

Data Science

Emiliano Actis Dato
Analytics Client Architect
actisdato@ar.ibm.com
@emiactisdato

Quién soy?

Emiliano Actis Dato

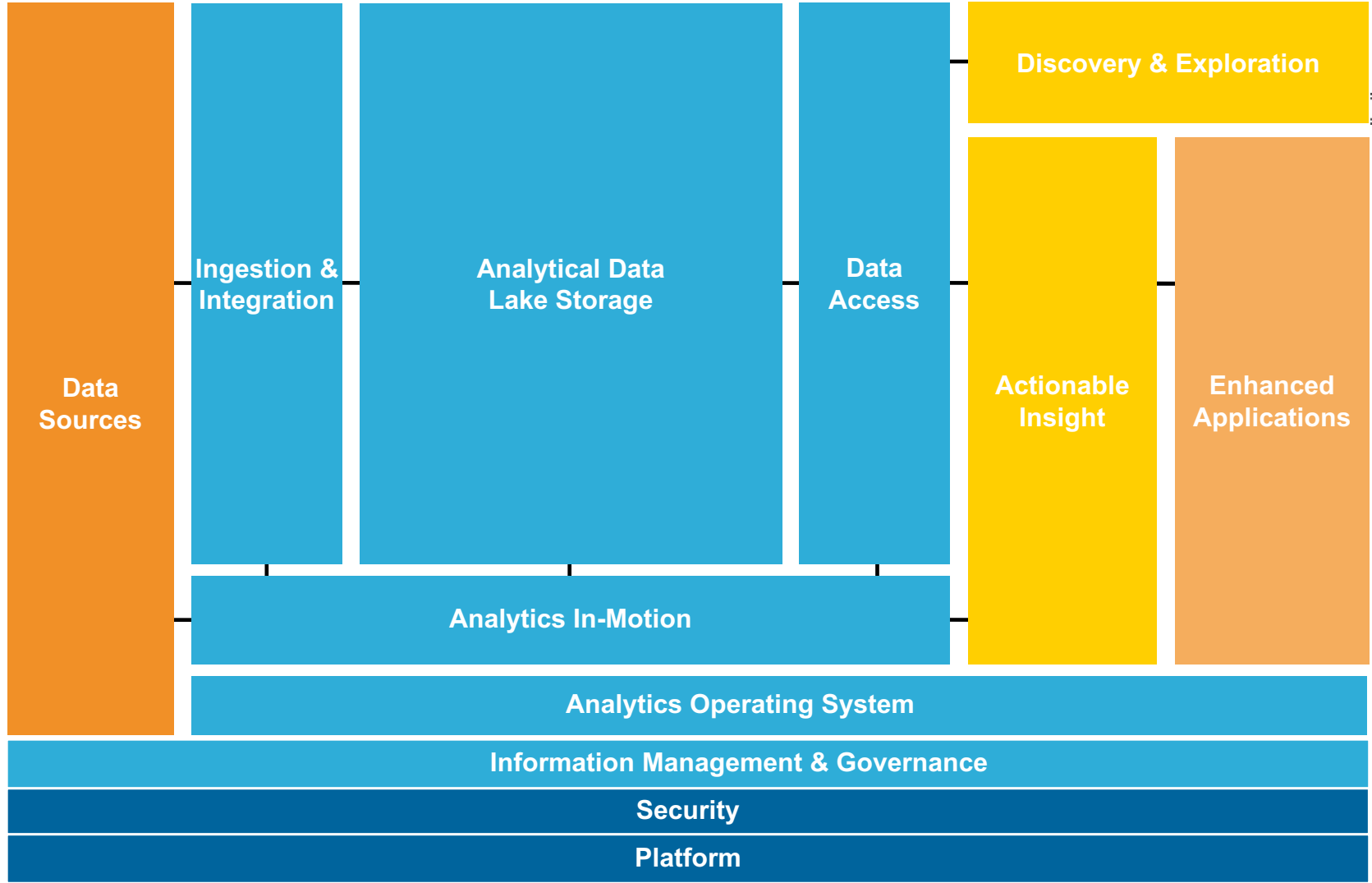
Ingeniero en Electrónica

