Hands on Introduction to IBM Data Science Experience

Power of data. Simplicity of design. Speed of innovation.

Asad Mahmood
Jean Bright
Charles Morrison

Agenda

Overview of DSX, Watson Data Platform and Use Case (Wednesday Feb. 28, 2018)

3:00 PM - 4:00 PM	**Introduction to IBM Data Science Experience and Watson Data Platform**
4:00 PM - 5:00 PM	**Overview of Use Case and Solution Approach**

Hands-On DSX Labs (Thursday March 1, 2018)

9:00 AM - 10:30 AM	**Lab 1 - Setting Up Your DSX Environment and Exploratory Data Analysis**
10:30 AM - 12:00 PM	**Lab 2 - Data Visualization with R Studio and Shiny**
12:00 PM - 12:30 PM	**Lunch**
12:30 PM - 1:30 PM	**Lab 3 - Building a Predictive Model with Watson Machine Learning**
1:30 PM - 2:00 PM	**Next steps regarding POC and project timeline**

Participant Background

Open Source

- R/Python/Scala
- Jupyter Notebook
- Spark
- Hadoop

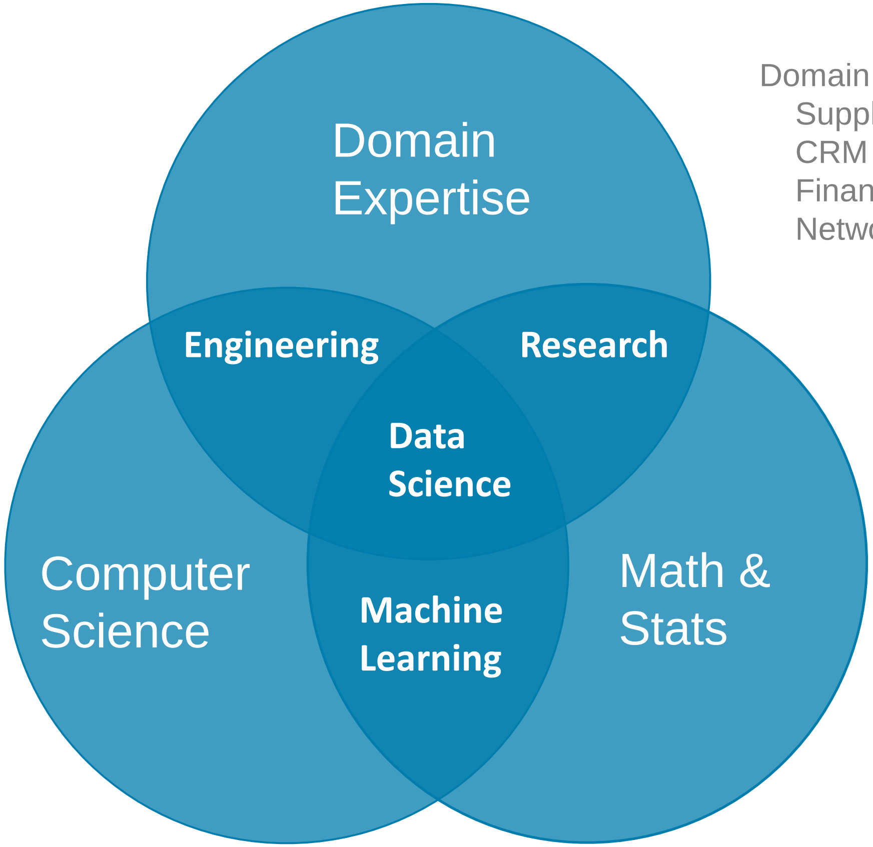
IBM

- Bluemix
- Data Science Experience
- Watson Machine Learning

What is Data Science?

Domain Knowledge
Supply Chain
CRM
Financials
Networking

Scripting, SQL
Python, R Scala
Data Pipelines
Big Data/ Apache Spark

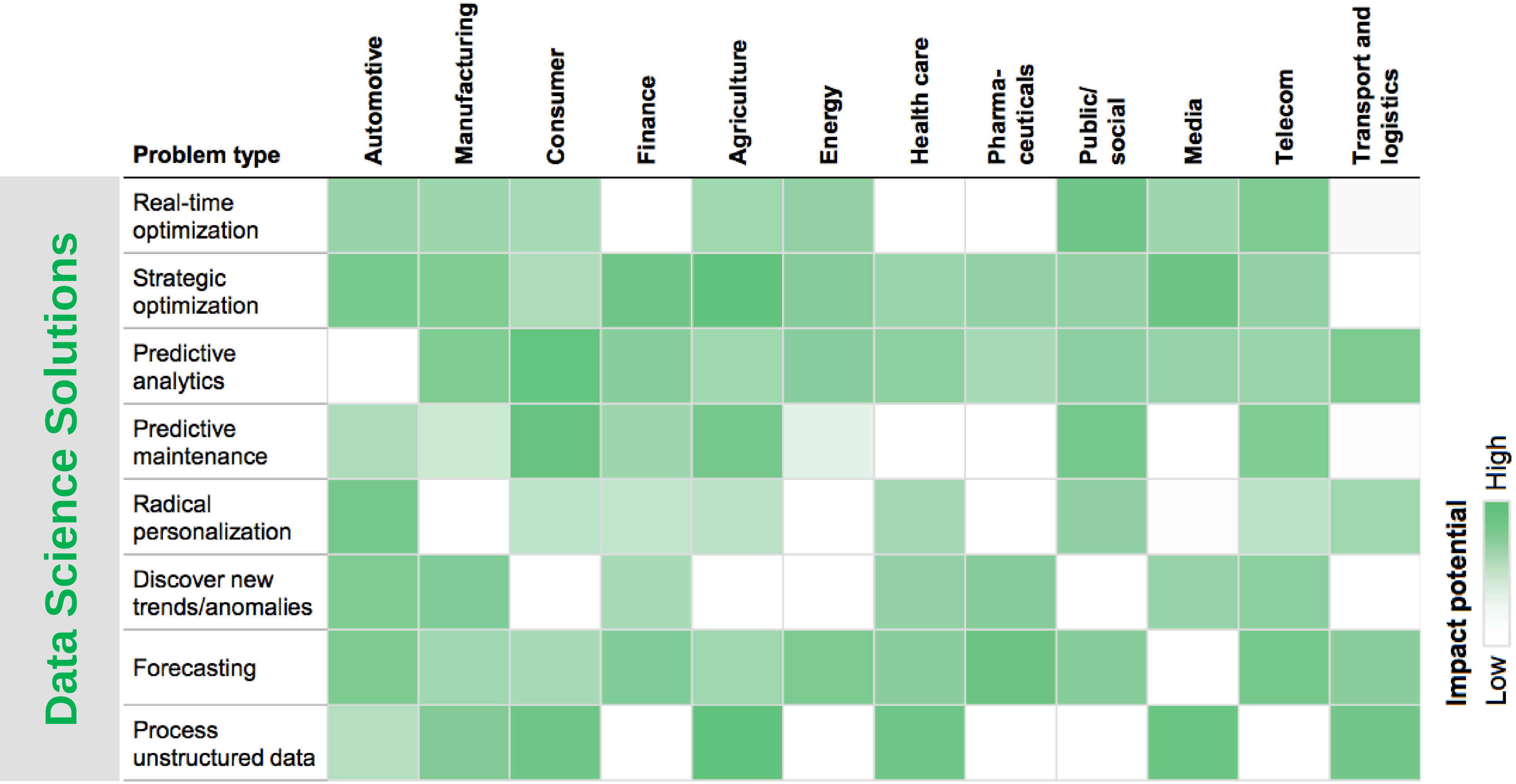


Mathematics
Computational

Data Science Projects Require Multiple Skills

Data Science Impact Across Industries and Use Cases

$10s of Billions in each industry and use case



SOURCE: McKinsey Global Institute analysis

Challenges in delivering value with Data Science

Data

- Data resides in silos and difficult to access
- Detailed data was never stored
- Unstructured and external data wasn't considered

Governance

- Self-service isn't a reality, if the data isn't secure
- Understanding lineage and getting to a system of truth

Skills

- Data Science skills are in low supply and high demand
- Nurturing new data professionals is challenging

Infrastructure

- Need an environment that enables collaboration and deployment to production
- Discrete tools present barriers to progress

Watson Data Platform

IBM Watson Data Platform

Mission: Make Data Simple and Accessible to All

Platform.

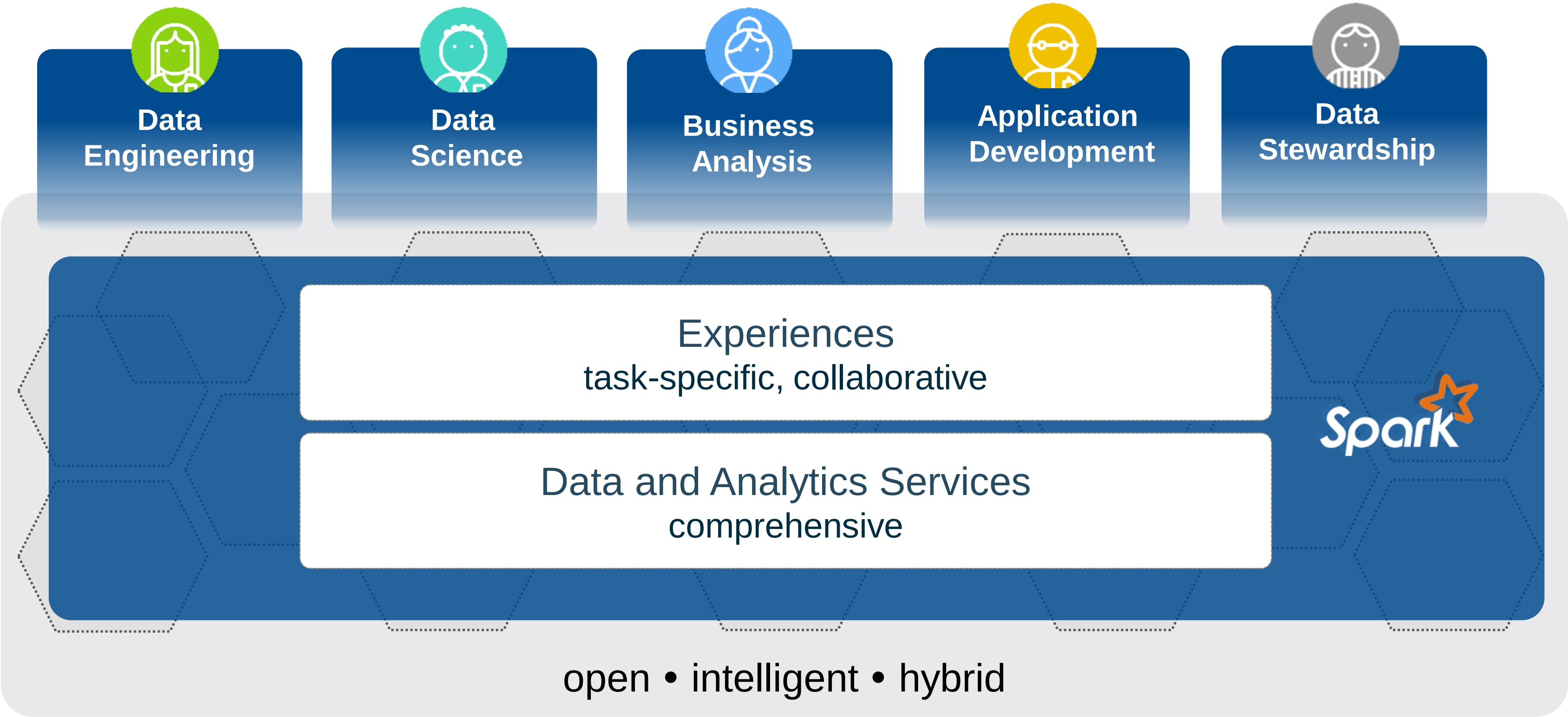
Method.

