



Data Mining, In Theory and In Practice

Quality and Productivity Research Conference:
From Data to Information to Decision Making
June 5, 2007

Valerie Peters
vapeter@sandia.gov
(505) 844-9490

Systems Sustainment and Readiness Technologies
Sandia National Laboratories



Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.





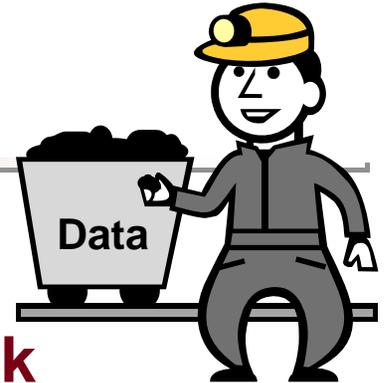
Overview



- **Introduction & Motivation**
- **Steps (for increasing real-world value)**
 - **Theory**
 - **Practice**
 - ◆ **Degree of Difficulty**
 - ◆ **Project Benefits**
 - ◆ **Organization & Data Benefits**
- **Closing & Key Takeaways**



Introduction



- **Data Mining, In Theory and In Practice**
 - **Differences between theoretical/textbook approach & real-life success**
 - **Discrepancies in**
 - ◆ **System Design vs. Data Collection**
 - ◆ **Data Collection vs. Output Use**
 - **Discrepancies cost**
 - ◆ **Money**
 - ◆ **Time**
 - ◆ **Confusion**





Experience

- **Corporate/Fortune 500**

- Intel Corporation
- Safeway
- Applied Materials

- **Government**

- Sandia National Laboratories
- State of New Mexico: Children, Youth, & Families Division

- **Other**

- Various Non-Profits
- Various Small Businesses



Common Thread: They ALL struggle with using data accurately and consistently



Motivation

- **Issue is NOT**
 - **Data Quality**
 - **Analysis Tools**
 - **Analyst Knowledge**
- **Issue is combining real raw data with analysis to create robust decisions**
 - **Somewhere, somehow, something is lost in translation**





Discrepancies

- **Design vs. Storage**
 - **Design is abstract**
 - **Data Storage is real**

- **Storage vs. Use**
 - **Data Storage is usually controlled**
 - **Data Use can be either controlled Or 'Ad-Hoc'**



Overview



- Introduction & Motivation
- **Steps (for increasing real-world value)**
 - **Theory**
 - **Practice**
 - ◆ Degree of Difficulty
 - ◆ Project Benefits
 - ◆ Organization & Data Benefits
- Closing & Key Takeaways



Step: Discuss with the Designer

● Theory

- First, Gather and Review Design Documents
- Then, Discuss Questions with the System Designer



● Practice

- Start with a Conversation
 - ◆ Ask open-ended questions about the system and its design
- Then move to serious documentation review
- Degree of Difficulty Very Low
- Project Benefits Medium
- Organization & Data Benefits Small





Step: Discuss with Other Designers

- **Theory**

- Experts from the data system you'll use are your most valuable resource



- **Practice**

- Speak with designers from related systems, too
 - ◆ “Upstream” (those providing data)
 - ◆ “Downstream” (those using the data output)

- Degree of Difficulty **Medium**
- Project Benefits **Medium**
- Organization & Data Benefits **Medium**





Step: Discuss with Data Recorders

- **Theory**

- Conversations with database designers are crucial



- **Practice**

- Speaking to those who enter the data is just as important as speaking to those who designed the data repository
 - ◆ Those entering the data know *a lot* about how the data can be interpreted
 - ◆ Also speak with those who maintain the data system

- Degree of Difficulty **Low**
- Project Benefits **Medium**
- Organization & Data Benefits **Medium**





Step: Make Suggestions

● Theory

- Unless part of your deliverables, making formal suggestions is not your job



● Practice

- Regardless of the contract or job definition, this can be one of your most valuable contributions
 - ◆ Analysis tends to have a shelf life; but good data systems can last decades

- | | |
|--------------------------------|---------------|
| – Degree of Difficulty | Medium |
| – Project Benefits | <i>Varies</i> |
| – Organization & Data Benefits | Large |





Valerie's Law of Data Maintenance

- **Those who enter the data should be the ones who gain when the data is correct**
 - **If feasible, assign data entry to those who need the data to be correct**
 - ◆ If not feasible, ensure data entry personnel get positive feedback for good data quality
 - **Implement a (quick!) data quality check into regular project updates**
 - ◆ Suggest that this data quality check is implemented as a long-term monitoring technique



Never underestimate the power of tracking something to highlight its quality level



Step: Utilize System Design

- **Theory**

- System Design and Use are aligned for most systems



- **Practice**

- Returning to system design, instead of current practices, can be quite beneficial
 - ◆ Often, system was well-designed, but isn't being used this way

- | | |
|--------------------------------|---------------|
| – Degree of Difficulty | Low |
| – Project Benefits | <i>Varies</i> |
| – Organization & Data Benefits | Large |





Step: Champion Changes

- **Theory**

- Data Improvement belongs to the data owners

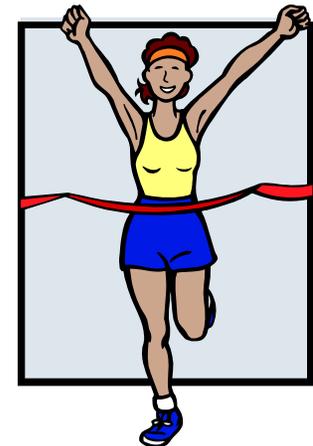


- **Practice**

- As a ‘Data Expert’, your fresh perspective is essential; champion the improvement suggestions

- ◆ Long-term Project: be accountable
- ◆ Medium-term Project: find owners
- ◆ Short-term Project: ensure continuity

- Degree of Difficulty *Varies*
- Project Benefits **Medium**
- Organization & Data Benefits **Large**





Overview



- Introduction & Motivation
- Steps (for increasing real-world value)
 - Theory
 - Practice
 - ◆ Degree of Difficulty
 - ◆ Project Benefits
 - ◆ Organization & Data Benefits
- **Closing & Key Takeaways**



Closing

- **Discuss, Discuss, Discuss**

- With the Designer
- With Other Designers
- With Data Recorders
- With Data Maintainers
- ...

- **Make Suggestions**

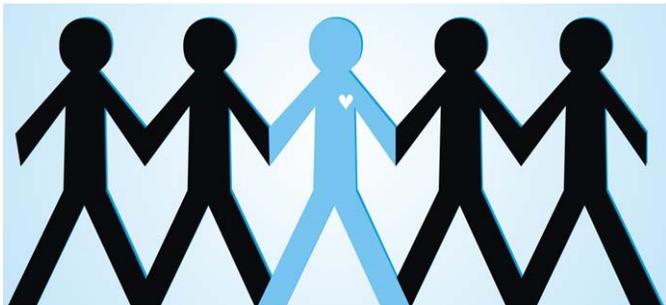
- and Follow Through to Implementation





Key Takeaways

- **Be a partner**
 - Your project's success or failure rests on your ability to glean knowledge from the experts
- **Talk to everyone involved**
 - Even if you're not sure what you're looking for
 - Ask open-ended Questions
- **Make (& Help Implement) Good Suggestions**



Many good ideas are lost or mal-implemented, when they come as directives from a party who walks away



Thank You!

Questions, Comments, ...?