Brian has master’s degrees in business administration (University of British Columbia) and statistics (Peking University). He was a mechanical engineer at Panasonic and Motorola before he got intoxicated by data magic using SAS in 2005. He is now experienced in DM, CRM, and OR in the telecom, entertainment, and health care industries, and was an executive committee member of the Vancouver SAS user group before relocating to Ontario and joining RIM as a senior data analyst.
“I, Robot”, Skyrocket Analytics with SAS Process Automation

Brian Sun
Jun 8, 2012

sun.gsm@hotmail.com
http://ca.linkedin.com/in/guanghuisun
What’s on the Table

• Ideas bigger than building a single SAS macro: Automation system, Business process optimization

• Ideas more than automation: Management, Collaboration

• An example which can be extended:
  - Small/medium size business/organization
  - Analysis team as shared service in large corporation
Agenda

• Why
• How
• Impact

Analytics

Statistician/Analyst Manager

Business Process
Why
Need for Transparency

Amy (Researcher) got an idea!!!

Jennifer (IT) extracted the raw data

Brian (Statistician) did some SECRET work on the data SILENTLY

Publication

Amy (researcher) wrote the paper
Managerial Concerns

- No idea of Brian’s work load
- No idea of the process in the black box
- When Brian is on vacation, Amy is automatically on vacation
- No idea whether Brian did a good job or not
Brian’s Nightmares

Being a “hotspot” in the office
Struggle to meet many deadlines

Always dynamic
Never clean

Repeated, and tedious!
Manage rules in the ocean of codes
Revisit your SAS code after 3 months

“what if…”
“Can you analyze the whole database?”
“Can you do statistical tests for everything?”
Consequence

Analytics run OUT of CONTROL

Boss would be very upset

Brian would evolve into a Coding Machine
Or......

What Doesn’t Kill You Makes You Stronger!
How
Glossary

Work Flow | Task | Rule
Process Automation - Roadmap

1. Process Mapping
2. Standardization
3. Repeated Tasks
4. Rules Identification
5. Automatic Decisions
6. Strategy
7. Ad hoc
8. Analytics
9. SAS Code
10. Translation
11. Centralization

Process Automation and Roadmap Overview:
- Process Mapping
- Standardization
- Repeated Tasks
- Rules Identification
- Automatic Decisions
- Strategy
- Ad hoc
- Analytics
- SAS Code
- Translation
- Centralization
Inside the “Dark Room” of Analytics
Process Mapping: Circle of the Analytics

25% Data Exploration
- Data Availability/Missing Check
- Distinct Value Calculation/Freq
- Identify Data Quality Issue
- Preliminary Analysis on Raw Data

5% Analysis Plan / Study Design
- Understand Research Questions
- Literature Review (Clinical/statistical)
- Data Preparation Plan
- Statistical Analysis Plan

40% Data Preparation
- Data Extract/Transform
- Cohort Selection
- Data Integration/Calculation Validation

5% Univariate Analysis
- Frequency Analysis
- Distribution Analysis
- Outlier/Extreme Value

15% Bivariate Analysis
- 2*2 Frequency Analysis
- Scatter Plot/Trend Analysis
- Statistical Tests for Correlation
- One-Way ANOVA

10% Multivariate Analysis / Statistical Computation
- Hypothesis Tests
- Multivariate Regression Analysis
- Longitudinal Data Analysis
- Ad hoc Model/Computation
Process Mapping - Details

1. **Researcher**
   - Research Protocol
     - Fill the data Request Form
   - Raw Master Table Build
     - Data Summary & Univariate Analysis
   - Rules for Data Recode, Transform, Derived Variable, etc.
   - Build Final Master Table
   - Update Data Request Form

2. **Statistician/Analyst**
   - Analysis Plan/Study Design
   - Multivariate Analysis/Modeling
     - Model Identification Plan
     - Bivariate Analysis
     - Revisit Data Summary & Univariate Analysis
   - Model Calibration/Selection
     - Update Multivariate Analysis/Modeling
     - Model Diagnosis/Validation
     - Analysis Report

3. **Paper Submission to Journal**
   - Research Paper Draft
   - Modify?

