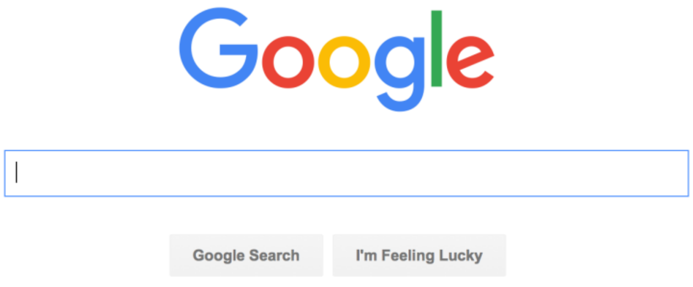
Introduction to Data Science and IBM's Data Science Experience

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

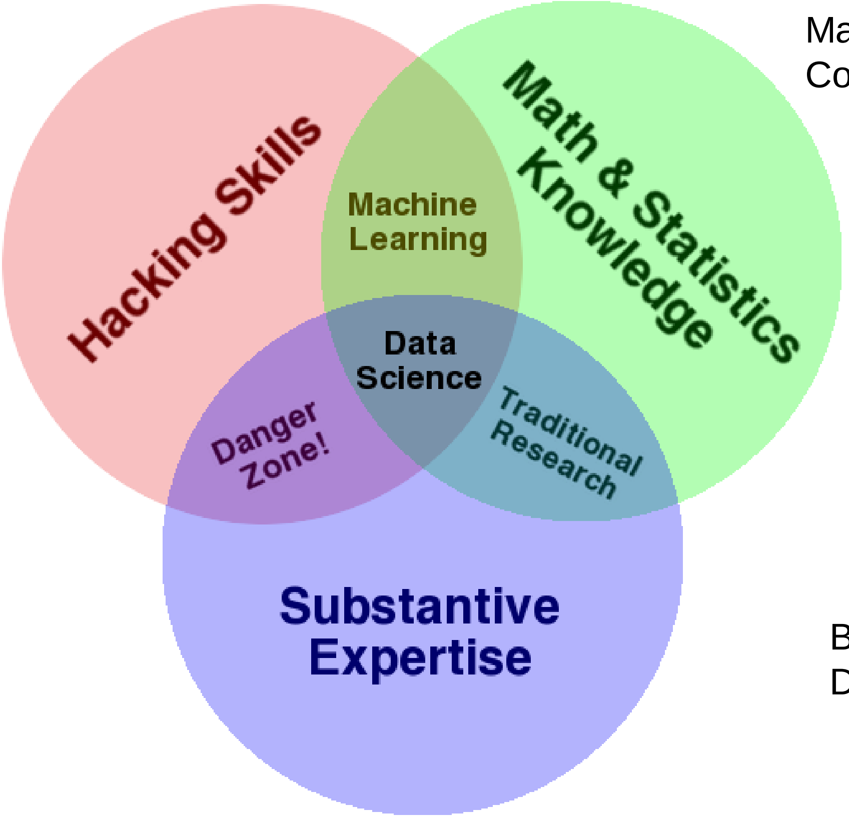
Joel Patterson
Bernie Beekman
Davin Shearer

The digital age is changing the way we live, play, learn and work…



What is the Data Scientist?

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



Mathematical Background
Computational Science

Business/Industry Expertise
Domain Knowledge
- Supply Chain
- CRM
- Financials
- Networking

