SAS AND OPEN SOURCE
MATT MALCZEWSKI, SAS CANADA
Your Trial, Your Data

Visual Analytics – Register for Trial
• Smart data exploration with self-services analytics makes this product usable for anyone. Interactive reporting makes it collaborative. Scalability and governance make it fit the needs of your organization, no matter the size.

Visual Statistics – Register for Trial
• Multiple users can explore and visualize data, then interactively create and refine descriptive and predictive models. Distributed, in-memory processing reduces model development time so you can run complex analytic computations – and get precise results – in minutes.
WHY OPEN SOURCE?

Why the drive to open source?

• Cost effective – considering total cost of ownership
• Flexible – customers can “build anything”
• Immediate access & easy to get started
• Latest technology and latest algorithms
• Strong community and online support
• Many new data scientists learn in open source

So why use SAS to extend open source?
SAS AS AN ENHANCEMENT

AND

SAS can augment open source

- Increase productivity
- Leverage your assets, people and platforms
- Bring the power of SAS to open source
- Create deployable analytics
- Goal is to ‘embrace’ and ‘extend’
THE ANALYTIC LIFECYCLE

Lots of Data
New Data
Experimentation
Fail Fast
Test & Learn
Interactive
Iterative
Innovation
Flexibility
Data Science

Discovery & Development of Analytics

Discovery

EXPLORE

PREPARE

MODEL

ASQ

INVENTOR

Execution of Analytics

Deployment

EXECUTE

MONITOR

Data

Data Science

Deployment

Regulated
Automated
Governed
Embed
Reliable
Decisions
Consistent
Documented
Actions
IT

Copyright © 2016, SAS Institute Inc. All rights reserved.
THE ANALYTIC LIFECYCLE: SAS AND OPEN SOURCE

**SAS**

- SAS embraces open source for Data Prep
- Open source and SAS work well for Discovery and Development
- SAS can extend open source
  - inventory, register and manage models
  - deploy and execute models in Hadoop and in database
  - enhance models and provide monitoring and reporting

---

**Discovery & Development of Analytics**

1. PREPARE DATA
2. EXPLORE
3. MODEL
4. INVENTORY
5. EXECUTE
6. MONITOR

**Deployment & Execution of Analytics**

**EMBRACE** + **EXTEND**
Enterprise Wish List

- Ability to connect to Hadoop
- Run natively in Hadoop
- Minimize data movement

How SAS Embraces…

- Optimized engine to access Hadoop
- Embedded engine so Hadoop can run SAS
HADOOP AS PROCESSING ENGINE

- Use Hadoop as the horsepower for analytics
- Run SAS in Hadoop - no data movement
- Expose Hadoop data to more people through a range of interfaces
- Predictive analytics and machine learning
- SAS for Model Deployment / Scoring

EMBRACE
THE ANALYTIC LIFECYCLE

Discovery & Development of Analytics

- PREPARE DATA
- EXPLORE
- MODEL
- INVENTORY
- EXECUTE
- MONITOR

Enterprise Wish List
- A way for users to interact with Hadoop
- Ability to create analytic views and tables
- Ability to assess data quality

Deployment & Execution of Analytics

How SAS embraces…
- A business user interface to facilitate:
  - Querying Hadoop
  - Adding data
  - Profiling data
  - Cleansing data
  - Transforming data
- With no data movement
SELF SERVE ACCESS TO HADOOP

What directive do you want to perform?

- Browse Tables
- Saved Directories
- Run Status
- Delete Rows
- Run SQL
- Profile Data
- Saved Profile Reports
- Copy Data into Hadoop
- Copy Data from Hadoop into a Database
- Profile Data
- Create Trusted Data

Business user UI

Profile Data

Create Trusted Data
THE ANALYTIC LIFECYCLE

Enterprise Wish List
- Best possible analytics
- Flexibility of tools
- Productivity
- Greater insights = models
- Trusted models

How SAS Extends…
- A variety of options to develop models
- Allows data scientist to code in language of choice
- Ability to scale to any data volume
- Handle complex graphics
USE SAS TO INTEGRATE R

Why?
• Model comparison
• Leverage R for new algorithms
• Generate score code
• Deploy R models
What if you coded this?

