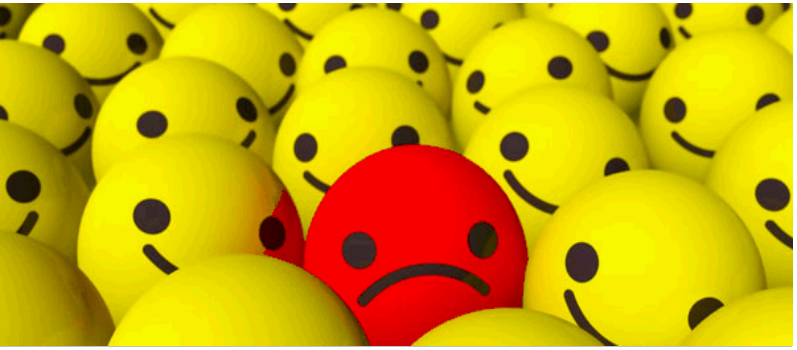
SAND2015-3115C

Exceptional service in the national interest

Sandia National Laboratories



Predicting Future Security Events to Better Manage High-Risk Situations

Presented to NLIT in May 2015

Gerry Giese, Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

The Business Problem

Sandia National Laboratories

Security Assurance – confidence in national security

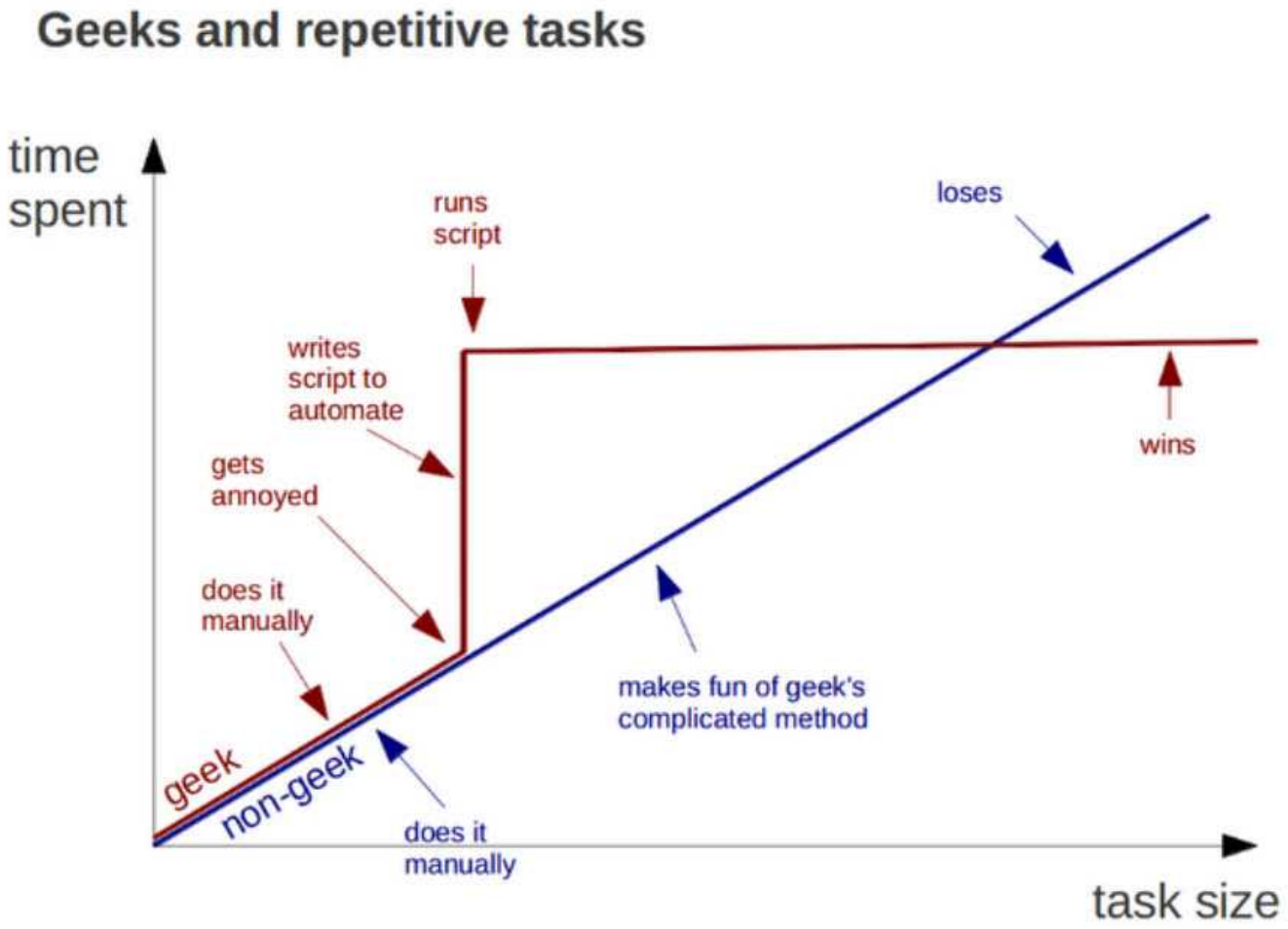
Anecdotal evidence of risk factors related to security incidents