Drew Conway's Data Science Venn Diagram

Google Trends – Data Science Languages

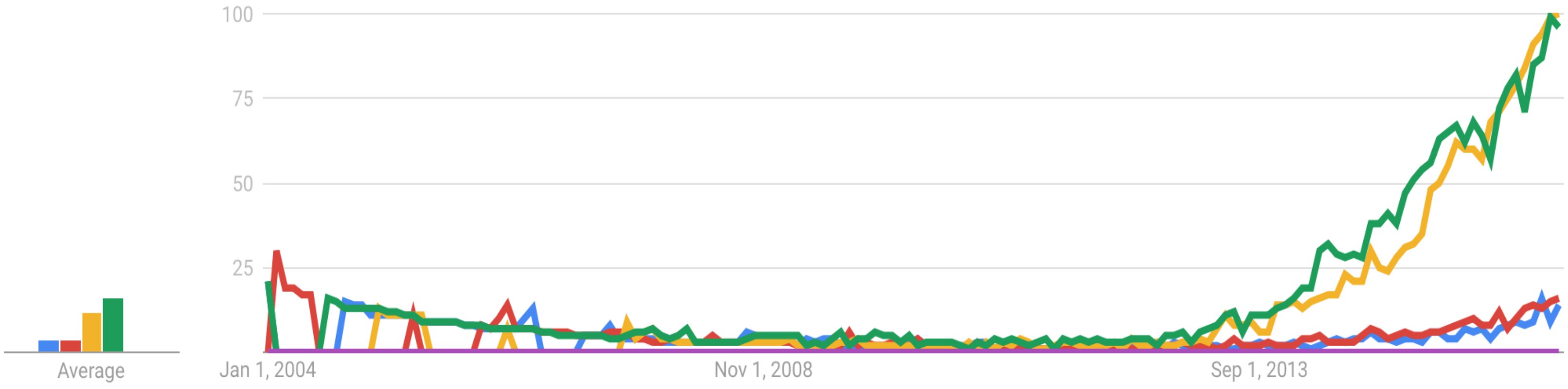
🟦 SPSS

🟥 SAS

🟩 Python

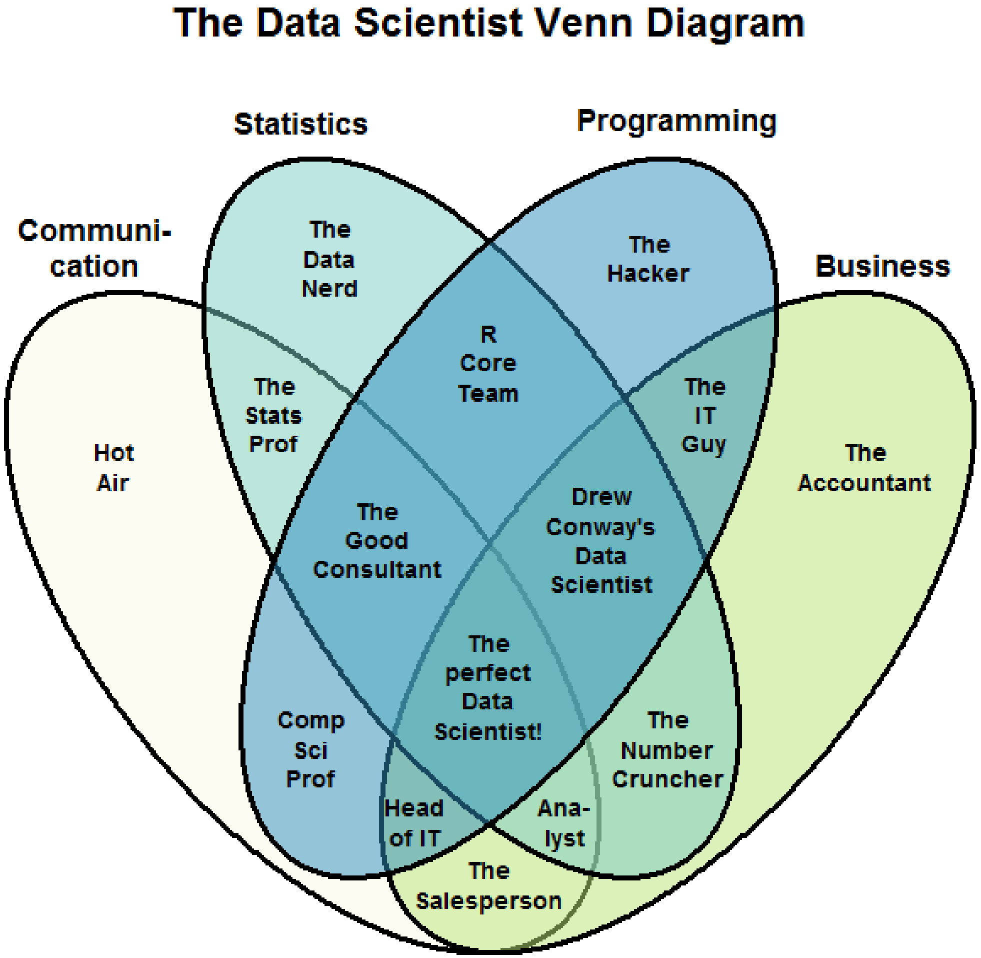
🟨 R

🟪 Scala



Trends in Google Searches (September 2nd 2016)

The perfect Data Science Team



Normally not all the skills are in one single person but rather in a data science team
In IBM Data Science Experience we include tools   to make the perfect Data Science Team
All in a collaborative, cloud environment that scales in demand

IBM Watson Data Platform

Mission: Make Data Simple and Accessible to All

Platform.

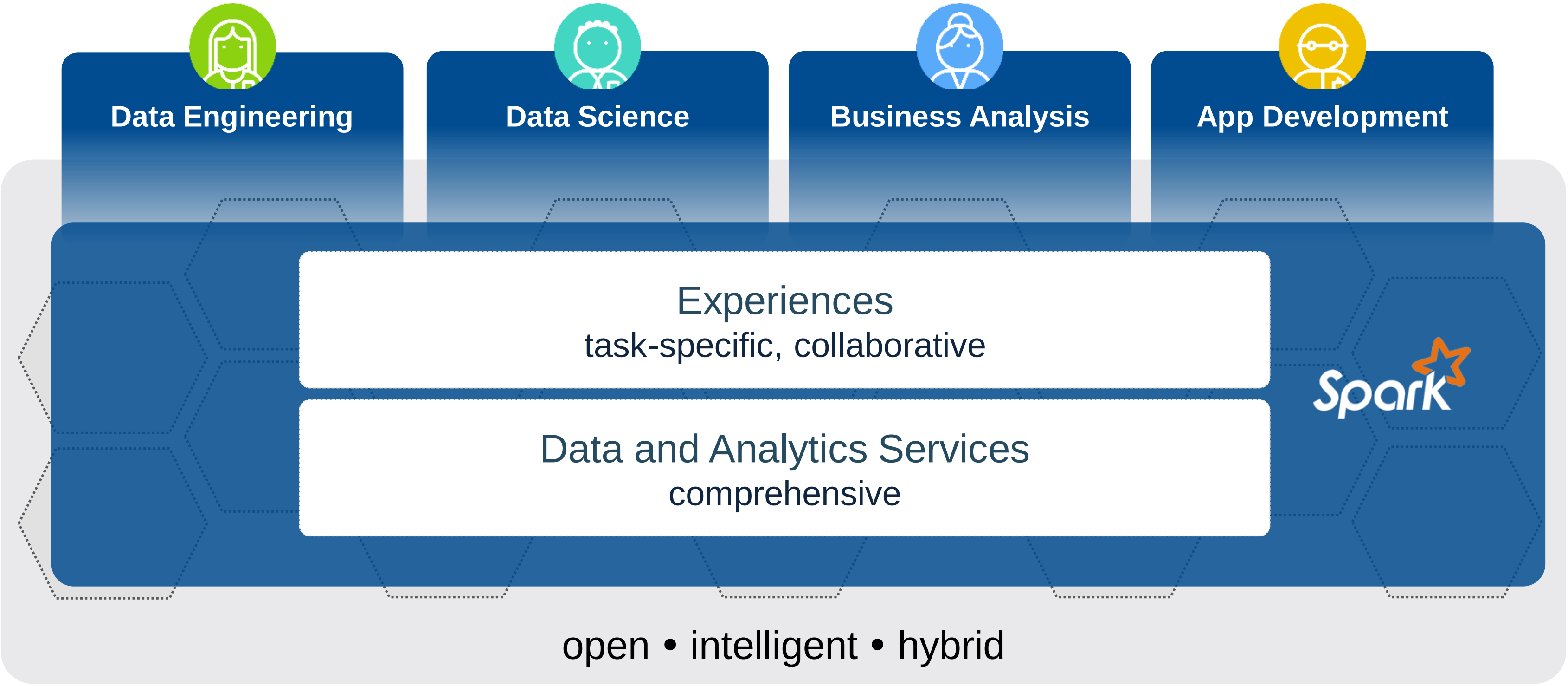
Method.

Ecosystem.

http://ibm.co/makedatasimple

IBM Watson Data Platform

Experience New Ways To Put Data To Work



Data Engineering

Data Science

Business Analysis

App Development

Experiences
task-specific, collaborative

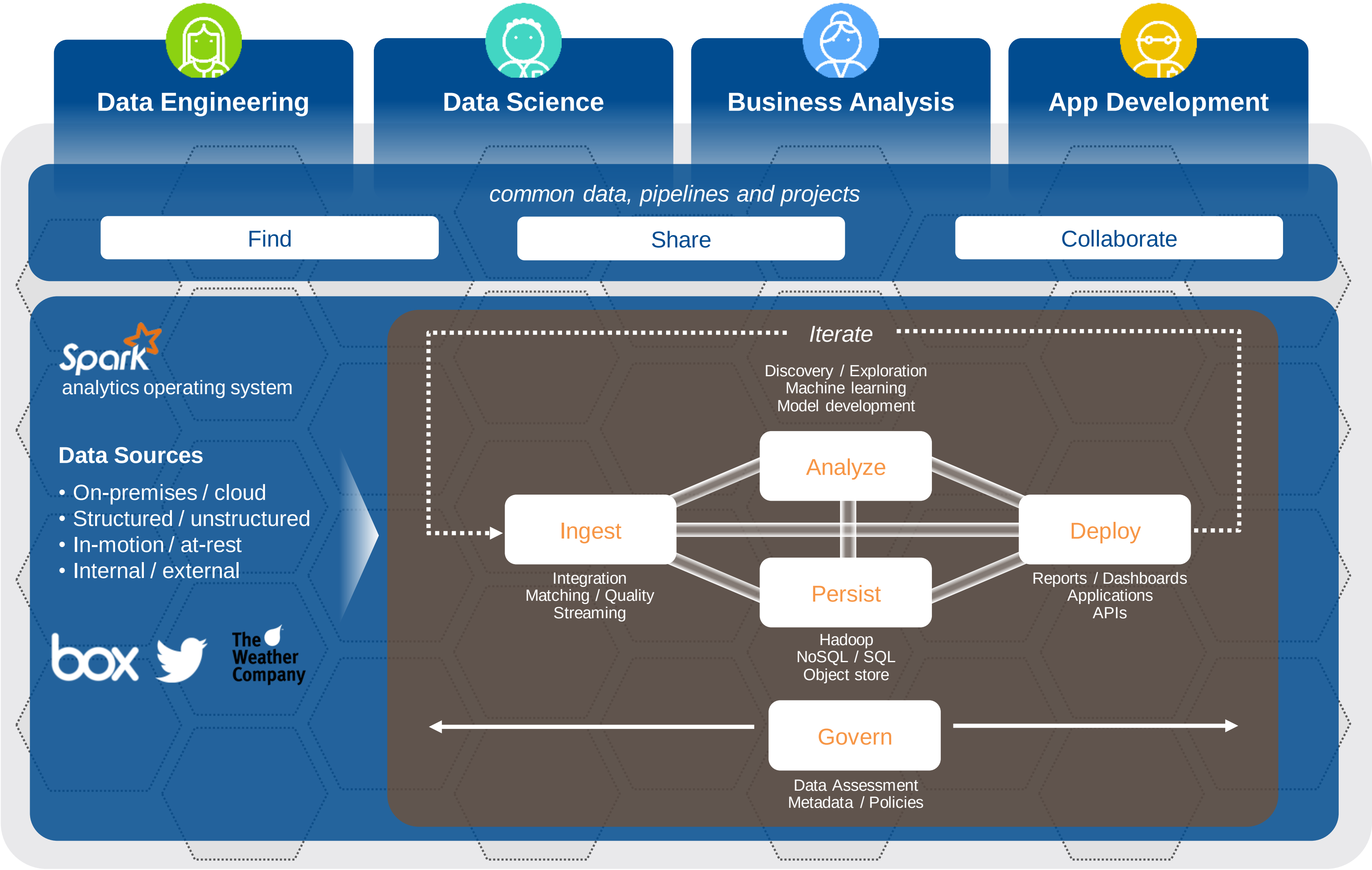
Data and Analytics Services
comprehensive

