



Blurring the Line Between Developer and Data Scientist

Notebooks with PixieDust

va barbosa | va@us.ibm.com

Developer Advocacy

IBM Watson Data Platform



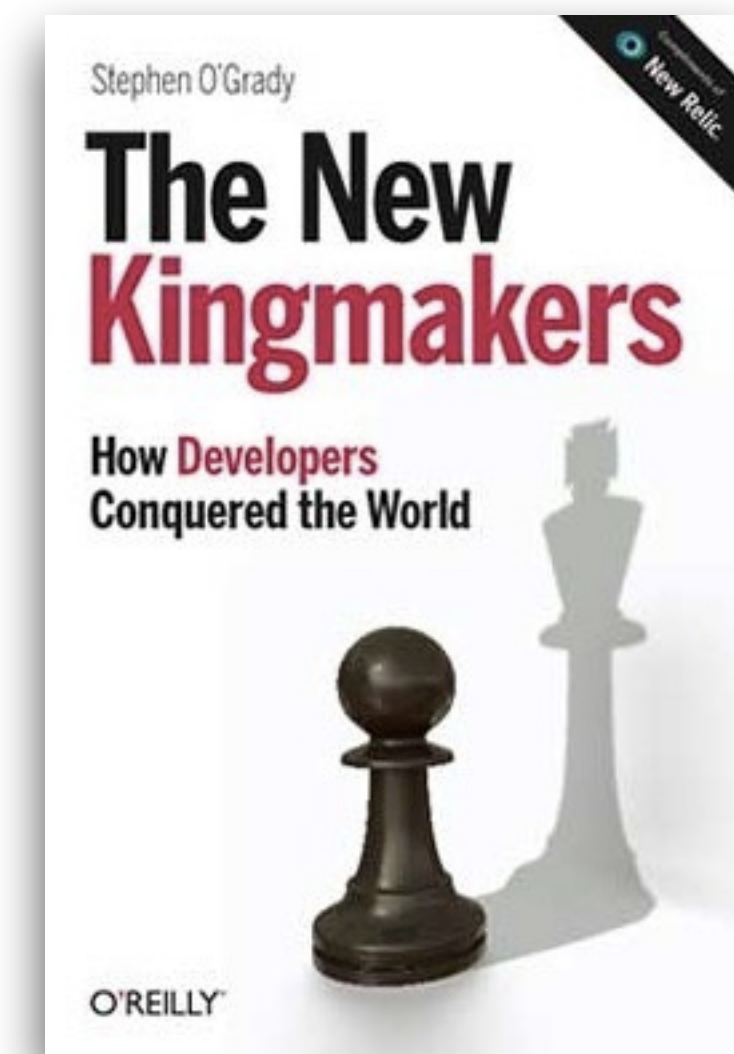
WHY ARE YOU HERE?

- More companies making **bet-the-business data driven decisions**
 - Good news: they are drowning in Data
 - Bad news: they are drowning in Data
- **Solving the Data problems of tomorrow** cannot be done by data scientists alone.
- Decision making has moved from the elite few to the empowered many: **Developers are getting more involved with Data Science**, moving from stovepipe applications to “data pipelines” that integrate data and analytics.

DEVELOPERS ARE THE “NEW BUILDERS”

The most successful companies today are those that understand the strategic role that developers will play in their success or failure. Not just successful technology companies – virtually every company today needs a developer strategy. There’s a reason that ESPN and Sears have rolled out API programs, that companies are being bought not for their products but their people. The reason is that developers are the most valuable resource in business.

<https://thenewkingmakers.com>






How do we blur the lines between developers and data scientists?

Let's start with a story... we all know too well.

Disclaimer: All characters and events depicted in this story are entirely fictitious. Any similarity to actual use cases, events or persons is actually intentional.



MEET BEN

THE DEVELOPER



- Holds a master degree in computer science
- 10 year experience, 6 years with the company
- Full stack Web developer
- Languages of choice: Java, Node.js, HTML5/CSS3
- Data: No SQL (Cloudant, Mongo), relational
- Protocols: REST, JSON, MQTT
- No major experience with Big Data

“The best line of code is the one I didn't have to write!”

MEET NATASHA

THE DATA SCIENTIST



- Holds a PHD in data science
- 5 year experience, 2 years with the company
- Experienced in Python and R
- Expert in Machine Learning and Data visualization
- Software engineering is not her thing

“In God we trust. All others bring data.”

— W. Edwards Deming



SURPRISE MEETING

With the VP of Development

“We have an urgent need for our marketing department to build an application that can provide real-time sentiment analysis on Twitter data.”

KEY CONSTRAINTS

- You only have 6 weeks to build the application
- Target consumer is the LOB user
 - Must be easy to use even for non technical people
- Web interface
 - Should be accessible from a standard browser
- It must scale out of the box
 - I want you to look at Apache Spark

