Data Mining for Process Performance Model Creation

Paul Below

paul_below@qsm.com

http://www.qsm.com/

Agenda

- Process Performance Models
- Data Mining and Model Creation Challenges
- Types of Data Mining Models and Examples
- Data Mining Issues

What is a Process Performance Model?

- "A description of relationships among attributes of a process and its work products that is developed from historical process-performance data and calibrated using collected process and product or service measures from the project and that are used to predict results by following a process" (SEI)
- **Whew! Let's back up…**

Process Performance

"A measure of the actual results achieved by following a process. It is characterized by both process measures and product service measures." (SEI)

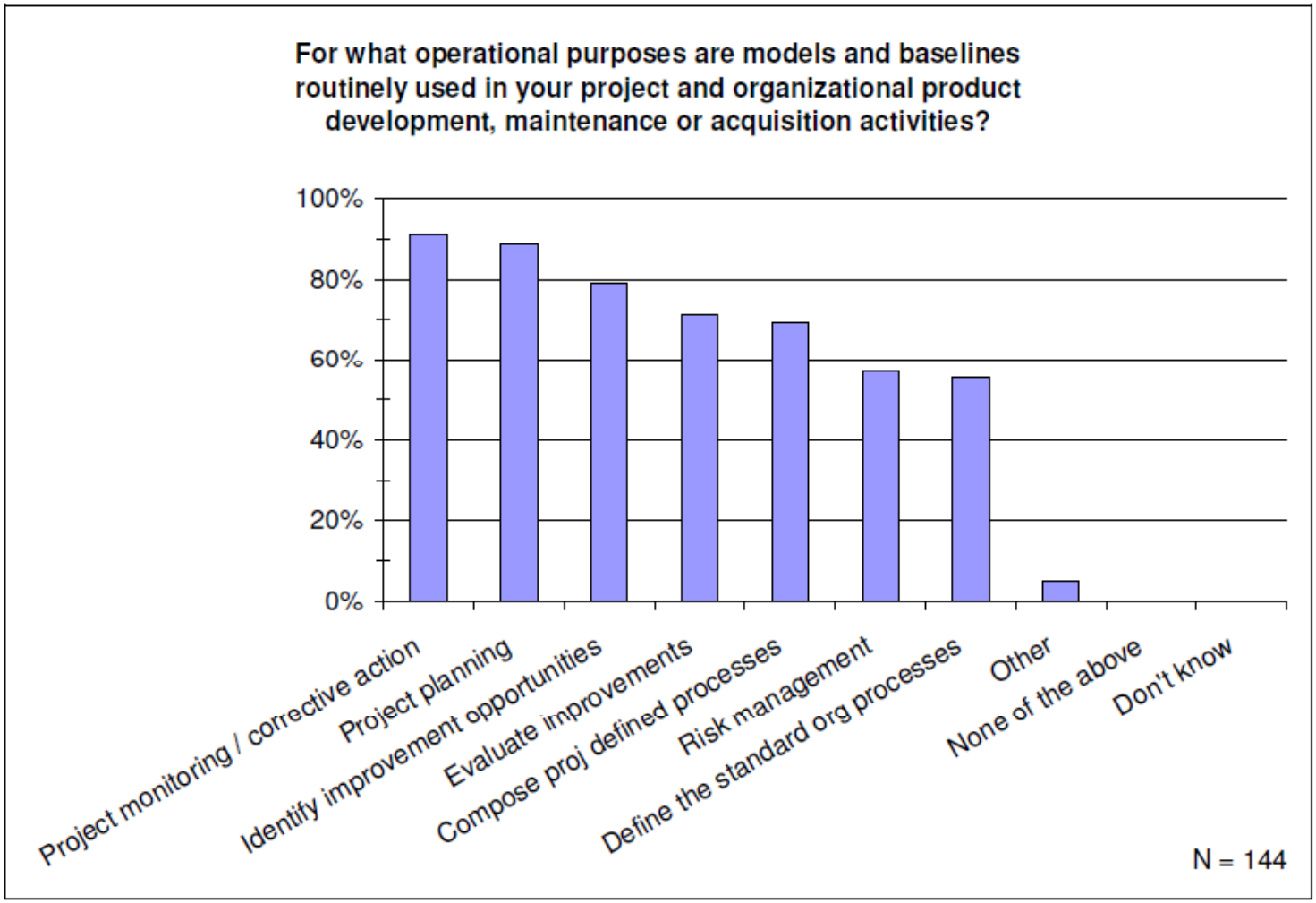
Example Process Measures: effort; cycle time; defect removal efficiency

Example Product Service Measures: reliability; defect density; response time

Process Performance Models for Prediction?