Spark

open • intelligent • hybrid

IBM Watson Data Platform

Connects Users to Data and Analytics



Data Engineering

Data Science

Business Analysis

App Development

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

Data Scientist Challenges

- **Rigid toolset**
  - Have to choose one and only one approach
  - Cannot easily connect all of the capabilities needed
  - Difficult to navigate between the various tools used
- **Fragmented and time consuming**
  - Using multiple disjointed environments
  - Separate on-ramp/community for each tool/environment
  - Does not have meta data or data lineage
- **Analytical Silo**
  - Difficult to maintain and version control project assets
  - Limited means of collaborating with team
  - Results are difficult to share

Data Science Experience

Brings together popular Data Science **Open Source tools** with
IBM value-add functionalities coupled with **community and social** features

Learn

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

Create

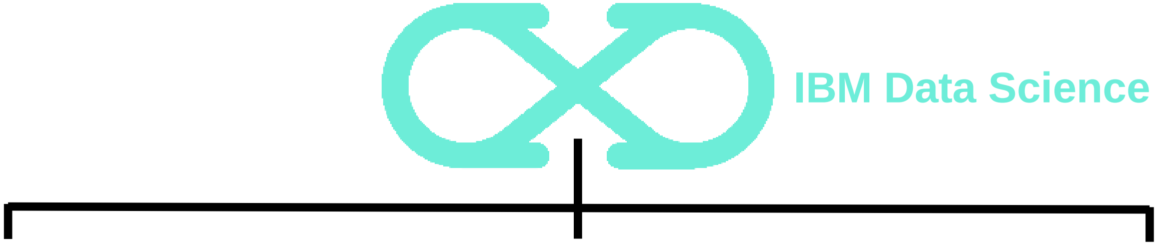
The best of open source and IBM value-add to create state-of-the-art data products