But these risk factors are "**lagging indicators**"

Lack of information useful for risk reduction

Facilitate communication between security and line operations

The Solution? It's Complicated…



(source unknown)

The Solution

Sandia National Laboratories

A behavioral predictive model was proposed and accepted in late FY2013

- To identify risk factors (predictive indicators)
- As a tool for having conversations with management

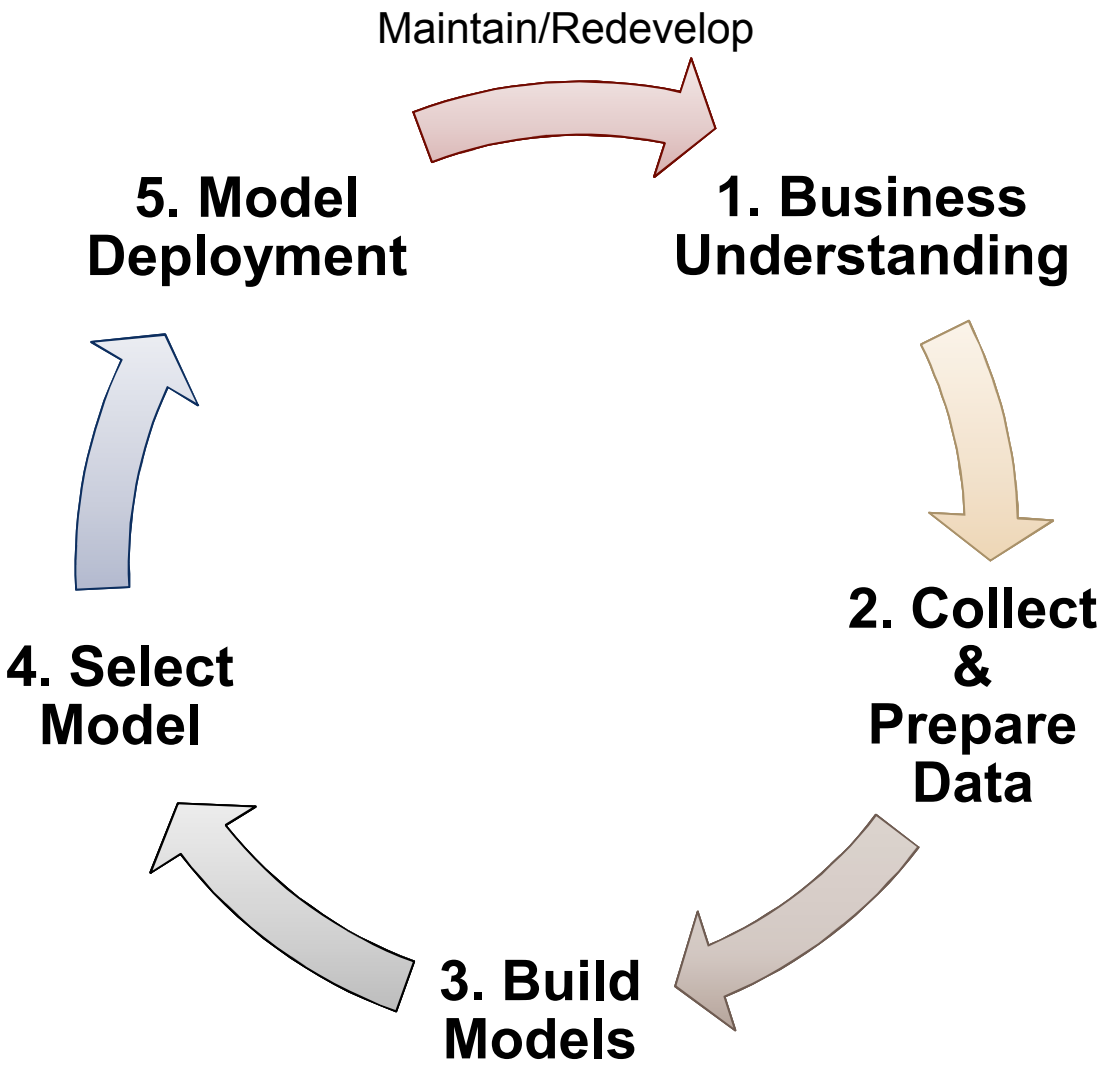
What is a predictive model?

