Software Code Growth: A New Approach Based on Historical Analysis of Actuals

2007 ISPA-SCEA Joint Conference
New Orleans, LA

12–15 June 2007

Robert R. Jones
Paul Hardin

Technomics
The Science of Informed Decision Making

Outline

1. Objective
2. Background
3. Analytical Hypothesis
4. Data Characteristics
5. Analysis
6. Results
7. Risk
8. Converting SLOC to ESLOC
9. Summary

Objective

Develop an improved method for estimating the growth of software source lines of code (SLOC)

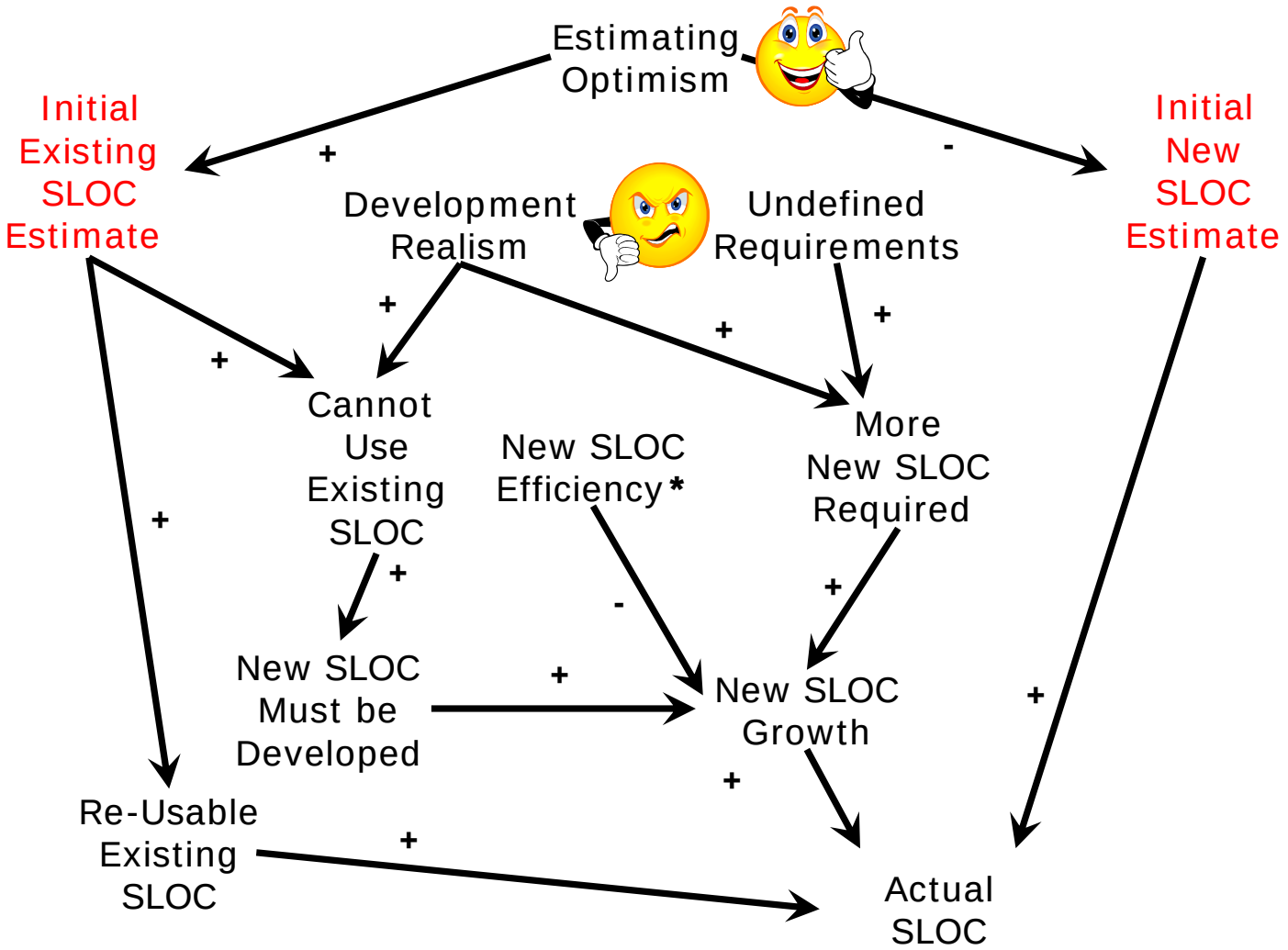
Background

Research shows that the actual SLOC and development effort is usually significantly greater than the initial estimates

We believe the difference between initial estimates and final, actual SLOC is driven by three factors

1. *Under-estimation* of the software complexity
2. *Under-estimation* of the amount of New SLOC, and
3. *Over-estimation* of the expected use of Existing SLOC, i.e., Modified and Unmodified SLOC

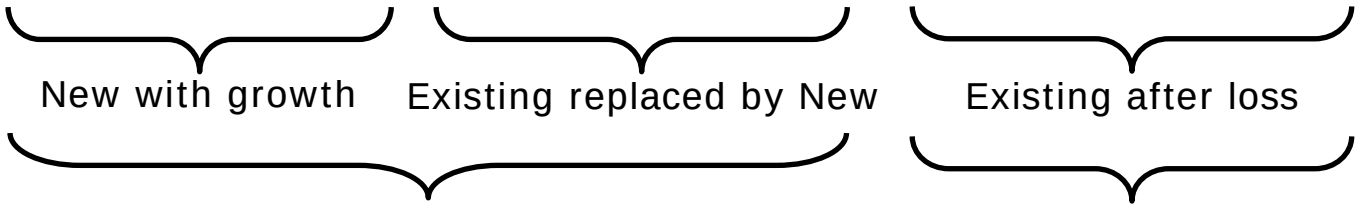
Analytical Hypothesis, 1 of 2



* Via tailoring to project-specific requirements, less New SLOC will be necessary to capture required functionality

Analytical Hypothesis, 2 of 2

$$SLOC_{w/growth} = [(1 + a) * New_{Est} + b * (c * Existing_{Est})] + [(1 - c) * Existing_{Est}]$$