Collaborate

Community and social features that provide meaningful collaboration

External URL: http://datascience.ibm.com

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

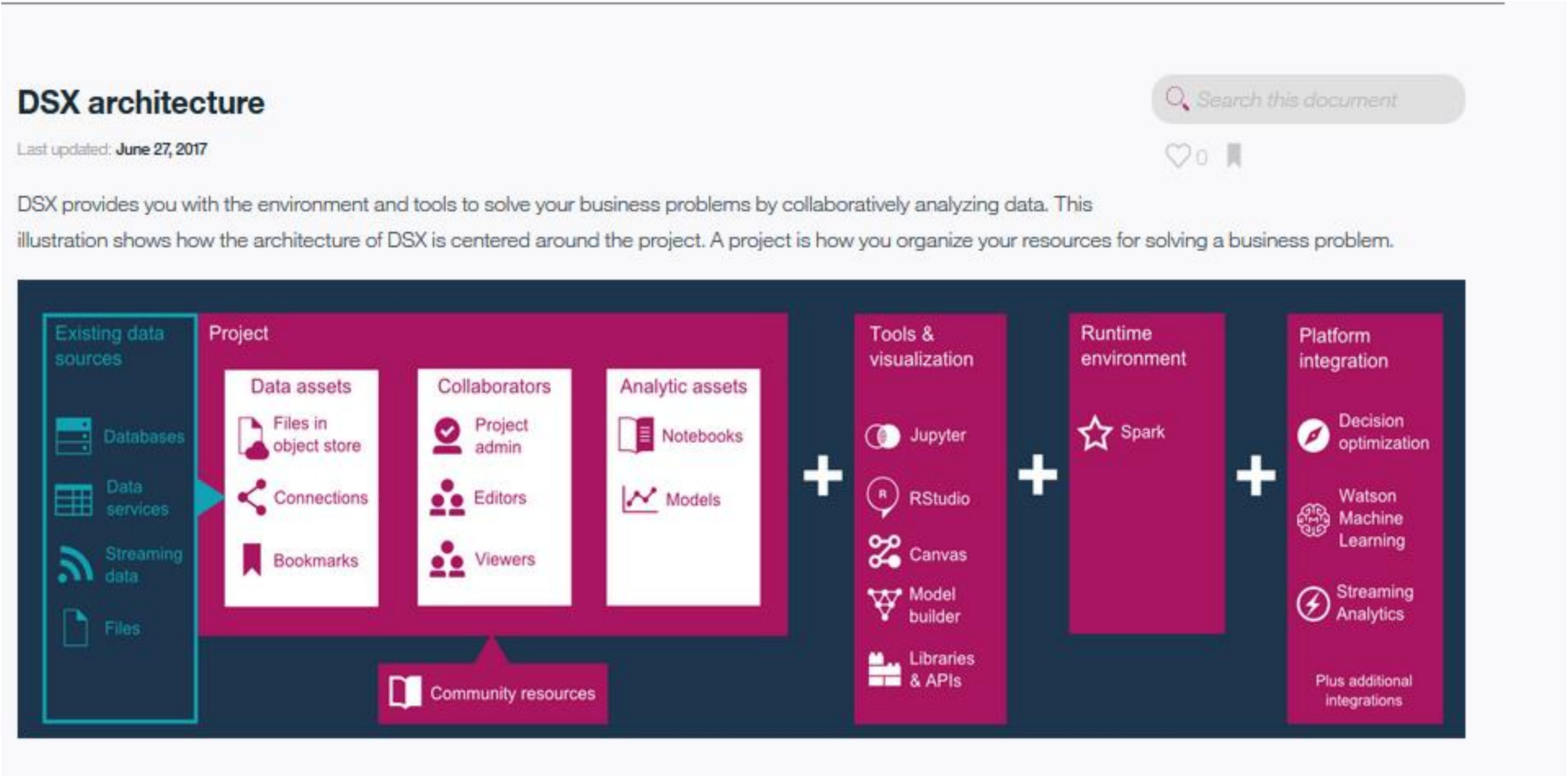
IBM Added Value

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

Powered by IBM **Watson Data Platform**

* Closed beta

DSX Architecture

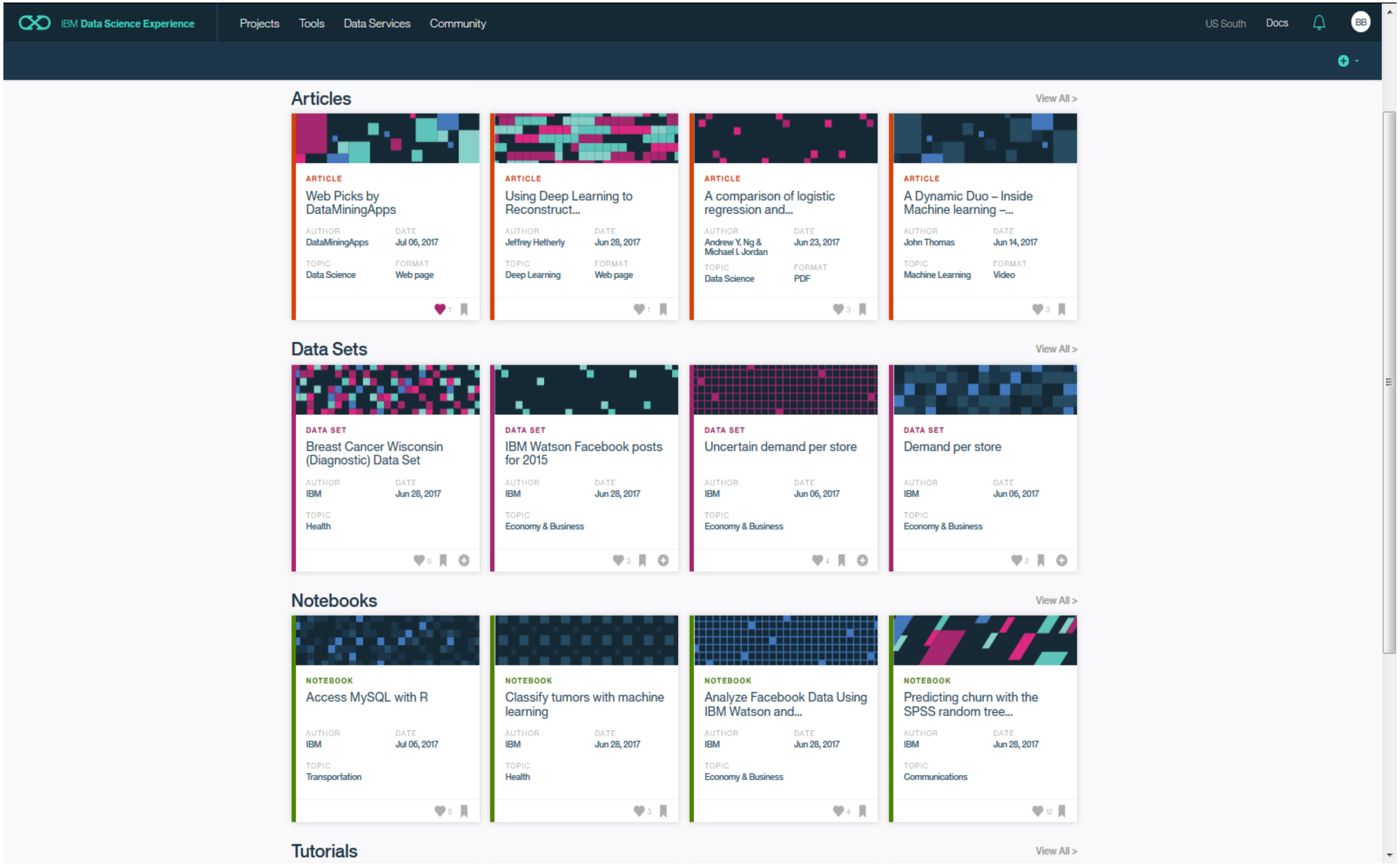


DSX architecture

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.

Community Cards provide in-context learning for users



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

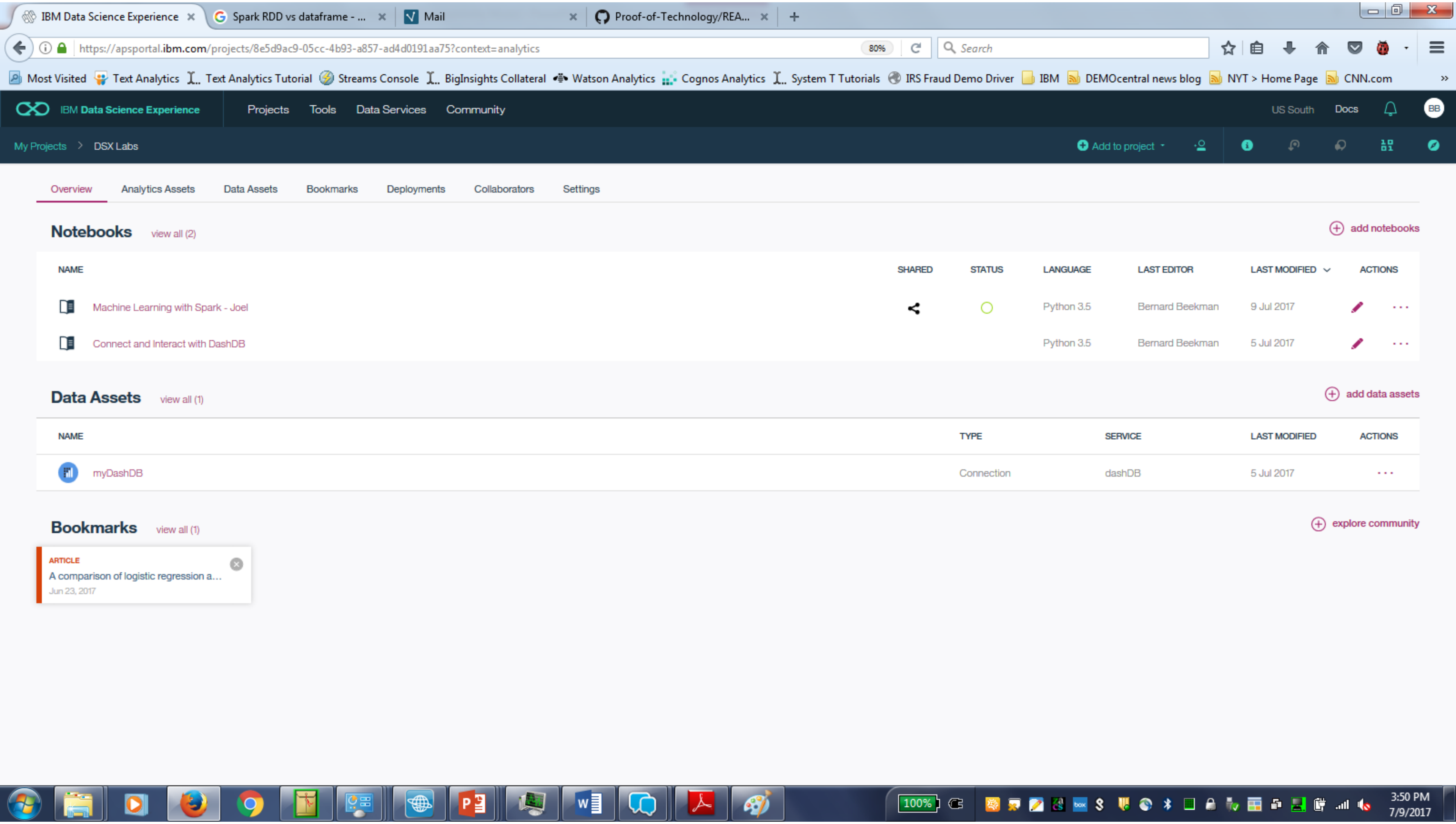
NOTEBOOK
Analyze Facebook Data Using IBM Watson and...
AUTHOR IBM
DATE Jun 28, 2017
TOPIC Economy & Business
4