Ecosystem.

http://ibm.co/makedatasimple

IBM Watson Data Platform

Experience New Ways To Put Data To Work



Data Engineering

Data Science

Business Analysis

Application Development

Data Stewardship

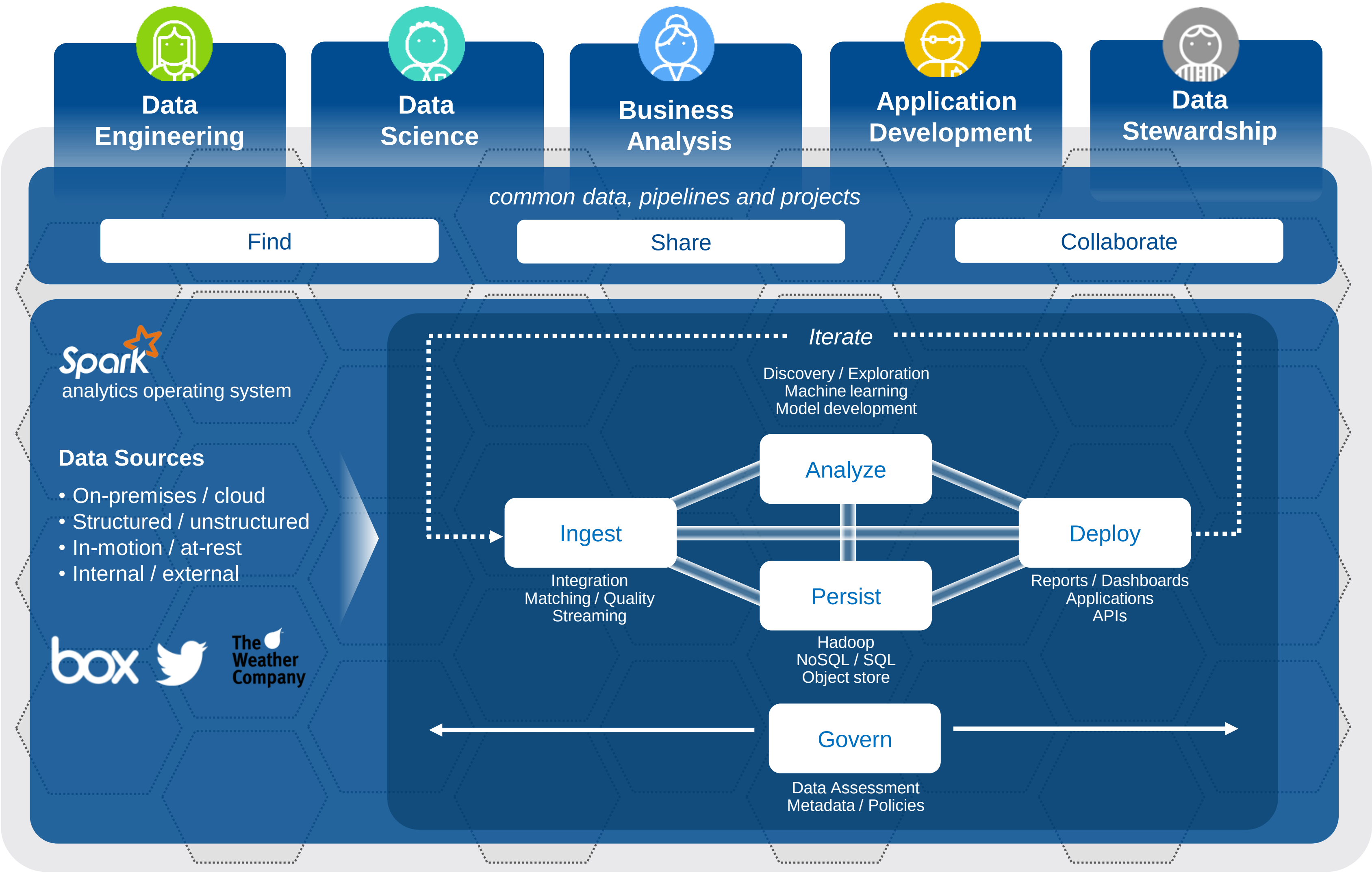
Experiences
task-specific, collaborative

Data and Analytics Services
comprehensive

open • intelligent • hybrid

IBM Watson Data Platform

Connects Users to Data and Analytics



Data Engineering

Data Science

Business Analysis

Application Development

Data Stewardship

common data, pipelines and projects

Find

Share

Collaborate

Spark analytics operating system

Data Sources

- On-premises / cloud
- Structured / unstructured
- In-motion / at-rest
- Internal / external

box

The Weather Company

Iterate

Discovery / Exploration
Machine learning
Model development

Analyze

Ingest

Integration
Matching / Quality
Streaming

Deploy

Reports / Dashboards
Applications
APIs

Persist

Hadoop
NoSQL / SQL
Object store

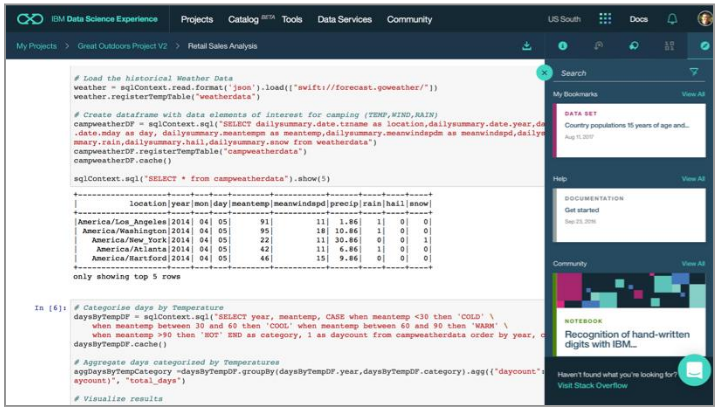
Govern

Data Assessment
Metadata / Policies

Data Science Experience

Brings together everything a Data Scientist needs to be successful

Data Scientist



```
# Load the historical Weather Data
weather = sqlContext.read.format('json').load(["swift://forecast.goweather/"])
weather.registerTempTable("weatherdata")

# Create dataframe with data elements of interest for camping (TEMP,WIND,RAIN)
campweatherDF = sqlContext.sql("SELECT dailysummary.date.tzname as location,dailysummary.date.year,da
.date.mday as day, dailysummary.meantempm as meantemp,dailysummary.meanwindspdm as meanwindspd,dailys
mmary.rain,dailysummary.hail,dailysummary.snow from weatherdata")
campweatherDF.registerTempTable("campweatherdata")
campweatherDF.cache()

sqlContext.sql("SELECT * from campweatherdata").show(5)
```

```
+-------------------+----+---+---+--------+-----------+------+----+----+----+
|           location|year|mon|day|meantemp|meanwindspd|precip|rain|hail|snow|
+-------------------+----+---+---+--------+-----------+------+----+----+----+
|America/Los_Angeles|2014| 04| 05|      91|         11|  1.86|   1|   0|   0|
| America/Washington|2014| 04| 05|      95|         18| 10.86|   1|   0|   0|
|   America/New_York|2014| 04| 05|      22|         11| 30.86|   0|   0|   1|
|    America/Atlanta|2014| 04| 05|      42|         11|  6.86|   1|   0|   0|
|   America/Hartford|2014| 04| 05|      46|         15|  9.86|   0|   0|   0|
+-------------------+----+---+---+--------+-----------+------+----+----+----+
only showing top 5 rows
```

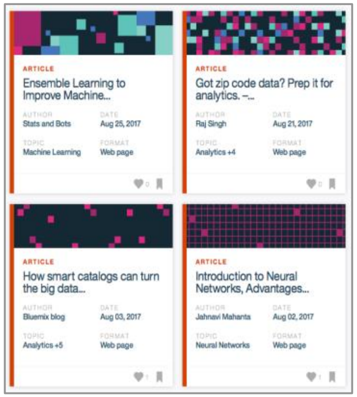
```
In [6]: # Categorise days by Temperature
daysByTempDF = sqlContext.sql("SELECT year, meantemp, CASE when meantemp <30 then 'COLD' \
    when meantemp between 30 and 60 then 'COOL' when meantemp between 60 and 90 then 'WARM' \
    when meantemp >90 then 'HOT' END as category, 1 as daycount from campweatherdata order by year, c
daysByTempDF.cache()

# Aggregate days categorized by Temperatures
aggDaysByTempCategory =daysByTempDF.groupBy(daysByTempDF.year,daysByTempDF.category).agg({"daycount":
aycount)", "total_days")

# Visualize results
```

DATA SET
Country populations 15 years of age and...
Aug 11, 2017

DOCUMENTATION
Get started
Sep 23, 2016

NOTEBOOK
Recognition of hand-written digits with IBM...

Haven't found what you're looking for?
Visit Stack Overflow

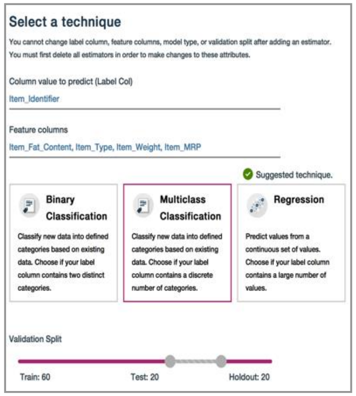


ARTICLE
Ensemble Learning to Improve Machine...
AUTHOR Stats and Bots
DATE Aug 25, 2017
TOPIC Machine Learning
FORMAT Web page

ARTICLE
Got zip code data? Prep it for analytics. –...
AUTHOR Raj Singh
DATE Aug 21, 2017
TOPIC Analytics +4
FORMAT Web page

ARTICLE
How smart catalogs can turn the big data...
AUTHOR Bluemix blog
DATE Aug 03, 2017
TOPIC Analytics +5
FORMAT Web page

ARTICLE
Introduction to Neural Networks, Advantages...
AUTHOR Jahnavi Mahanta
DATE Aug 02, 2017
TOPIC Neural Networks
FORMAT Web page



Select a technique

You cannot change label column, feature columns, model type, or validation split after adding an estimator. You must first delete all estimators in order to make changes to these attributes.

Column value to predict (Label Col)
Item_Identifier

Feature columns
Item_Fat_Content, Item_Type, Item_Weight, Item_MRP

✓ Suggested technique.

Binary Classification
Classify new data into defined categories based on existing data. Choose if your label column contains two distinct categories.

Multiclass Classification
Classify new data into defined categories based on existing data. Choose if your label column contains a discrete number of categories.

Regression
Predict values from a continuous set of values. Choose if your label column contains a large number of values.

Validation Split
Train: 60 Test: 20 Holdout: 20

Learn

Built-in learning to get started or go the distance with advanced tutorials

Create

The best of open source and IBM value-add to create state-of-the-art data products

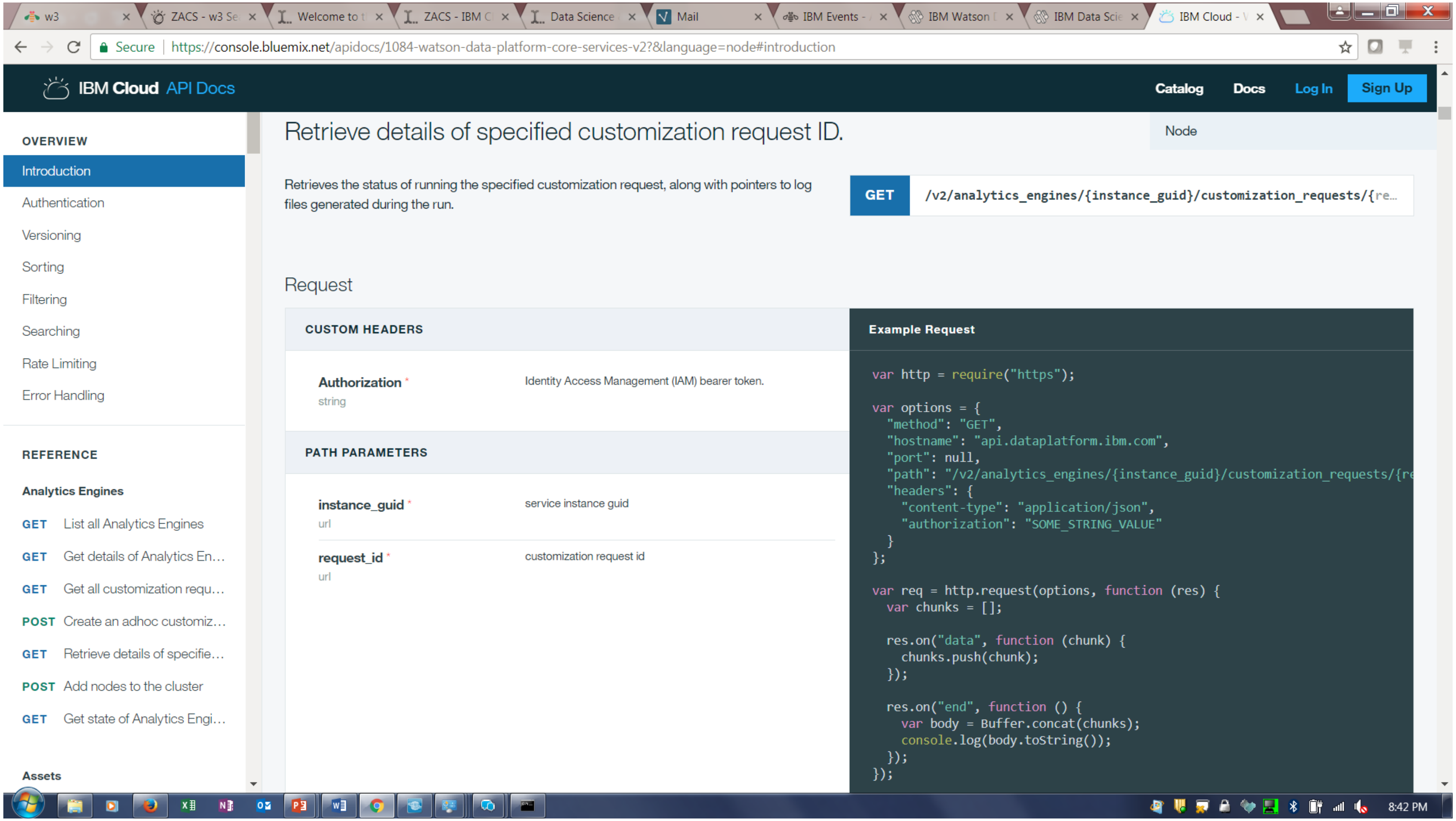
Collaborate

Data and Analytic assets are contained within projects which can be shared with other users.