- Statistical data-mining technique that analyzes historical and current data to generate a model that predicts future behavior
- Used successfully in the financial, insurance and retail industries

How can it be used?

- Indicates probability, does not indicate causality
- Identifies behaviors or conditions related to increased risk

Simplified Modeling Process



Step 1 - Business Understanding

Sandia National Laboratories

Question: What is the risk that a Sandia organization will have a Security Issue in the Next 6 Months?

Brainstorming Indicators

- Use a diverse team
- Prioritize the list – can't afford to get everything

How will the predictive indicators be used?

- Avoid privacy and legal issues
- Imagine discussions with mgt. on practices/changes to make

How can the model be used?

- Indicates probability, does not indicate causality
- Identifies behaviors or conditions related to increased risk
- Never totally accurate – behaviors affected by personal factors

Predictive Model Indicators

Sandia National Laboratories

Fake Example – predicting risk using buying patterns to indicate mindfulness

Buying lots of vegetables & fruit decreases risk

Predictive Model Indicators

Sandia National Laboratories

Fake Example – predicting risk using buying patterns to indicate mindfulness

Buying lots of vegetables & fruit decreases risk

Buying some vegetables & fruit means no change in risk

Predictive Model Indicators

Sandia National Laboratories

Fake Example – predicting risk using buying patterns to indicate mindfulness

Buying lots of vegetables & fruit decreases risk

Buying some vegetables & fruit means no change in risk

Buying no vegetables & fruit increases risk

Predictive Model Indicators

Sandia National Laboratories

Fake Example – predicting risk using buying patterns to indicate mindfulness

Buying lots of vegetables & fruit decreases risk

Buying some vegetables & fruit means no change in risk

Buying no vegetables & fruit increases risk

KEY POINT

Not buying vegetables & fruit does not *cause* security issues!

Step 2 – Collect and Prepare Data

Sandia National Laboratories

Data was gathered (slowly) from multiple data sources

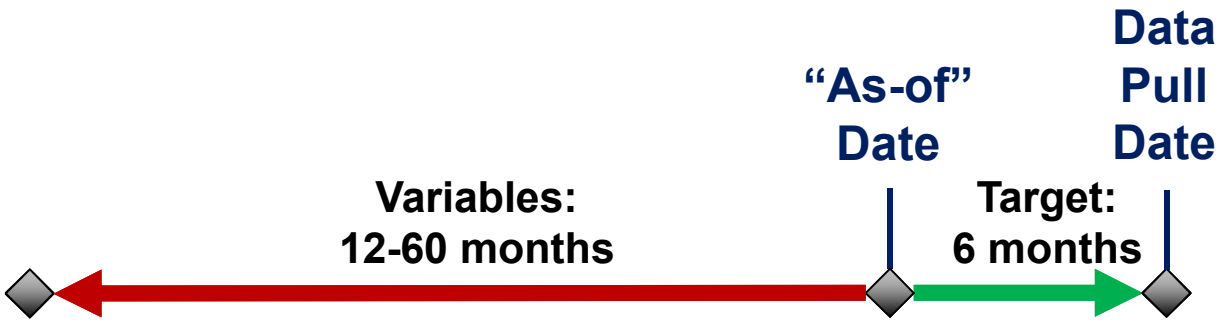
- Human Resources systems
- IT systems management records
- Safety & Health records
- Training records
- And more…

Data Governance Needed

- Commonly heard: *That's MY data! What are you going to do with it?*
- Had to brief 4 Directors, 1 VP, and 3 Lawyers
- As a result, Sandia is implementing a corporate data governance process to enable faster/safer data collection