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

Tutorials

View All >

Collaborate Using Projects



Notebooks

view all (2) | add notebooks

NAME	SHARED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS
Machine Learning with Spark - Joel	⋖	○	Python 3.5	Bernard Beekman	9 Jul 2017	✎ ···
Connect and Interact with DashDB			Python 3.5	Bernard Beekman	5 Jul 2017	✎ ···

Data Assets

view all (1) | add data assets

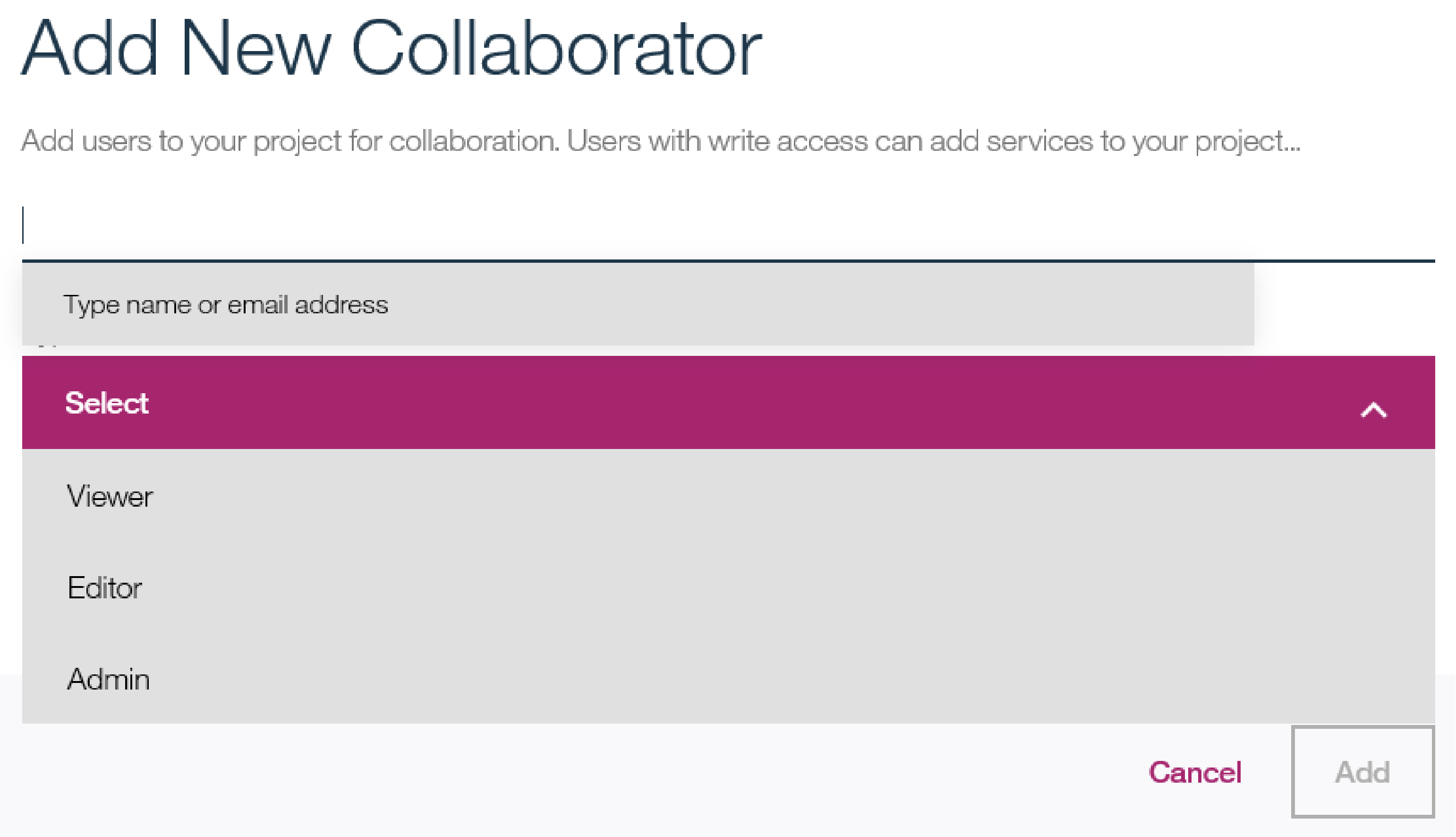
NAME	TYPE	SERVICE	LAST MODIFIED	ACTIONS
myDashDB	Connection	dashDB	5 Jul 2017	···