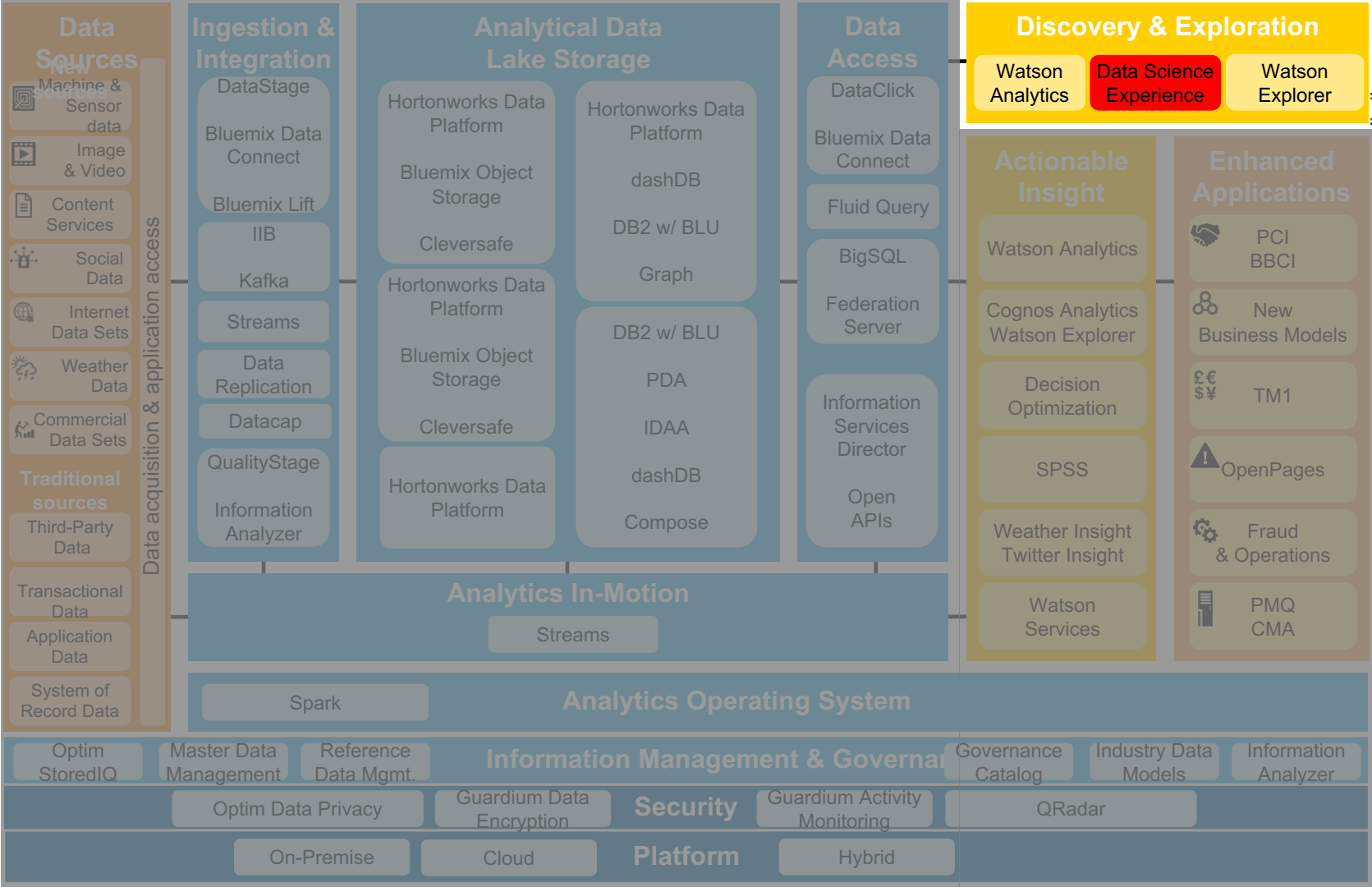
UNLP 1995-2001

11 años en IBM

Analytics Client Architect

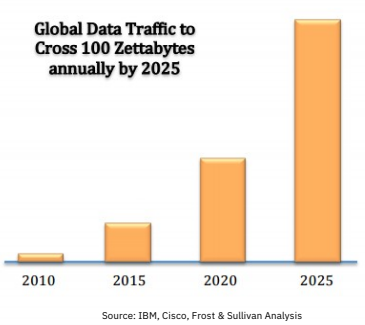




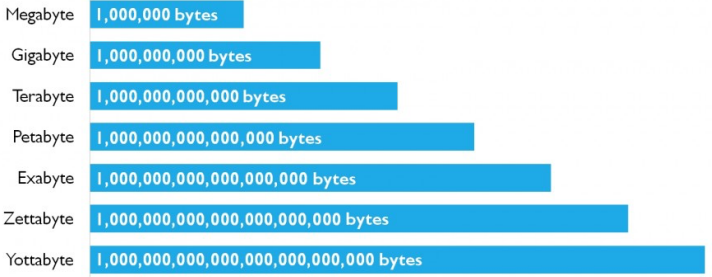


Por qué es Data Science Importante?

