Using Natural Language Processing (NLP) and Predictive Analytics, to Monitor Grant and Program Performance
Summary

Situation

- Effective grants monitoring involves trade-offs and challenges in predicting performance, in an environment of limited data.
- While more data is desirable, increased data collection may be neither feasible nor allowable.

Solution

- Through Natural Language Processing (NLP), grantor agencies can leverage new, innovative data – to better understand:
  - the wider context and milieu, supporting program evaluation
  - key variables and dynamics, influencing outcomes
  - predictors of crises, failures and fraud
- NLP gathers unstructured data from the web, assesses relationships and correlations.
- When transformed to structured formats, such data can be used in predictive and prescriptive analytics.
The federal grants landscape is large - and poised for performance management enhancements through analytics

- Many government organizations (26) accomplish their work through grant programs (1,000+)
- Analytics contributes value where grant-making agencies have opportunities for discretionary action based on forecasts (i.e. programmatic grants vs. block grants)

*Source: Grants.gov, USASpending.gov query*
Grant monitoring balances investment, risk and reward throughout the lifecycle

- Typical challenges and trade-offs:
  - Fund an array of programs, including innovative (“risky”) grants
  - Predict and withdraw funding as quickly as possible from failing grants
  - Intercede with assistance at critical junctures
Within the analytic framework, NLP augments data sources for insights

...transformed by analytics......for decision makers.

Structured Data & Unstructured Content...

...made consumable and assessable to provide business insights through:

Descriptive Analytics, Predictive Analytics, and Prescriptive Analytics, ...

to provide business insights.
NLP ("content analytics") feeds predictive and prescriptive analytics

<table>
<thead>
<tr>
<th>Descriptive Analytics</th>
<th>Use Natural Language Processing (NLP) “fact extraction” with descriptive statistical formulae to make use of the wealth of unstructured data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Analytics</td>
<td>Predictive Analytics aids the discovery of relationships, as well as potential repeatable statistical patterns in large data sets – both supervised and non-supervised.</td>
</tr>
<tr>
<td>Prescriptive Analytics</td>
<td>Understanding relationships, key variables and past outcomes, take better actions (monitoring, technical assistance, etc.) to facilitate successful performance and detect fraud.</td>
</tr>
</tbody>
</table>

*The combination of multiple statistical approaches drives actionable and accurate insight.*
An illustrative example of key questions in grantmaking & oversight

**Big Data Management**

- Use IBM Content Analytics to demonstrate the value of descriptive data analysis on newly acquired, unstructured data

**Advanced Analytics**

- Use SPSS to demonstrate how advanced analytics yield grant performance insights

**Descriptive Analytics**

- What factors increase likelihood of grant selection?
- What is happening with grant execution?
- Is grant execution data correlated with other factors?

**Predictive Analytics**

- To what extent do applications & reports describe “how” grants are executed?

**Prescriptive Analytics**

- Given this level of risk, what are the next actions to take (monitoring, tech assistance)?
- How should these actions be allocated across grants with constrained resources?
- Are grant recipients similar in some way?
ICA Text Analytics: An Overview of the Process

1. ICA acquires, combines multiple sources of unstructured data with structured data

   +

   Unstructured Data

   Structured Data

   - ID
   - Zip
   - Award Date
   - Size

2. Within ICA, these are organized into meaningful facets and collections by the user and queried

3. Search terms are input, revealing textual correlations

   - All Documents
   - Applications
   - Grantee Reports
   - External Delivery
   - Performance Audit Findings

   Public NHTSA Data
ICA Text Analytics: How does it reveal strength of correlation?

Q: Is use of an external grant delivery correlated with grant awards?

\[
\frac{\#(A \cap B)/\#A}{\#B/\#D} = \frac{\#(A \cap B)/\#D}{(#A/\#D)(#B/\#D)}
\]

When the documents containing the keyword “Awarded” are examined, the percentage of grantees with “External Delivery” are examined, the percentage of grantees with “External Delivery”

A = {documents that contain a keyword “Awarded”}
B = {documents that contain a keyword “External Delivery”}
D = All Documents

\[
= \frac{20/40}{30/100} = 1.66
\]
Is there a correlation between “Jeep Wrangler” and “Fuel System” issues?