IBM Cloud PaaS

Application Developer

Rich Platform and Service APIs for your developers


w3
ZACS - w3 Se
Welcome to t
ZACS - IBM C
Data Science
Mail
IBM Events -
IBM Watson
IBM Data Scie
IBM Cloud -
Secure
https://console.bluemix.net/apidocs/1084-watson-data-platform-core-services-v2?&language=node#introduction
IBM Cloud API Docs
Catalog
Docs
Log In
Sign Up
OVERVIEW
Introduction
Authentication
Versioning
Sorting
Filtering
Searching
Rate Limiting
Error Handling
REFERENCE
Analytics Engines
GET List all Analytics Engines
GET Get details of Analytics En…
GET Get all customization requ…
POST Create an adhoc customiz…
GET Retrieve details of specifie…
POST Add nodes to the cluster
GET Get state of Analytics Engi…
Assets
Node
8:42 PM

Retrieve details of specified customization request ID.

Retrieves the status of running the specified customization request, along with pointers to log files generated during the run.

GET /v2/analytics_engines/{instance_guid}/customization_requests/{re…

Request

CUSTOM HEADERS

Authorization * string — Identity Access Management (IAM) bearer token.

PATH PARAMETERS

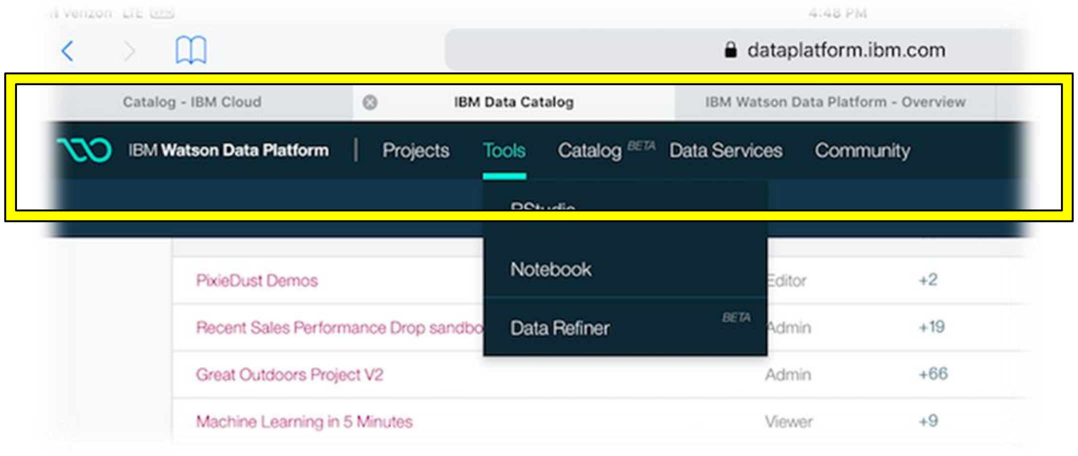
instance_guid * url — service instance guid

request_id * url — customization request id

Example Request

```
var http = require("https");

var options = {
  "method": "GET",
  "hostname": "api.dataplatform.ibm.com",
  "port": null,
  "path": "/v2/analytics_engines/{instance_guid}/customization_requests/{re
  "headers": {
    "content-type": "application/json",
    "authorization": "SOME_STRING_VALUE"
  }
};

var req = http.request(options, function (res) {
  var chunks = [];

  res.on("data", function (chunk) {
    chunks.push(chunk);
  });

  res.on("end", function () {
    var body = Buffer.concat(chunks);
    console.log(body.toString());
  });
});
```

Intelligent data fabric provides consistent platform experience



PixieDust Demos	Editor	+2
Recent Sales Performance Drop sandbo	Admin	+19
Great Outdoors Project V2	Admin	+66
Machine Learning in 5 Minutes	Viewer	+9

This fabric remains consistent throughout the Watson Data Platform experience – regardless if you are ingesting data, shaping data, building algorithms, deploying models and more…

How does WDP help fulfill the promise of your data?

Data

Puts every important data source at the fingertips of the teams that need it wherever resides

Governance

Enforces your policies without getting in the way of delivering insights

Skills

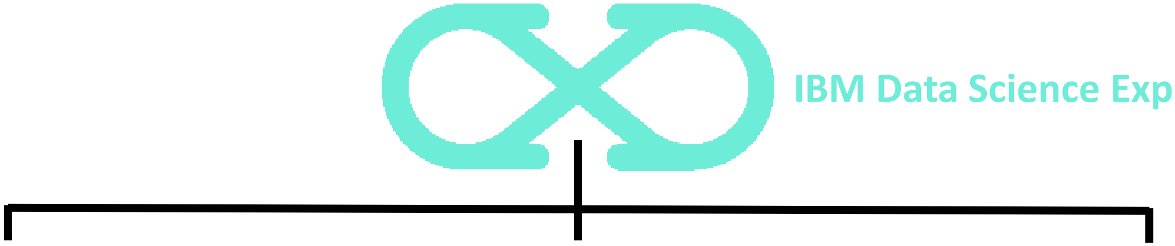
Makes the most of the data professionals you have and helps them grow and learn from each other as a team

Infrastructure

Delivers the foundation for your first data project through to the complete transformation of your business

Data Science Experience

Core Attributes of the Data Science Experience



IBM Data Science Experience

Community

- Find tutorials and datasets
- Read articles and papers
- Connect with Data Scientists
- Share comments
- Copy and share notebooks

Open Source

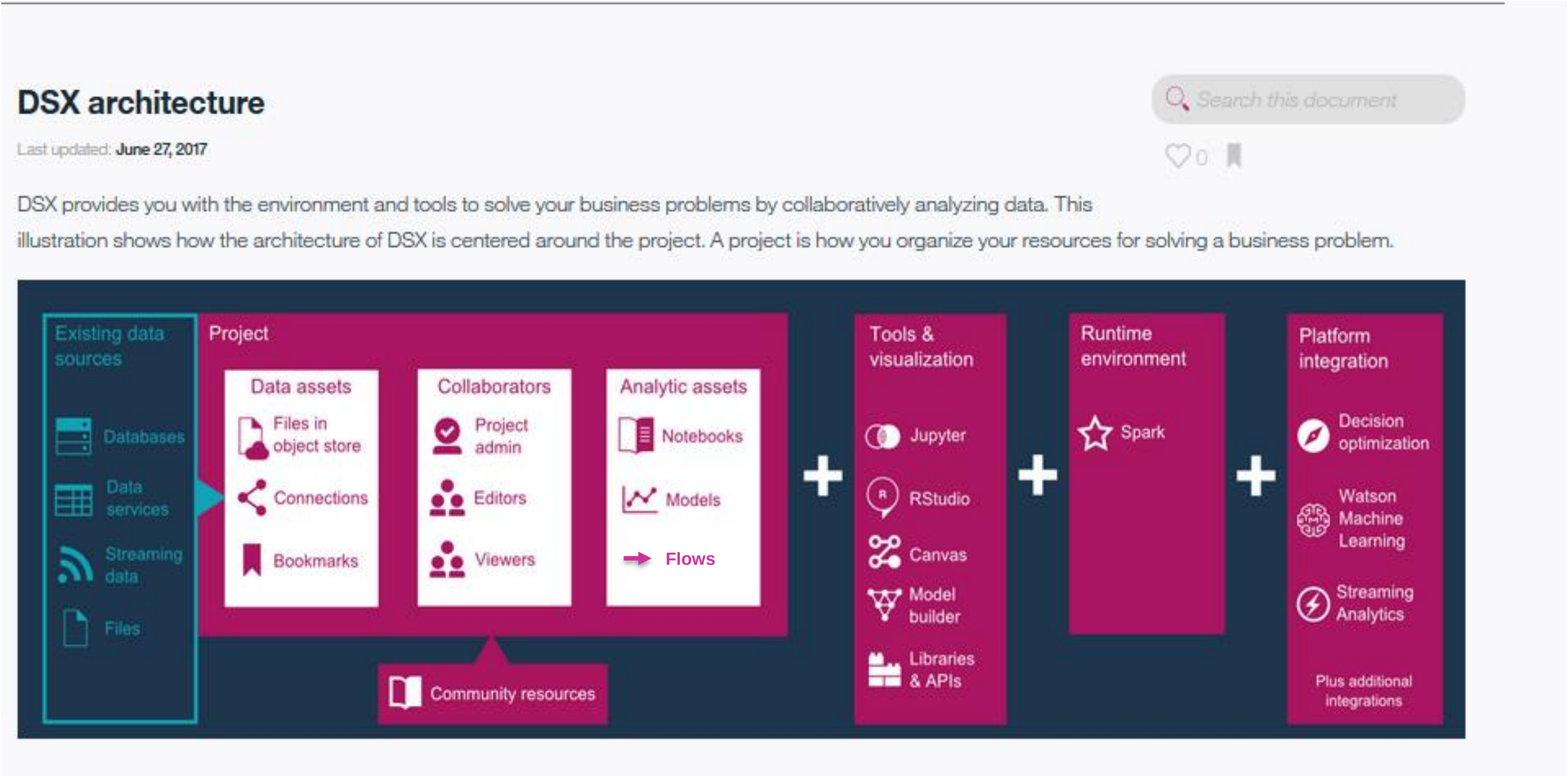
- Code in Scala/Python/R/SQL
- Jupyter Notebooks
- RStudio IDE and Shiny
- Apache Spark
- Your favorite libraries

IBM Added Value

- IBM Machine Learning
- SPSS Modeler Canvas
- Prescriptive Analytics - DOcplexcloud
- Projects and Version Control
- Managed Spark Service

Powered by IBM **Watson Data Platform**

DSX Architecture



DSX architecture

Last updated: June 27, 2017

DSX provides you with the environment and tools to solve your business problems by collaboratively analyzing data. This illustration shows how the architecture of DSX is centered around the project. A project is how you organize your resources for solving a business problem.