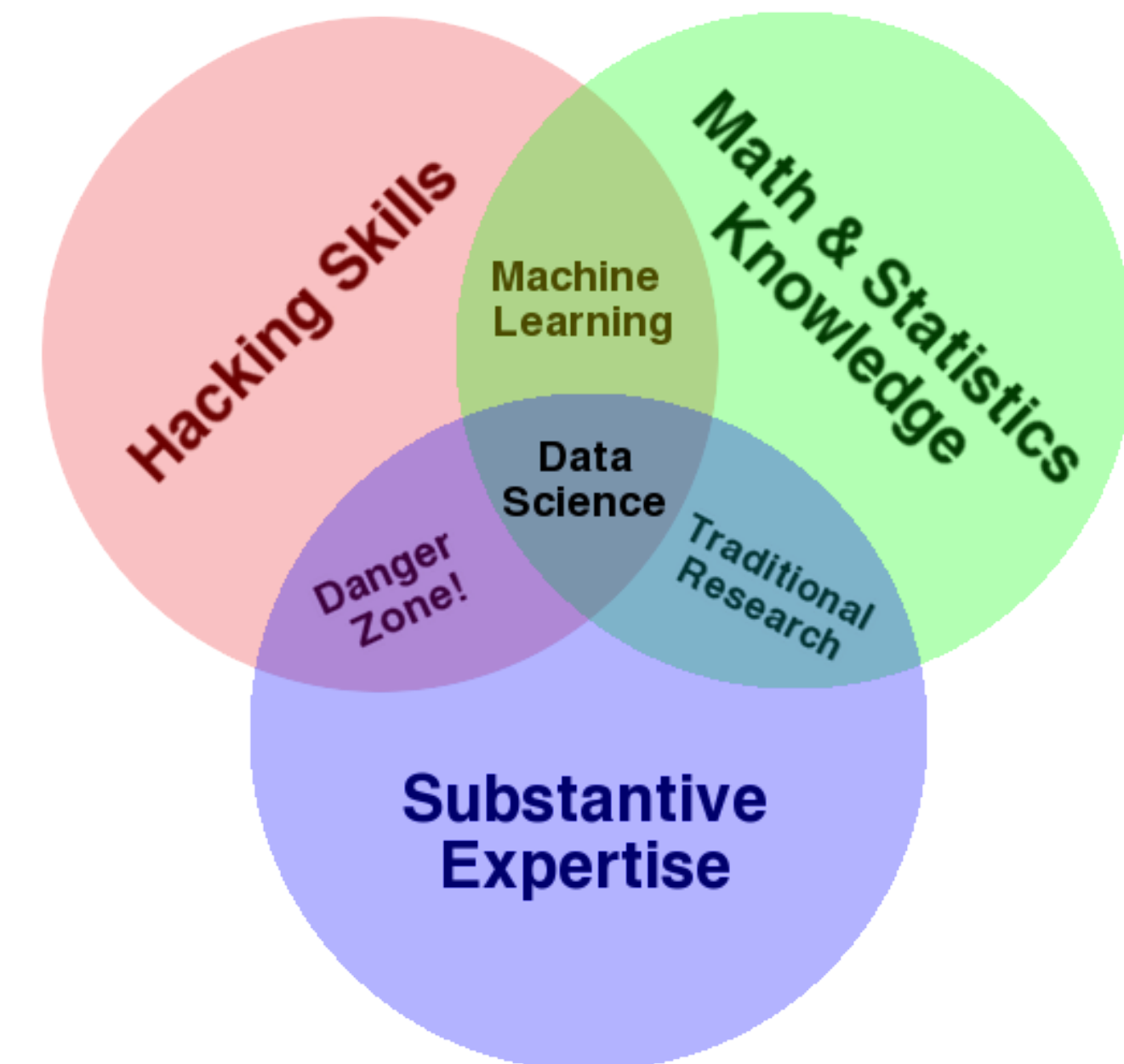
SOME LEARNING TO DO...

“What exactly is Data Science?”

— BEN 

WHAT IS DATA SCIENCE?

Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to Knowledge Discovery in Databases (KDD).



SOME LEARNING TO DO...

“What exactly is Apache Spark?”

— NATASHA



WHAT IS APACHE SPARK?

Spark is an open source in-memory computing framework for distributed data processing and iterative analysis on massive data volumes.



SPARK CORE LIBRARIES

Spark
SQL

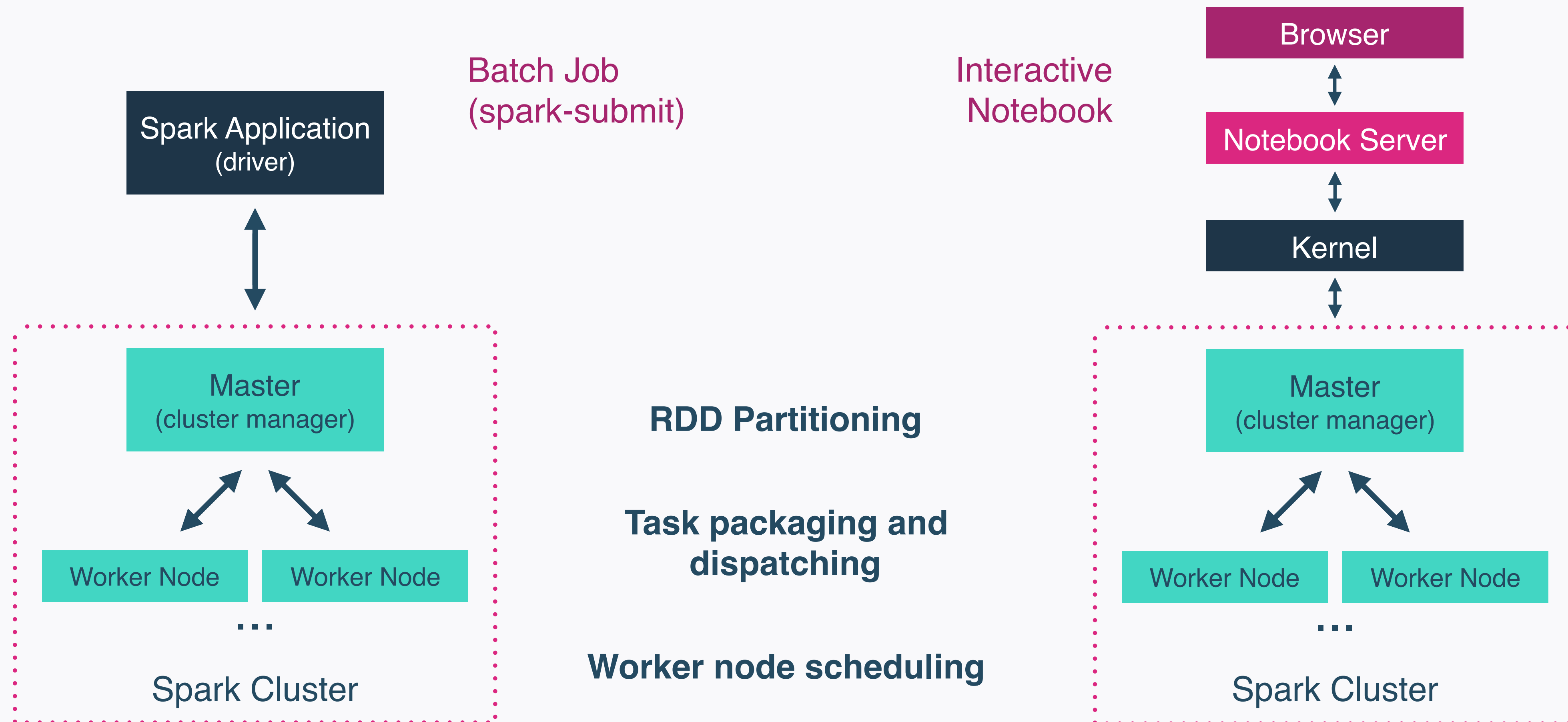
Spark
Streaming

MLlib
(machine learning)