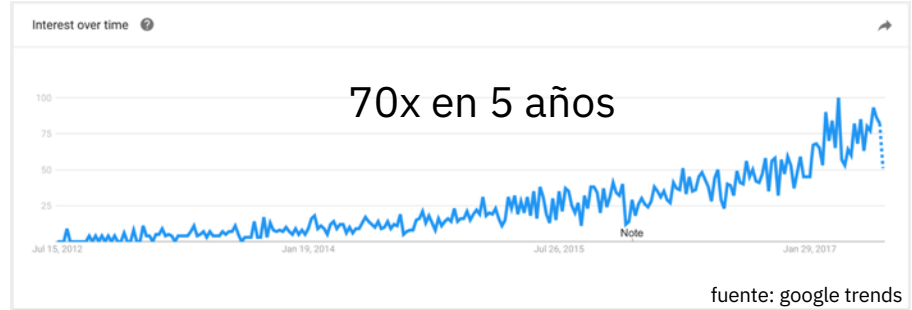
Cantidad de datos



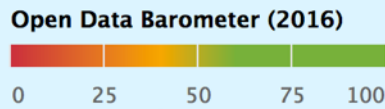
Qué hacen las empresas con los datos?



Data Science: Tendencia

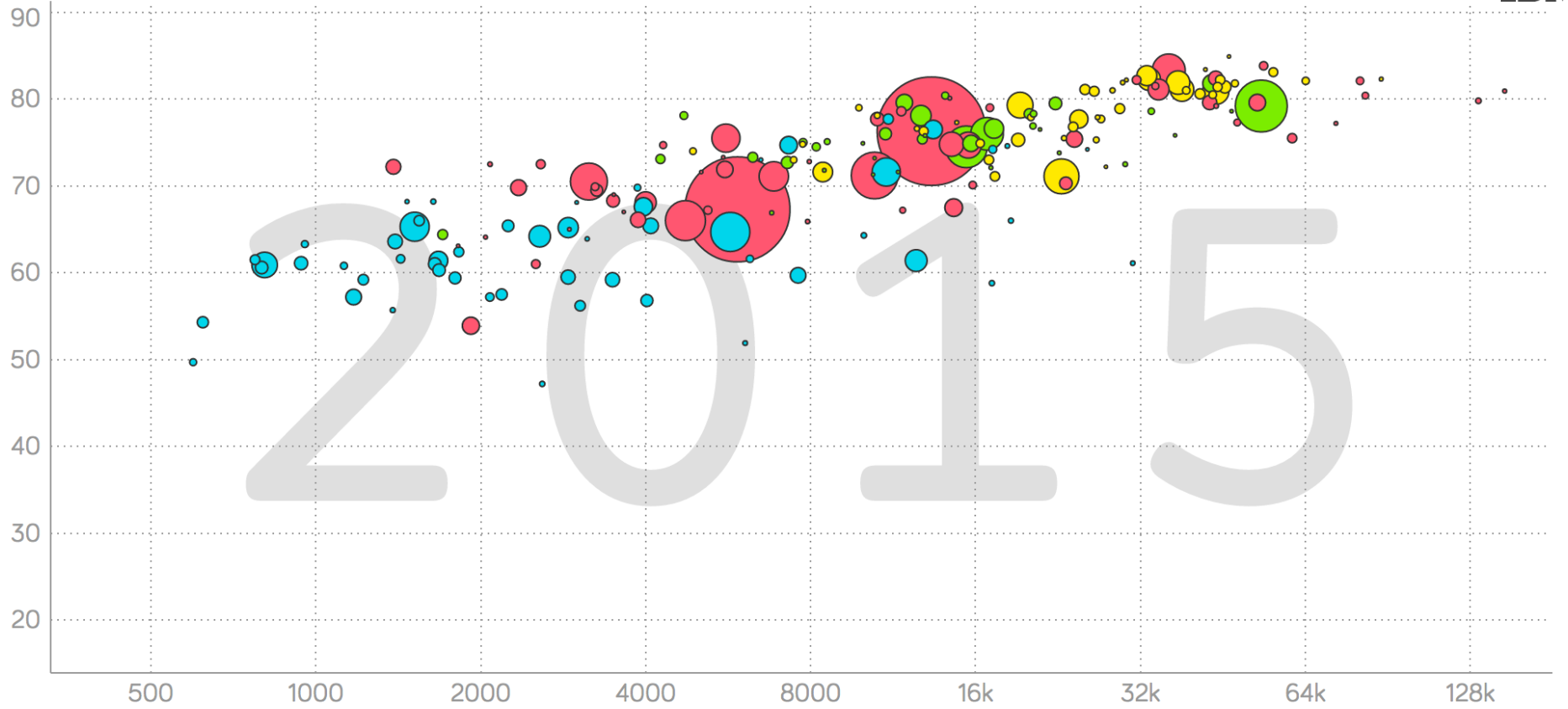


Argentina
Open Data Barometer
Value: 37.51 | Rank: 38



The Open Data Barometer

Life expectancy, years ?