where

New_{Act} $Existing_{Act}$

a = New SLOC Growth Factor

b = Coding Efficiency Factor (Efficiency of replacing un-usable Existing SLOC with New SLOC)

c = Existing SLOC Loss Factor (Fraction of Existing SLOC that is not re-usable)

(1-c) = Existing SLOC Reuse Factor (Fraction of Existing SLOC that is re-usable)

Act = Actual

Est = Initial estimate

$SLOC_{Est} = New_{Est} + Existing_{Est} = New_{Est} + (Modified_{Init} + Unmodified_{Est})$

New = newly developed SLOC
Modified = SLOC obtained from another project; reused with minor modifications
Unmodified = SLOC obtained from another project; reused without modifications; aka “Reuse” SLOC

Data Characteristics

50 Programs

11 Mission types: C2 (31); Database, Diagnostic, Mission Plans, Simulation, Utilities (5); Office Automation, Software Tools, Signal Processing (6); Operating System (3); Testing (5)

3 Development processes: Waterfall (15); Incremental (12); Spiral (8); Undefined (15)

4 Complexities: Simple, Routine, Moderate, Difficult

Values in parentheses = Number of data points

Complexity of the SLOC	Number of Data Points	Product Line	Environment	State of the Art
Simple	2	Existing	Existing	Current
Routine	10	New	Existing	Current
Moderate	14	New	New	Current
Difficult	24	New	New	New