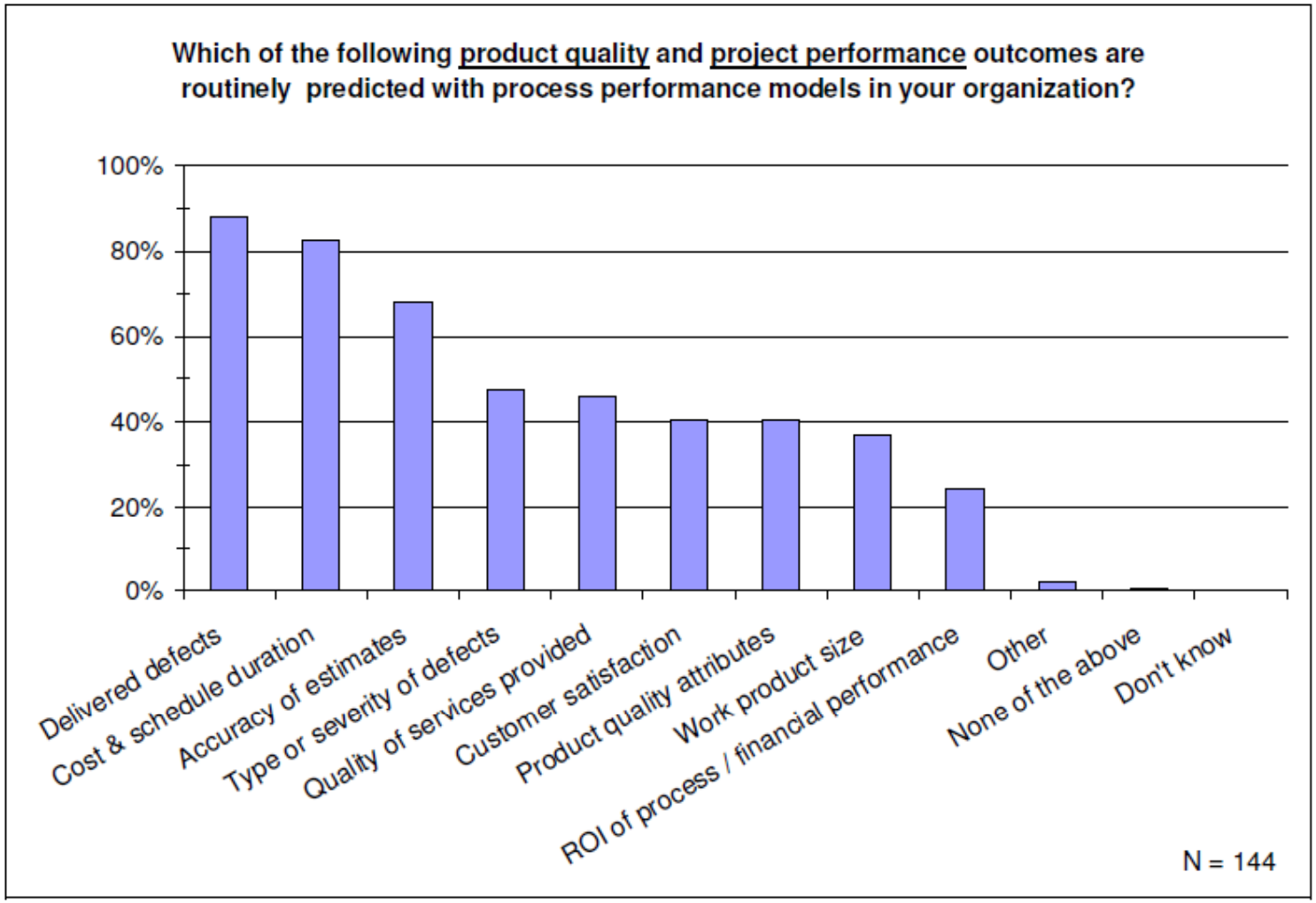
- The "healthy ingredients" of process performance models include:
  - ***Modeling uncertainty*** in the model's predictive factors
  - Ensuring that the model has ***controllable factors*** in addition to the possible uncontrollable factors
  - Identifying factors to construct models that are directly associated with sub-processes
  - ***Predicting*** final and interim project outcomes
  - Using ***confidence intervals*** to provide a range of expected outcome behaviors enabling what-if analysis using the model
  - Enabling projects to identify and implement mid-course corrections to help ensure project success

Why Model Process Performance?



From CMU/SEI-2008-TR-024

What Product/Project Output Variables are Predicted?



From CMU/SEI-2008-TR-024

Model Creation Challenge: Getting Started

- I want to predict an output variable…
- Dozens of input variables usually available, which should I use? Which are vital?
- What about colinearity?
- Which relationships do I explore first?

Data Mining Can Help

- Data mining can aid in conducting hypothesis testing or getting started with exploratory analysis
- For example, classification trees can be useful for exploratory analysis:
  - Which variables does the tool split on first?
  - Which variable does the tool think is most important?
  - What variables does it pick for the first 5 or 6?
- Some data mining techniques are supported in basic statistical tools (e.g., SPSS, SPLUS, JMP, SAS, Minitab)
- Many data mining specific tools exist in the marketplace

Thin out the forest, so we can examine the important trees

What is Data Mining?

Every book has a different definition, but the common themes are:

- Use of very large databases
- Use of automated tools and a process
- Results have to be useful

The hard thing is not figuring out which algorithm to use,

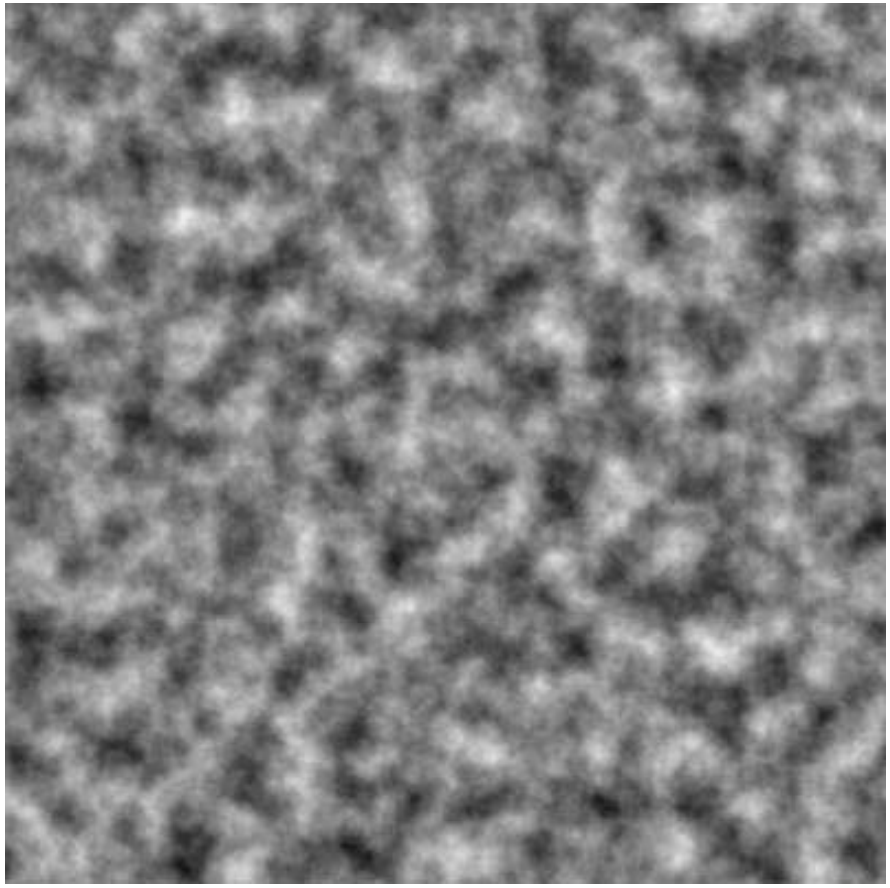
the hard thing is to figure out what to do with the results.