Income per person, GDP/capita in \$/year adjusted for inflation & prices ?

El Data Scientist

- Talento
- Set de herramientas rígido
- Fragmentado y perdiendo tiempo
- Silo analítico

Data Science Experience (DSX)



Aprender



Crear



Colaborar



Artículos, Ejemplos, DataSets

ARTICLE ⋮

How can data scientists collaborate to build...

SOURCE
IBM

DATE
Jun 24, 2016

ARTICLE ⋮

What is machine learning?

SOURCE
IBM

DATE
Jun 24, 2016

NOTEBOOK ⋮

Insights from Twitter data about car makers

SOURCE
IBM

DATE
Jun 22, 2016

NOTEBOOK ⋮

Insights from New York car accident reports

SOURCE
IBM

DATE
Jun 16, 2016

DATA SET ⋮

Country Surface Area (sq. km)

SOURCE
IBM

DATE
Jun 16, 2016

NOTEBOOK ⋮

Improved Flight delay prediction

SOURCE
IBM

DATE
Jun 06, 2016

NOTEBOOK ⋮

Load data from different sources

SOURCE
IBM

DATE
Jun 02, 2016

NOTEBOOK ⋮

Learn basics about notebooks and Apache Spark

SOURCE
IBM

DATE
Jun 02, 2016

NOTEBOOK ⋮

Analyze precipitation data

SOURCE
IBM

DATE
Jun 02, 2016

GitHub



☰ ∞ Data Science Experience ▾ Settings

Integrations

Profile Services **Integrations**

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.

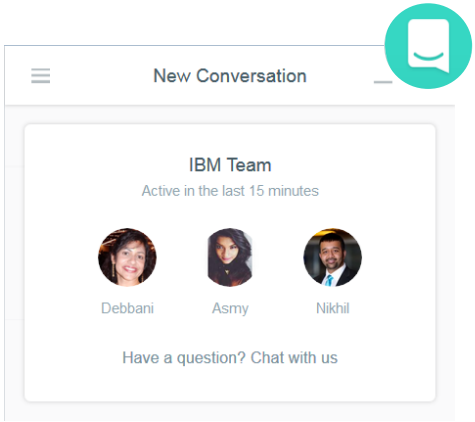
Paste generated personal access token here

40

Clear




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

Chat Online



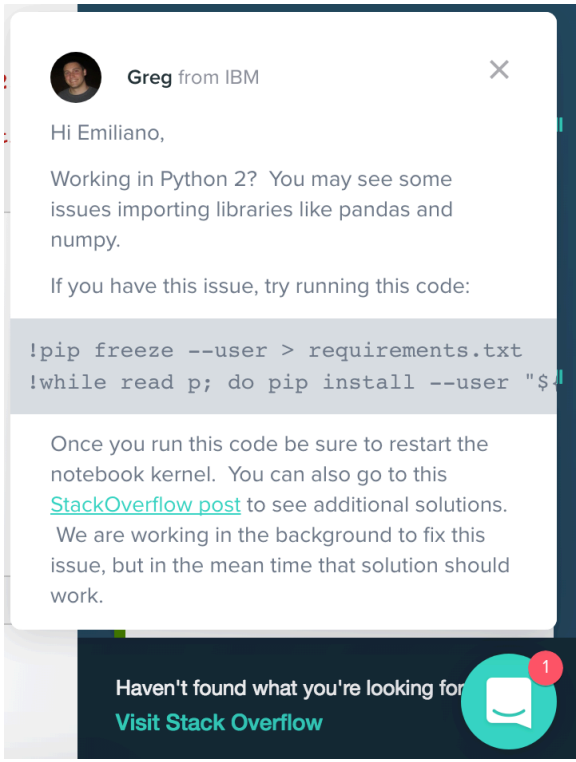
New Conversation

IBM Team
Active in the last 15 minutes

Debbani Asmy Nikhil

Have a question? Chat with us



Greg from IBM

Hi Emiliano,

Working in Python 2? You may see some issues importing libraries like pandas and numpy.

If you have this issue, try running this code:

```
!pip freeze --user > requirements.txt  
!while read p; do pip install --user "$p
```

Once you run this code be sure to restart the notebook kernel. You can also go to this [StackOverflow post](#) to see additional solutions. We are working in the background to fix this issue, but in the mean time that solution should work.