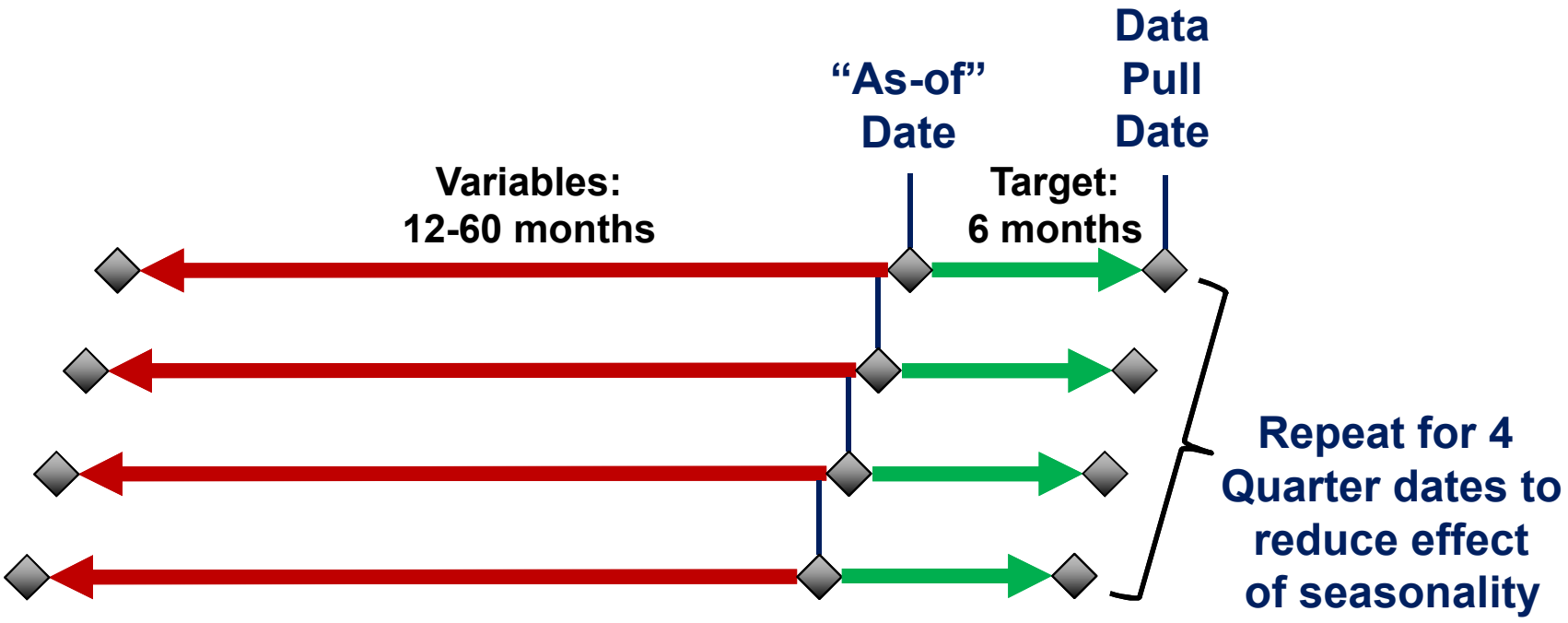
Step 2 – Collect and Prepare Data

Data was integrated and summarized over multiple time periods



Step 2 – Collect and Prepare Data

Data was integrated and summarized over multiple time periods



Step 3 – Build Models

Sandia National Laboratories

Tools Used

- Data Exploration & Visualization - IBM SPSS, SAS JMP
- Model Development - SAS Enterprise Miner (Desktop)

Modeling

- Customer desire to know the predictors limited model choices
- Explored data with bi-variate and cluster analysis
- Model algorithms used included Decision Trees (CART, CHAID, C5.0), Linear Regression, Logistic Regression, and Neural Network

Model Training and Testing

- With a very low number of security issues as a percentage of the population, we used 10-fold cross-validation to compensate

Step 4 – Select Model

Sandia National Laboratories

How effectively do the models separate two populations?