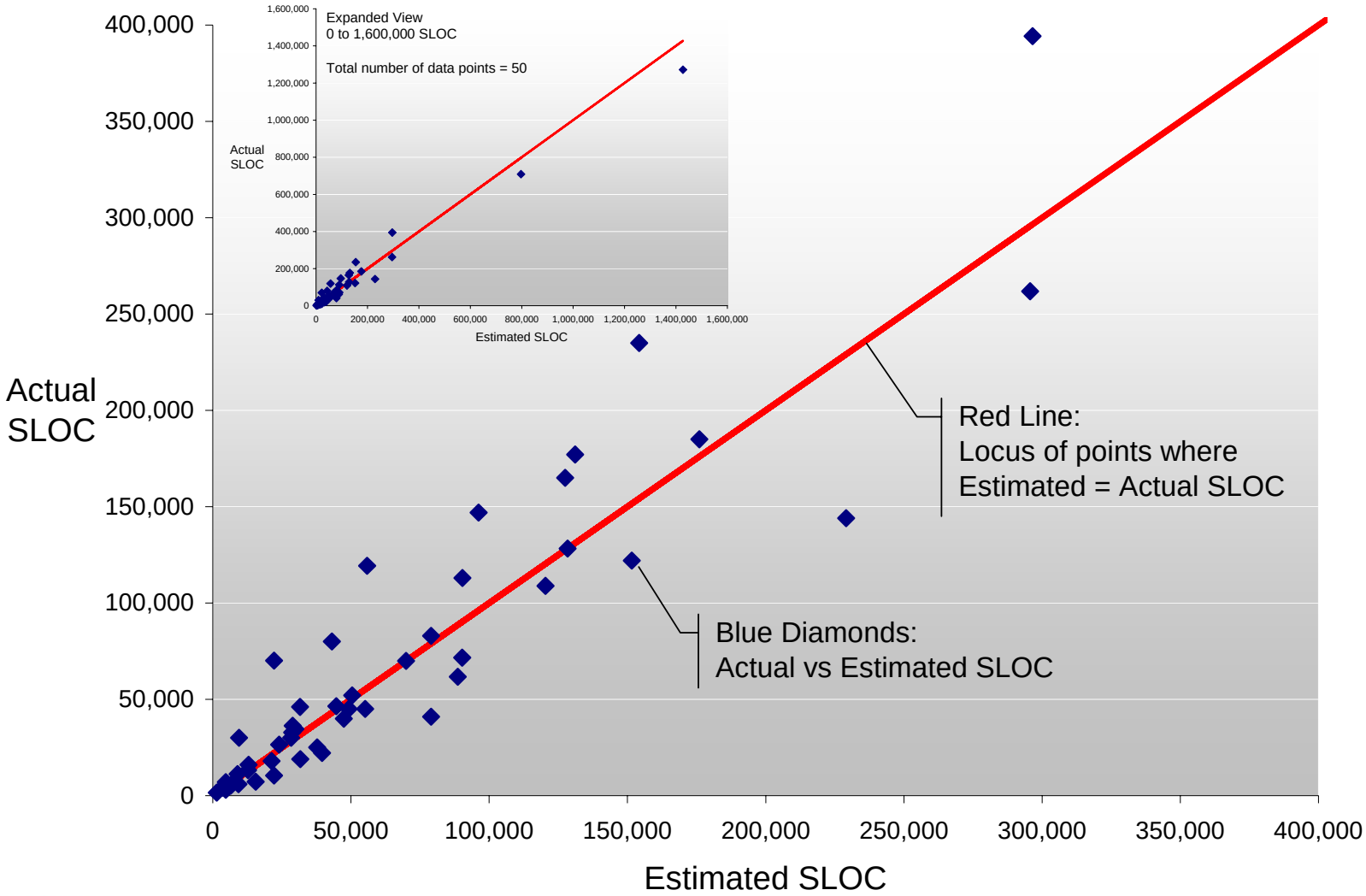
SLOC breakout:

Category	Source Lines of Code (SLOC)							
	Initial Estimate				Final Actual			
	Minimum	Maximum	Average	Median	Minimum	Maximum	Average	Median
New SLOC	3,000	517,071	94,482	20,000	3,000	519,600	129,507	33,325
Existing SLOC	0	729,201	59,643	1,055	0	752,500	59,721	0
Total SLOC	3,000	1,246,272	154,125	30,227	3,000	1,272,200	189,234	45,034
% New SLOC	25.7%	100.0%	86.6%	90.0%	33.7%	100.0%	87.4%	100.0%

Existing SLOC = Modified SLOC + Unmodified SLOC

Analysis

The equation parameters, a, b, and c, are estimated using MS Excel Solver and residual-minimization techniques



Results, 1 of 3

The equation parameter values are

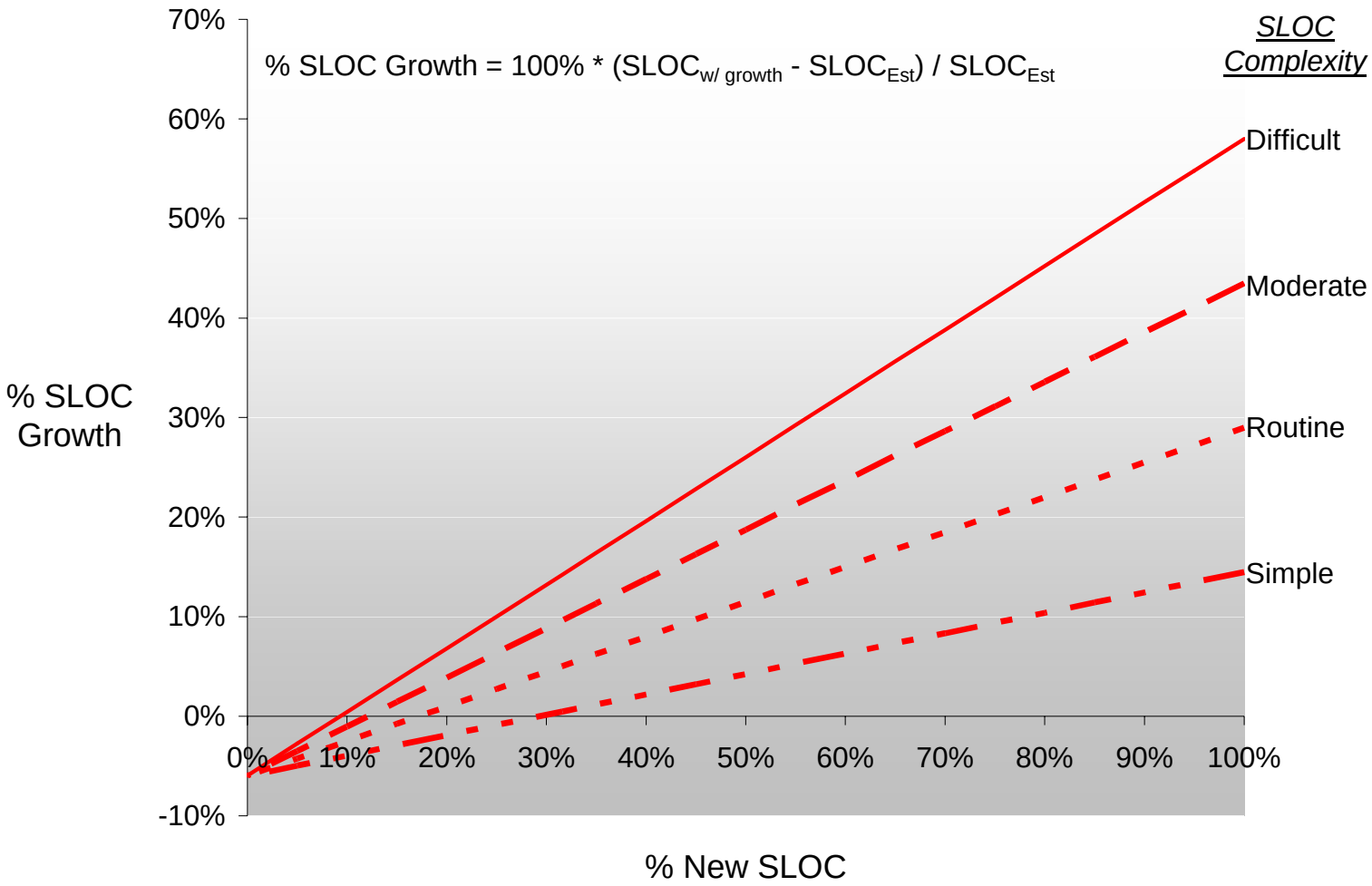
a = New SLOC Growth Factor
= 0.145 if the complexity of the SLOC = Simple
= 0.290 if the complexity of the SLOC = Routine
= 0.435 if the complexity of the SLOC = Moderate
= 0.580 if the complexity of the SLOC = Difficult
b = Coding Efficiency Factor = 0.70
c = Existing SLOC Loss Factor = 0.20
(1 - c) = Existing SLOC Reuse Factor = (1 - 0.20) = 0.80

