ISSUES in DATA COLLECTION

Barry Boehm, TRW
May 1986
CONCLUSIONS

- Data collection can be very cost effective
- Need care to make it cost effective
  - defining data items
  - asking for data
  - collecting data
  - conditioning data
SIZING DATA DEFINITIONS

- COBOL DSI : 1/3 x data declarations

- ADA DSI : lines of semicolons or...
  - multiline statements
  - array aggregates

- Function Points : counting rules
  \[
  \text{FP} = 4 \times \text{no. inputs} + 5 \times \text{no. outputs} + 4 \times \text{no. inquiries} + 10 \times \text{no. master files} + 7 \times \text{no. interfaces}
  \]
How Many "Inputs" in 10x8 Matrix?

<table>
<thead>
<tr>
<th>Project</th>
<th>Product</th>
<th>Computer</th>
<th>Personnel Attributes</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDSI</td>
<td>AAF</td>
<td>RELY</td>
<td>DATA</td>
<td>CPLX</td>
</tr>
<tr>
<td>TIME</td>
<td>VIRT</td>
<td>ACAP</td>
<td>PCAP</td>
<td>VERP</td>
</tr>
<tr>
<td>LEXP</td>
<td>AI</td>
<td>MODP</td>
<td>SCED</td>
<td>EAF</td>
</tr>
<tr>
<td>MM NOM</td>
<td>MM DEV/AM</td>
<td>EDSI/MM ACT</td>
<td>EDSI</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Process</td>
<td>7,000</td>
<td>1.0</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>Opeps</td>
<td>5,000</td>
<td>1.0</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>I/O</td>
<td>10,000</td>
<td>1.0</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Total EDSI**

| 22,000 |

**Development Mode:** Original

**FIGURE 9-1** COCOMO software cost model: Component-level estimating form (CLEF)
How Many Inputs?

Different possible answers,
1 screen managed by forms management system
10 component vectors
18 attribute vectors
19 items being limit-checked
28 entities being sequenced through
180 individual inputs

Find out what works well for
data processing systems — may
be a long, difficult process of
trial and error.
ASKING FOR DATA

- Management-relevant data
- Hypothesis-driven data collection
  - hypoth:
    "COCOMO fits my organization"
- Minimal extra project effort
  - T-shaped data collection program
  - USAF/ESD-Mitre reporting metrics
MANAGEMENT-RELEVANT, HYPOTHESIS-DRIVEN DATA:
TRW DATA COLLECTION POLICY

- Identify major project milestones

- At each milestone:
  - re-do cost model inputs
  - re-run cost model
  - collect actuals corresponding to cost model outputs
  - compare results

Easier for managers to provide actuals
the "cost office" has taken over.
T-SHAPED DATA COLLECTION PROGRAM

Top-level, High-leverage Data on all projects

Extensive data collection on a few projects
1. Size
2. Personnel
3. Complexity
4. Development progress
5. Test progress
6. Computer utilization
7. Software volatility
8. Increment content
COLLECTING DATA:
SOME TRW EXPERIENCE

- Ensure management commitment
- Establish responsible collection agent
- Tie to major milestones
- Follow "Asking for data" guidelines
- Follow up
- Understanding definitions
  - 95% storage constraint
- Global vs. local context
  - what's "nominal"?
  - personal vs. project ratings
- Uniform follow-up
  - not just on model mismatches
- Consistency checks
  - average no. of people
  - comparison between components
REWORK COSTS CONCENTRATED IN A FEW HIGH-RISK ITEMS
SOFTWARE ESTIMATION ACCURACY VS. PHASE

- SIZE (PSI)
- COST ($)

RELATIVE SIZE RANGE

COMPLETED PROGRAMS

USAF/ESD PROPOSALS

CONCEPT OF OPERATION
ROTS. SPEC.
PRODUCT DESIGN SPEC.
DETAIL DESIGN SPEC.
ACCEPTED SOFTWARE

PHASES AND MILESTONES

FEASIBILITY PLANS AND RQTS
PRODUCT DESIGN
DEVELOPMENT AND TEST
CONCLUSIONS

- Data collection can be very cost effective
- Need care to make it cost effective
  - defining data items
  - asking for data
  - collecting data
  - conditioning data