Existing data sources
- Databases
- Data services
- Streaming data
- Files

Project
- Data assets: Files in object store; Connections; Bookmarks
- Collaborators: Project admin; Editors; Viewers
- Analytic assets: Notebooks; Models; Flows

Community resources

+

Tools & visualization
- Jupyter
- RStudio
- Canvas
- Model builder
- Libraries & APIs

+

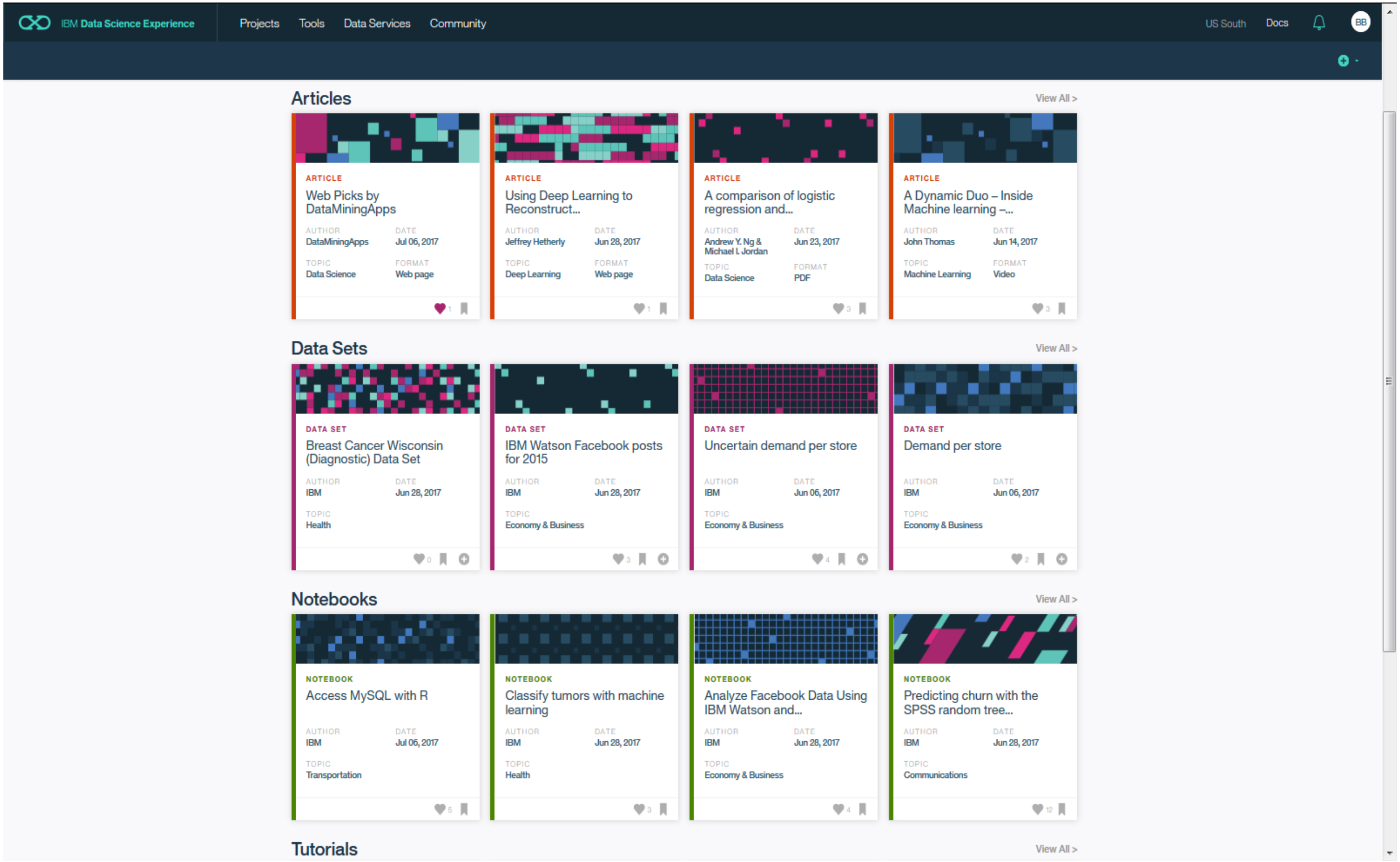
Runtime environment
- Spark

+

Platform integration
- Decision optimization
- Watson Machine Learning
- Streaming Analytics

Plus additional integrations

Community Cards provide in-context learning



Articles

View All >

ARTICLE
Web Picks by DataMiningApps
AUTHOR DataMiningApps · DATE Jul 06, 2017
TOPIC Data Science · FORMAT Web page
♥ 1

ARTICLE
Using Deep Learning to Reconstruct...
AUTHOR Jeffrey Hetherly · DATE Jun 28, 2017
TOPIC Deep Learning · FORMAT Web page
♥ 1

ARTICLE
A comparison of logistic regression and...
AUTHOR Andrew Y. Ng & Michael I. Jordan · DATE Jun 23, 2017
TOPIC Data Science · FORMAT PDF
♥ 3

ARTICLE
A Dynamic Duo – Inside Machine learning –...
AUTHOR John Thomas · DATE Jun 14, 2017
TOPIC Machine Learning · FORMAT Video
♥ 3

Data Sets

View All >

DATA SET
Breast Cancer Wisconsin (Diagnostic) Data Set
AUTHOR IBM · DATE Jun 28, 2017
TOPIC Health
♥ 0

DATA SET
IBM Watson Facebook posts for 2015
AUTHOR IBM · DATE Jun 28, 2017
TOPIC Economy & Business
♥ 3

DATA SET
Uncertain demand per store
AUTHOR IBM · DATE Jun 06, 2017
TOPIC Economy & Business
♥ 4

DATA SET
Demand per store
AUTHOR IBM · DATE Jun 06, 2017
TOPIC Economy & Business
♥ 2

Notebooks

View All >

NOTEBOOK
Access MySQL with R
AUTHOR IBM · DATE Jul 06, 2017
TOPIC Transportation
♥ 5

NOTEBOOK
Classify tumors with machine learning
AUTHOR IBM · DATE Jun 28, 2017
TOPIC Health
♥ 3

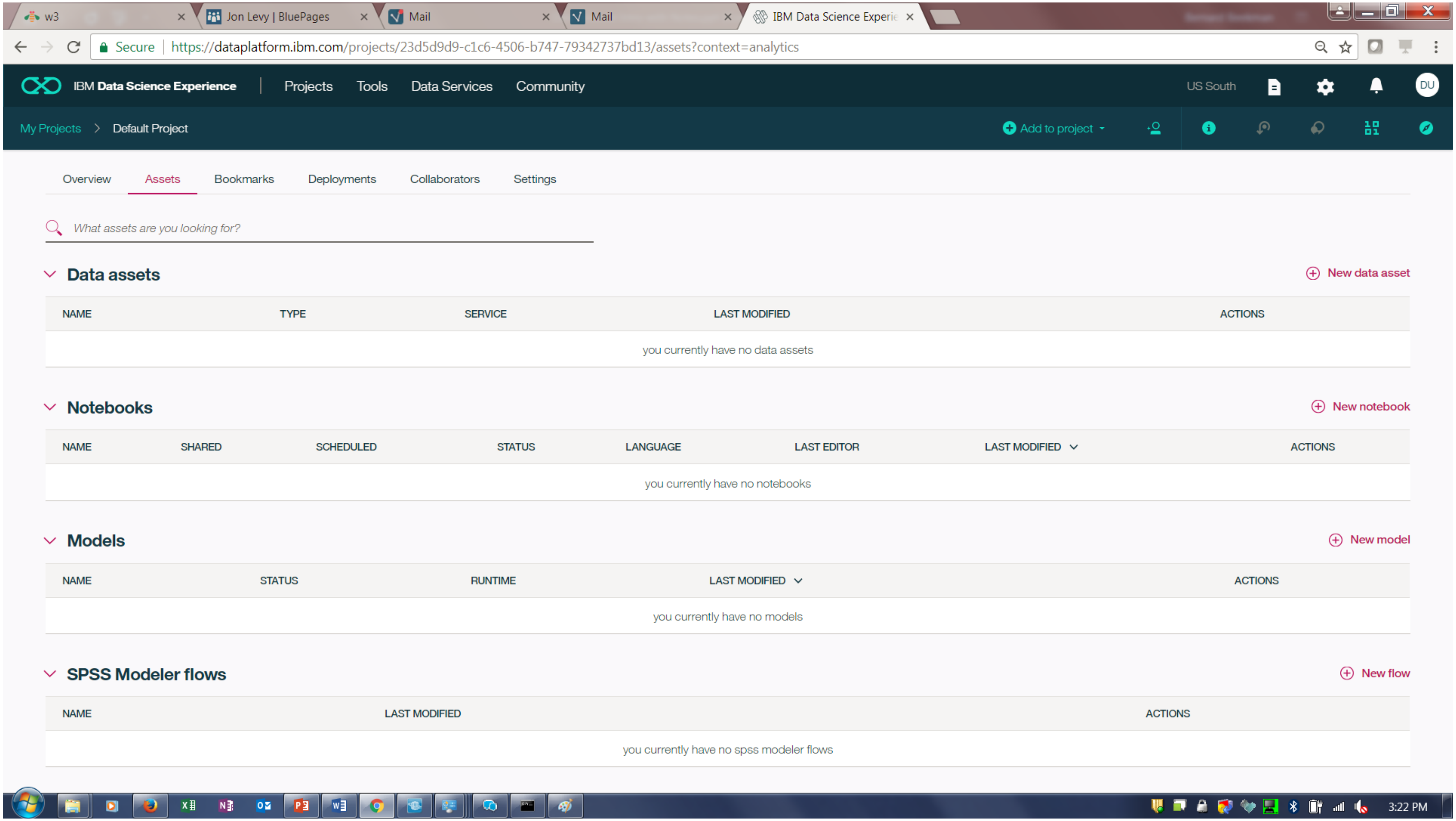
NOTEBOOK
Analyze Facebook Data Using IBM Watson and...
AUTHOR IBM · DATE Jun 28, 2017
TOPIC Economy & Business
♥ 4

NOTEBOOK
Predicting churn with the SPSS random tree...
AUTHOR IBM · DATE Jun 28, 2017
TOPIC Communications
♥ 12

Tutorials

View All >

Collaborate Using Projects



Data assets

New data asset

NAME	TYPE	SERVICE	LAST MODIFIED	ACTIONS
you currently have no data assets				

Notebooks

New notebook

NAME	SHARED	SCHEDULED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS
you currently have no notebooks							

Models

New model

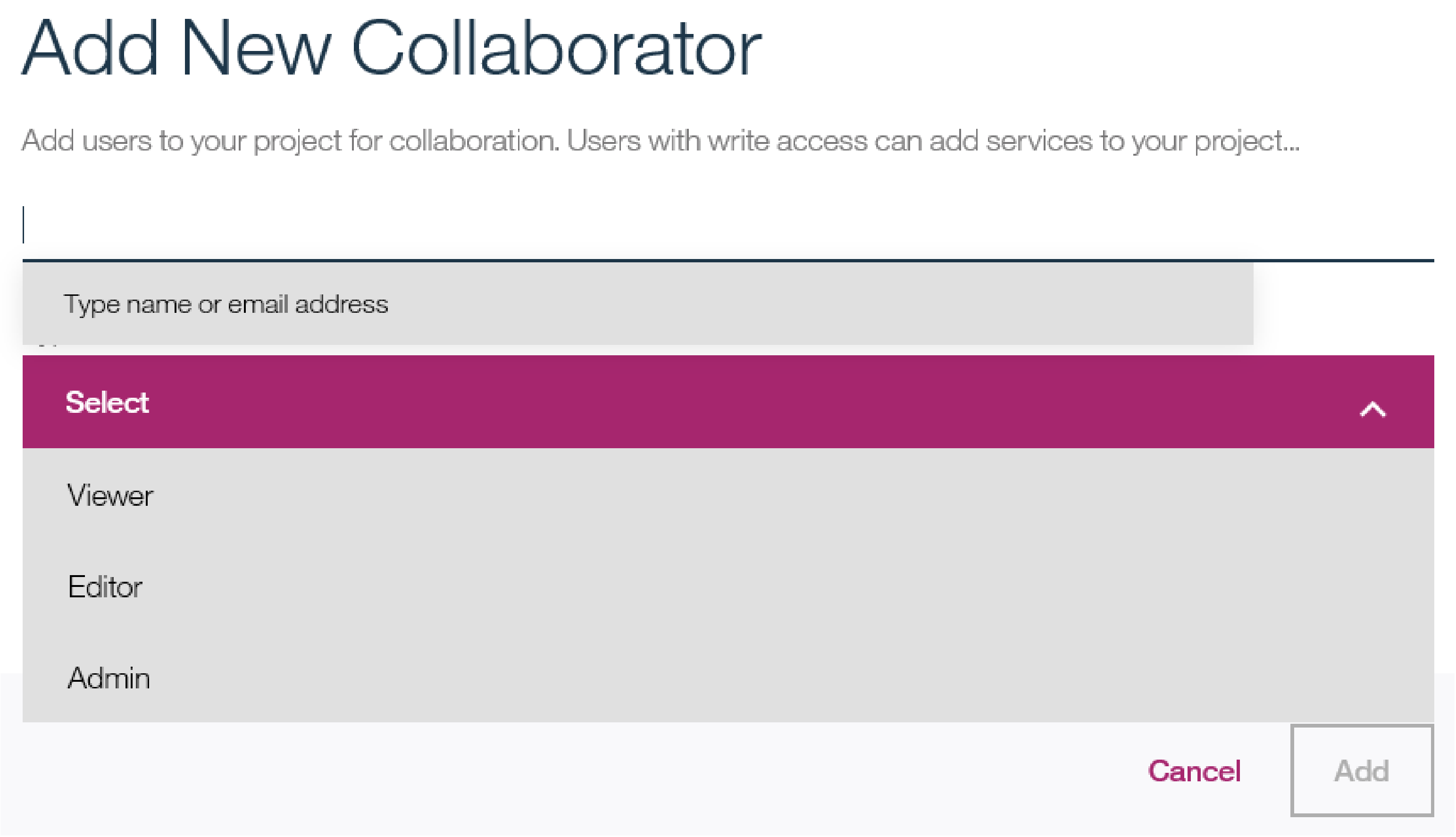
NAME	STATUS	RUNTIME	LAST MODIFIED	ACTIONS
you currently have no models				