Data Mining Myths

- Find answers to unasked questions
- Continuously monitor your data for interesting patterns
- Eliminate the need to understand your business
- Eliminate the need to collect good data
- Eliminate the need to have good data analysis skills

“On two occasions I have been asked, 'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?' I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.” Charles Babbage

Model Creation



"Statisticians, like artists, have the bad habit of falling in love with their models."

George Box

People love to interpret noise.

Model Creation Challenges

- Databases are already built, and not designed for our purposes
- Databases were designed by committee, and everything anyone thought of is in there
- Database may have been used inconsistently by different user groups or at different times
- Data structure is often horrible, keys not appropriate

The model is no better than the data

Types of Data Mining Models

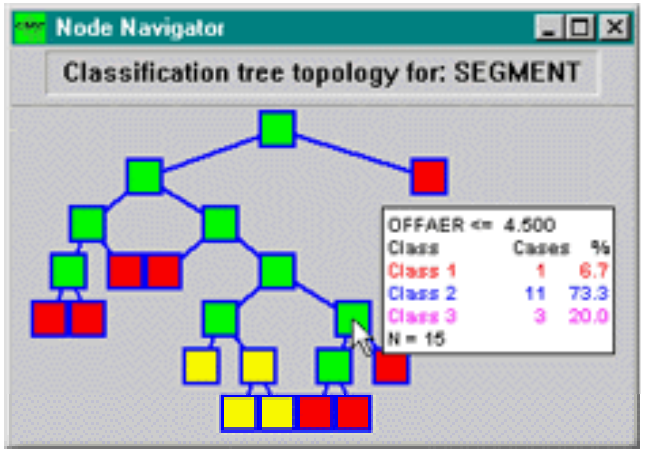
Category	Description	Purpose	Primary Data Type
Classification	Split the data to form homogenous subsets	Predict response variable	Discrete is best
Regression	Best fit to estimating model	Predict response variable	Continuous (ratio or interval)
Clustering	Group cases that are similar based on selected variables	Identify homogeneous groups of cases	Any
Association	Group variables that are similar	Determine colinearity, identify factors that explain correlations	Ratio or interval (not categorical)

Types of Models

- Some data mining techniques are black box, others are white box
- Black box is used for prediction (examples are neural networks and k nearest neighbors)
- White box is used for interpretation (classification trees and regression are examples)
- Users generally dislike black box because they cannot see how the model works

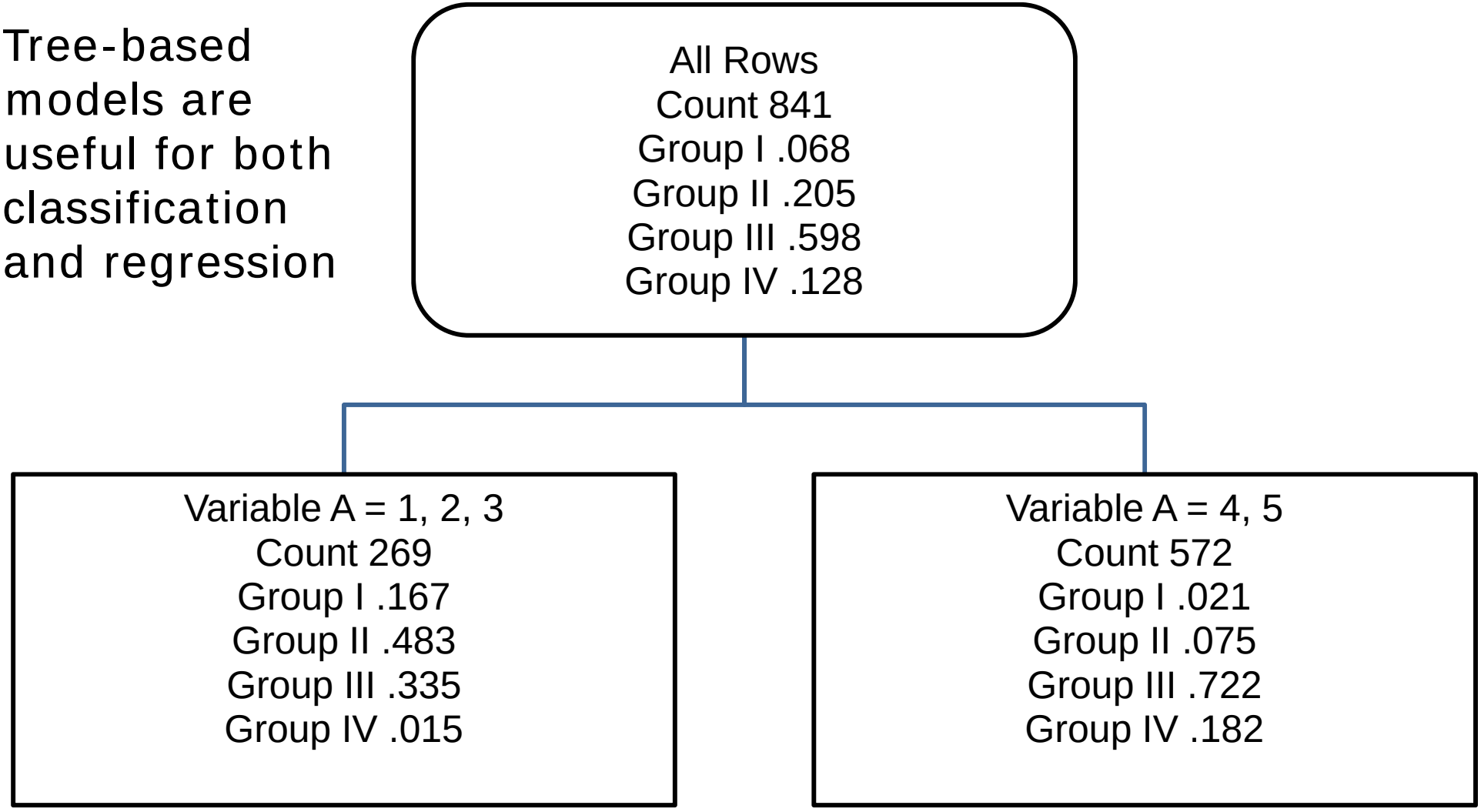
Example: Classification

- Decision trees can be explanatory tools to distinguish between objects of different classes (for example, high quality and low quality systems)
- Classification can be used to predict the class of cases
- For process improvement, trees identify the most important (vital few) factors



Example Output: Classification

Tree-based models are useful for both classification and regression



Example Output: Regression

- Stepwise Regression enters variables one by one and tests them for removal
- Good method when independent variables are correlated
- This example went through 9 steps to build a model
- There were 20 variables excluded from the final model

Model	R	R Square	Adjusted R Square	Std. Error
9	.840	.706	.691	330.332

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	47883321.563	9	5320369.063	48.757	.000
Residual	19968796.365	183	109119.106		
Total	67852117.927	192			

Predictors: (Constant), Effective SLOC, Life Duration (Months), MB Time Overrun %, MB Effort (MM), Life Peak Staff (People), Data Complexity, MBI, MB Effort %, Mgmt Eff.