Bookmarks

view all (1) | explore community

ARTICLE
A comparison of logistic regression a...
Jun 23, 2017

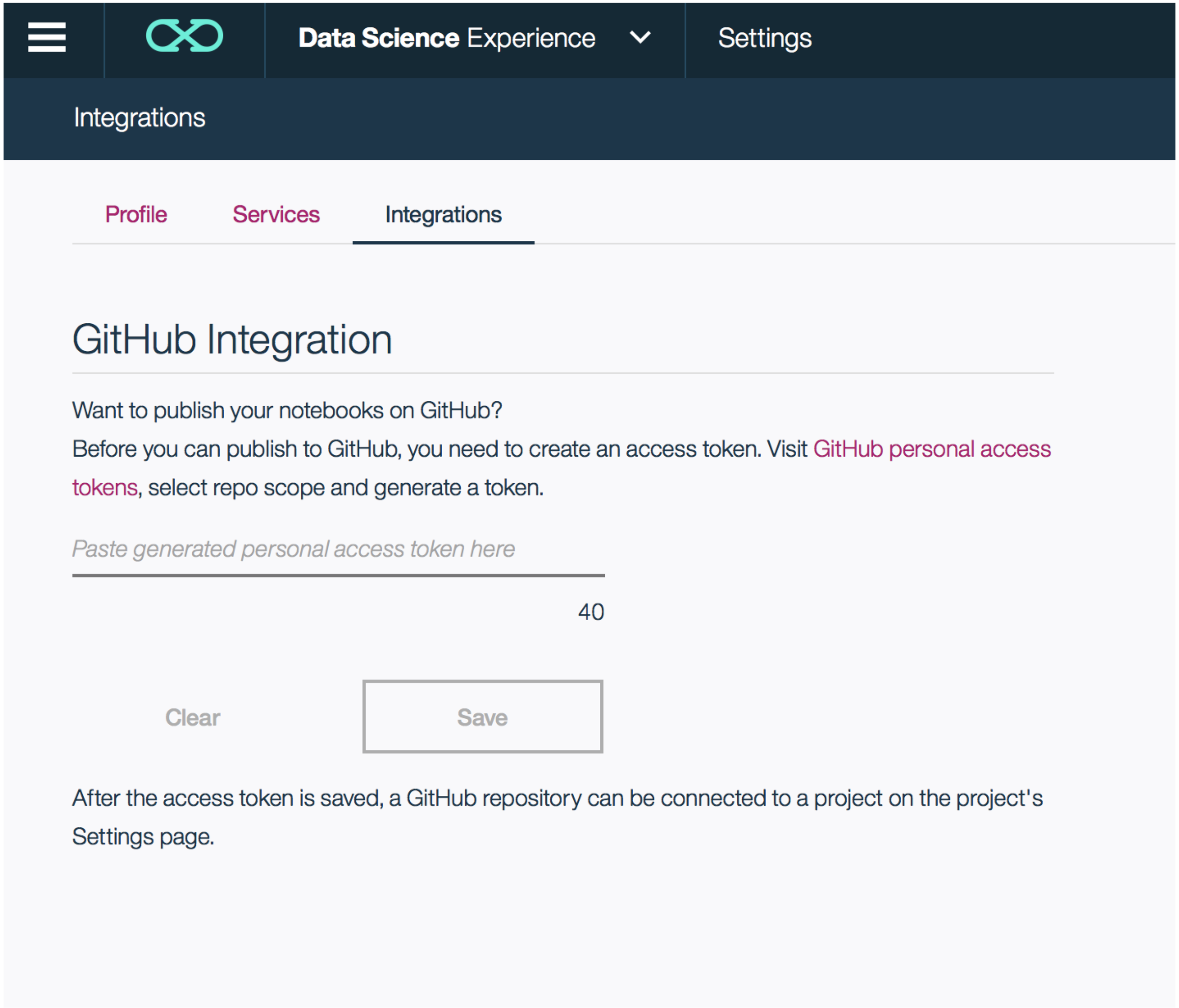
Add Collaborators to a Project



Add New Collaborator

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

GitHub Integration



GitHub Integration

Want to publish your notebooks on GitHub?

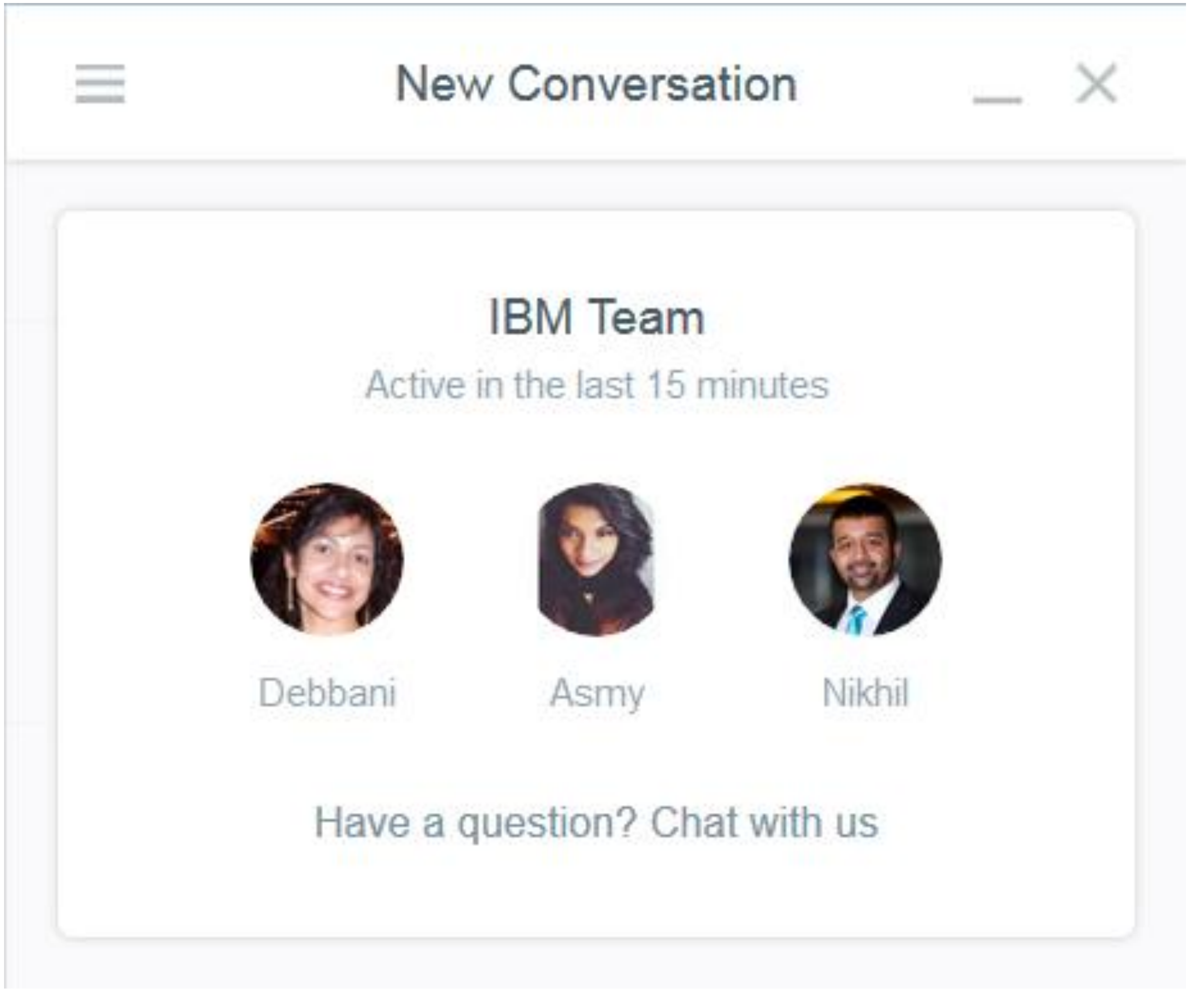
Before you can publish to GitHub, you need to create an access token. Visit GitHub personal access tokens, select repo scope and generate a token.

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 DSX



IBM Team

Active in the last 15 minutes

Debbani

Asmy

Nikhil

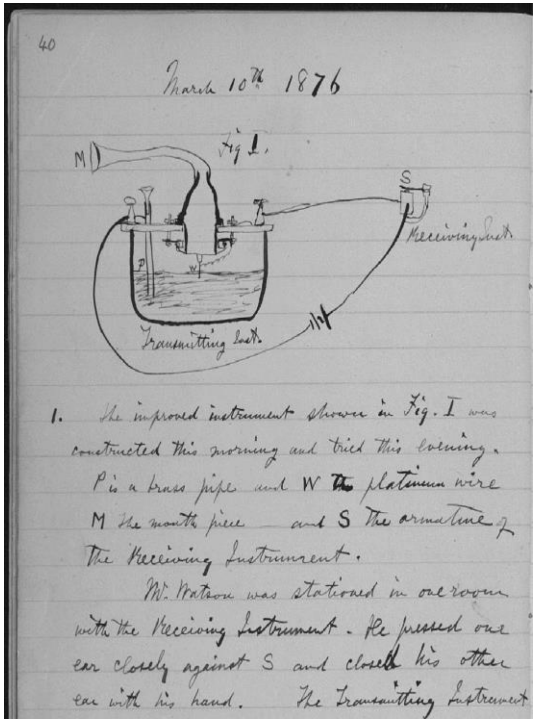
Have a question? Chat with us

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
Fig 1.
M
P
W
S
Receiving Instt.
Transmitting Instt.