SPSS Modeler flows

New flow

NAME	LAST MODIFIED	ACTIONS
you currently have no spss modeler flows		

Add Collaborators to a Project

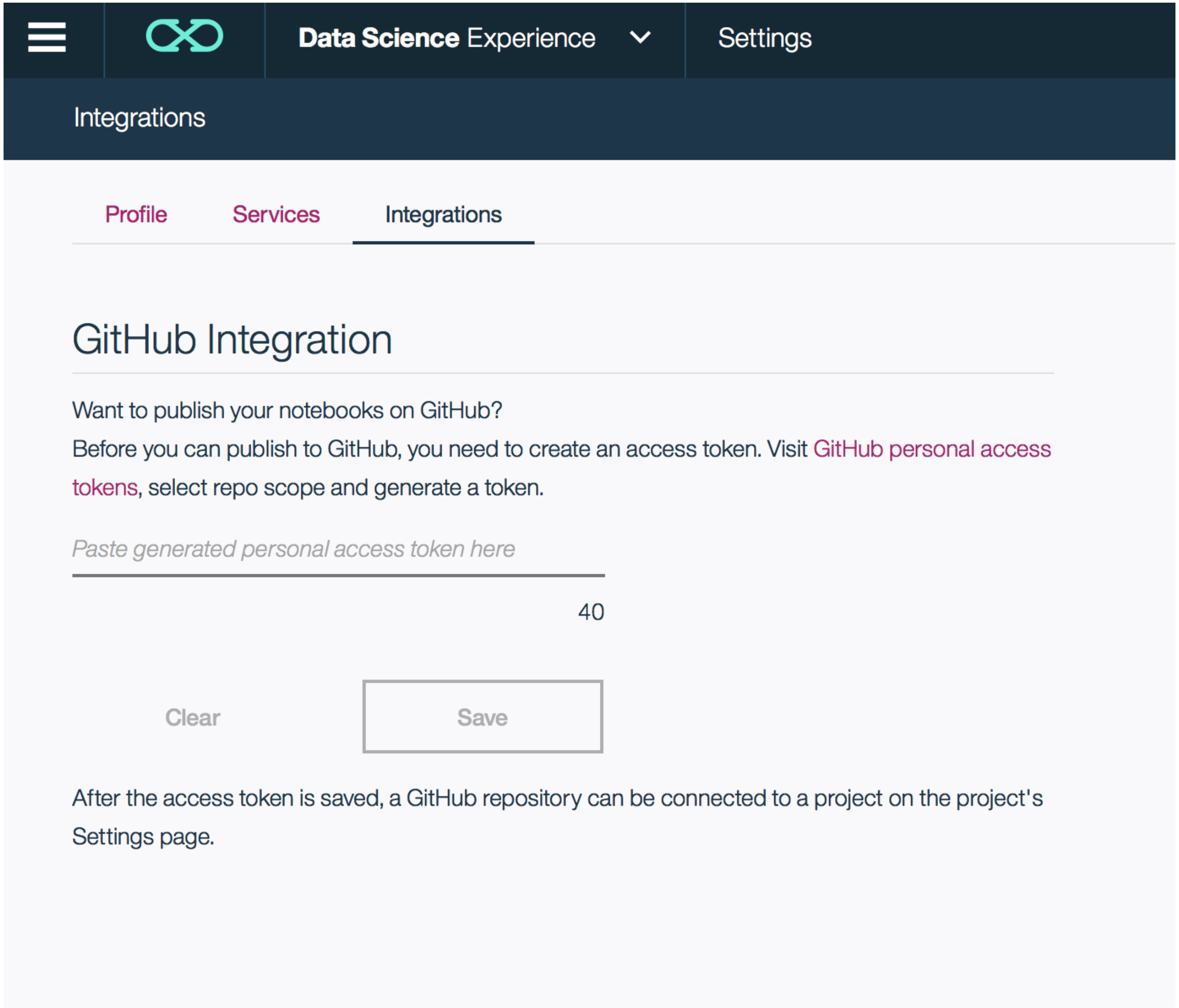


Add New Collaborator

Add users to your project for collaboration. Users with write access can add services to your project...

Type name or email address

GitHub Integration



GitHub Integration

Want to publish your notebooks on GitHub?

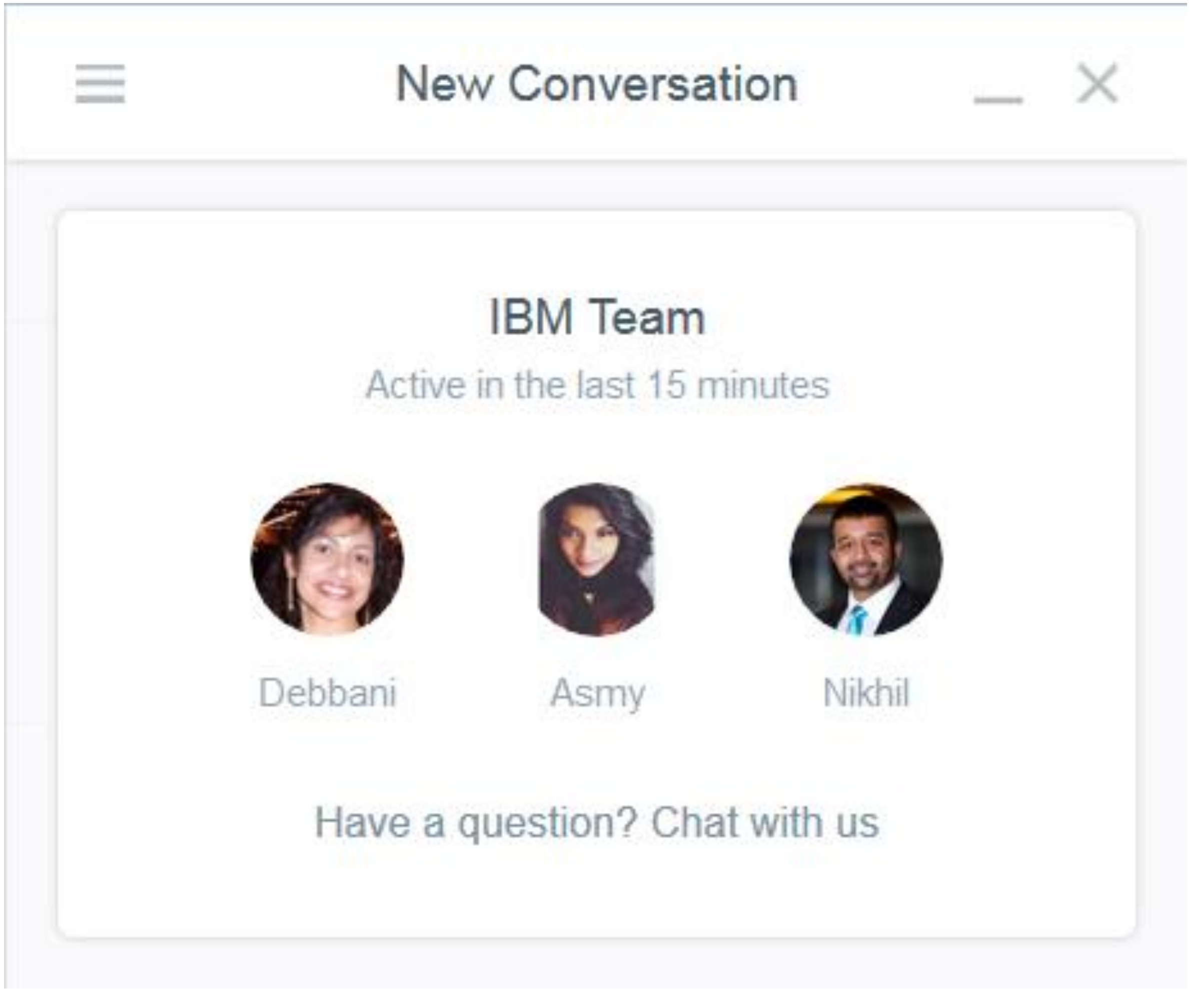
Before you can publish to GitHub, you need to create an access token. Visit GitHub personal access tokens, select repo scope and generate a token.

Paste generated personal access token here

40

After the access token is saved, a GitHub repository can be connected to a project on the project's Settings page.

Live chat on Intercom for support from the IBM team and to provide your feedback on how we can improve



New Conversation

IBM Team

Active in the last 15 minutes

Debbani

Asmy

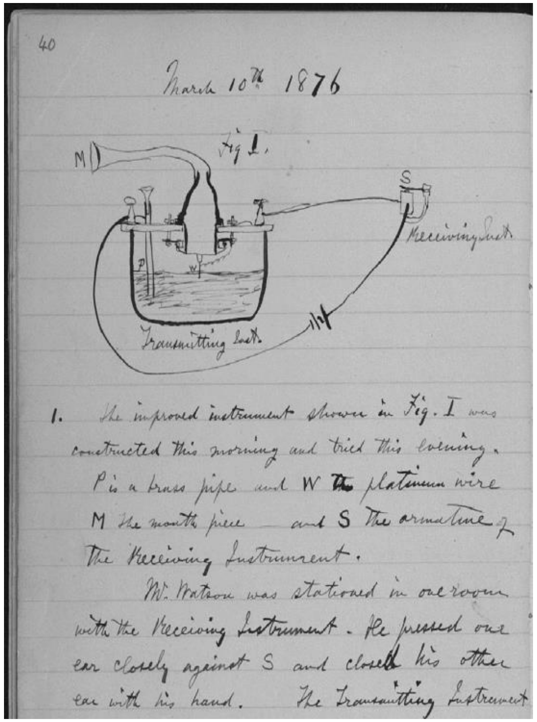
Nikhil

Have a question? Chat with us

What is a “Notebook”?

Pen and Paper

- Pen and paper has long provided the rich experience that scientists need to document progress through notes and drawings:
  - Expressive
  - Cumulative
  - Collaborative



40

March 10th 1876

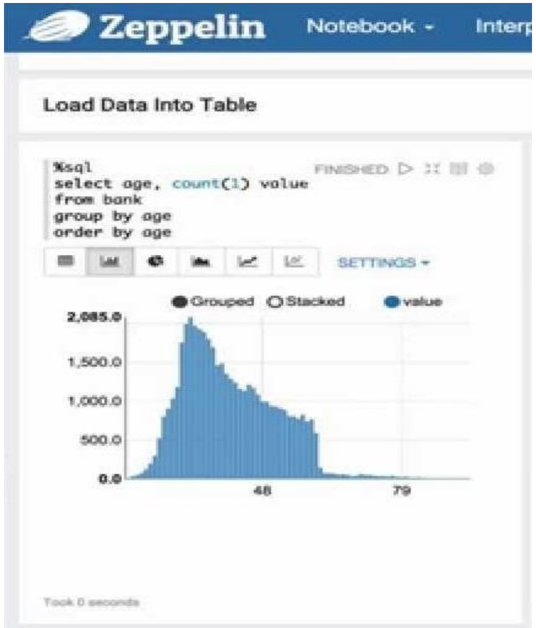
Fig I.
M
S
P
W
Receiving Inst.
Transmitting Inst.

1. The improved instrument shown in Fig. I was constructed this morning and tried this evening. P is a brass pipe and W the platinum wire M the mouth piece — and S the armature of The Receiving Instrument.