- Organizations more likely to have security issue in next 6 months
- Organizations less likely to have security issue in next 6 months

Kolmogorov Smirnov (K-S) Score

- Measures the degree of separation between positive and negative distributions
- A common standard used for Credit Scoring in the financial industry
- Logistic Regression algorithm scored best in our case

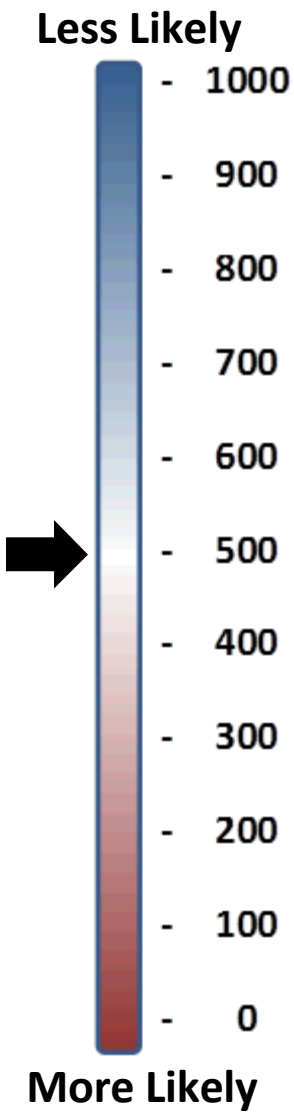
Bivariate Analysis Revisited

- Analysis found strong data relationships that the model did not select
- Chose a set of these to include in customer results

Example Model Scorecard

Sandia National Laboratories

Characteristic	Data Value	Points	SCORE
Constant		R	**R**
Number of X	0	0	**R**
	1 or More	-n	
Number of Y	7 or Less	0	**R-n**
	8 or More	-n	
Number of Z	14 or Less	+n	**R-m**
	14 to 19	0	
	20 or More	-m	
Number of A	1 or Less	0	**R-n**
	2 to 5	-n	
	6 or More	-m	
Number of B	2 or Less	0	**R-n**
	3 or More	-n	
Number of C	0	0	**R-m**
	1 to 3	-n	
	4 or More	-m	
		TOTAL	**T**



Points are factors in the equation produced by the modeling algorithm, converted for purposes of communication.

Step 5 – Model Deployment

Sandia National Laboratories

Implementation constraints

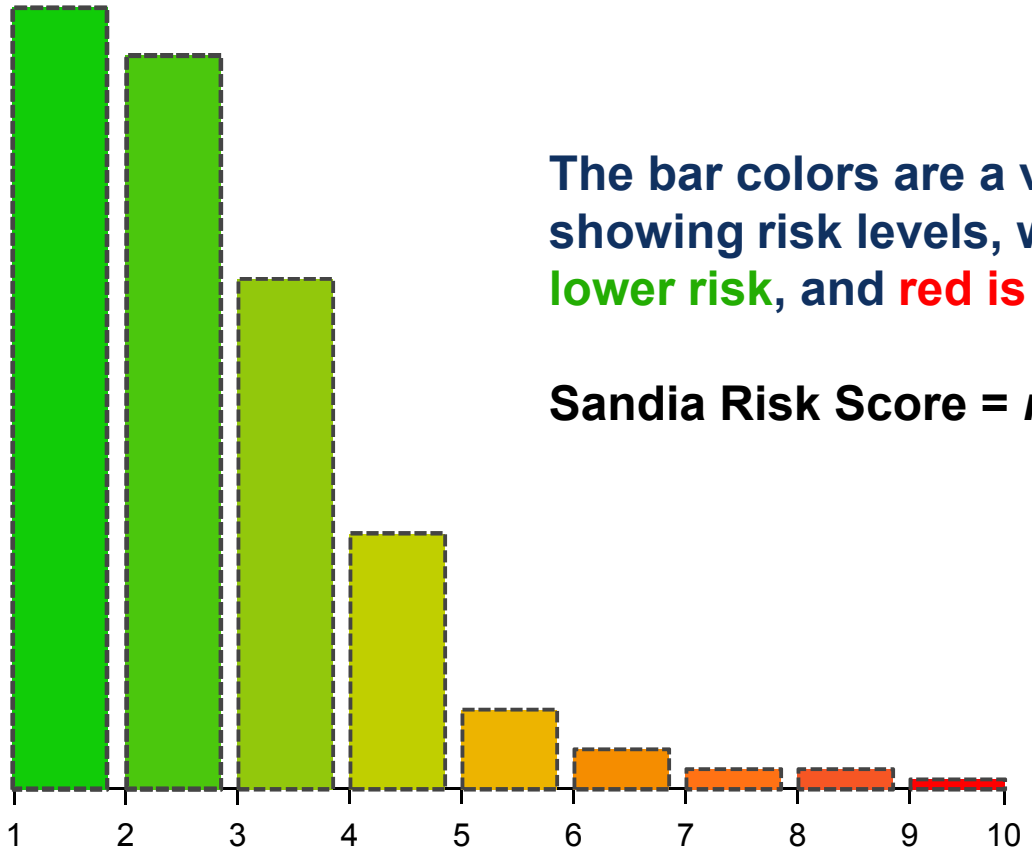
- Ensure privacy of employees
- “Roll up” the results to higher organization levels
- Only show organizations if they include at least 20 employees
- Provide a “number”