Dependent Variable: Errors (SysInt-Del)

Example Output: Regression

	Unstandardized Coefficients		Sig.	95% Confidence Interval for B	
	B	Std. Error		Lower Bound	Upper Bound
(Constant)	-580.411	239.656	.016	-1053.255	-107.568
Effective SLOC	.001	.000	.000	.001	.001
Life Duration (Months)	27.633	5.832	.000	16.126	39.139
MB Time Overrun %	.026	.006	.000	.015	.037
MB Effort (MM)	1.535	.326	.000	.892	2.177
Life Peak Staff (People)	-7.438	1.905	.000	-11.197	-3.679
Data Complexity	66.840	18.269	.000	30.795	102.886
MBI	33.683	14.609	.022	4.859	62.507
MB Effort %	3.924	1.552	.012	.862	6.987
Mgmt Eff.	-50.012	22.775	.029	-94.948	-5.076

Dependent Variable: Errors detected integration through deployment

Example Output: Regression

- Stepwise regression creates a linear equation. Use graphical or other techniques to look for nonlinear relationships.
- Each of the predictors should be examined individually to learn about the relationship with the dependent variable.
- A negative number is a negative correlation.

Example: Correlation

- Independent variables can be correlated with a dependent variable
- Types of correlation include:
  - Nominal: chi-square on crosstabs
  - Ordinal: Kendall's Tau-B correlation
  - Ratio: Pearson correlation
- Thin out list of variables, identify and examine those that show significant correlation
- Remember that correlation might be non-linear
  - If so, consider transformation or non linear methods

Example: Clustering

- Clustering detects groupings in data
- Two types:
  - K-Means iteratively moves from initial to final cluster centers, used with large number of cases
  - Hierarchical finds the closest pair of objects then continues iteratively until all objects are in one cluster, results in clusters with sub clusters
- Could stratify the projects into the clusters prior to doing additional analysis. May result in the creation of multiple prediction models.

Example Output: Clustering

K-Means example output with centers for each cluster

Look for the variables with differing means

	Cluster		
	1	2	3
Project Count	5	22	166
Life Effort (MM)	750.7	617.8	89.1
Errors (SysInt-Del)	1898	1030	186
Errors First Month	138	117	8
Total FP	37167	26533	2648
Effective SLOC	1272194	298791	26444
Life Duration (Months)	21.3	18.4	9.3
Life Peak Staff (People)	56.5	61.1	15.4
Life Avg Staff (People)	23.8	26.5	7.1
MB Eff Overrun %	.0	62.0	45.8
SLOC/MB MM	2384.5	1606.4	910.9
Putnam's PI	24.4	21.5	14.1

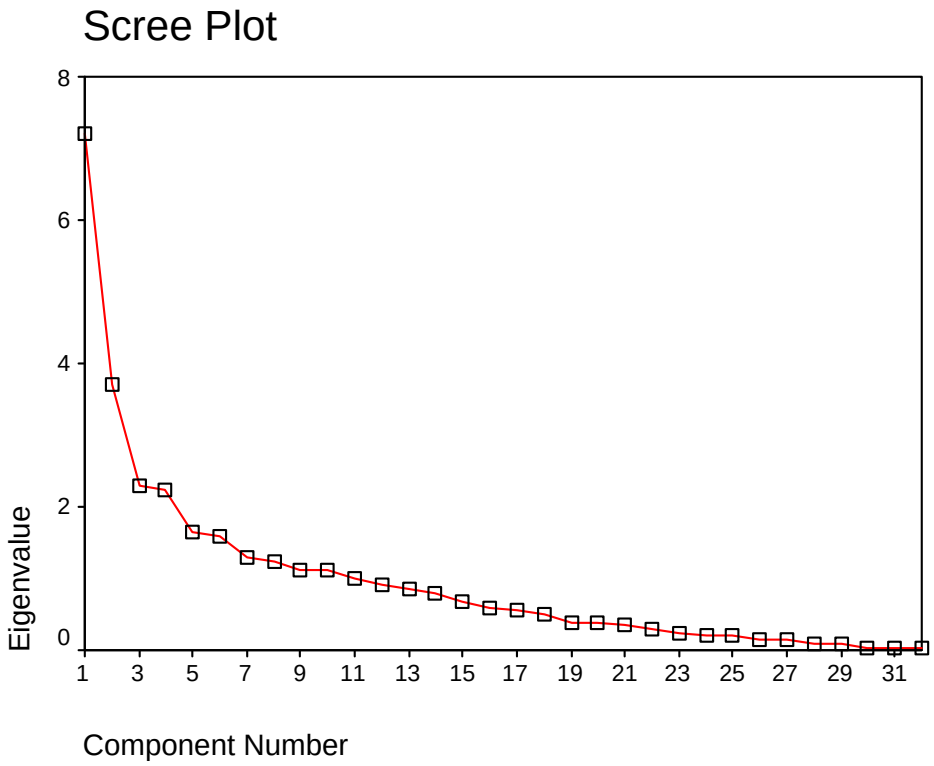
Example Output: Association

Study correlations between large number of interrelated quantitative variables by grouping the variables into factors

One type of association is principal components

Interpret each factor according to the meaning of the variables

Advantage of association is the ability to summarize many variables by a few factors



Example Output: Association

- In this example, Lifecycle effort is correlated with Component 1, not with the other 3 Components
- The table is also useful for naming the Components
- Variables that are closely related should be combined or one selected as representative when searching for root causes