Equation (1) below estimates the median value for SLOC with growth included

$$SLOC_{w/growth} = [(1 + a) * New_{Est} + 0.70 * (0.20 * Existing_{Est})] + [0.80 * Existing_{Est}]$$

Results, 2 of 3

SLOC growth can be positive or negative…it depends on the % New SLOC and complexity of the SLOC



Results, 3 of 3

1. New SLOC
   a. Growth is always positive
   b. Growth has two components
      1. New SLOC growth
      2. Replacement of Existing SLOC with New SLOC
   c. Growth is dependent on the complexity of the SLOC, i.e., Difficult, Moderate, Routine, or Simple
   d. Growth ranges from 14.5% to 58.0%

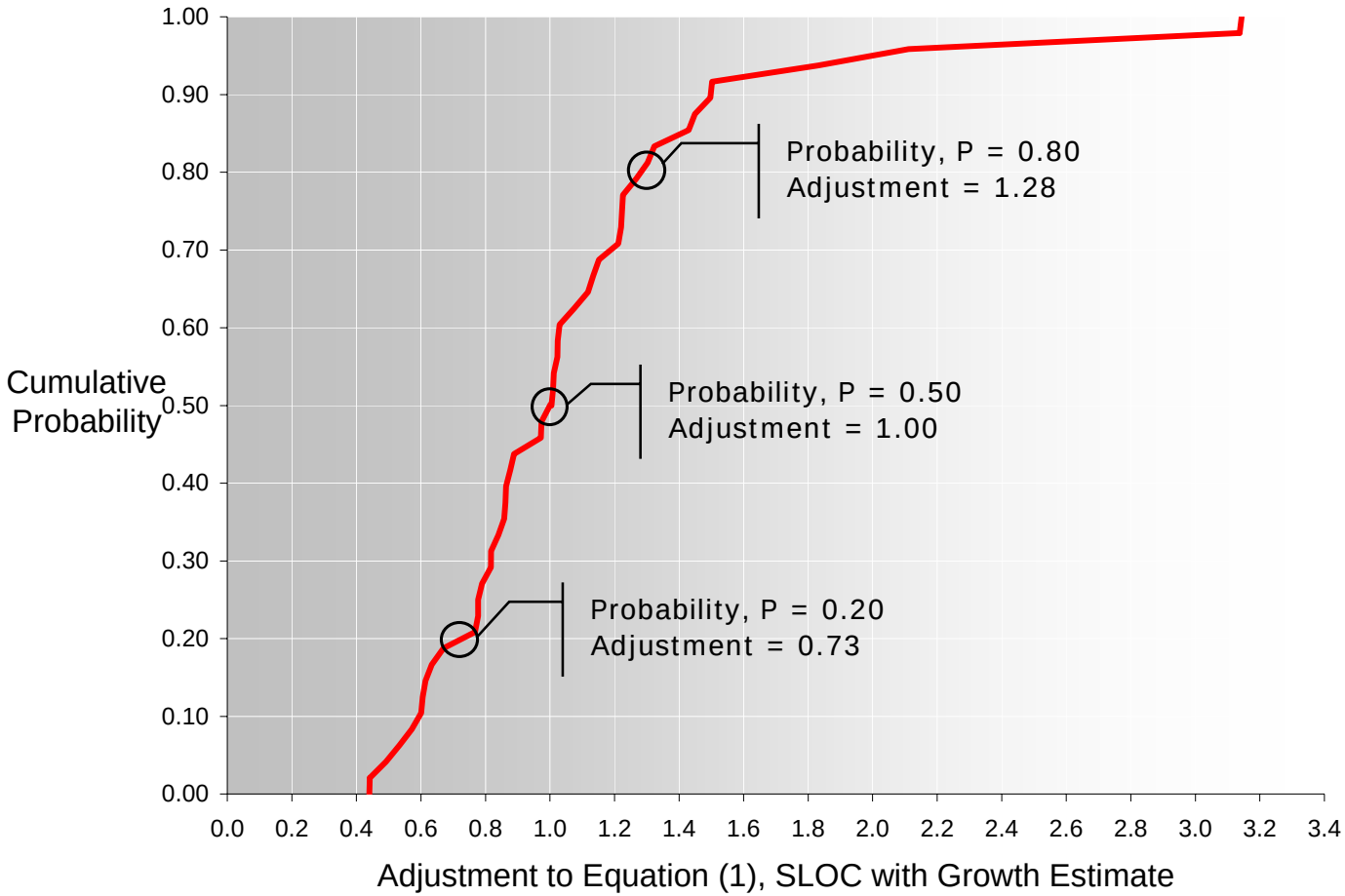
2. Existing SLOC
   a. Reusability is 20% less than expected
   b. Non-reusable Existing SLOC is replaced by New SLOC using 30% fewer lines of code