Mr. Watson was stationed in one room with the Receiving Instrument. He pressed one ear closely against S and closed his other ear with his hand. The Transmitting Instrument

Notebooks

- Notebooks are the digital equivalent of the “pen and paper” lab notebook, enabling data scientists to document reproducible analysis:
  - Markdown and visualization
  - Iterative exploration
  - Easy to share

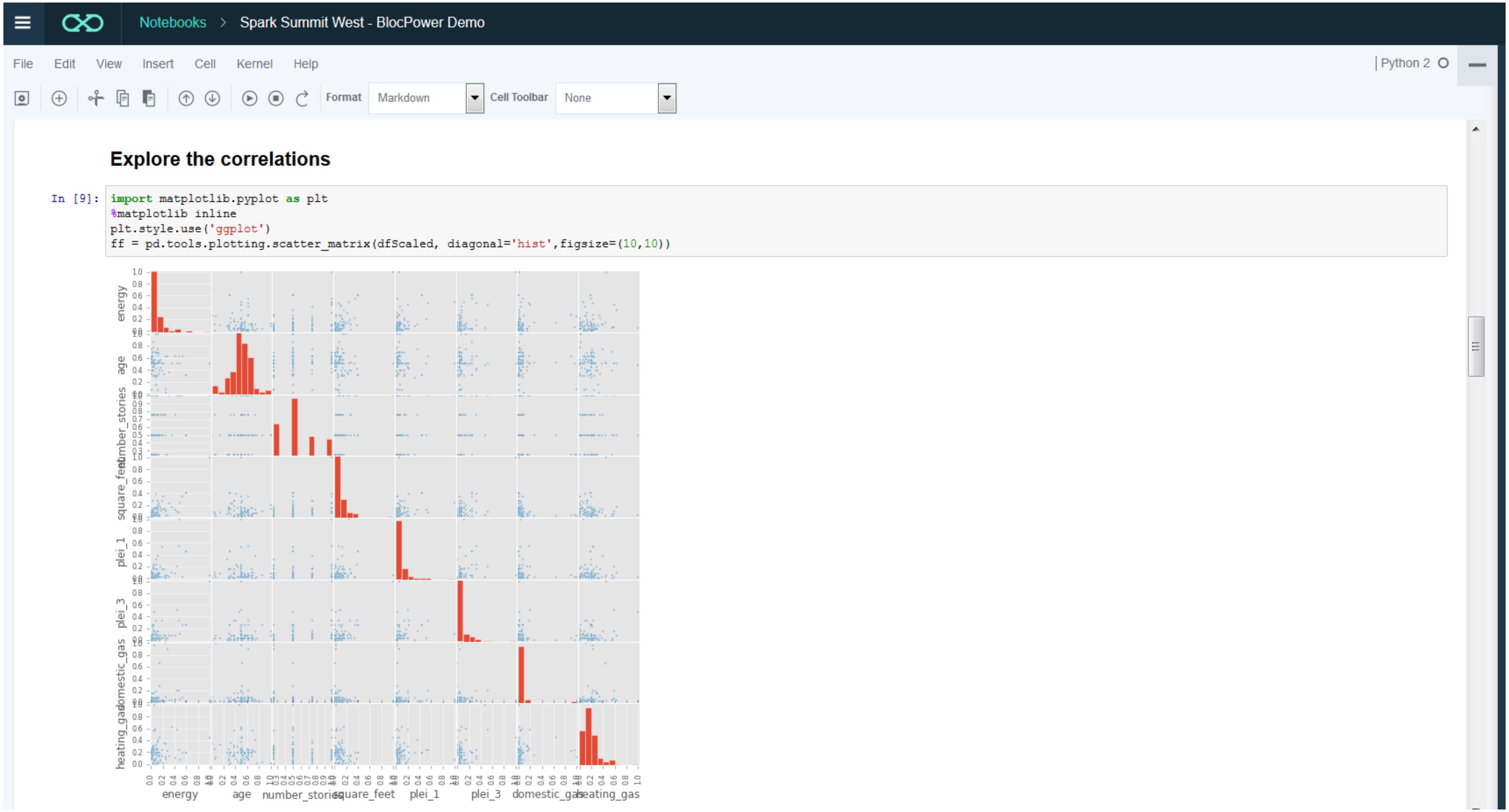


Load Data Into Table

```
%sql
select age, count(1) value
from bank
group by age
order by age
```

Took 0 seconds

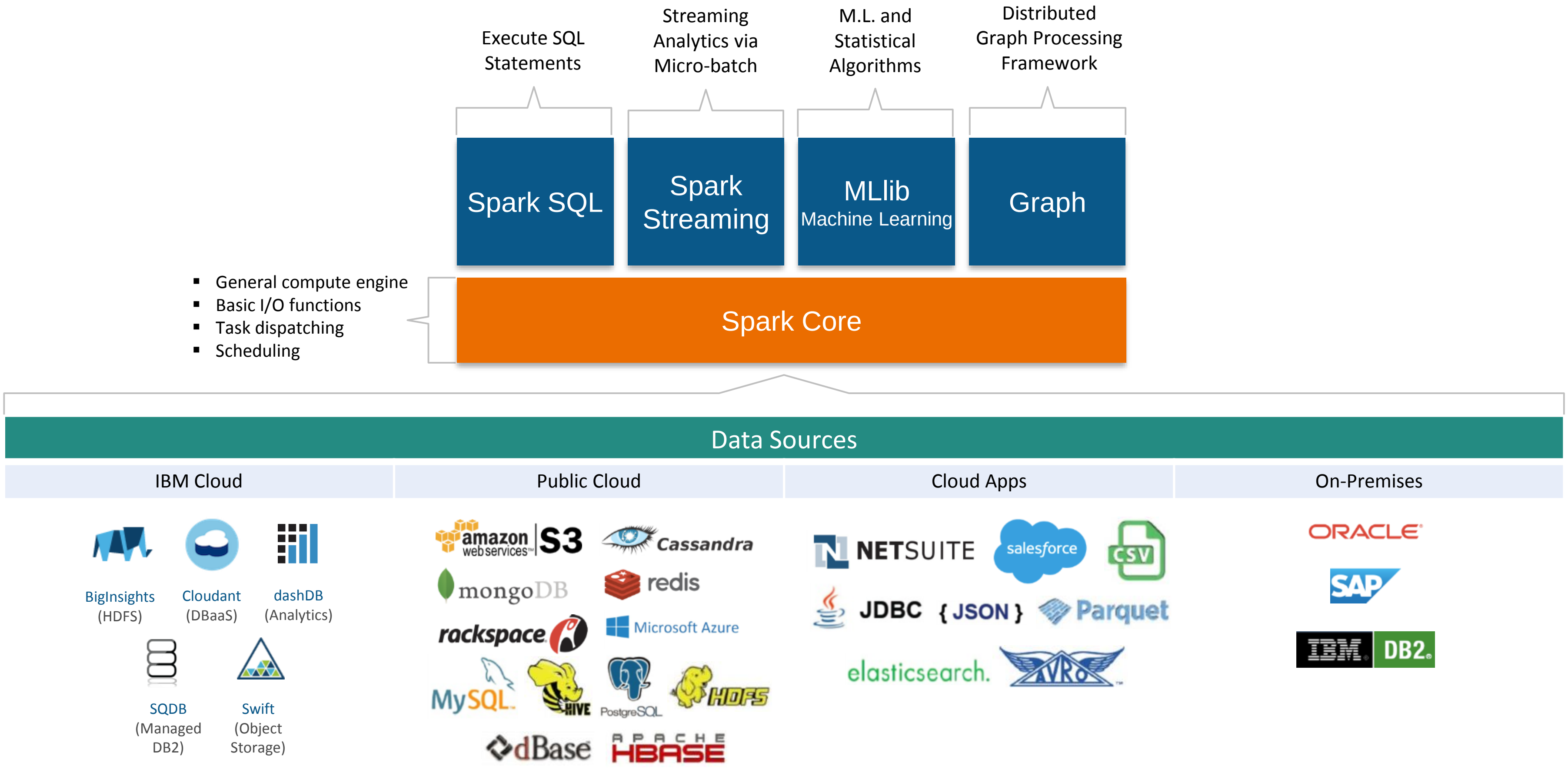
Integrated Jupyter Notebooks for interactive and collaborative development - seamless execution on Spark



Explore the correlations

In [9]:
```
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('ggplot')
ff = pd.tools.plotting.scatter_matrix(dfScaled, diagonal='hist',figsize=(10,10))
```

From a Notebook in DSX you can use IBM's managed Spark Service to blend multiple data types, sources, and workloads



Execute SQL Statements

Streaming Analytics via Micro-batch

M.L. and Statistical Algorithms

Distributed Graph Processing Framework

Spark SQL

Spark Streaming

MLlib Machine Learning

Graph

- General compute engine
- Basic I/O functions
- Task dispatching
- Scheduling

Spark Core

Data Sources

IBM Cloud

BigInsights (HDFS)

Cloudant (DBaaS)

dashDB (Analytics)

SQDB (Managed DB2)

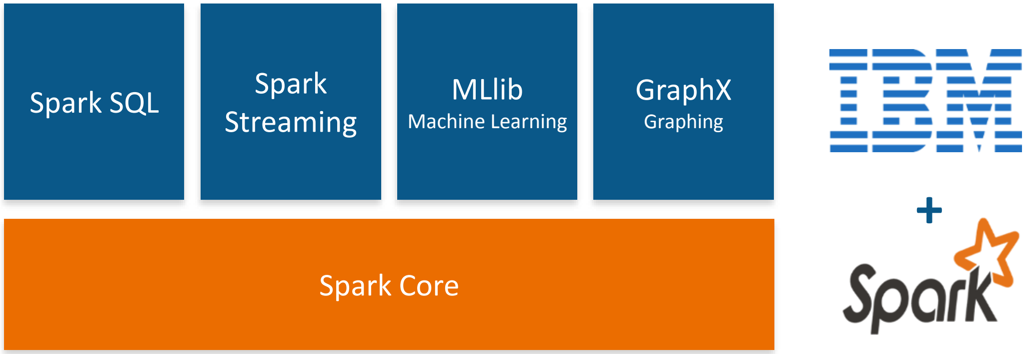
Swift (Object Storage)

Public Cloud

Cloud Apps

On-Premises

Benefits of Spark for Data Science



Spark SQL

Spark Streaming

MLlib
Machine Learning

GraphX
Graphing

Spark Core

IBM + Spark

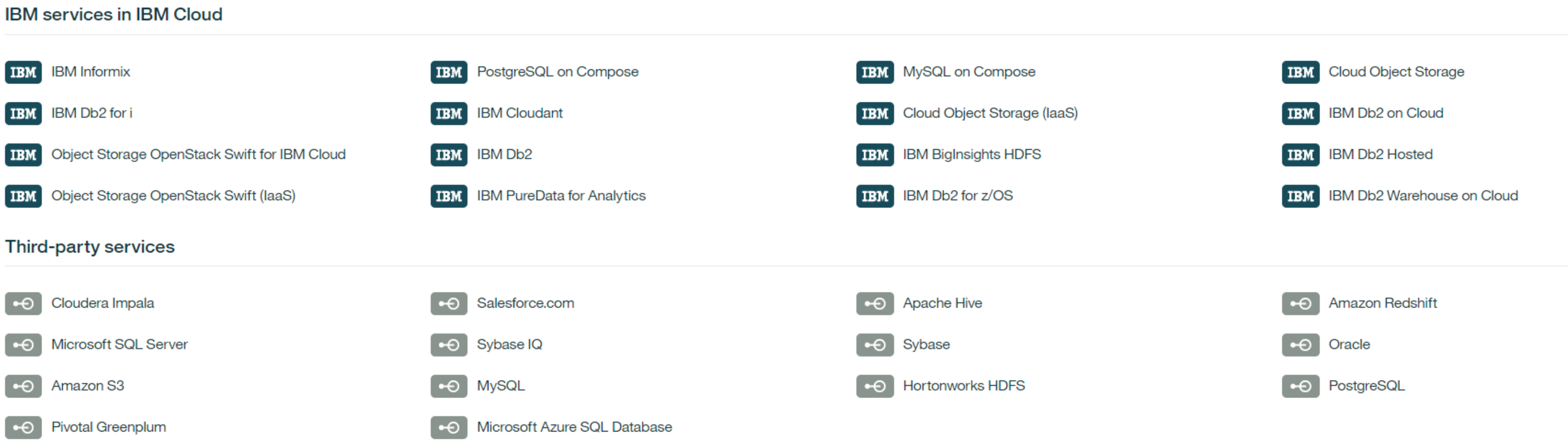
- General compute engine
- Basic I/O functions
- Task dispatching
- Scheduling

- Allows Data Scientists to code at scale
  - In-Memory processing that scales in a distributed architecture
- Supports multiple programing interfaces (Scala, Python, Java and R)
- Provides unified APIs (SQL, Streaming, Machine Learning, etc.)