Haven't found what you're looking for
[Visit Stack Overflow](#)

Jupyter Notebooks

RStudio

Explore the correlations

```
In [9]: import matplotlib.pyplot as plt
import matplotlib
matplotlib.rcParams.update({'font.size': 10})
ff = pd.tools.plotting.scatter_matrix(dfScaled, diagonal='hist', figsize=(10,10))
```

The plot displays a scatter matrix for variables: energy, age, number_somewhere_feet, ple_1, ple_3, domestic_gasboating_gas, and heading_sadnessmic_gaps. The diagonal shows histograms, and the off-diagonal cells show scatter plots of pairs of variables.

```
library(ggmap)
# import some example data
bloombury <- importKML(site = "b30", year = 2005-2010, net = TRUE)
# have a look at the data
summaryPlot(bloombury)
# trend in o3 by wd
smoothTrend(bloombury, pollutant = "o3", desseason = TRUE, type = "wd")
# polarPlot of raw
polarPlot(bloombury, pollutant = "o3", type = "daylight")
# calendar plot
calendarPlot(bloombury, pollutant = "o3", )
```

Console

```
library(ggmap)
# import some example data
bloombury <- importKML(site = "b30", year = 2005-2010, net = TRUE)
# have a look at the data
summaryPlot(bloombury)
# trend in o3 by wd
smoothTrend(bloombury, pollutant = "o3", desseason = TRUE, type = "wd")
# polarPlot of raw
polarPlot(bloombury, pollutant = "o3", type = "daylight")
NOTE - mass units are used
ug/m3 for ND6, ND2, SO2, O3; ng/m3 for CO
PM2.5 raw is raw data multiplied by 1.3
Warning message:
In importKML(site = "b30", year = 2005-2010, net = TRUE) :
Some of the more recent data may not be fulfilled.
      date      desc      desc      desc      desc      desc      desc      desc      desc      desc
      code      wd      solar      rain      temp      tp      rhum      site
"character" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "factor"
```

Workspace

bloombury 50004 obs. of 18 variables

Files **Plots** **Packages** **Help**

O₃ in 2006

	January	February	March	April
1	1	2	3	4
2	5	6	7	8
3	9	10	11	12
4	13	14	15	16
5	17	18	19	20
6	21	22	23	24
7	25	26	27	28
8	29	30	31	
9		1	2	3
10	4	5	6	7
11	8	9	10	11
12	12	13	14	15
13	16	17	18	19
14	20	21	22	23
15	24	25	26	27
16	28	29	30	31
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				
38				
39				
40				
41				
42				
43				
44				
45				
46				
47				
48				
49				
50				
51				
52				
53				
54				
55				
56				
57				
58				
59				
60				
61				
62				
63				
64				
65				
66				
67				
68				
69				
70				
71				
72				
73				
74				
75				
76				
77				
78				
79				
80				
81				
82				
83				
84				
85				
86				
87				
88				
89				
90				
91				
92				
93				
94				
95				
96				
97				
98				
99				
100				

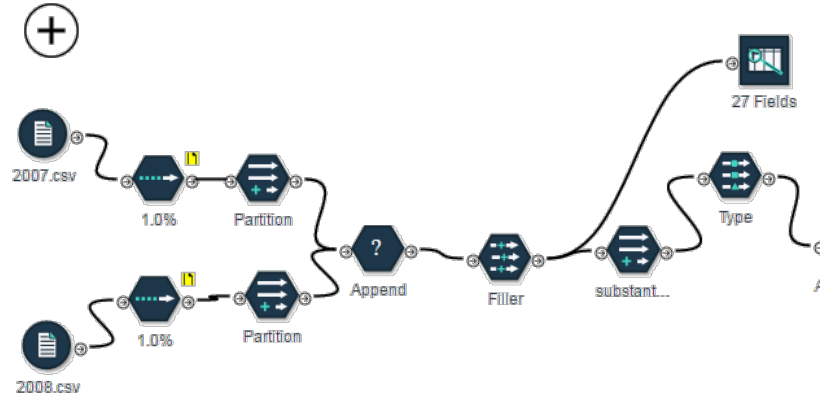


Complementos

Model Builder

- Select Data
- Prepare
- Train
- Select Model
- Evaluate
- Deploy
- Predict

Flow



Cómo consigo DSX?

Tres opciones:

- DSX en Bluemix: <http://datascience.ibm.com>
- DSX Desktop: <http://datascience.ibm.com/desktop>
- DSX Local: <https://datascience.ibm.com/local>

Pricing Plans Monthly prices shown are for country or region: [Argentina](#)

PLAN	FEATURES	PRICING
✓ Free	5 GB Object Storage 2 Spark executors Support via the community In-app chat support	Free

The Free plan for Data Science Experience offers everything you need to become a better data scientist in a collaborative environment.