	Component			
	1	2	3	4
Life Effort (MM)	.920	-.152	.196	-.006
Effective SLOC	.652	.111	-.475	.106
Life Duration (Months)	.658	-.198	-.429	.066
Life Peak Staff (People)	.865	-.115	.338	-.137
Life Avg Staff (People)	.823	-.157	.381	-.156
FUNC Effort (MM)	.880	-.169	.151	.098
MB Effort (MM)	.925	-.160	.065	-.122
Func Effort %	-.241	.088	.236	.719
MB Effort %	-.059	-.072	-.247	-.765
Knowledge	.186	.770	.161	-.076
Staff Turnover	.083	-.717	.049	.110
Dev Team Skill	.133	.746	.029	-.225
MBI	-.006	-.011	.640	-.200

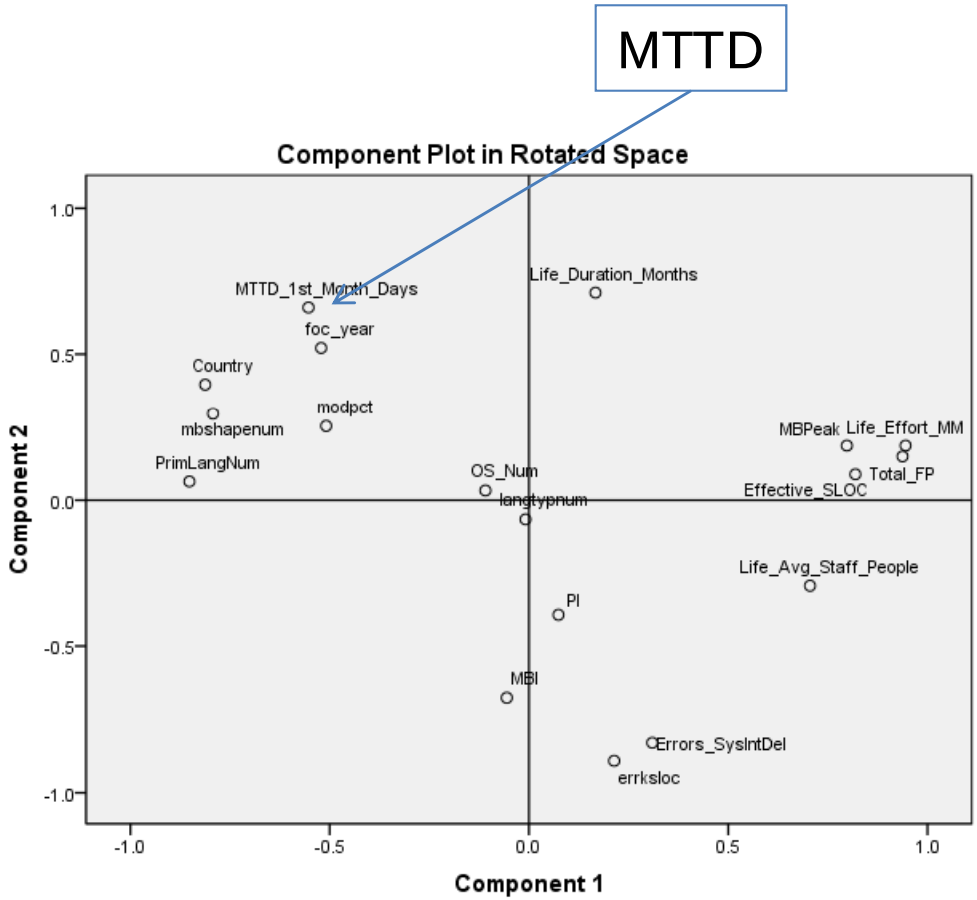
Factor Analysis Dimension Reduction: Two Component Plot

Reduction of 18 key variables into two factors accounted for 58% of variance in the 11 projects.

MTTD most closely associated with Year, next with Country and the shape of the staffing curve and the percent of modified code.

This quality relationship is verified on the next slide with a three factor analysis.

Improvements in quality at XYZ are likely to come through examination and improvements of processes used within and between Countries as well as staffing processes and plans



Factor Analysis: Three Factor Matrix

Component 1 (30% of variance): Size, Primary Language, Effort, Peak Staff

Component 2 (27% of variance): **Errors, MTTD**, Country, Year, Staffing Curve Shape

Component 3 (17% of variance): Duration, PI, MBI

Rotated Component Matrix[a]

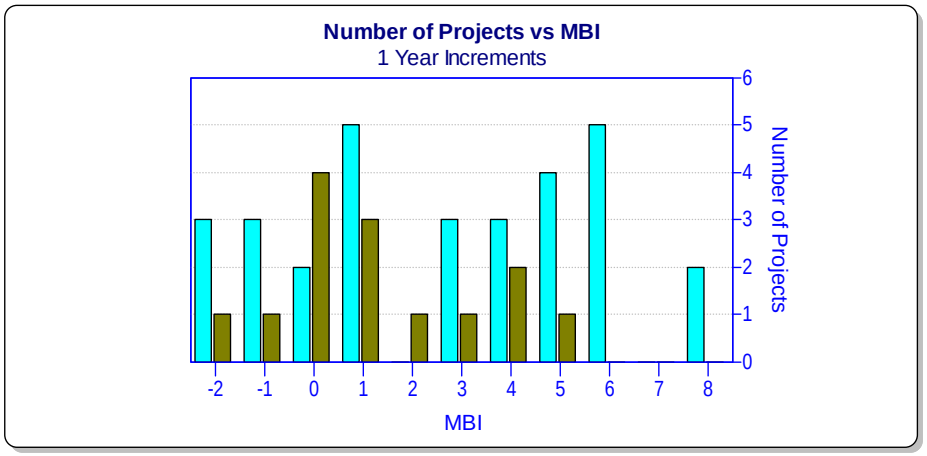
	Component		
	1	2	3
ESLOC	.917	-.006	.239
Total FP	.967	-.121	-.012
Primary Language	-.792	.329	.003
Life Duration (mths)	.154	.220	-.893
Life Effort (mm)	.922	-.194	-.202
Life Avg Staff (People)	.675	-.335	.345
Errors (Sys Int thru Del)	.000	-.902	.204
MTTD 1st Mth (Days)	-.283	.831	-.126
PI	.235	.086	.933
MBI	-.038	-.226	.876
Country	-.596	.702	.025
FOC Year	-.242	.782	.071
OS	-.037	.156	.146
MB Peak Staff	.859	-.021	.017
MB Staffing Shape	-.546	.708	.235
Mod SLOC Pct	-.331	.511	.126
Errors per KESLOC	-.120	-.943	.205
Language Type	-.223	-.369	-.507