GraphX
(graph)

Spark Core

CONSUMING SPARK



BEN and NATASHA

START BRAINSTORMING



- I'll work on data acquisition from Twitter and enrichment with sentiment analysis scores using Spark Streaming
- I know Java very well, but I don't have time to learn Python.
- However, I am willing to learn Scala if that helps improve my productivity

I'll need to do some data exploration too.



- I'll perform the data exploration and analysis
- I know Python and R, but I am not familiar enough with Java or Scala
- I like pandas and numpy. I'm ok to learn Spark but expect the same level of apis
- I need to work iteratively with the data

I'll need APIs to access my data.

CAN WE COLLABORATE USING NOTEBOOKS?

“What exactly is a Notebook?”

— NATASHA



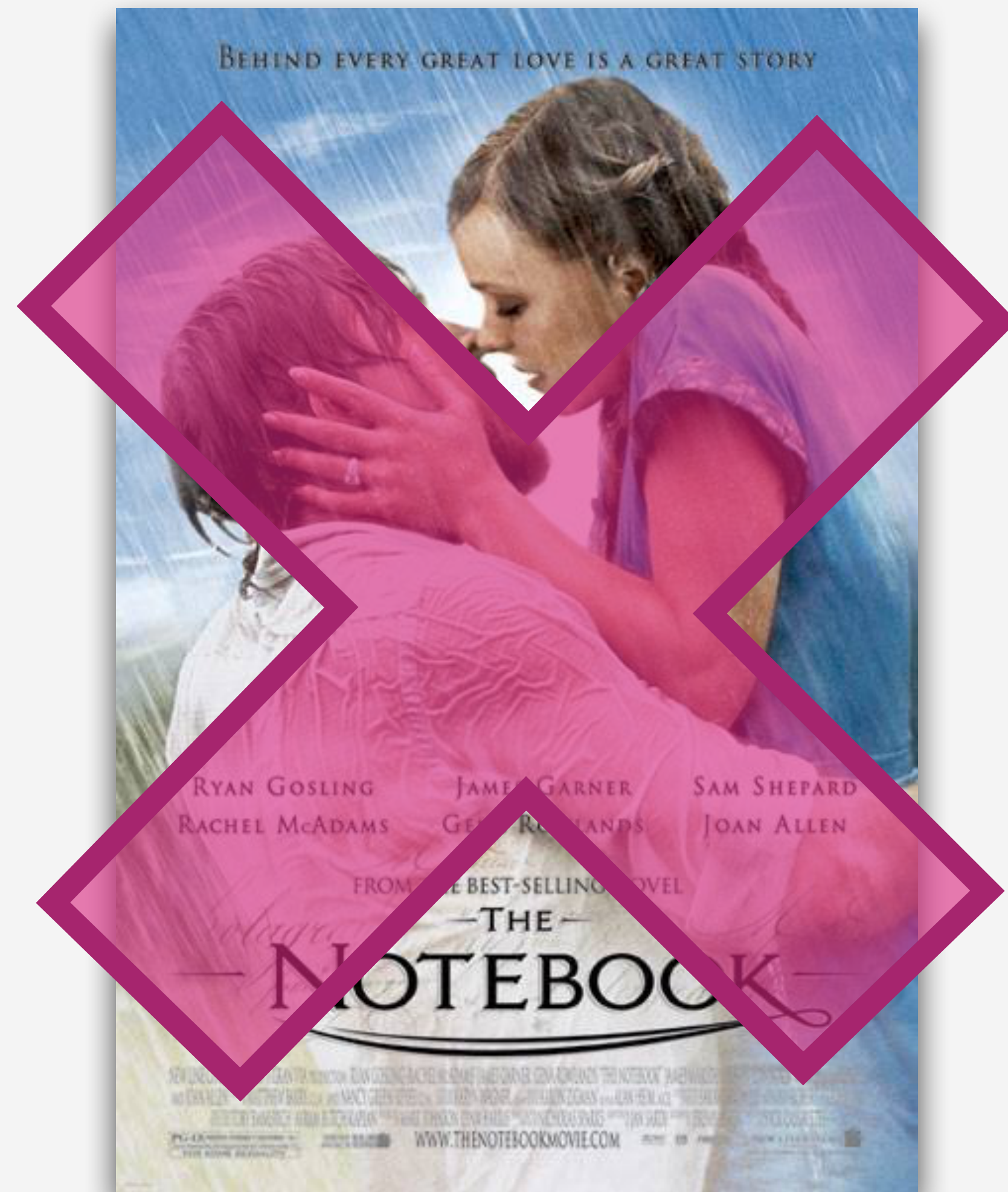
— BEN



BOOK FOR NOTES?



RYAN GOSLING MOVIE?