- Compare 7 models
- Choose champion
- Inventory Model
- Generate score code
- Deploy in database/Hadoop
**SAS FROM JUPYTER**

TRAIN a random forest model on customer transaction data to predict which ones can be expected to be repurchased.

```python
trainForest(data = "shoptrain", numTrees = 250, numVarsToTry = 20)
```

<table>
<thead>
<tr>
<th>240</th>
<th>1120</th>
<th>0.150</th>
<th>0.163</th>
<th>0.210</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>130950</td>
<td>0.150</td>
<td>0.163</td>
<td>0.210</td>
</tr>
</tbody>
</table>

**Loss Reduction Variable Importance**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of Rules</th>
<th>Gain</th>
<th>2016 Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITEM_FACTOR20</td>
<td>257</td>
<td>0.000425</td>
<td>0.000425</td>
</tr>
<tr>
<td>ITEM_FACTOR18</td>
<td>208</td>
<td>0.004394</td>
<td>0.004394</td>
</tr>
</tbody>
</table>
Enterprise Wish List
- Model management platform
- Inventory ALL models
- Know who’s working on what
- Ability to deploy models
- Auditable models

How SAS Extends…
- Central model management platform
- Repository for SAS models and open source (R, Python, PMML)
- Model history
- Version control
- Model and data lineage
- Model governance
MODEL INVENTORY

- Model inventory and search
- SAS and Open Source models
- Model Metadata

SAS and Open Source models
Enterprise Wish List
• Deployable analytics
• Automation
• Faster time to model execution
• In Hadoop/database model execution

How SAS Extends…
• Model execution platform
• Execute models as database functions
• No language conversion
• Purpose built model execution engines
MODEL EXECUTION

Model Publishing and automation

Create a Scoring Output Table

Specify the name, library, and the variables to include in the scoring outputtable.

Name:
Library: MNLID

Input variables:

Output variables:

Model Score Code Creation

In Hadoop/database deployment
How SAS Extends…

- Model performance platform to keep models “fresh”
- Compare multiple models at once
- Assess model accuracy (Lift, ROC, K-S)
- Champion/challenger modeling
- Model retraining including open source

Enterprise Wish List

- Best possible models
- Model tournaments
- Visibility into performance
- Easy retraining
- Champion/challenger modelling
MODEL PERFORMANCE

Retrain models

Model performance reports

Monitor data drift

Model comparisons
THE FUTURE IS NOW...
SAS VIYA

SUPPORTING CURRENT INDUSTRY TRENDS

- Elastic
- Scalable
- Charge-back capable
- Micro-services architecture
- Easy installs
- Backward compatible
- Analytics lifecycle support
- Advanced machine learning
- Multi-threaded hyper-computing
- Memory spillover
- Integrated solutions
- End-to-end
- RESTful API's
- 'Any data, any platform'
- Python, Java, Lua support
- Plug n’ play
SAS AND OPEN SOURCE

EMBRACE
open source by including it and leveraging it where we can

EXTEND
open source by improving its interoperability and utility for the enterprise
THANK YOU
MATT MALCZEWSKI
MATT.MALCZEWSKI@SAS.COM
FOR MORE INFORMATION

Empowering the SAS/IML user with the functionality of R

**Documentation:** *IML User’s Guide - Calling Functions in the R Language*

**Video:** *Calling R Procedures from SAS/IML® Software*
https://www.youtube.com/watch?v=rUaTTTre24kl

**Video:** *SAS/IML and R: Using Them Together*
https://www.youtube.com/watch?v=nmRQ3MtkG6A

**Blogs:** *The DO Loop – R tags*
http://blogs.sas.com/content/iml/tag/r/

**Paper (p 14-17):** *Rediscovering SAS/IML® Software: Modern Data Analysis for the Practicing Statistician*

**Article:** *Versions of R that are supported by SAS/IML*
http://blogs.sas.com/content/iml/2013/09/16/what-versions-of-r-are-supported-by-sas.html
FOR MORE INFORMATION - EXTENDING R

Video: Using R in SAS Enterprise Miner
https://www.youtube.com/watch?v=TbXo0xQCqDw

Blogs: Spectral Clustering in SAS® Enterprise Miner™ Using Open Source Integration Node
https://communities.sas.com/docs/DOC-8011

Blogs: How to execute a Python script in SAS® Enterprise Miner™
https://communities.sas.com/docs/DOC-10832

Blogs: Open Source Integration Using the Base SAS Java Object
https://communities.sas.com/docs/DOC-10746

Article: The Open Source Integration node installation cheat sheet
https://communities.sas.com/docs/DOC-9988

Usage Notes:
FOR MORE INFORMATION MATERIALS ON GITHUB

Sas integration and sample code
Integration with R, Python
https://github.com/sassoftware/enlighten-integration

Integration with Jupyter Notebook and Python
https://github.com/sassoftware/sas_kernel
https://github.com/sassoftware/saspy

Sample codes of SAS Machine Learning methods
https://github.com/sassoftware/enlighten-apply

SAS Enterprise Miner process flow diagrams
https://github.com/sassoftware/dm-flow