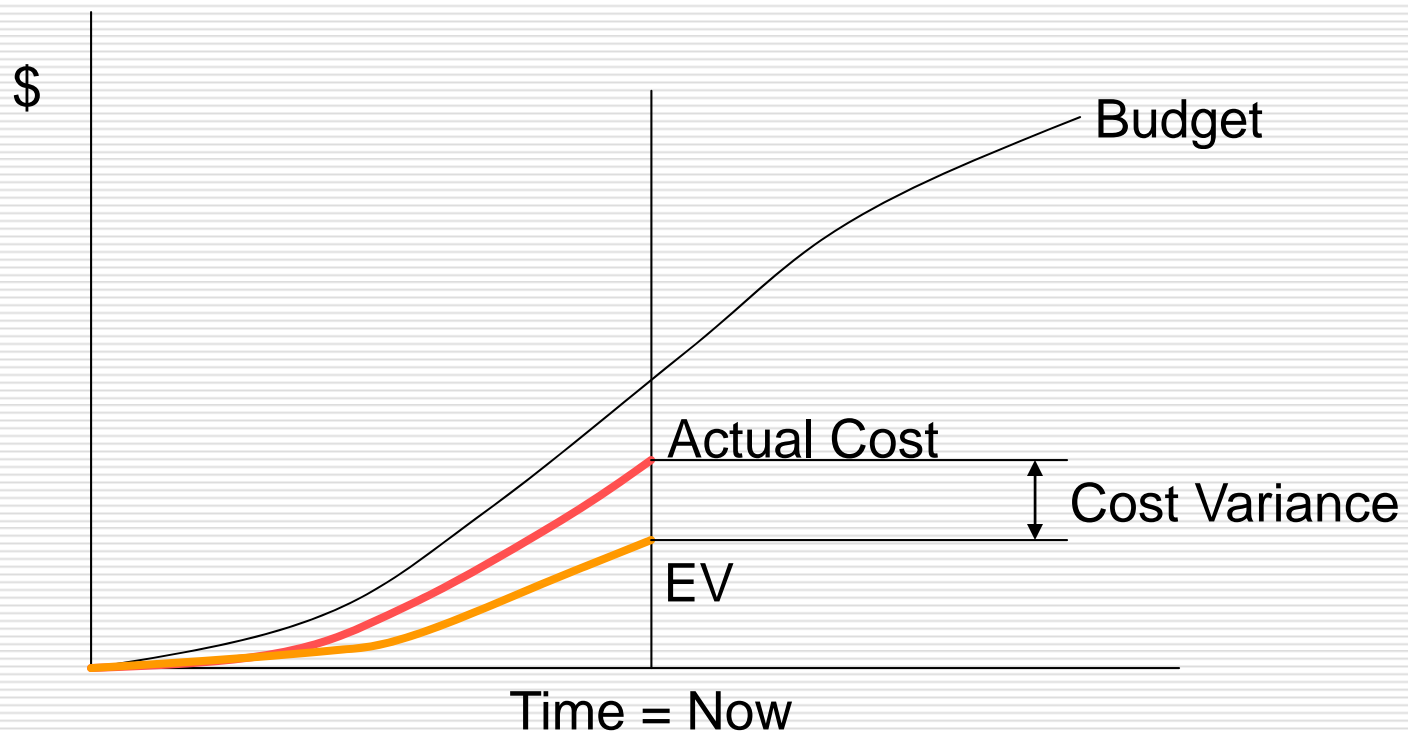
Normalization Examples

- Technology changes
 - Model inputs could be adjusted to account for certain improved technologies
 - Some analysts may deduct estimated savings from the bottom-line estimate