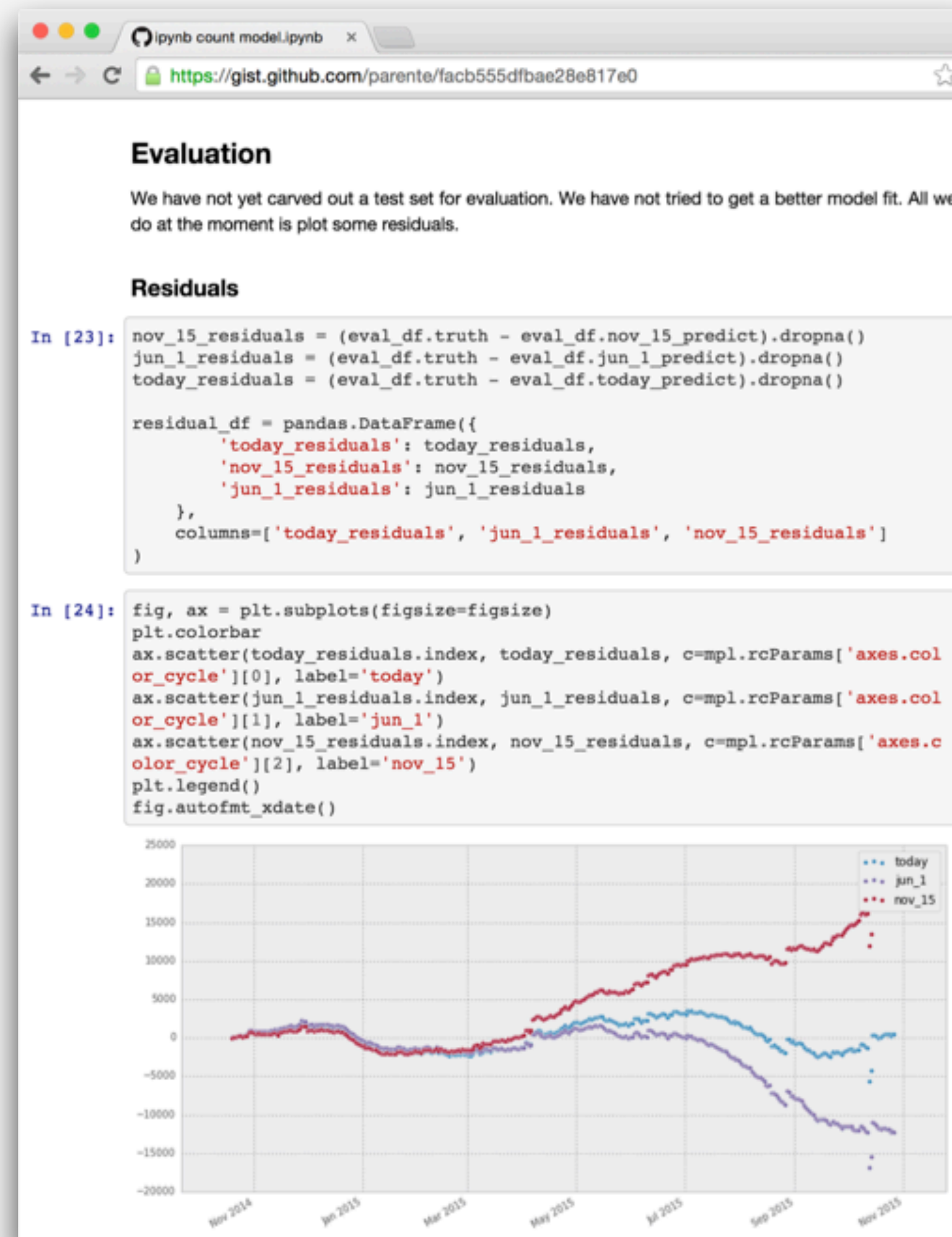


WHAT IS A NOTEBOOK?