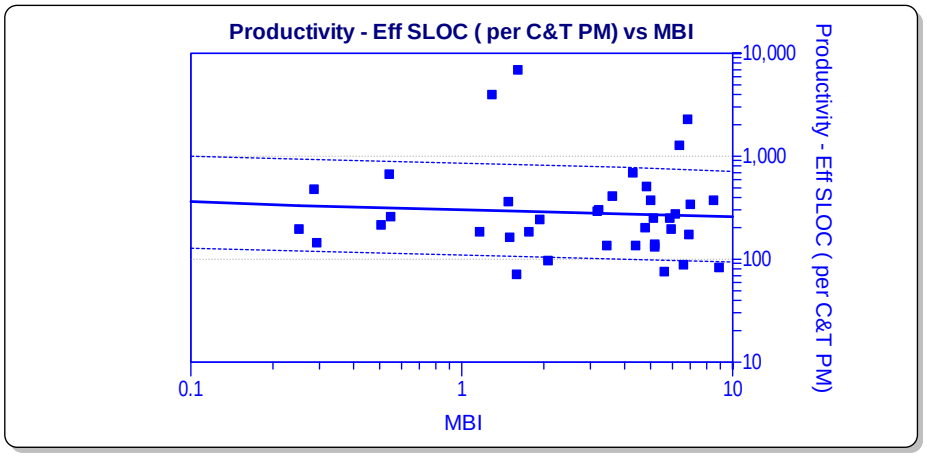
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

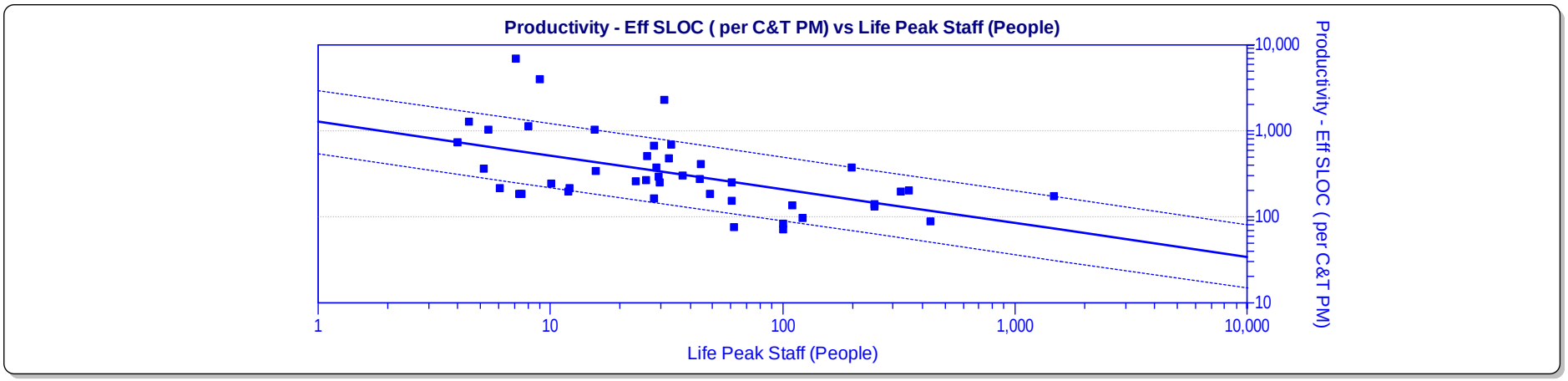
a. Rotation converged in 5 iterations.

Counterexample: MBI and Peak Staff

MBI Trends







Projects being Assessed | Engineering | Real Time | Avg. Line Style | 1 Sigma Line Style

Causation: Post Hoc

Will the estimation model continue to work?

Retrospective studies (in absence of DOE) must meet these criteria to make a good case for causality:

- Association
- Temporal Priority
- Non-spuriousness
- Theoretical Adequacy

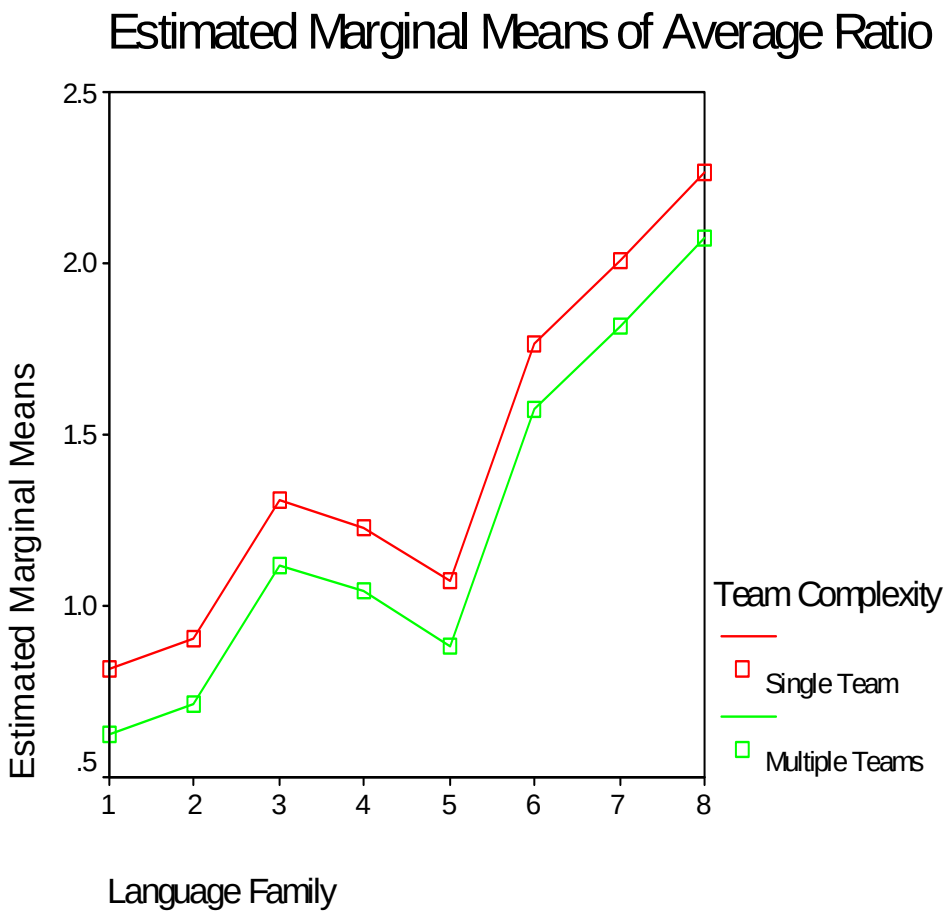
There are two clocks that keep perfect time.
When “a” points to the hour, “b” strikes.
Did “a” cause “b” to strike?

