Machine Learning and Apache Spark: A Dynamic Duo

John J Thomas
Distinguished Engineer & Director
IBM Analytics
@johnjaithomas
I love Machine Learning… What is it?

“Computers that learn without being explicitly programmed”
“Using algorithms to understand patterns in data”

\[ f(x) = \sum_{i=0}^{n} \alpha_i y_i x_i^T x + b \]
Top 10 Use Cases for Data Science & Machine Learning

- **Healthcare**: Patient Diagnosis
- **Finance**: Fraud Detection
- **Manufacturing**: Anomaly Detection
- **Retail**: Inventory Optimization
- **Insurance**: Client Risk Scoring
- **Transportation**: Demand Forecasting
- **Networks**: Intrusion Detection
- **E-Commerce**: Recommender Systems
- **Marketing**: Customer Segmentation
- **Energy**: Demand Forecasting
Transforming Manual, Expensive Processes...
...into new engagement models

Hi there! I am the new BlocPower Bot, here to help you make your church as energy efficient as possible.

Please take or upload a picture of your building from the outside for me so that I can take a look!

Great, thanks for that picture! Looks like your building is made of stone and has large windows.

I estimate your building has a high energy usage intensity (EUI), with a 97.01% probability...
Providing Celebrity Experiences!

Transaction History
Customer data

Since you’re going during the rainy season, I found a good offer for you:

Marmot Minimalist Rain Jacket - Men’s
$139.73 $200.00 Save $60.27 (30%)

Before you leave, we have something special just for you...

In appreciation of your loyalty, we will waive your foreign transaction fees for this trip!
Machine Learning itself is not new…

So what exactly is different now?
Forces driving massive interest in Machine Learning

- A wave of Open languages, frameworks
- Advances in compute performance
- Availability of different types of data
Buyer preferences favoring open vendors with flexible deployment options...

- Designed for experts
- Proprietary technology
- On-premise

- Designed for teams
- Open source w/ Extensions
- Deployment choice (public cloud, private cloud, desktop)
Machine learning technology is advancing quickly thanks to open source
<table>
<thead>
<tr>
<th>Commitment to contribute</th>
<th>Commitment to adopt</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM is one of the largest contributors to Spark</td>
<td>Spark is IBM’s ‘Analytics DNA’</td>
</tr>
<tr>
<td>Spark Technology Center</td>
<td>Off-premises and On-premises</td>
</tr>
<tr>
<td>Contribution to Machine Learning</td>
<td>Spark is integrated and embedded across the IBM analytics portfolio</td>
</tr>
<tr>
<td>Spark ML</td>
<td>Spark as a Service</td>
</tr>
<tr>
<td>Spark MLLib</td>
<td>Data Connect</td>
</tr>
<tr>
<td>R4ML</td>
<td>dashDB Analytics</td>
</tr>
<tr>
<td>Online Retraining</td>
<td>Data Science Analytics</td>
</tr>
<tr>
<td>Apache Arrow</td>
<td>z Systems</td>
</tr>
<tr>
<td>SystemML</td>
<td>Deep Learning</td>
</tr>
</tbody>
</table>
Commitment to Contribute

38,500 Spark LOC
863 Spark JIRAs
253 SystemML JIRAs
422 Commits in Spark 2.0

Take a closer look at the Top 3
Commitment to Adopt Cloud Services with Spark Connectors to data sources Platforms that run Spark

Spark embedded across portfolio
Making Data Science a Team Sport

A Growing Set of Data Science Tools

Jupyter Notebooks
Create and collaborate on Python, R, and Scala notebooks that contain code and visualizations.

RSstudio
Jumpstart your R experience with a free, open-source RStudio tool.

Configured estimators

- Logistic Regression
  - Trained
- Decision Tree Classifier
  - Trained
- Gradient Boosted Tree Classifier
  - Trained
- Naive Bayes
  - Trained
- Random Forest Classifier
Spark, Machine Learning and Blockchain

Blockchain
Utilize IBM's Blockchain within Bluemix

Cloudant NoSQL DB
Cloudant NoSQL DB is a fully managed data layer designed for IBM

Spark
Apache Spark and IBM Watson

Combine the power of Spark Machine Learning with IBM Watson’s Deep Learning

Create compelling Cognitive Applications
Try Machine Learning with Apache Spark!

Questions?
jothomas@us.ibm.com
@johnjaithomas

ML Hub:
MLHub@us.ibm.com

Machine Learning :
http://ibm.com/machine-learning/

Data Science :
http://datascience.ibm.com