Text
Annotations

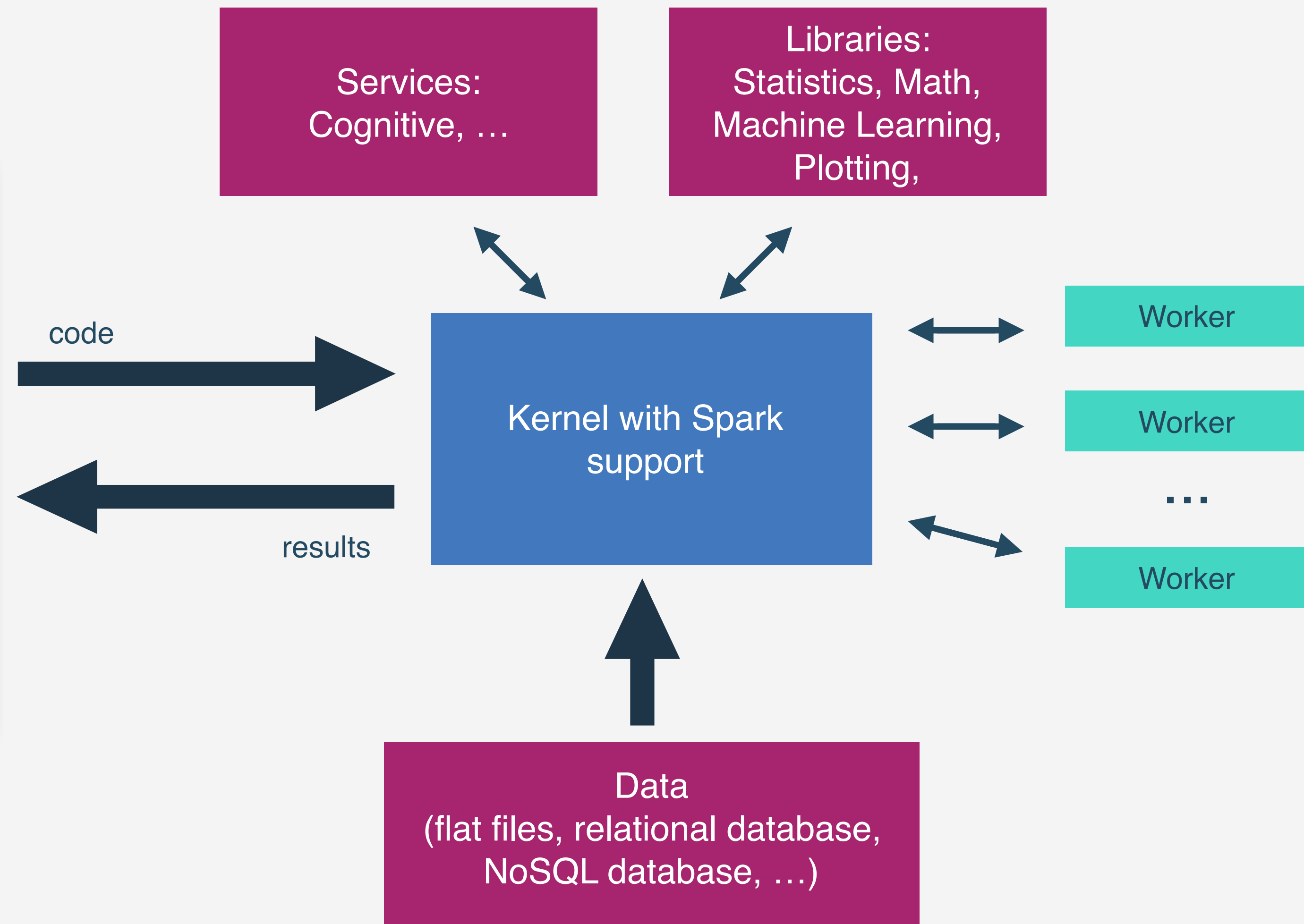
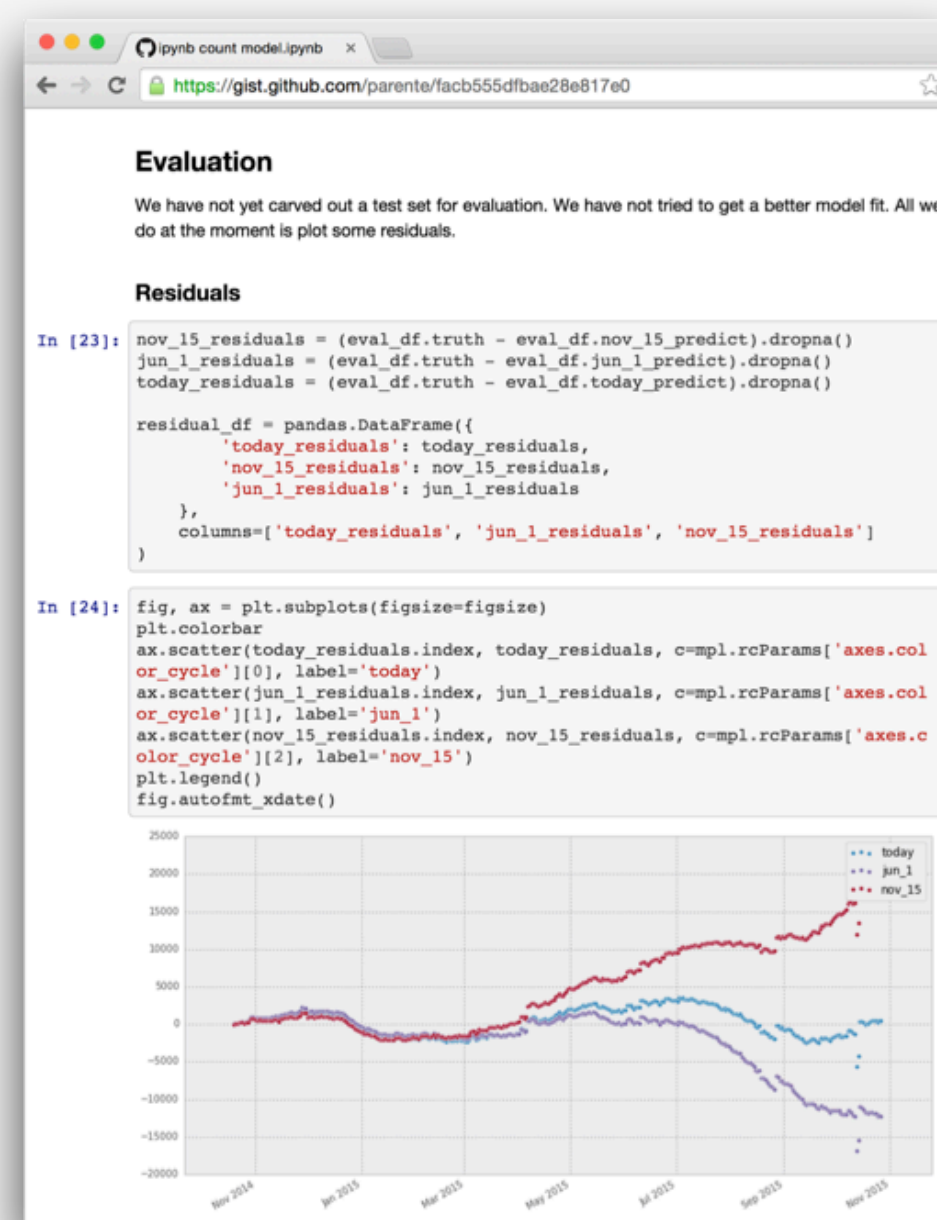
Code
Data

Visualizations
Widgets
Output



- Web based UI for running Apache Spark console commands
- Easy, no install spark accelerator
- Best way to start working with spark
- Multiple flavors
 - Jupyter
 - Zeppelin
- Local or cloud hosted
 - IBM Data Science Experience
 - Databricks

BIG DATA ANALYSIS



NOTEBOOKS ARE POWERFUL TOOLS FOR DATA SCIENTISTS

“But they seem complicated for developers like me”

— BEN



ENTER PIXIEDUST

Python helper library for Jupyter Notebooks



- Visualize data (e.g., Table, Charts, Map, etc)
- Download/export data (e.g., File, Cloudant, etc.)
- Use Scala directly in a Python notebook
- Install Spark packages into Python notebook
- Spark job progress monitor
- Extensible