3. Total SLOC: Growth range is -6% to 58%

Risk Analysis

Via analysis of the residuals, risk-based adjustments can be made to the SLOC with growth estimate

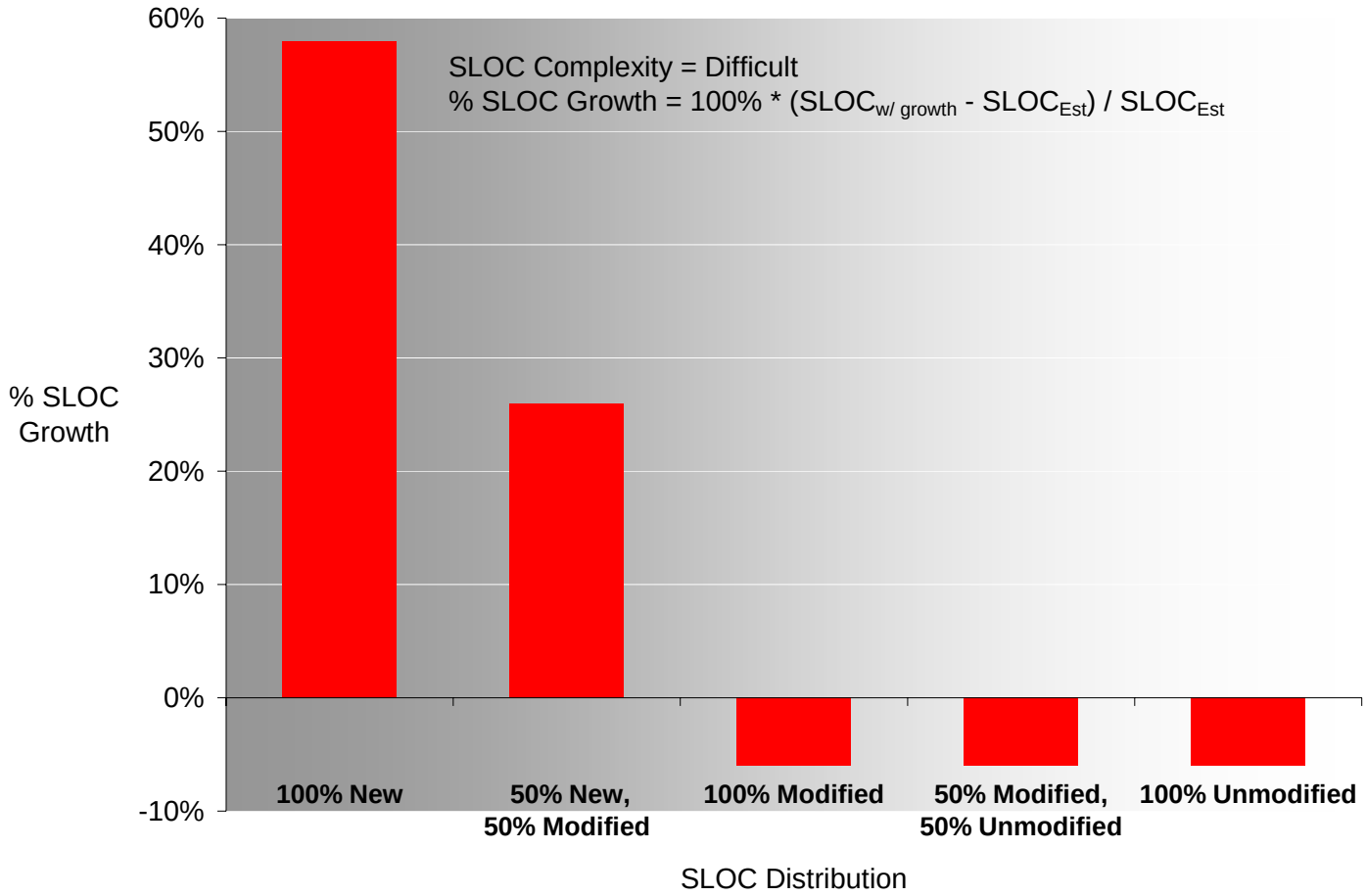
For example: P = 0.80, SLOC with growth <= 1.28 * Equation (1)