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Frequency</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WRANGLER 2-DR. 4X4</td>
<td>96</td>
<td>30.8</td>
</tr>
<tr>
<td>WRANGLER 4-DR. 4X4</td>
<td>45</td>
<td>45.5</td>
</tr>
<tr>
<td>DURANGO 4X4</td>
<td>66</td>
<td>30.2</td>
</tr>
<tr>
<td>ASPEEN</td>
<td>3</td>
<td>20.8</td>
</tr>
<tr>
<td>S50X1</td>
<td>6</td>
<td>25.8</td>
</tr>
<tr>
<td>ASPEEN 4X4</td>
<td>7</td>
<td>20.9</td>
</tr>
<tr>
<td>NITRO 4X4</td>
<td>10</td>
<td>24.9</td>
</tr>
<tr>
<td>WRANGLER SAHARA</td>
<td>1</td>
<td>18.4</td>
</tr>
<tr>
<td>WRANGLER</td>
<td>34</td>
<td>20.9</td>
</tr>
<tr>
<td>NITRO</td>
<td>11</td>
<td>18.4</td>
</tr>
<tr>
<td>WRANGLER 4-DR. 4X2</td>
<td>1</td>
<td>6.9</td>
</tr>
<tr>
<td>M1 100S</td>
<td>9</td>
<td>10.7</td>
</tr>
<tr>
<td>FIERCE</td>
<td>1</td>
<td>6.9</td>
</tr>
</tbody>
</table>
Is there a correlation between “Jeep Wrangler” and “Fuel System” issues?

I bought a used 2007 Jeep Wrangler in June of 2010. A couple months after owning the vehicle, I noticed that if I left the gasoline pump on automatic, the pump would cut off, and gasoline would gush out from the tank, spilling onto the vehicle and ground below. This happened at various stations and every time I went to fill my tank to full, to prevent it from happening, I only fill my tank to 3/4 full.
Is there a correlation between “Jeep Wrangler” and “Fuel System” issues?

**Top Product Ratings:** Tires | Sedans | SUVs | Small Cars | GPS

**Consumer News**

JEEP

**Jeep Wrangler: Owners complain about fuel spill problem, TSB issued**

In an original investigation into consumer complaints to the government, Consumer Reports has found an unusually high number of complaints related to the fuel system of the Jeep Wrangler, starting in the 2007 model year. The Wrangler problems were highlighted by our statistics department due to the spike in complaints.

One owner of a 2007 two-door Wrangler complained, "My fuel tank overflows when filling tank. Every time. Every filling station we have tried."

Several owners also complained that their vehicle stalled or wouldn't start when facing downhill.
Best Practice: Integrated unstructured data into iterative cycle of program design & exploratory evaluation

**Challenge:**
- Minimize Grantee Burden while Acquiring Maximum Decision-Relevant Data

**Reality:**
- Design of data collection can be a long, inflexible process (ex. OMB Clearance, etc)

**Best Practice:** Design forms/formats to facilitate structured collection of unstructured text

**Analytics:**
- Use NLP to understand facets of unanticipated interest
- Exploratory Mining of unstructured data “What we didn’t know we were interested in” (ex. External Delivery, institutional climate)
Predictive Analytics: An Overview of Integration

- Predictive modeling uses historical data to identify relationships between predictors (e.g., attributes of organizations delivering services, customers) and outcomes (e.g., performance, safety, risk).
  
  A strong, supportable hypothesis is formed and tested using descriptive analytics techniques:
  - indicate a predictive target
  - demonstrate that certain variables in the data predict the target’s future performance

Target Outcome

Array of Independent Predictor Variables

Examples:
- Funding
- Grantee performance in prior years
- Grantee organizational characteristics
- External delivery organization presence

Examples of Targets
- Grant Performance
- Grant Failure (Y/N)
- Likelihood of Fraud

SPSS helps determine and measure which variables are statistically valid predictors…and in what functional form.
The individual and combined strengths of the relative predictors are revealed.

Model rules with frequencies and confidences

- Predictor importance
- Not related to model accuracy
- Represents each field’s relative contribution to the model
Clustering is another predictive analytics method for understanding patterns and forecasting.

The ‘distance’ between grantees measures similarity across an array of meaningful characteristics (predictors).
Based on proximity, we ascertain risk and performance profiles.
Next Steps: Prescriptive Analytics maximizes grant outcomes

- Optimization-type models maximize over a set of criteria or goals, given limited resources
  - Maximize ROI based on project investment, duration and return
Concluding Insights

- Structured and unstructured data CAN be “discovered” and used from external sources.
- This forms a basis for predicting performance characteristics of interest (ex. grantee-level performance success from grant performance characteristics).
- Insights can help grantor organizations better understand the success factors for grants.

Questions for participants:

- What information is missing from your grants portfolio?
- What is the value of being able to better predict performance?