Causation: Confounding Factors

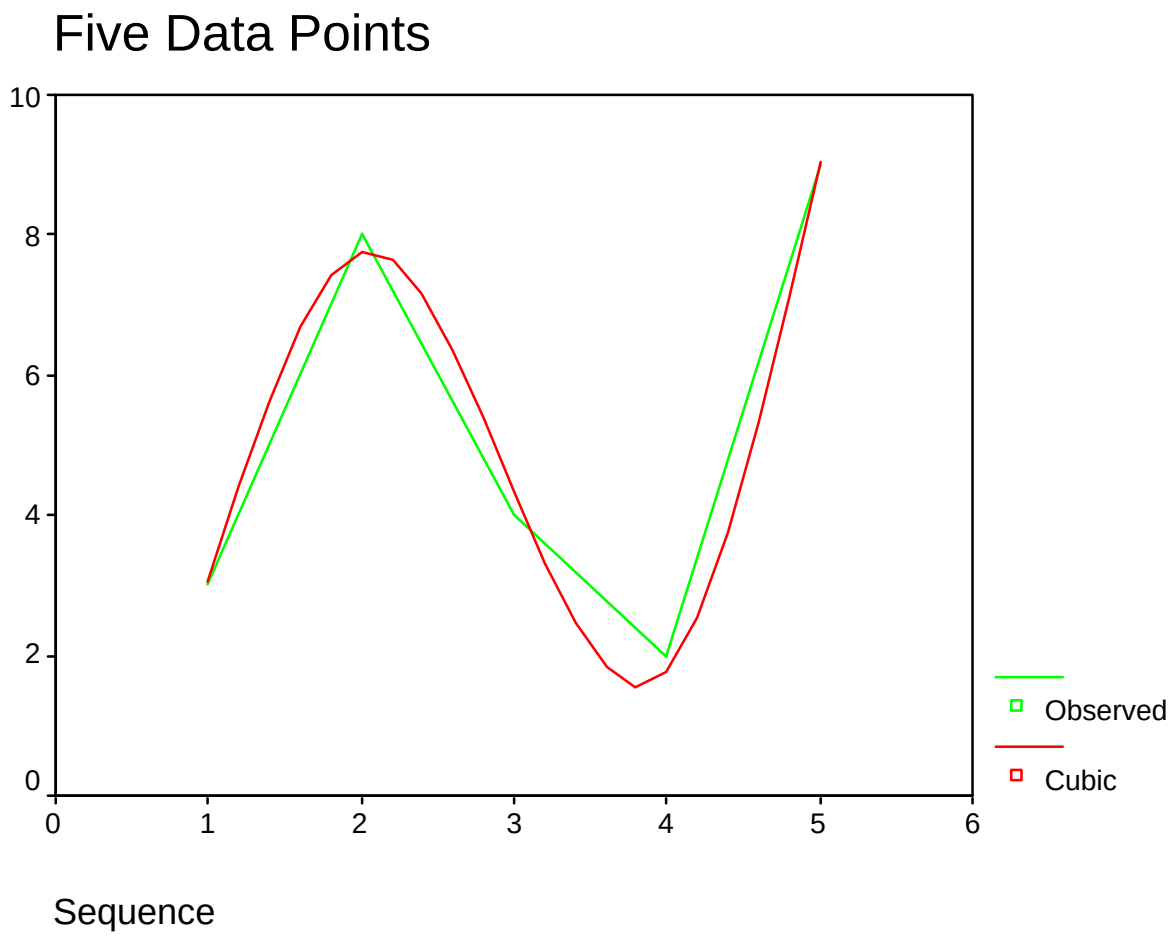
Apparent causation could be due to:

- A third factor, correlated to the supposed cause
- Interaction between two or more factors (higher order effects)