The Spark service uses Bluemix Object Storage as its preferred data store for building performant applications

- Object storage provides inexpensive, scalable and self-healing retention of massive amounts of unstructured data
- Every object exists at the same level in a flat address space
- Bluemix Object Storage has a drag-and-drop upload and Swift API for programmatic access

Supported Data Sources via on- premises and cloud Connectors



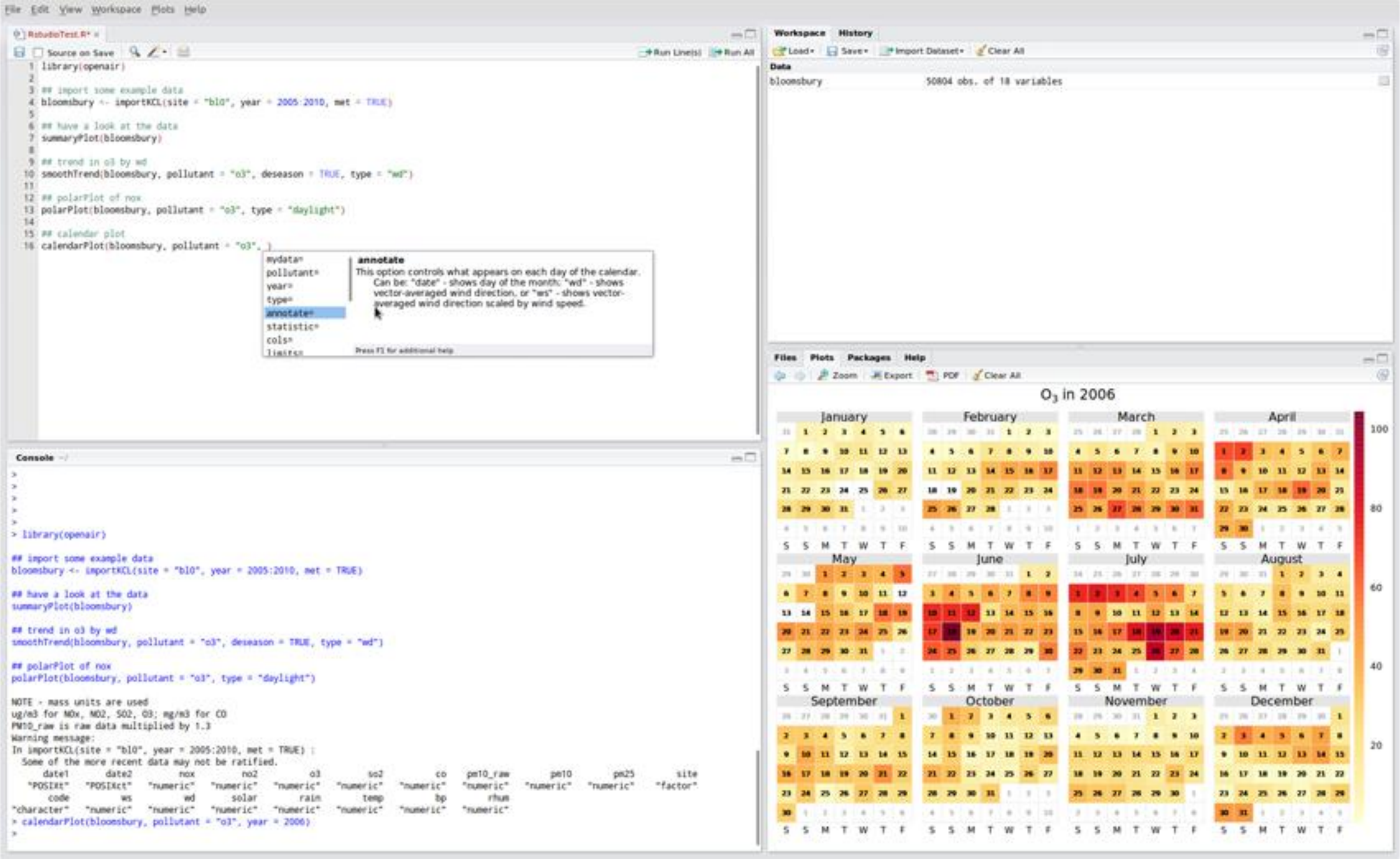
IBM services in IBM Cloud

- IBM Informix
- IBM Db2 for i
- Object Storage OpenStack Swift for IBM Cloud
- Object Storage OpenStack Swift (IaaS)
- PostgreSQL on Compose
- IBM Cloudant
- IBM Db2
- IBM PureData for Analytics
- MySQL on Compose
- Cloud Object Storage (IaaS)
- IBM BigInsights HDFS
- IBM Db2 for z/OS
- Cloud Object Storage
- IBM Db2 on Cloud
- IBM Db2 Hosted
- IBM Db2 Warehouse on Cloud

Third-party services

- Cloudera Impala
- Microsoft SQL Server
- Amazon S3
- Pivotal Greenplum
- Salesforce.com
- Sybase IQ
- MySQL
- Microsoft Azure SQL Database
- Apache Hive
- Sybase
- Hortonworks HDFS
- Amazon Redshift
- Oracle
- PostgreSQL

DSX has RStudio built into the experience thanks to our strategic partnership



```
library(openair)

## import some example data
bloomsbury <- importKCL(site = "bl0", year = 2005:2010, met = TRUE)

## have a look at the data
summaryPlot(bloomsbury)

## trend in o3 by wd
smoothTrend(bloomsbury, pollutant = "o3", deseason = TRUE, type = "wd")

## polarPlot of nox
polarPlot(bloomsbury, pollutant = "o3", type = "daylight")

## calendar plot
calendarPlot(bloomsbury, pollutant = "o3", )
```

```
mydata=
pollutant=
year=
type=
annotate=
statistic=
cols=
limits=
```

annotate

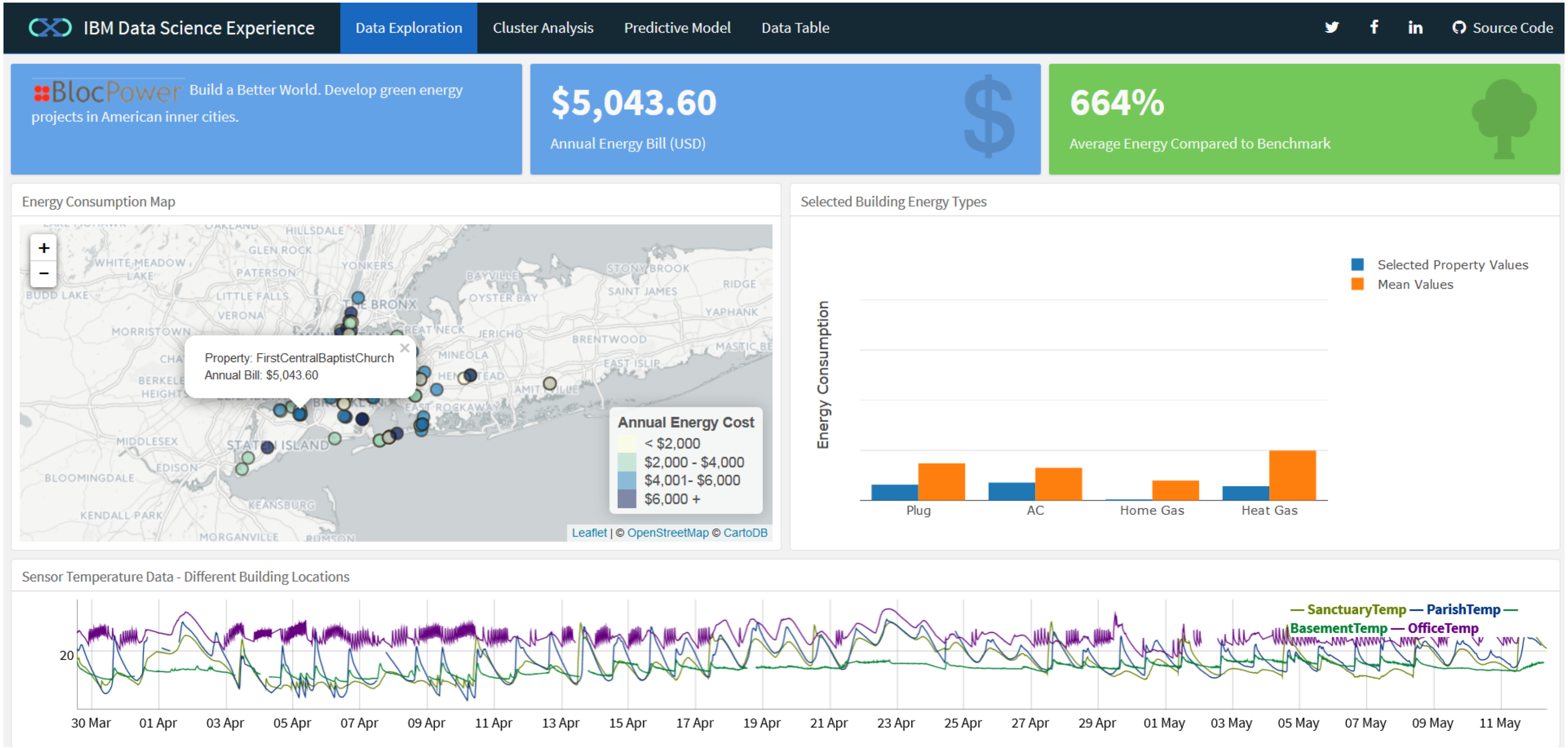
This option controls what appears on each day of the calendar. Can be: "date" - shows day of the month; "wd" - shows vector-averaged wind direction, or "ws" - shows vector-averaged wind direction scaled by wind speed.

Press F1 for additional help

```
bloomsbury                    50804 obs. of 18 variables
```

```
>
>
>
>
>
> library(openair)

## import some example data
bloomsbury <- importKCL(site = "bl0", year = 2005:2010, met = TRUE)

## have a look at the data
summaryPlot(bloomsbury)

## trend in o3 by wd
smoothTrend(bloomsbury, pollutant = "o3", deseason = TRUE, type = "wd")

## polarPlot of nox
polarPlot(bloomsbury, pollutant = "o3", type = "daylight")

NOTE - mass units are used
ug/m3 for NOx, NO2, SO2, O3; mg/m3 for CO
PM10_raw is raw data multiplied by 1.3
Warning message:
In importKCL(site = "bl0", year = 2005:2010, met = TRUE) :
  Some of the more recent data may not be ratified.
      date1       date2         nox         no2          o3         so2          co    pm10_raw        pm10        pm25        site
  "POSIXt"   "POSIXct"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"    "factor"
       code          ws          wd       solar        rain        temp          bp        rhum
"character"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"
> calendarPlot(bloomsbury, pollutant = "o3", year = 2006)
>
```

With RStudio you can create Shiny web applications to make your analysis accessible to the business



BlocPower Build a Better World. Develop green energy projects in American inner cities.

$5,043.60

Annual Energy Bill (USD)

664%

Average Energy Compared to Benchmark

Energy Consumption Map

Property: FirstCentralBaptistChurch
Annual Bill: $5,043.60

Annual Energy Cost

🟨	< $2,000
🟩	$2,000 - $4,000
🟦	$4,001- $6,000
🟪	$6,000 +

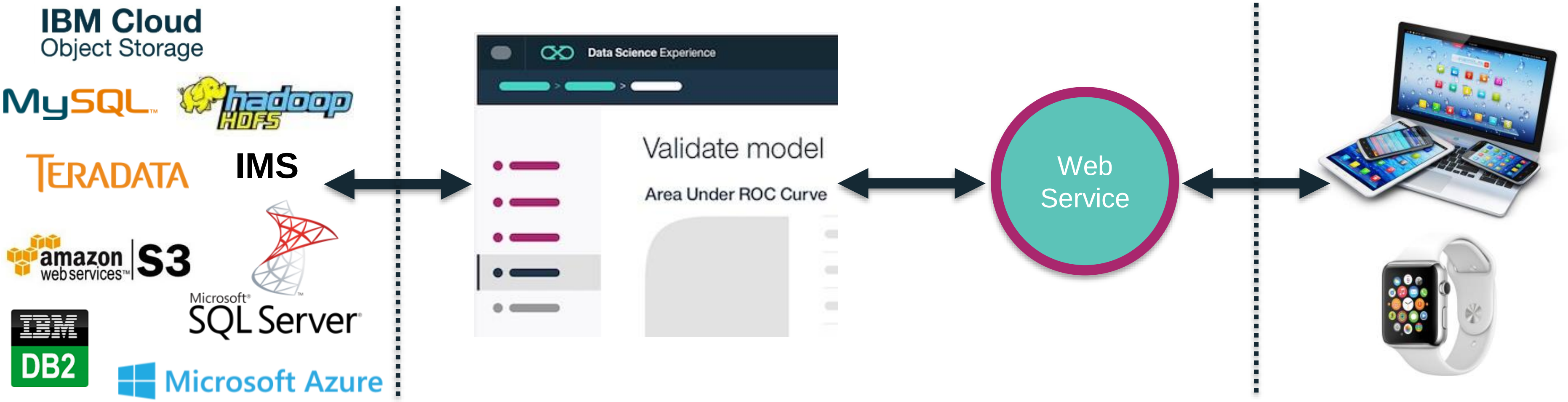
Leaflet | © OpenStreetMap © CartoDB

Selected Building Energy Types

Sensor Temperature Data - Different Building Locations

Operationalize insights with Machine Learning

Watson Machine Learning



Data Access:

- Easily connect to Behind-the-Firewall and Public Cloud Data
- Catalogued and Governed Controls through Watson Data Platform

Creating Models:

- Single UI and API for creating ML Models on various Runtimes
- Auto-Modeling and Hyperparameter Optimization

Web Service:

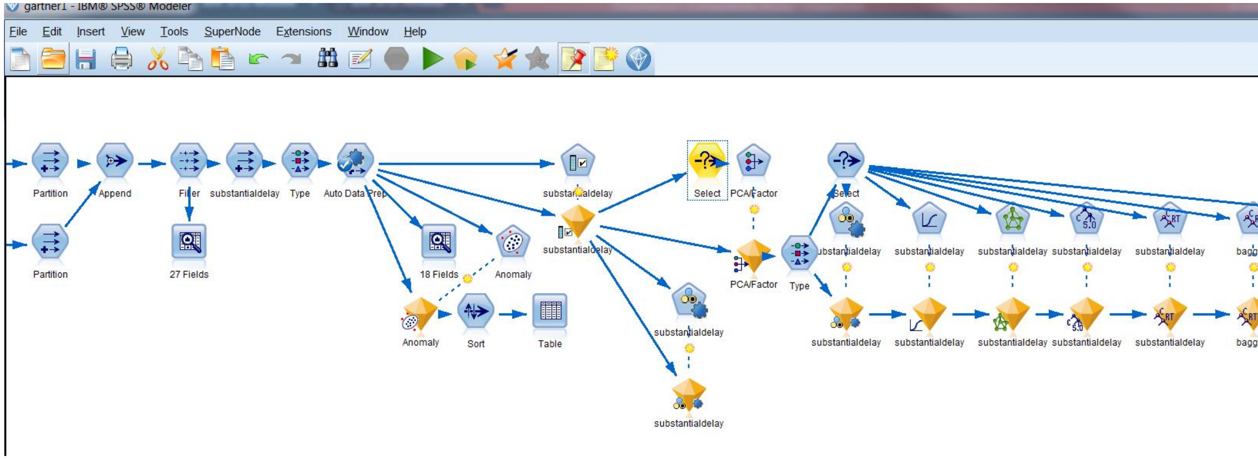
- Real-time, Streaming, and Batch Deployment
- Continuous Monitoring and Feedback Loop

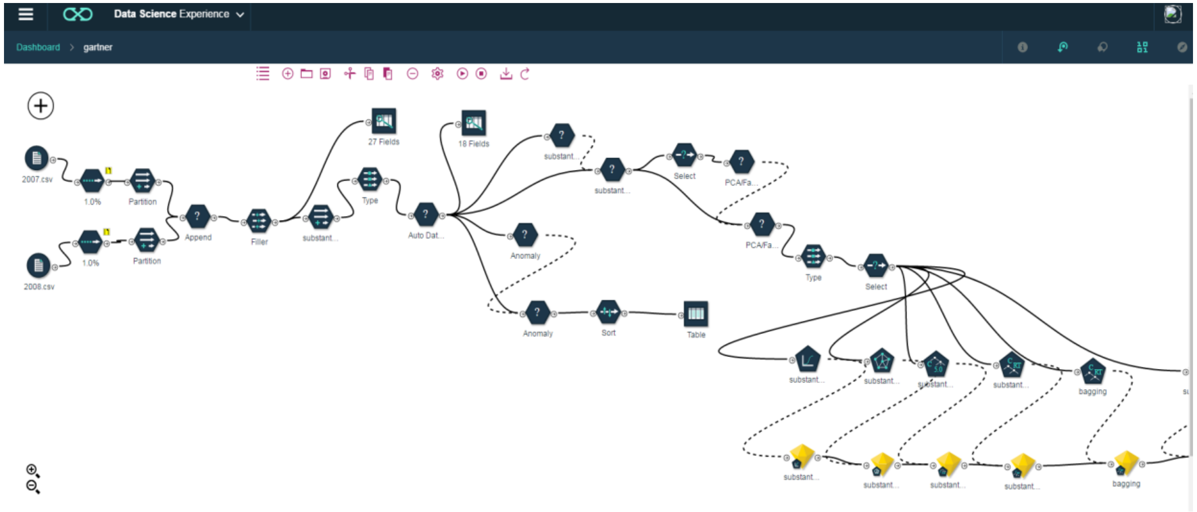
Intelligent Apps:

- Integrate ML models with apps, websites, etc.
- Continuously Improve and Adapt with Self-Learning

Use SPSS Modeler to Visually Create ML Flows

- This DSX Canvas will have compatibility with legacy SPSS Modeler streams
- Multiple execution runtimes: SPSS Modeler, SparkML
- Planned support for R/Python/SQL code





- Pipeline deployment from DSX Canvas (left) via SPSS Modeler

DSX Local

- **Very similar to the public cloud version of DSX**

- **Runs on hardware that is provided by the customer**
  - The DSX Local software and hardware are managed by the customer

- **DSX Local comes with all the software it needs to run, although it can integrate with existing customer systems such as**
  - Databases and HDFS storage
  - LDAP servers for authentication

Labs

Lab Overview

Use IBM's Data Science Experience (DSX) and IBM cloud services to create a working cloud-based application from start to finish. Participants will be led through a series of three labs. The three labs build upon one another so it is important that they are completed in order.

- Lab 1 - The first lab will begin with loading raw delimited data into DB2 Warehouse for Cloud and interacting with that data from a Jupyter notebook in DSX with python.
- Lab 2 - The second lab will guide participants in creating an R notebook and Shiny UI in DSX using RStudio.
- Lab 3 - The third lab will show how to use the Watson Machine Learning capability to create a machine learning model based on the supply chain data set. The machine learning model, deployed in the IBM Cloud, will be used to predict the severity level of each discrepancy based on action request characteristics.

Lab 1

This lab will begin with loading raw delimited data into DB2 Warehouse for Cloud and interacting with that data from a Jupyter notebook in DSX with Python.

Objectives:

- Upon completing the lab, you will know how to:
  - Create a Jupyter IPython notebook from a URL
  - Establish a connection to DB2 Warehouse on Cloud
  - Use a dataframe to read and manipulate tables
  - Use Spark to explore and analyze the dataset
  - Write the modified dataset back to DB2 Warehouse on

Lab 2

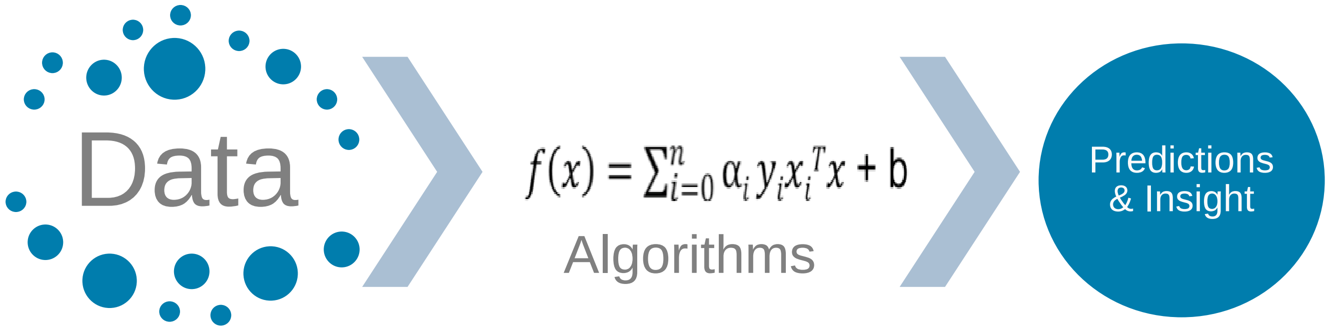
In this lab, you will learn some of the fundamentals of using RStudio and Shiny in DSX to work and interact with data in DB2 Warehouse and then create a fully operational "reactive" web application that you can enhance further.

Objectives:

- Upon completing the lab, you will:
  - Create an RStudio project from a Git repository
  - Establish a connection to DB2 Warehouse
  - Query, explore and visualize data in an R notebook
  - Use ggplot2 to create bar plots of several columns in an R dataframe
  - Close the database connection
  - Leverage shiny to create and run a web application
  - Interact with the shiny web application by running it externally

What is Machine Learning?

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

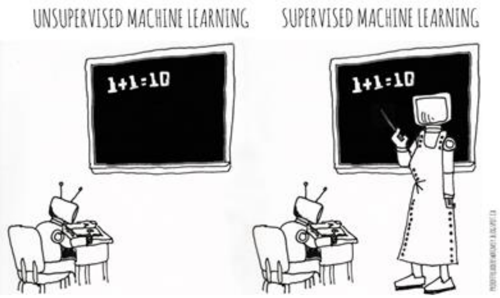


$$f(x) = \sum_{i=0}^{n} \alpha_i y_i x_i^T x + b$$

Categories of Machine Learning

- **Supervised learning**
  - The program is “trained” on a pre-defined set of “training examples”, which then facilitate its ability to reach an accurate conclusion when given new data
  - The algorithm is presented with example inputs and their outcomes (labels)
  - The goal is to learn a general rule that maps inputs to outputs

- **Unsupervised learning**
  - No labels are given to the learning algorithm, leaving it on its own to find structure (patterns and relationships) in its input



Categories of Machine Learning

Technique	Usage	Algorithms
Classification (or prediction)	• Used to predict group membership (e.g., will this employee leave?) or a number (e.g., how many widgets will I sell?)	• Decision Trees • Logistic Regression • Random Forests • **Naïve Bayes** • Linear Regression • Lasso Regression etc
Segmentation	• Used to classify data points into groups that are internally homogenous and externally heterogeneous. • Identify cases that are unusual	• K-means • Gaussian Mixture • Latent Dirichlet allocation etc
Association	• Used to find events that occur together or in a sequence (e.g., market basket)	• FP Growth

Training, testing, & validation sets

- **During the model development process, supervised learning techniques employ training and testing sets and sometimes a validation set.**
  - Historical data with known outcome
  - Data is randomly split into training, testing, and/or validation sets (mutually exclusive records)
- **Why?**
  - Training set
    - Build the model
    - Tune the parameters
  - Testing set
    - Assess model quality during training/tuning process
    - Avoid overfitting the model to the training set
  - Validation set
    - Estimate accuracy or error rate of model after tuning
    - Used to compare multiple models

Spark ML

- **Spark ML is Spark's machine learning (ML) library**

- **Goal is to make machine learning scalable and easy**
  - No need to understand the detailed math!

- **Divides into two packages:**
  - spark.mllib contains the original API built on top of RDDs
  - spark.ml provides higher-level API built on top of DataFrames for constructing ML pipelines
  - A pipeline is a series of stages where each stage either transforms, or runs through a machine learning algorithm.

- **Using spark.ml is recommended because with DataFrames the API is more versatile and flexible**
  - spark.mllib will continue to be supported

Spark ML Pipeline Terminology

Spark ML standardizes APIs for machine learning algorithms to make it easier to combine multiple algorithms into a single pipeline, or workflow

- **DataFrame**: Spark ML uses DataFrame from Spark SQL as an ML dataset, which can hold a variety of data types
- **Transformer**: A Transformer is an algorithm which can transform one DataFrame into another DataFrame
- **Estimator**: An Estimator is an algorithm which can be fit on a DataFrame to produce a Transformer
- **Pipeline**: A Pipeline chains multiple Transformers and Estimators together in a sequence to specify an ML workflow
- **Parameter**: All Transformers and Estimators share a common API for specifying parameters

Lab 3 - Watson Machine Learning

In this lab, you will use IBM's Watson Machine Learning GUI to train, evaluate, and deploy a Watson Machine Learning model based on the modified supply chain dataset.

Objectives:

- Upon completing the lab, you will:
  - Become familiar with the Watson Machine Learning GUI.
  - Train/Evaluate a machine learning model
  - Deploy a machine learning model.
  - Use the deployed machine learning model to make predictions.