(4) Earned Value Reports-3

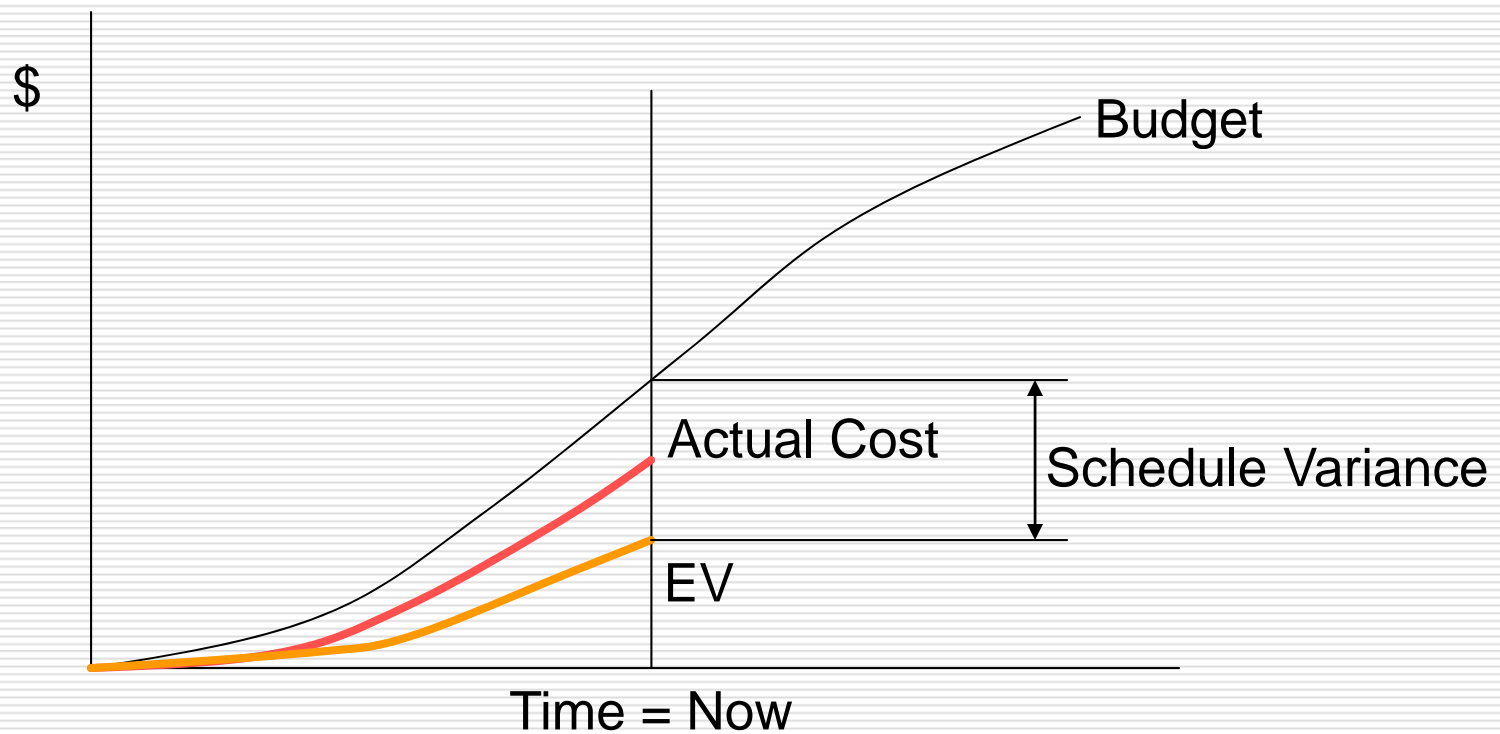
- Cost variance concept of EVM





(4) Earned Value Reports-4

- Schedule variance concept of EVM





Normalization

Economic

- Inflation
- Constant Dollars
- Then-Year Dollars
- Calendar Year vs. Fiscal Year
- Inflation Indexes
 - Historical vs. Projected
 - Simple vs. Weighted
 - Raw vs. Composite



Normalization Examples

Software languages

- Software lines of code convey two general types of information
 - Executable – causes a computer to perform certain actions
 - Declarations - used by the executable code to establish database structures, headers, and other structuring or labeling produced during the program execution
 - Comments – contains descriptive information about the executable code
- Most software models use the sum of executable and declaration code as basis of estimate
- Different languages, i.e., Fortran, C, C++, Java, etc. have different computing power per line of code.
 - The newer the language the more computing power
 - Know what language a particular model is based upon



Normalization

Quantity

- Basis: Learning (Cost Improvement) Curve
 - $Y = AX^b$
- Where:
 - A = First Unit Cost
 - X = Quantity (Cumulative)
 - b = Slope
- Selection of Normalized Quantity
(T_1 VS. T_{100} VS. T_{1000})
- Lot Size can be a factor



(b) Cost Drivers-1

- Cost driver: a parameter that tends to increase or decrease cost or is highly correlated with cost
- To qualify as a cost driver, a parameter must be a significant predictor of cost at least in the statistical sense

$$\text{Price} = 105A^{0.9} - 1200B^{-1.1}$$

More bedrooms would likely increase the price.
Multicollinearity is the likely culprit.