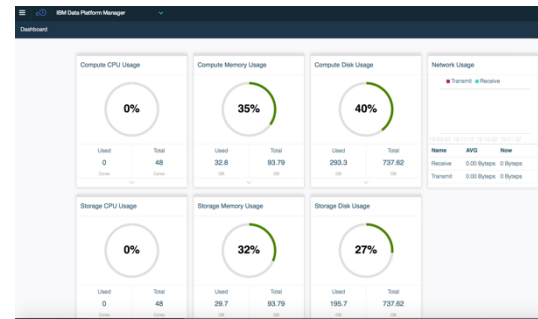
Desktop

Master the art of data science from your desktop

Download Beta for Windows

Download Beta for Mac

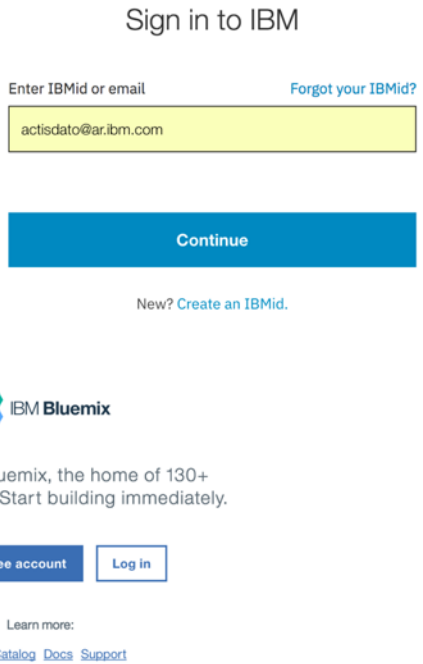
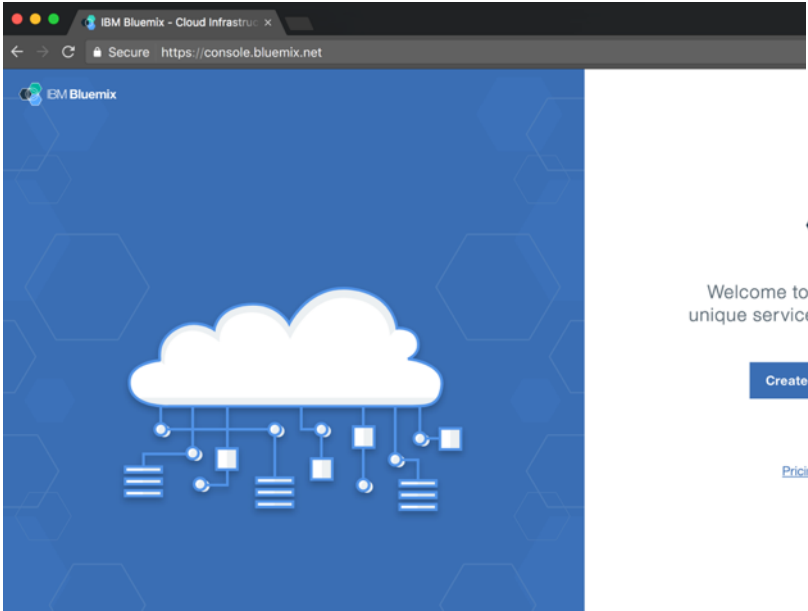
Download Beta for Linux



