Measurement & Analysis in the Real World
Tools for Cleaning Messy Data
Will Hayes – SEI
Robert Stoddard – SEI
Rhonda Brown – SEI
Software Solutions Conference 2015
November 16–18, 2015
Copyright 2015 Carnegie Mellon University

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Department of Defense.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN “AS-IS” BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

DM-0003055
Agenda

Introduction
Matching Information Needs
Getting to the Data
Cumulative Flow Diagrams
Tool Demonstration
Predictive Modeling
Complementary but Different Focus

Government Program Office
- Assess forecasted risk
- Manage to outcomes
- Responsible for total cost of ownership (and current cost)
- Obliged to seek out and communicate user needs

Development Contractor
- Predict performance
- Control performance drivers
- Responsible for meeting current commitments
- Subject to re-direction based on user needs

Strive to avoid directing the contractor on HOW to work

Influence on WHAT to build may be constrained by contract
Different Audiences for Metrics and Status

Program office personnel who interact directly with contractors
- Generally need insight at a finer level of detail
- Must maintain visibility/continuity over time

Stakeholders in the program, beyond Program Management
- May focus on specific topics to the exclusion of all else
- Participate in less frequent status discussions, perhaps

Senior leadership who oversee the program office
- Focus on performance of the program, not just this contract
- Frame of reference may be broader and more long term
Matching Information with Needs

Re-Casting Metrics for the Target Audience
Time-Horizon and Specificity

The chart below shows the trend in estimated size, with thresholds for potential corrective action.

The simplified version above shows only 9 weeks, focusing only on variance from original estimate.

Choose time-horizon and specificity to meet audience needs
Converging Indicators

Some information is visible only when you combine data
Useful Graphical Tool

Cumulative Flow Diagram
Constructing a Cumulative Flow Diagram

Here we have a Pie Chart showing the status of 30 defects across the four stages of the defect handling life-cycle.

This is a snapshot for a single point in time.
Constructing a Cumulative Flow Diagram

Same data, but presented in a stacked column chart

For a single point in time.
Constructing a Cumulative Flow Diagram

... adding the next 7 times
Constructing a Cumulative Flow Diagram

… now we are looking at the flow from “identified”… to “Closed”,…
This view starts to show patterns a little easier…
Tell-Tale Signals

- Backlog
- Cycle Time

Legend:
- Red: Waiting
- Green: In Process
- Blue: Done
Exercise: What is Going on Here?
Exercise: What MIGHT BE Happening

At time 2, and then again at time 4, the number of items “In Process” goes to zero.

- Have we lost the resource(s) that were preparing the items in the “Waiting” state?
- Is this intentional, due to limited resource(s) who can work on items in the “In Process” state?
Exercise: What *MIGHT BE* Happening?

The number of items that are “In Process” is growing over time.

- The rate at which things enter “In Process” is greater than the rate at which things leave “In Process.”
- Are people moving onto new items without completing their work?
- Are new resources being added, who start new work at each time period?
- Are things moving into the “Done” state quickly enough?
Getting to the Data

Mining a Configuration Management Database

or Application Lifecycle Management Tool
Activity Flow: Mining the Database

Weekly analysis activities comprised of these steps:

• Data pulled directly from Configuration Management system
• Inventory ‘change records’ to verify completeness and accuracy
• Tabulate by pre-defined time intervals and validate totals
• Chart data using Cumulative Flow Diagrams to analyze progress
Details: Process View

Process flow for a defect being worked:
- Entry/exit criteria for each step
- Individual assigned to work each one
- Progress through the process tracked
- Database fields used to record:
  - Current state in the process
  - History of progression through the states
  - Date/time stamp for each state change
  - … and lots of other information
### Details: Raw Data

#### Main Data Table

<table>
<thead>
<tr>
<th>Defect ID</th>
<th>Title</th>
<th>Description</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000001</td>
<td>Dropped data…</td>
<td>Message traffic is overwritten when buffer size not specified in…</td>
<td>1</td>
</tr>
<tr>
<td>1000002</td>
<td>Missing header…</td>
<td>File never read at initialization due to missing pointer in…</td>
<td>2</td>
</tr>
<tr>
<td>1000003</td>
<td>Unpredictable close…</td>
<td>Process XYZ terminates while opening file…</td>
<td>1</td>
</tr>
</tbody>
</table>

#### Change Auditing Table

<table>
<thead>
<tr>
<th>ID</th>
<th>Old State</th>
<th>New State</th>
<th>TimeStamp</th>
<th>… LOTS of other data …</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000001</td>
<td>New</td>
<td>Open</td>
<td>mm/dd/yy hh:mm:ss</td>
<td>- - -</td>
</tr>
<tr>
<td>1000001</td>
<td>Open</td>
<td>Assign</td>
<td>mm/dd/yy hh:mm:ss</td>
<td>- - -</td>
</tr>
<tr>
<td>1000001</td>
<td>Assign</td>
<td>Test</td>
<td>mm/dd/yy hh:mm:ss</td>
<td>- - -</td>
</tr>
<tr>
<td>1000002</td>
<td>New</td>
<td>Open</td>
<td>mm/dd/yy hh:mm:ss</td>
<td>- - -</td>
</tr>
<tr>
<td>- - -</td>
<td>- - -</td>
<td>- - -</td>
<td>- - -</td>
<td>- - -</td>
</tr>
</tbody>
</table>
Details: Mining the Change Auditing Table

This database table provides:

- Date and time when each item entered a given state
- History of all such transitions since the record was created

Using that information, we can derive:

- How many records are in each state at a given time
- How long each item stayed in any particular state

This allows us to:

- Draw Cumulative Flow Diagrams to show flow
- Model the state-transition activity with a predictive model
Leveraging Excel and Access with VB
Tool Demonstration
Useful Statistical Tool

Predictive Modeling
Building Models

Data derived from time stamps
- Duration associated with each state in the sequence
- Information about range of time seen in the past
- Benchmarks for durations can aid in planning

A variety of modeling techniques can be applied
Predicting Change Request Closure

Although 80% of closures occurred by Day 200 on the prior release, we will need 430 days to close 80% of changes on the current release!
Predicting Remaining Changes to Close

Although only 5% of closures remained by Day 460 of the prior release, we will need 1,375 Days to reach 5% of closures remaining on the current release!
Time in State Compared to Past Release

For each state that change requests may be in, we can compare time in state to a previous release, identifying if change requests are unexpectedly lingering longer in states than they should.
Tracking Software Quality Trends

Using results of software inspections to track trends in appraising software quality
Using Discrete Event Simulation, we can create simulations of the flow of software change requests, and Conduct what-if analysis of various strategies to work off the change requests including staff assignments.
Contact Information

Presenters
Will Hayes
Client Technical Solutions Division
Telephone: +1 412.268.6398
Email: wh@sei.cmu.edu

Robert Stoddard
Software Engineering
Measurement and Analysis
Telephone: + 1 412.268.1121
Email: rws@sei.cmu.edu

Rhonda Brown
Software Engineering
Measurement and Analysis
Telephone: + 1 412.268.3963
Email: rbrown@sei.cmu.edu