PIXIEDUST

<https://github.com/ibm-cds-labs/pixiedust>

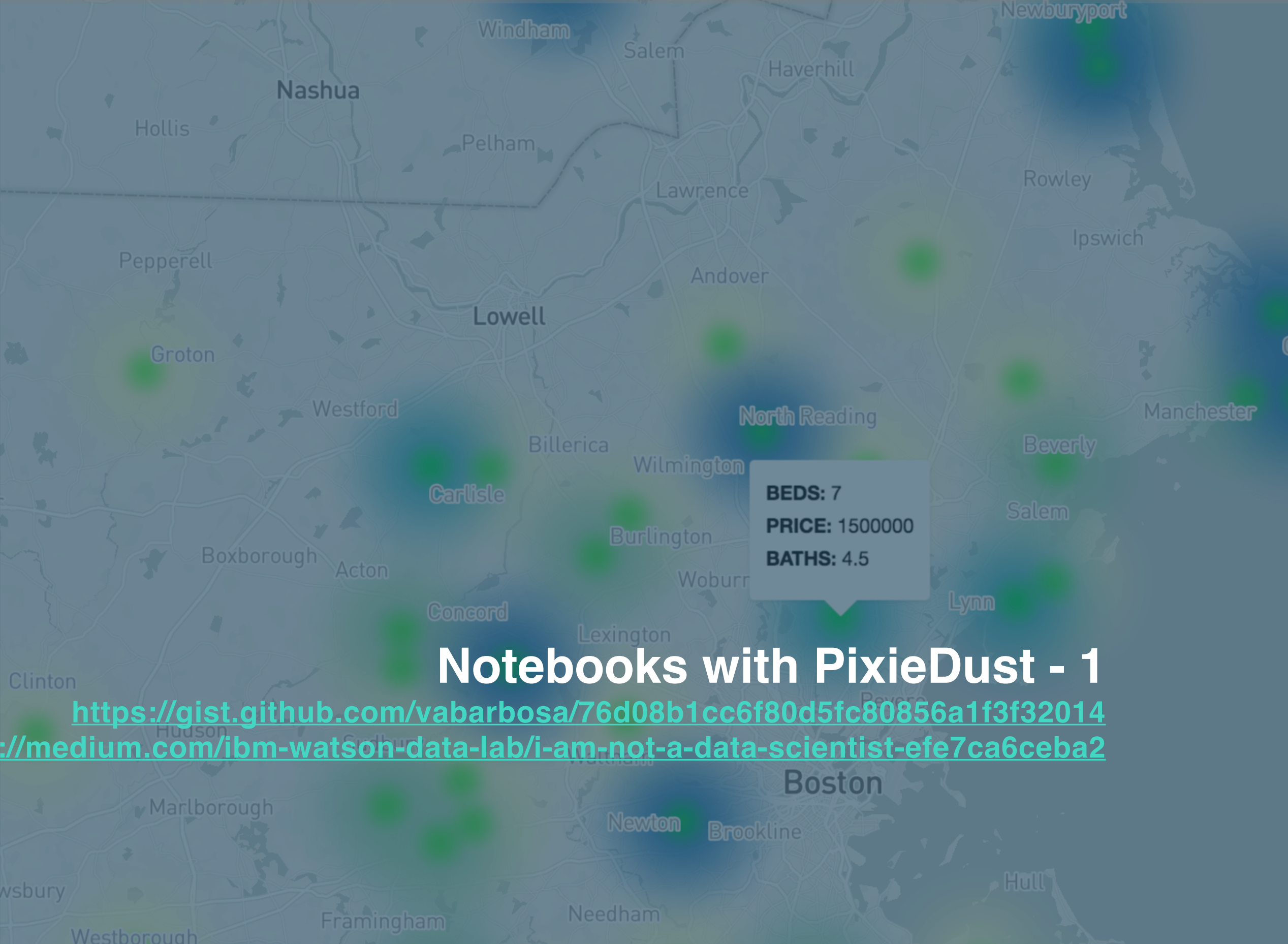
Developers can:

- Visualize my data, without having to learn Matplotlib, Bokeh, etc
- Explore my data in an interactive interface
- Use Spark without having to learn it
- Save my data locally or on the cloud with a simple menu
- Extend PixieDust with new visualizations

Data Scientists can:

- Use Python and Scala in the same notebook
 - Share variables between Scala and Python
 - Access Spark Libraries written in Scala from Python Notebooks
 - Monitor the progress of my Spark Jobs
- 
- 


```
import pixiedust
d1 = sqlContext.createDataFrame(
[(2010, 'Camping Equipment', 3, 200),(2010, 'Camping Equipment', 10, 200),(2010, 'Golf Equipment', 1, 240),
(2010, 'Mountaineering Equipment', 1, 348),(2010, 'Outdoor Protection',2,200),(2010, 'Personal Accessories', 2, 200),
(2011, 'Camping Equipment', 4, 489),(2011, 'Golf Equipment', 5, 234),(2011, 'Mountaineering Equipment',2, 123),
(2011, 'Outdoor Protection', 4, 654),(2011, 'Personal Accessories', 2, 234),(2012, 'Camping Equipment', 5, 876),
(2012, 'Golf Equipment', 5, 200),(2012, 'Mountaineering Equipment', 3, 156),(2012, 'Outdoor Protection', 5, 200),
(2012, 'Personal Accessories', 3, 345),(2013, 'Camping Equipment', 8, 987),(2013, 'Golf Equipment', 5, 434),
(2013, 'Mountaineering Equipment', 3, 278),(2013, 'Outdoor Protection', 8, 134),(2013, 'Personal Accessories',4, 200)],
["year", "zone", "unique_customers", "revenue"])
display(d1)
```



Notebooks with PixieDust - 1

<https://gist.github.com/vabarbosa/76d08b1cc6f80d5fc80856a1f3f32014>
<https://medium.com/ibm-watson-data-lab/i-am-not-a-data-scientist-efe7ca6ceba2>

WHAT IF PIXIEDUST DOESN'T HAVE THE
VISUALIZATION I WANT?

“Can I write my own?”