Converting SLOC to ESLOC, 1 of 2

Remember, for SLOC…

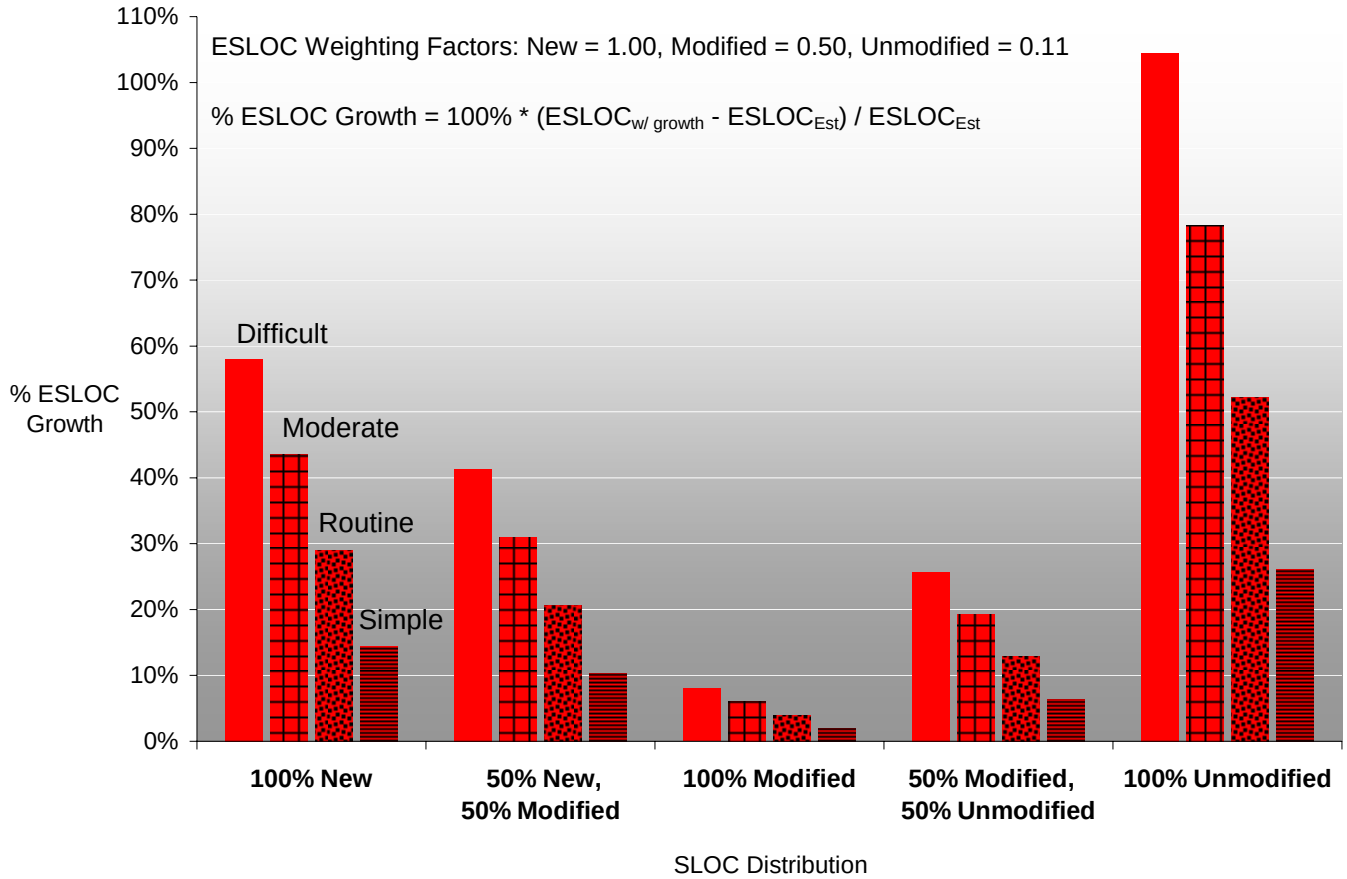
- Growth can be positive or negative depending on the % New SLOC and the complexity of the SLOC



Converting SLOC to ESLOC, 2 of 2

ESLOC and effort growth will always be positive due to

1. New SLOC growth is always positive
2. Replacement of Existing SLOC with New SLOC
3. New SLOC weighting factor > Modified SLOC weighting factor
4. New SLOC weighting factor >> Unmodified SLOC weighting factor



Summary

A. Analysis conclusions

1. SLOC growth is driven by
   a. the complexity of the SLOC
   b. the distribution of New, Modified, and Unmodified SLOC
2. New SLOC is usually under-estimated
3. Re-usability of Existing SLOC is usually over-estimated
4. SLOC growth can be positive or negative
5. ESLOC and effort growth are always positive

B. SLOC with growth and % SLOC growth can be estimated using Equation (1)

$$SLOC_{w/growth} = [(1 + a) * New_{Est} + 0.70 * (0.20 * Existing_{Est})] + [0.80 * Existing_{Est}]$$

$$\% \text{ SLOC Growth} = 100\% * (SLOC_{w/growth} - SLOC_{Est}) / SLOC_{Est}$$

C. Probabilities can be assigned to the SLOC with growth estimate (Slide 12)

D. Using weighting factors, a SLOC with growth estimate can easily be transformed into % ESLOC or % effort growth (Slide 14)

1. Growth will always be positive
2. Initial estimates with 100% Unmodified SLOC are the highest-risk (growth) option