Our visualization solution

- Represent proportion of the distribution of employees
- No “Y” axis – showing values could break privacy rule
- Security Risk Level Chart
- Predictive Indicator Charts
- Side-by-side Dashboard

Overall Risk Level

On a 1 to 10 scale, show the proportion of all Sandia organizations who fall into each risk level



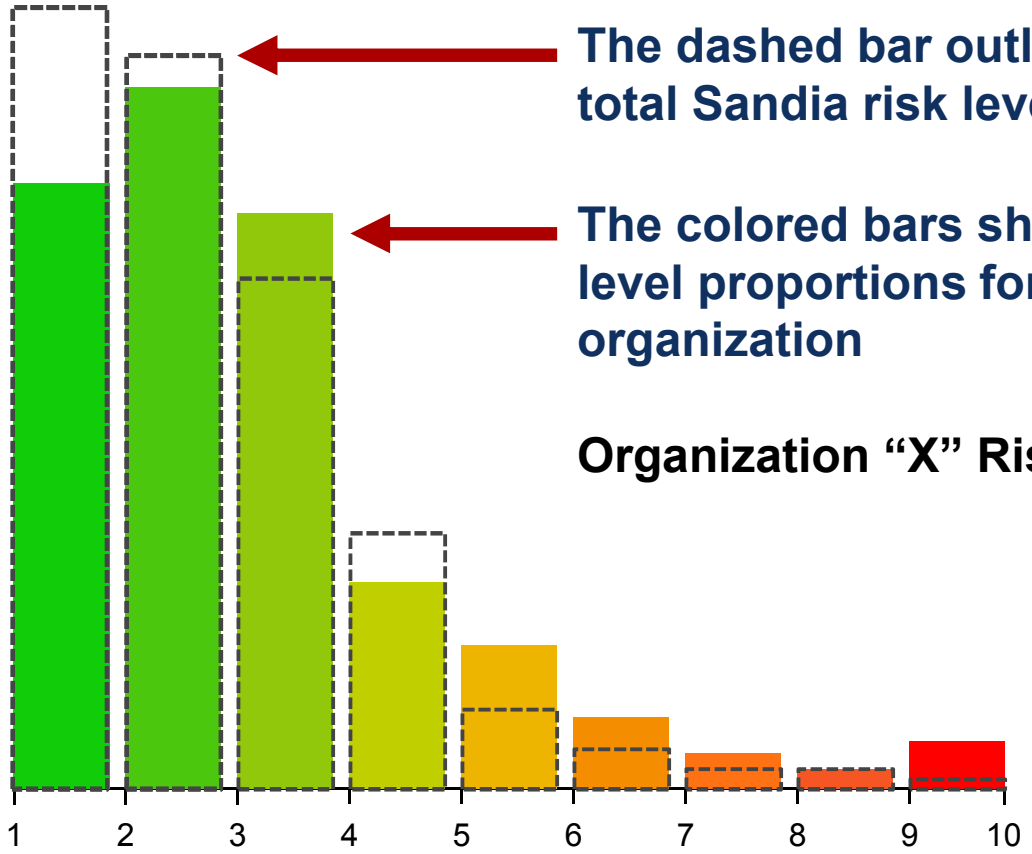
Not Actual Results

The bar colors are a visual way of showing risk levels, where green is lower risk, and red is higher risk

Sandia Risk Score = *n*

Organization “X” Risk Level

On a 1 to 10 scale, show how Organization “X” compares with all Sandia organizations for each risk level