DSX en Bluemix en 3 pasos

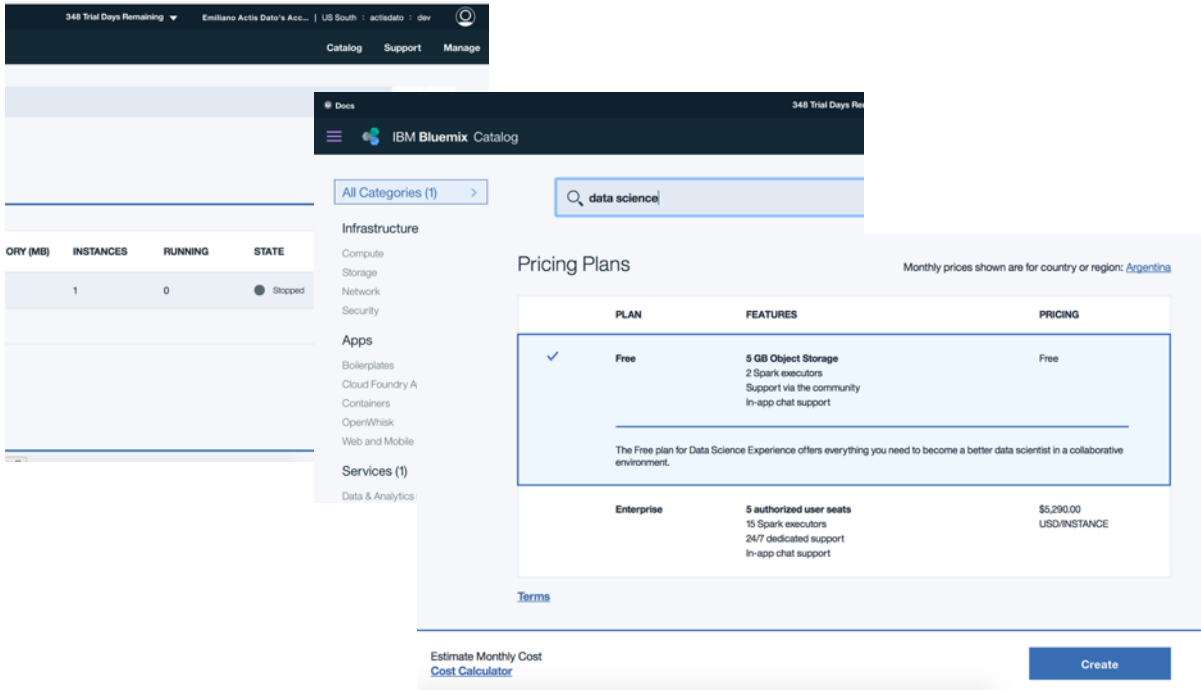
Paso 1: Abrir cuenta en Bluemix

- <http://bluemix.net>



DSX en Bluemix en 3 pasos

Paso 2: Crear Servicio Data Science Experience



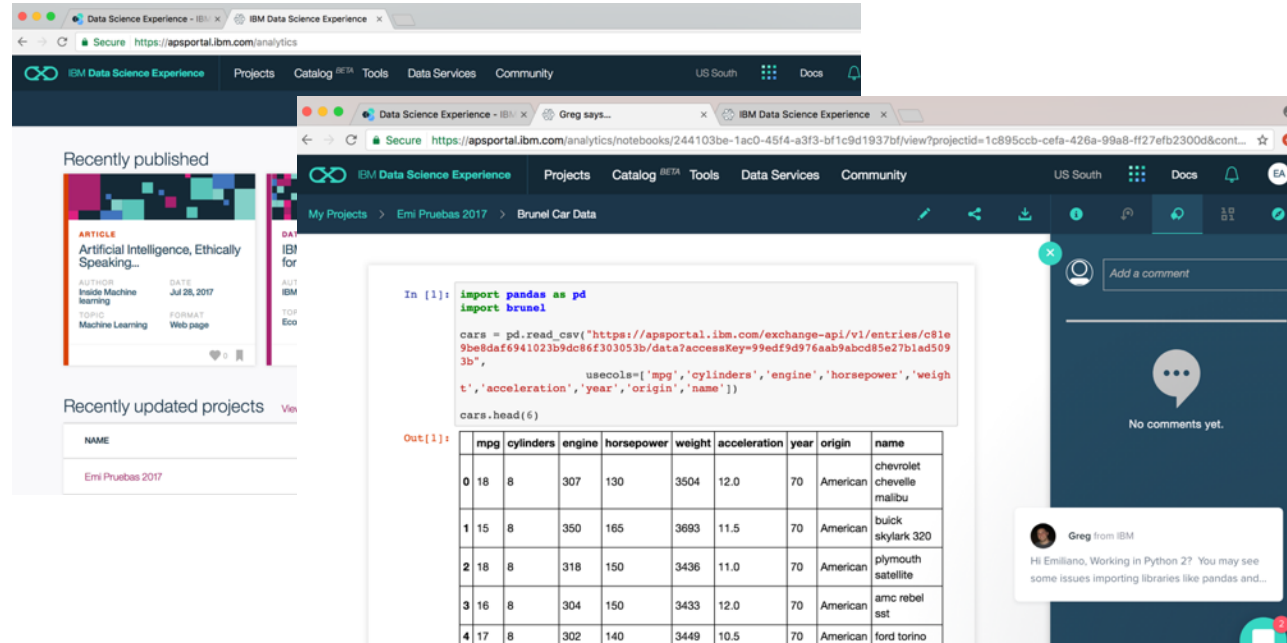
The screenshot shows the IBM Bluemix Catalog interface. At the top, there is a dark navigation bar with 'Catalog', 'Support', and 'Manage' options. Below this, a search bar contains the text 'data science'. The main content area is divided into two columns. The left column contains a table with columns 'ORY (MB)', 'INSTANCES', 'RUNNING', and 'STATE'. The right column displays 'Pricing Plans' for 'data science'. A note indicates that monthly prices are shown for the country or region of Argentina. The pricing table has three columns: 'PLAN', 'FEATURES', and 'PRICING'. The 'Free' plan is selected with a checkmark and includes 5 GB Object Storage, 2 Spark executors, community support, and in-app chat support. The 'Enterprise' plan costs \$5,290.00 USD/INSTANCE and includes 5 authorized user seats, 15 Spark executors, 24/7 dedicated support, and in-app chat support. At the bottom, there is a 'Cost Calculator' link and a blue 'Create' button.