4. **Discuss the Analysis Result**

Collaborative Processes
Standardization of Work Flows and Tasks

Can not be Standardized /Automated

Can be Fully Standardized /Automated
Standardize the Work Flow - Data Preparation

External Database/Data

Data Extraction
- From Access database
- From Excel data
- From Ad hoc Format data

Data Transformation
- Variable Format Conversion
- Variable Rename
- Value Recode
- Category Regroup

Data Selection
- Selection by Logic Expression
- Selection in Multiple Records
- Selection for Distinct Value

Data Integration
- Vertical Merge
- Horizontal Merge
- Information Consolidation

Data Calculation
- Simple New Variable Creation
- Scoring
- Ad hoc Variable Creation/Calculation

Data Validation
- Range Check
- Missing Value
- Consistency

Master Table

Can not be Standardized/Automated
Can be Fully Standardized/Automated
Standardize Task and Rule Identification

**Data Extraction**
- From Excel data

**Data Transformation**
- Value Recode

### Automated Rules (Decisions) for Correlation Test

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Correlation test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>N of distinct values</td>
<td>Format</td>
</tr>
<tr>
<td>Numeric</td>
<td>any</td>
<td>Numeric</td>
</tr>
<tr>
<td>Char</td>
<td>&lt;=N_thread</td>
<td>Char</td>
</tr>
<tr>
<td>Numeric</td>
<td>any</td>
<td>Char</td>
</tr>
<tr>
<td>Char</td>
<td>&gt;N_thread</td>
<td>Char</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>

N_thread=10 (Control parameter=> set by user)
### Opportunity for Process Automation

<table>
<thead>
<tr>
<th>Overall</th>
<th>Research Analytics</th>
<th>Work Flow</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Data Exploration</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>0</td>
<td>Analysis Plan / Study Design</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>Data Preparation</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>Univariate Analysis</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>Bivariate Analysis</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>Multivariate Analysis / Statistical Computation</td>
<td>3</td>
<td>8</td>
</tr>
</tbody>
</table>

0 = Can not be Standardized/Automated  
9 = Can be Fully Standardized/Automated
“SAS Robot” System

- Gear to business Circle
- Highly Flexible – independent with dataset/database
- Super Time Saving
- Smart (Automatic Decision Making)
- Control Box (Centralized Rules & Decisions)
- Customized Output

- Data Exploration
- Analysis Plan/Study Design
- Data Preparation
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis/Statistical Computation
- Automatic Database Extractor
- Power Master Table Builder
- Analytics Report Robot
- Regression Robot
Example 1: “Can you analyze the whole database?”

A Complete Analysis of SASHelp Library

Total N of Tables: 120
Total N of Variables: 2050

Output:
1) A full PDF report of **1227** pages with all univariate plots and tables included.
2) A **150**-page PDF executive summary report is automatically summarized from the full report, with **index** created for navigating the full report.
Example 2: “Can you do the test for everything?”

An Exhaustive Pair-wise Correlation Analysis

Note: Detailed snapshots not available due to data confidential

Total N of variables: 116
Total N of variable pairs for correlation test: 116*115/2=6670

Output:
1) A full PDF report of 9613 pages with all bivariate plots and statistical test results included.
2) A 325-page PDF executive summary report (main body only 10 pages) is automatically summarized from the full report, with index created for navigating the full report.
### Collaboration Process:

**Model Identification Plan**

<table>
<thead>
<tr>
<th>Response</th>
<th>Independent</th>
<th>Pearson Coeff</th>
<th>P Value</th>
<th>Model No</th>
<th>Model Title</th>
<th>Model Type</th>
<th>Include/Exclude in Model</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverse_Event_A_Yes</td>
<td>age</td>
<td>0.45</td>
<td>0.009</td>
<td>1-1</td>
<td>Adverse Event (A) Predictive Model</td>
<td>Binary Logistic Regression</td>
<td>Include</td>
<td>BS: it makes sense that elder people has higher risk getting adverse events</td>
</tr>
<tr>
<td>Adverse_Event_A_Yes</td>
<td>BMI</td>
<td>0.26</td>
<td>0.02</td>
<td>1-1</td>
<td>Adverse Event (A) Predictive Model</td>
<td>Binary Logistic Regression</td>
<td>Exclude</td>
<td>BS: I decided to exclude this variable from analysis because the significant collinearity issue we found between BMI_Overweight25 and the high blood pressure.</td>
</tr>
<tr>
<td>Adverse_Event_A_Yes</td>
<td>smoking_dummy</td>
<td>-0.26</td>
<td>0.004</td>
<td>1-1</td>
<td>Adverse Event (A) Predictive Model</td>
<td>Binary Logistic Regression</td>
<td>Exclude</td>
<td>AN: No clinical explanation</td>
</tr>
<tr>
<td>Adverse_Event_A_Yes</td>
<td>EAD24_ad_Yes</td>
<td>0.30</td>
<td>0.006</td>
<td>1-1</td>
<td>Adverse Event (A) Predictive Model</td>
<td>Binary Logistic Regression</td>
<td>Exclude</td>
<td>AN: I don’t think that we need both trach and EAD as these are both types of tubes used to treat a patient</td>
</tr>
<tr>
<td>Adverse_Event_A_Yes</td>
<td>Gender_Female</td>
<td>0.22</td>
<td>0.015</td>
<td>1-1</td>
<td>Adverse Event (A) Predictive Model</td>
<td>Binary Logistic Regression</td>
<td>Include</td>
<td>AN: Literature support that female is more likely to get adverse event</td>
</tr>
<tr>
<td>Adverse_Event_A_Yes</td>
<td>Mortality_Yes</td>
<td>-0.32</td>
<td>0.006</td>
<td>1-1</td>
<td>Adverse Event (A) Predictive Model</td>
<td>Binary Logistic Regression</td>
<td>Exclude</td>
<td>BS: Does not make logical sense. Mortality should be an outcome variable (Model 3)</td>
</tr>
</tbody>
</table>

**Notes**
- Blue: Clinical Researcher
- Green: Statistician

**Collaboration Process:**

1. **Input**
   - Bivariate Analysis

2. **Model Identification Plan**

3. **Output**
   - Multivariate Analysis
New Collaboration Model

- **Researcher**
  - Run stored process to view/update analysis report
- **SAS office Add-in**
  - Input, manage rules from domain point of view
  - Rule Input Tables
- **Statistician/Analyst**
  - Input, modify rules from data and model point of view
  - Build, maintain SAS stored processes
- **SAS Stored Processes**
  - Output
- **SAS EG**
- **SAS BI Server**
Impact
Impact on the Analytics Life Circle

**Before**
- Data Preparation: 40%
- Analysis Plan / Study Design: 25%
- Bivariate Analysis: 15%
- Multivariate Analysis / Statistical Computation: 10%
- Data Exploration: 5%

**After**
- Data Preparation: 10%
- Analysis Plan / Study Design: 8%
- Bivariate Analysis: <1%
- Multivariate Analysis / Statistical Computation: 30%
- Data Exploration: 50%
Impact on Manager and Researcher

Manager:
- Process Visibility
- Quality Control
- Project Planning
- Take control of business rules
- Run/update standard analytics report by themselves
- Easy modification of hypotheses/Test “what if” question by themselves
- Analytics Mining

Project management improved

Empowered and more engaged in analytics

Researchers:
Impact on Statistician/Analyst

Brian in 2011 → Brian in 2050

Brian in 2011 → Brian in 2050
Impact on the Business Process

Amy (Researcher) got an idea!!!
Jennifer (IT) extracted the raw data
Brian (Statistician) did some SECRET work on the data SILENTLY
Amy (Researcher) wrote the paper
Publication

SAS office Add-in
SAS Stored Process

Rule Input Tables
Input, manage rules from subject-matter point of view
Run stored process to update analysis report

Build, maintain SAS stored procedures
Input, modify rules from data and model point of view
SAS EG

Input, output
Impact on Organization/Business

Process Automation → Productivity → Time → Money

Management → Quality

Funding From Government and Public

Global Reputation → Publication
Conclusions

By building a highly functional and flexible SAS process automation system, we can

- **Skyrocket** the analytics capacity.
- **Re-engineer** the business process.
- **Improve** management and collaboration.
- **Save** significant time and money.

**This approach can be applied in other businesses!**
**The success can be duplicated!**
MY DREAM IS TO DO LESS WORK WHILE ALLEGEDLY BEING MORE VALUABLE.