The dashed bar outlines show the total Sandia risk level proportions

The colored bars show the risk level proportions for a selected organization

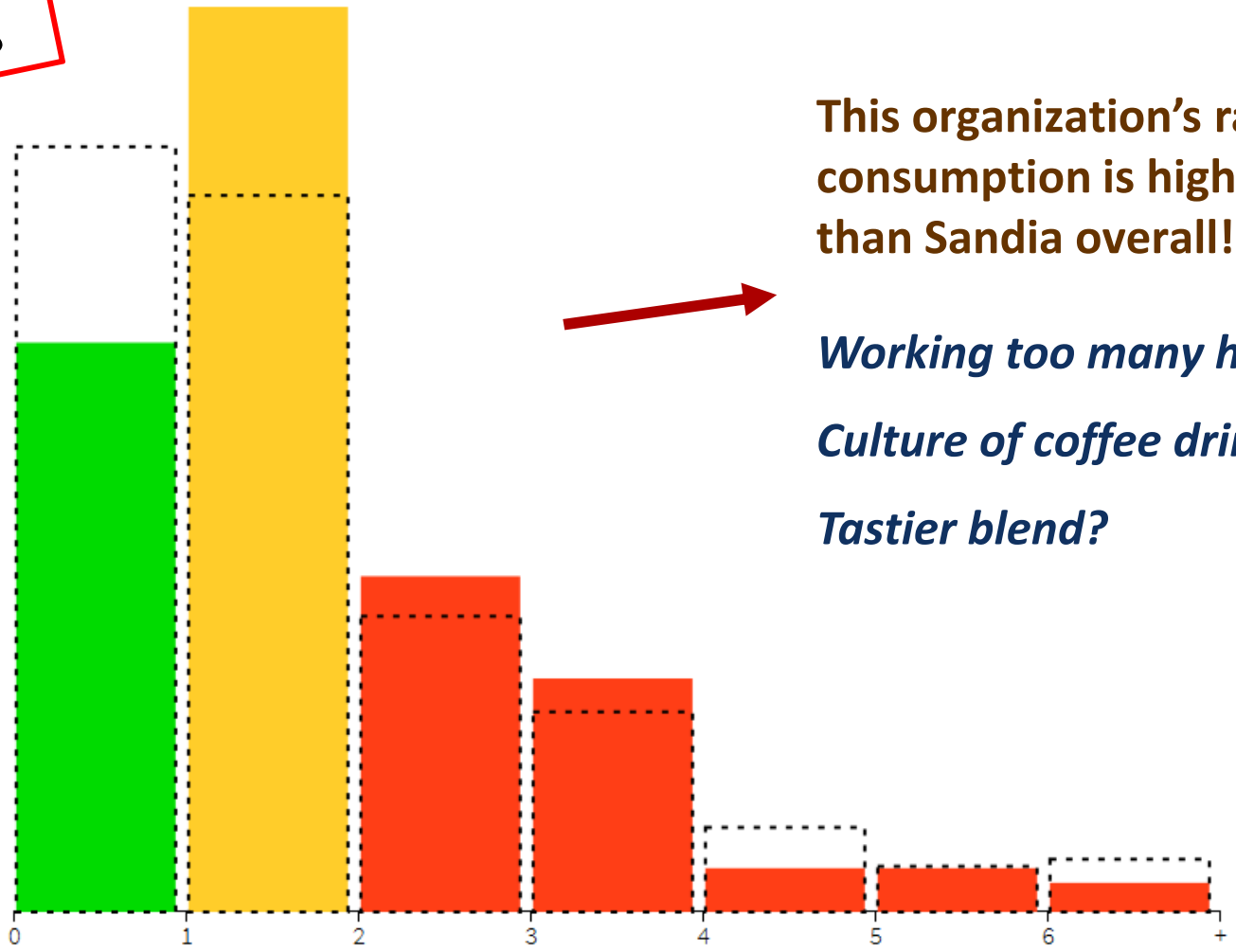
Organization “X” Risk Score = *m*

Not Actual Results

Cups of Coffee/day, Past 12 Months

Sandia National Laboratories

Not Actual Results



This organization's rate of consumption is higher than Sandia overall!

Working too many hours?