PLAN	FEATURES	PRICING
Free	5 GB Object Storage 2 Spark executors Support via the community In-app chat support	Free
Enterprise	5 authorized user seats 15 Spark executors 24/7 dedicated support In-app chat support	\$5,290.00 USD/INSTANCE

DSX en Bluemix en 3 pasos

Paso 3: Usar Data Science Experience

- <http://datascience.ibm.com>



The screenshot shows the IBM Data Science Experience (DSX) interface. The top navigation bar includes 'EM Data Science Experience', 'Projects', 'Catalog BETA', 'Tools', 'Data Services', and 'Community'. The main content area displays a notebook titled 'Brunel Car Data' with the following Python code and output:

```
In [1]: import pandas as pd
import brusel

cars = pd.read_csv("https://apeportal.ibm.com/exchange-api/v1/entries/c81e9be8daf6941023b9dc86f303053b/data?accessKey=99edf9d976aab9abcd85e27blad5093b",
                  usecols=['mpg', 'cylinders', 'engine', 'horsepower', 'weight', 'acceleration', 'year', 'origin', 'name'])
cars.head(6)
```

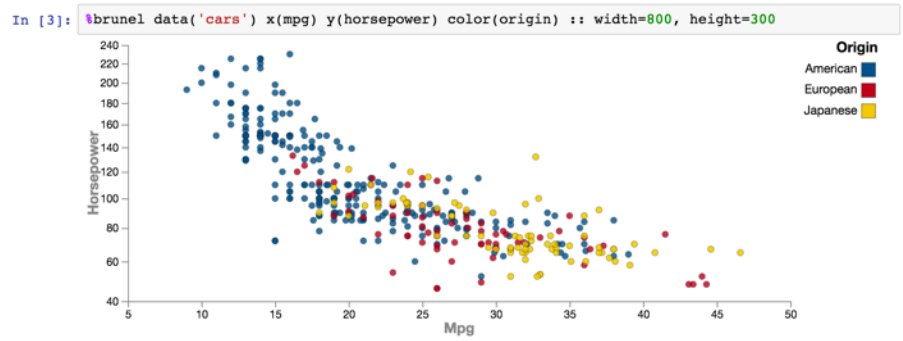
Out[1]:

	mpg	cylinders	engine	horsepower	weight	acceleration	year	origin	name
0	18	8	307	130	3504	12.0	70	American	chevrolet chevelle malibu
1	15	8	350	165	3693	11.5	70	American	buick skylark 320
2	18	8	318	150	3436	11.0	70	American	plymouth satellite
3	16	8	304	150	3433	12.0	70	American	amc rebel sst
4	17	8	302	140	3449	10.5	70	American	ford torino

The interface also shows a 'Recently published' section with an article titled 'Artificial Intelligence, Ethically Speaking...' and a 'Recently updated projects' section with a project named 'Emi Pruebas 2017'. A comment box on the right shows a comment from 'Greg from IBM'.

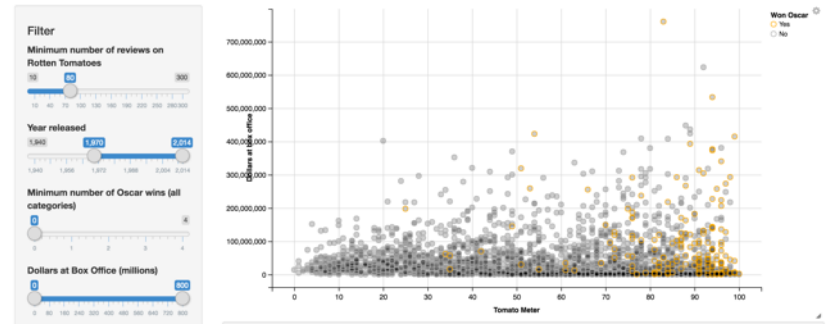
Demo

- Demo 1: Brunel Visualization



- Demo 2: Shiny Movie Explorer

Movie explorer



1. Hay muchos datasets disponibles
2. DSX Cloud gratis por 30 días (<http://datascience.ibm.com>)
 - DSX Desktop gratis sin límite de tiempo
3. El momento es ahora!