March 10th 1876

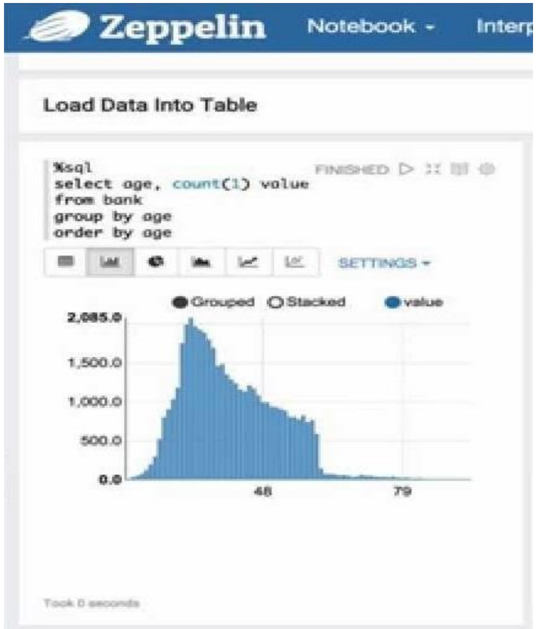
1. The improved instrument shown in Fig. I was constructed this morning and tried this evening. P is a brass pipe and W the platinum wire M the mouth piece and S the armature of The Receiving Instrument.

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

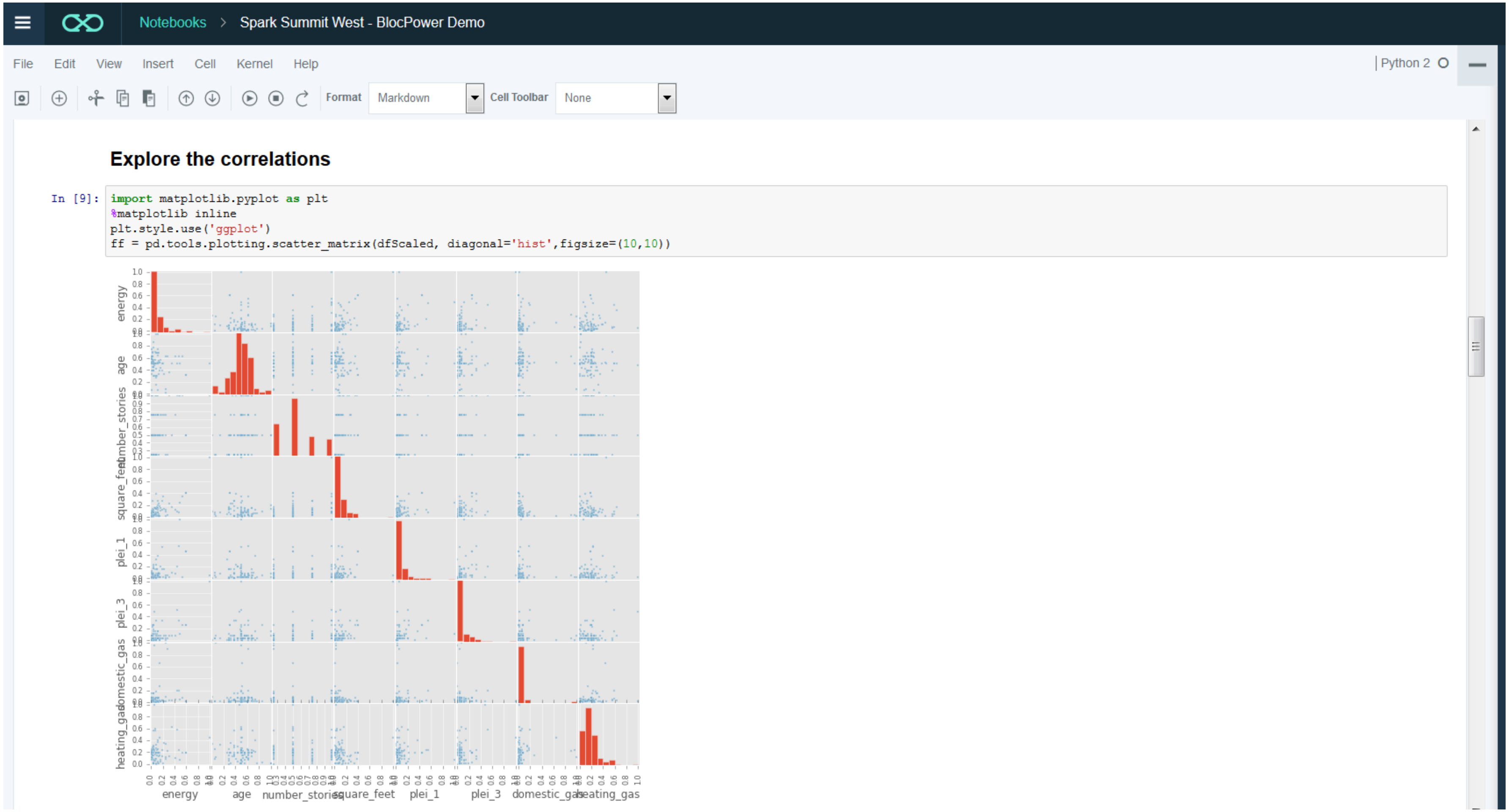


Load Data Into Table

```
%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))
```