— BEN



PIXIEDUST EXTENSIBILITY APIs

Create your own visualizations by

- Defining the HTML fragment
 - Jinja2- Python templating engine
 - Bootstrap
 - Font Awesome
- Specify the metadata
 - Menu info

I AM OK TO USE PYTHON

“But I am really more
comfortable with Scala”

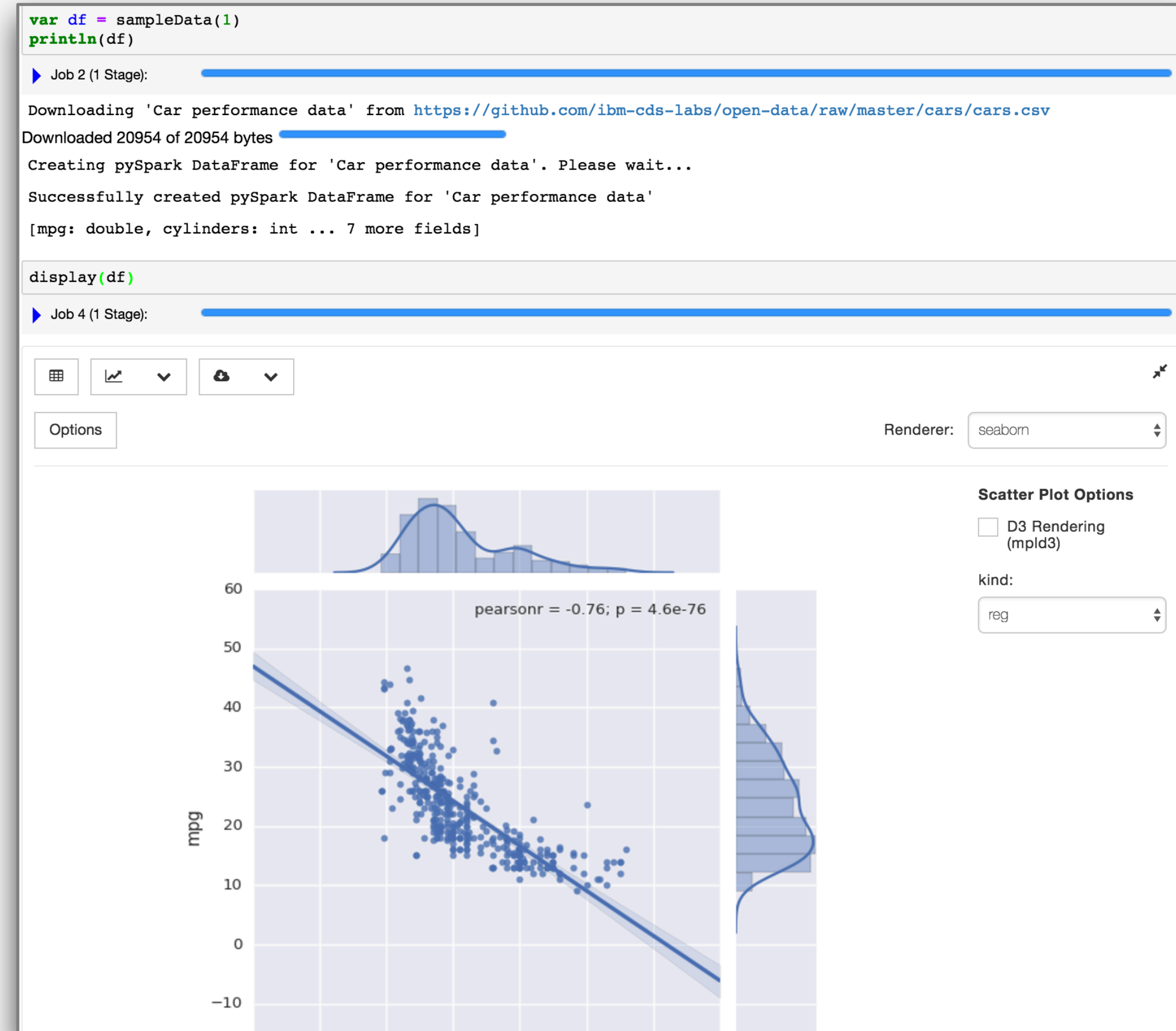
— BEN



SCALA NOTEBOOKS

PixieDust also works with
Scala Notebooks

Same PixieDust Scala APIs as in
Python








DEMO

```
@PixiedustDisplay()
class TestPluginMeta(DisplayHandlerMeta):
    @addId
    def getMenuInfo(self, entity, dataHandler):
        if entity.__class__.__name__ == "DataFrame":
            return [
                {"categoryId": "Table", "title": "NewSample Table",
                 "icon": "fa-table", "id": "newsampleTest"}
            ]
        else:
            return []
    def newDisplayHandler(self, options, entity):
        return TestDisplay(options, entity)
```

```
class TestDisplay(Display):
    def doRender(self, handlerId):
        self._addHTMLTemplateString(
            """
            <div>NewSample Plugin</div>
            <table class="table table-striped">
                <thead>
                    {%for field in entity.schema.fields%}
                    <th>{{field.name}}</th>
                    {%endfor%}
                </thead>
                <tbody>
                    {%for row in entity.take(100)%}
                    <tr>
                        {%for field in entity.schema.fields%}
                        <td>{{row[field.name]}}</td>
                        {%endfor%}
                    </tr>
                    {%endfor%}
                </tbody>
            </table>
            """
```

display(d1)

		
 DataFrame Table		
 NewSample Table	zone	
2010	Camping Equipment	
2010	Golf Equipment	
2010	Mountaineering Equipment	
2010	Outdoor Protection	
2010	Personal Accessories	