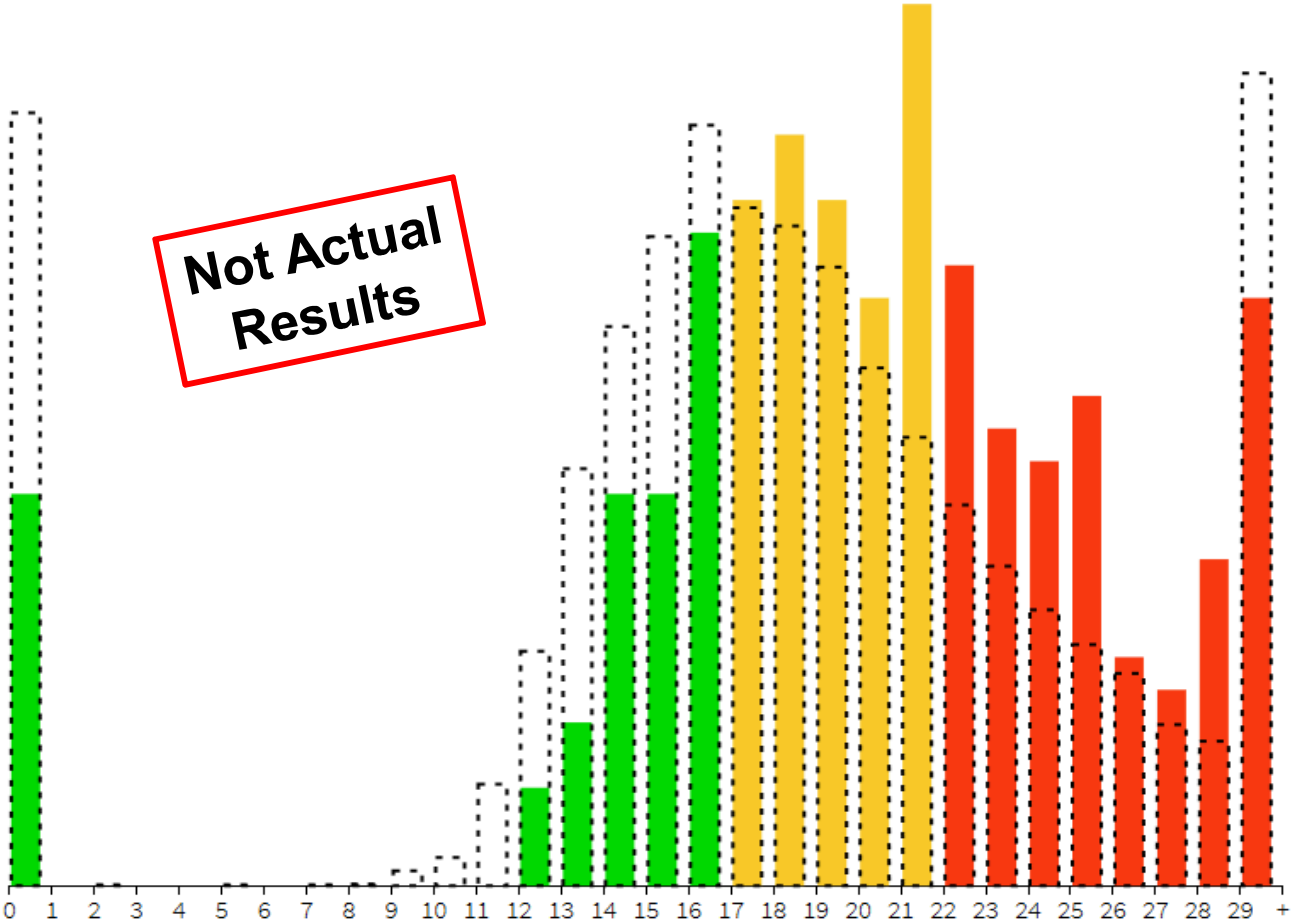
Culture of coffee drinking?

Tastier blend?

Times Car Keys Lost, Past 36 Months

Sandia National Laboratories

People in this organization lose their car keys more than Sandia overall. ***Lack of process? Management doesn't stress enough the importance of keeping track of car keys?***



Hours of Paperwork, Past 3 Months

Sandia National Laboratories

This organization tracks fairly closely with most Sandia organizations.

However, there are still people in the "red zone", so it's worth investigating if all that paperwork is truly necessary…

Step 5 – Model Deployment

Sandia National Laboratories

Document and automate the probability calculation

- This is the secret sauce – we keep it under wraps

Collect new data periodically, automate model updates

User Experience is important

- Evaluate how each user type will interact with the software
- If the user doesn't understand the visualization/results, you failed!

Monitor usage and results

- How well are users adopting it into their process?
- Evaluate the model for "drift"

Risk Dashboard

Sandia National Laboratories

Showing two independent graphs side-by-side can support comparison and trending…

- Across organizations
- Across dates
- Across indicators

Not Actual Results