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

Object Storage

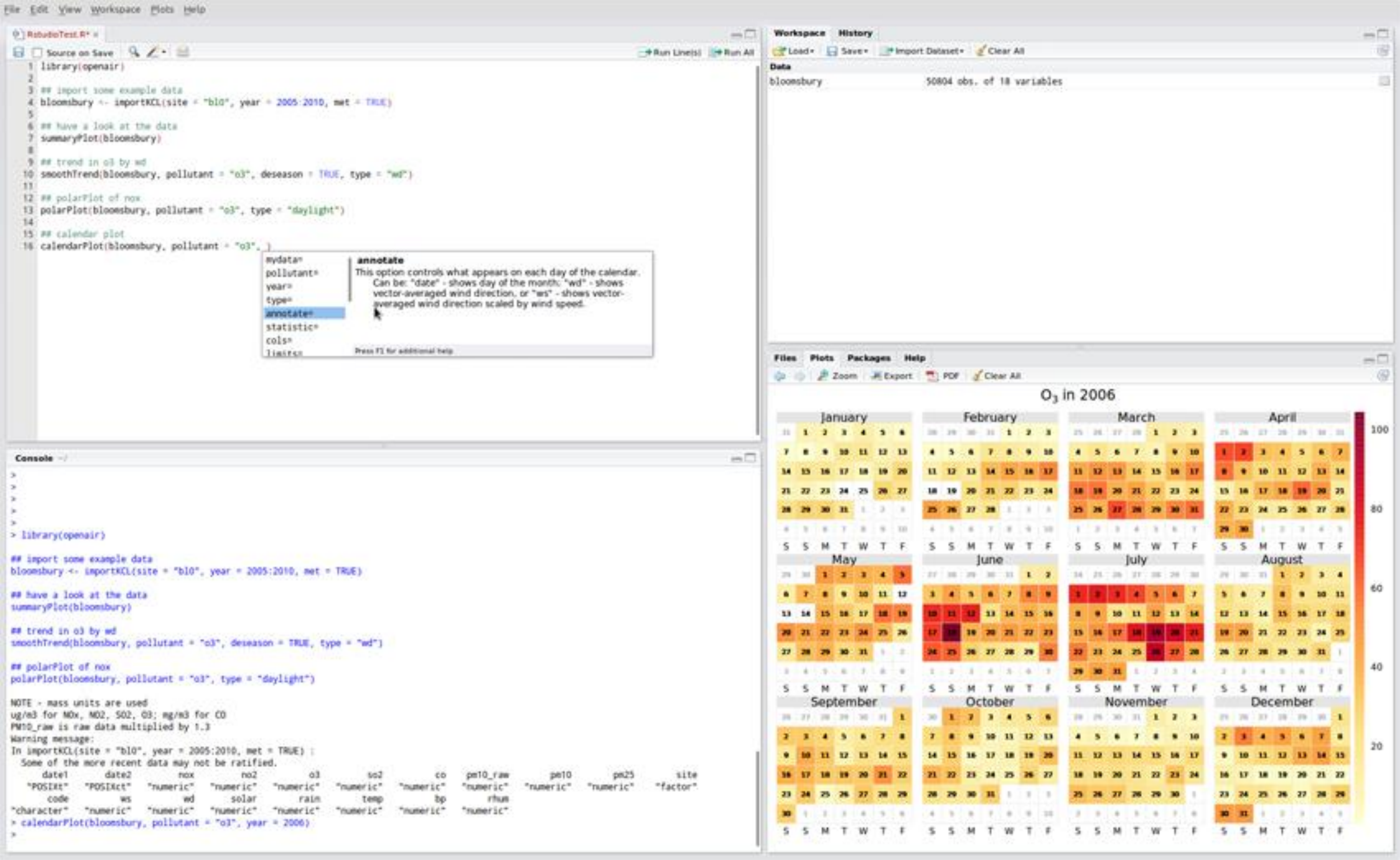
IBM

Supported Data Sources/Targets for DSX via on- premises and cloud Connectors

Cloud Sources	On-Premises Sources	Cloud Targets	On-Premises Targets
Amazon Redshift	Apache Hive	Amazon S3	IBM DB2® LUW
Amazon S3	Cloudera Impala	Bluemix Object Storage	IBM Pure Data for Analytics®
Apache Hive	IBM DB2® LUW	IBM Cloudant™	Teradata
Bluemix Object Storage	IBM Informix®	IBM dashDB	
IBM BigInsights™ on Cloud *	IBM Pure Data for Analytics®	IBM BigInsights™ on Cloud *	
IBM Cloudant™	Microsoft SQL Server	IBM DB2® on Cloud	
IBM dashDB	MySQL Enterprise Edition	IBM SQL Database	
IBM DB2® on Cloud	Oracle	IBM Watson™ Analytics	
IBM SQL Database	Pivotal Greenplum	PostgreSQL on Compose	
Microsoft Azure	PostgreSQL	SoftLayer Object Storage	
PostgreSQL on Compose	Sybase		
Salesforce	Sybase IQ		
SoftLayer Object Storage	Teradata		

All of the supported targets are compatible with each source

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.

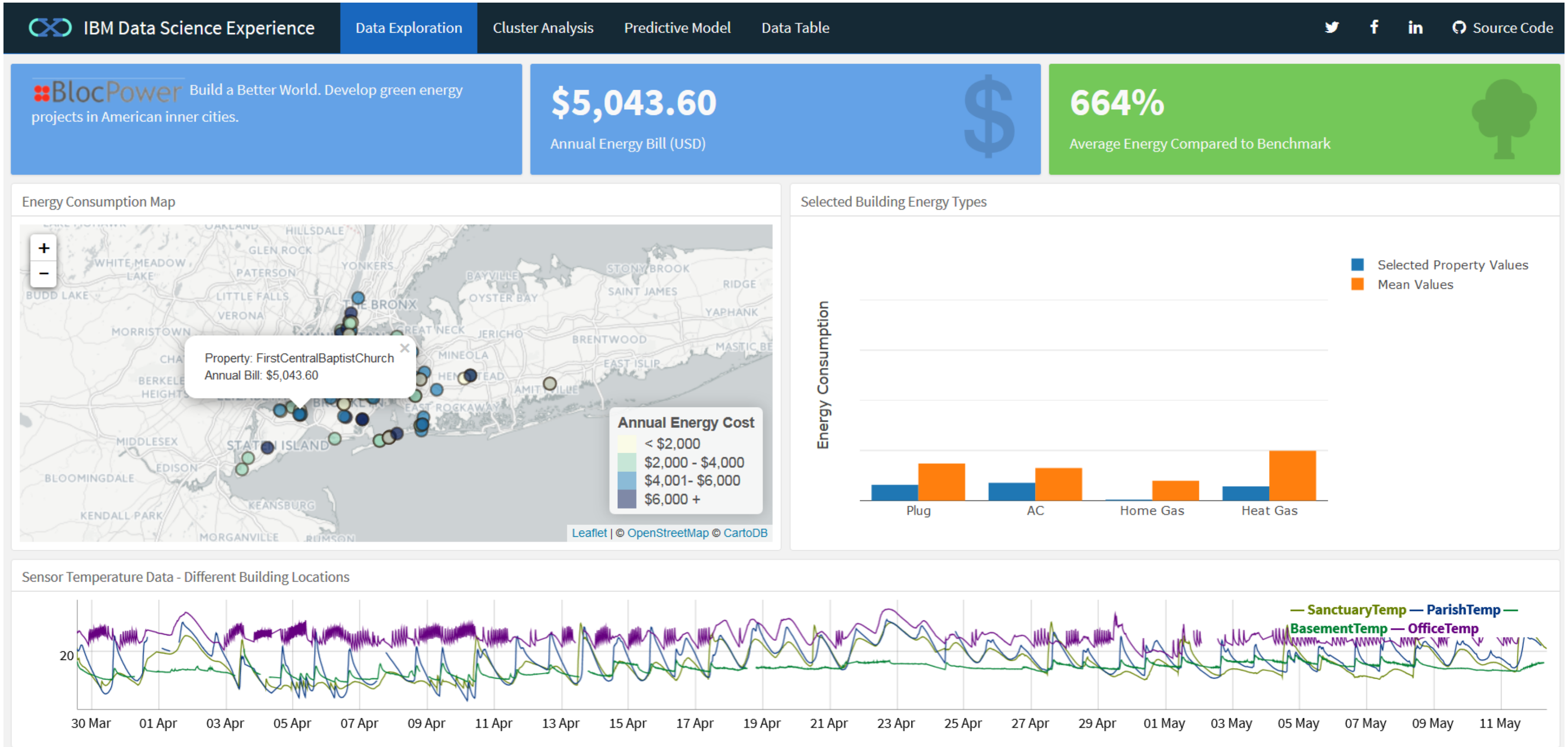
Press F1 for additional help

Data

```
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

Property: FirstCentralBaptistChurch
Annual Bill: $5,043.60

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

Get Started with Data Science Experience Today!

Calling all Data Scientists!

- Our mission is to win the **hearts and minds** of Data Scientists

- IBM Data Science Experience is a **freemium model** with value-add features, pricing and up-sell in development

- **Sign up** and encourage your colleagues to do so at **datascience.ibm.com**

IBM Data Science Experience

https://www.youtube.com/watch?v=1HjzkLRdP5k&t=29s