Therefore, potential confounding factors must be investigated



Data Mining Issue: Overfitting



Data Mining Issue: the Data

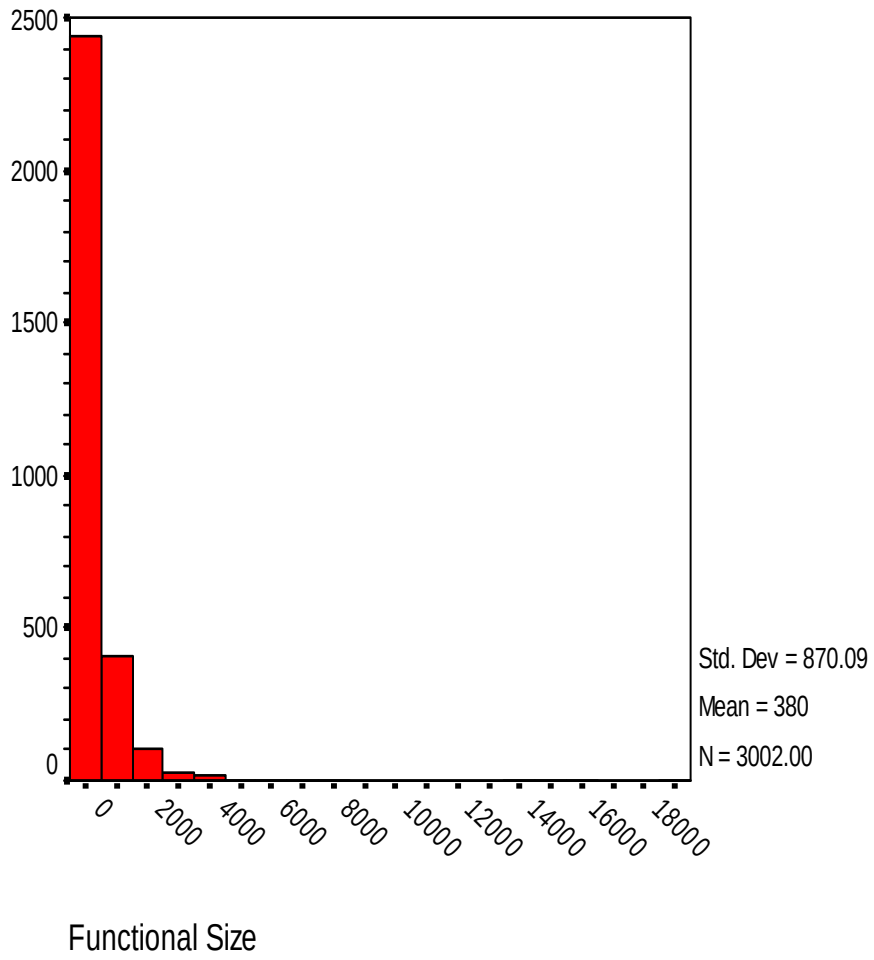
Check for normality

Check for outliers

Check for missing values

Consider transformation

Know how the tool is dealing with missing data



Data Mining Issue: the Data

- 80/20 rule: roughly 80% of effort and duration in Data Mining is in preparing the data
- Training for Lean Six Sigma does not provide sufficient coverage of data collection (although it does include Gage R&R) and data challenges

"You better hurry. Management wants the data cleaned up by tomorrow morning."

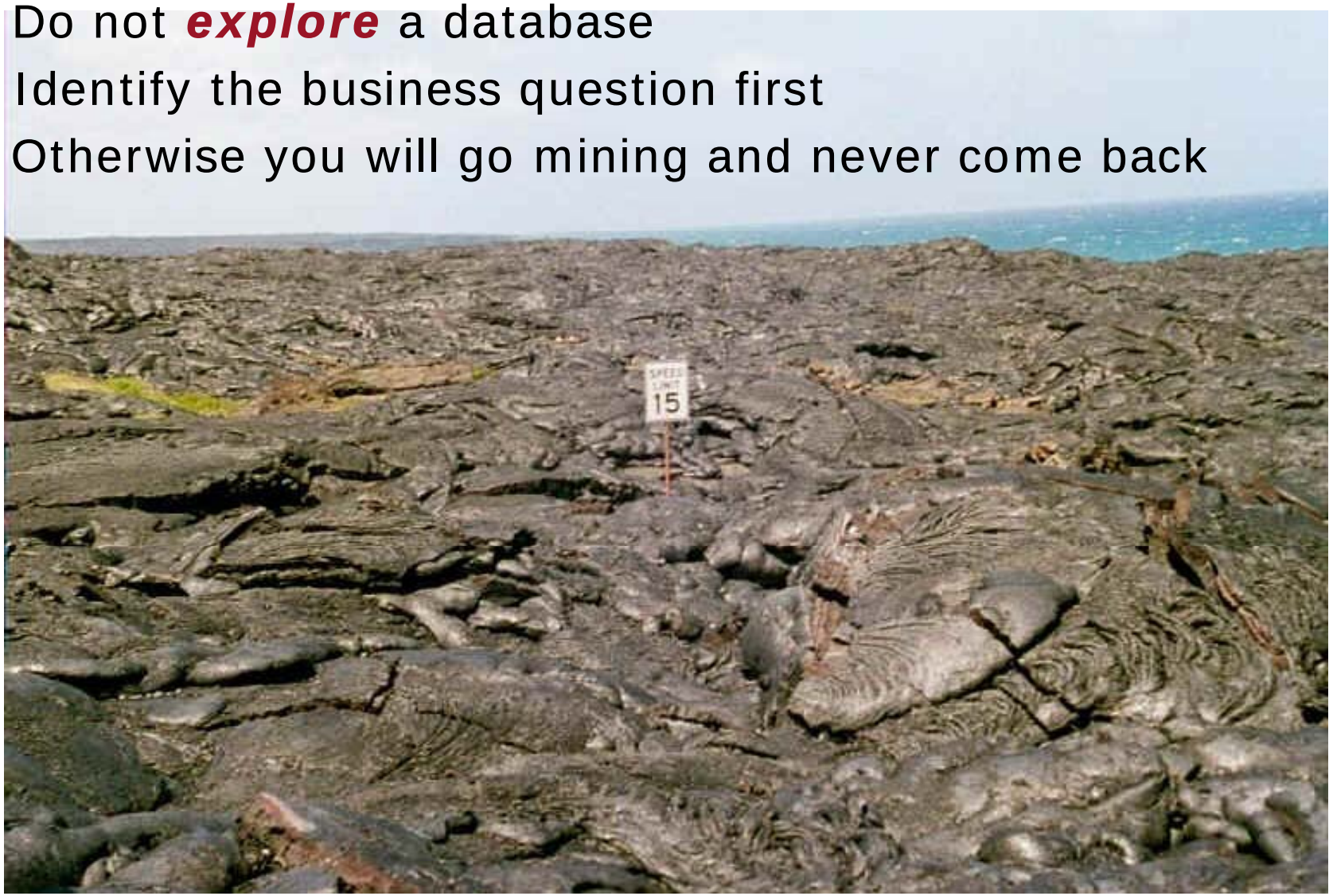
J.B. Landers ©

Data Mining Issue: Exploration

Do not ***explore*** a database

Identify the business question first

Otherwise you will go mining and never come back



Summary

Topics covered:

- Process Performance
- Data mining and model creation challenges
- Types of data mining models and examples
- Data mining issues

Consider the use of data mining to aid in filtering many variables down to a vital few to for process improvement and model based estimates.

Final Data Mining Issue: The Laugh Test

Software cannot discriminate between an important strong association and something that is obvious and trivial.

Your conclusions will have to pass the "laugh test" with the project team.

Twyman's Law: If it looks interesting, it is probably wrong.

Resources

- www.twocrows.com (free 36 page introductory booklet in Adobe format)
- www.kdnuggets.com (extensive data mining industry website including links to free evaluation software)
- *Introduction to Data Mining*, by Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison-Wesley, 2006
- *Principles of Data Mining*, by David Hand, Keikki Mannila and Padhraic Smyth, MIT Press, 2001.
- *Data Mining – Concepts, Models, Methods and Algorithms*, by Mehmed Kantardzic, IEEE Press, 2003.
- *The QSM Software Almanac: Application Development Series*, QSM, 2006.
- QSM High Performance Benchmark Consortium, http://www.qsm.com/