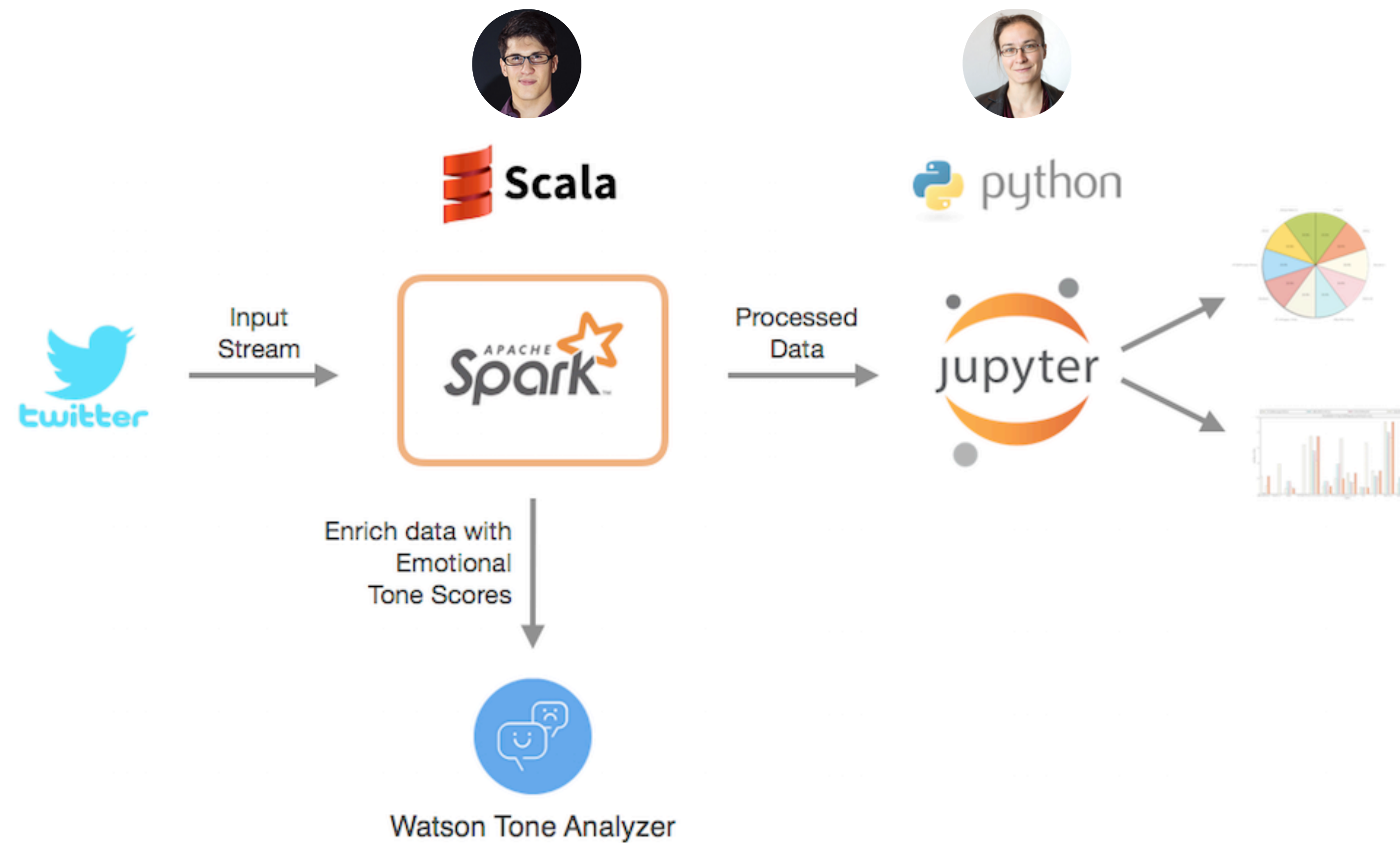
Notebooks with PixieDust - 2

<https://gist.github.com/vabarbosa/dca176c3a68f0c101cbe475571e56bf7>

<https://medium.com/ibm-watson-data-lab/you-too-can-make-magic-in-jupyter-notebooks-with-pixiedust-505d20f4fd13>

OK, I'M SOLD...

LET'S AGREE ON THE ARCHITECTURE



BEN and NATASHA

DIVIDING THE TASKS



- Implement a Spark Streaming connector to Twitter
- Call Watson Tone Analyzer for each tweets
- Return a Spark DataFrame with the tweets enriched with Tone scores
- Code written in Scala, delivered as a Jar



- Works in a Python Notebook
- Using PixieDust PackageManager, install the Scala library delivered by Ben to load the twitter data with Tone scores
- Using PixieDust display() api, perform the data exploration and analysis: trending hashtags and sentiments
- Produce visualizations to LOB Users

WATSON TONE ANALYZER

<http://www.ibm.com/watson/developercloud/tone-analyzer.html>

- Uses linguistic analysis to detect 3 types of tones
 - Emotion
 - Social Tendencies
 - Language Styles
- Available as a cloud service on IBM Bluemix

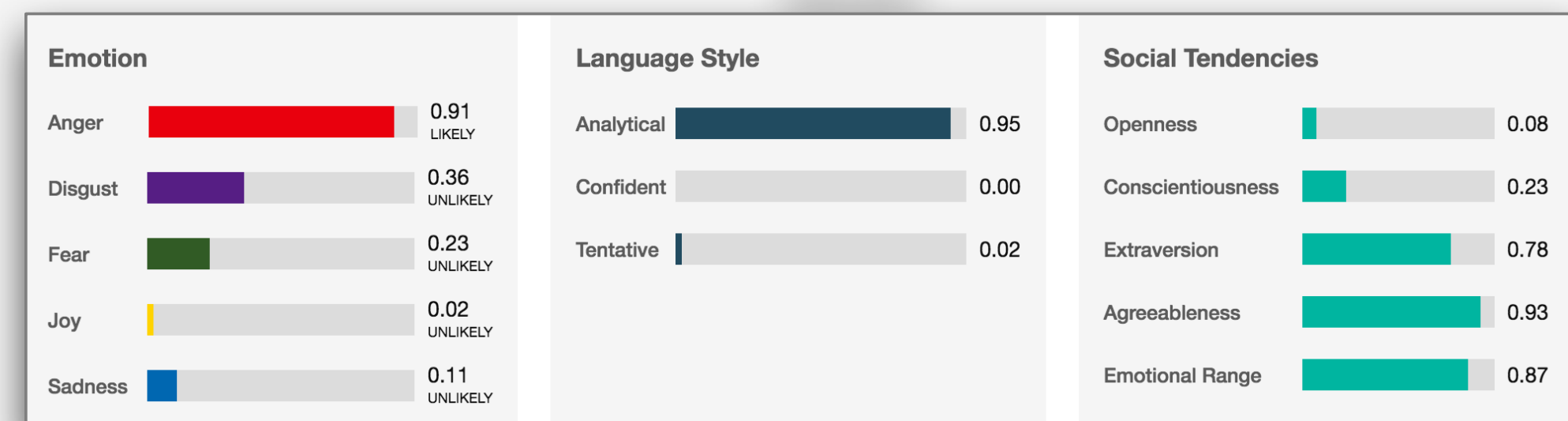
Input

Hi Team,

The times are difficult! Our sales have been disappointing for the past three quarters for our data analytics product suite. We have a competitive data analytics product suite in the industry. But we are not doing a good job at selling it.

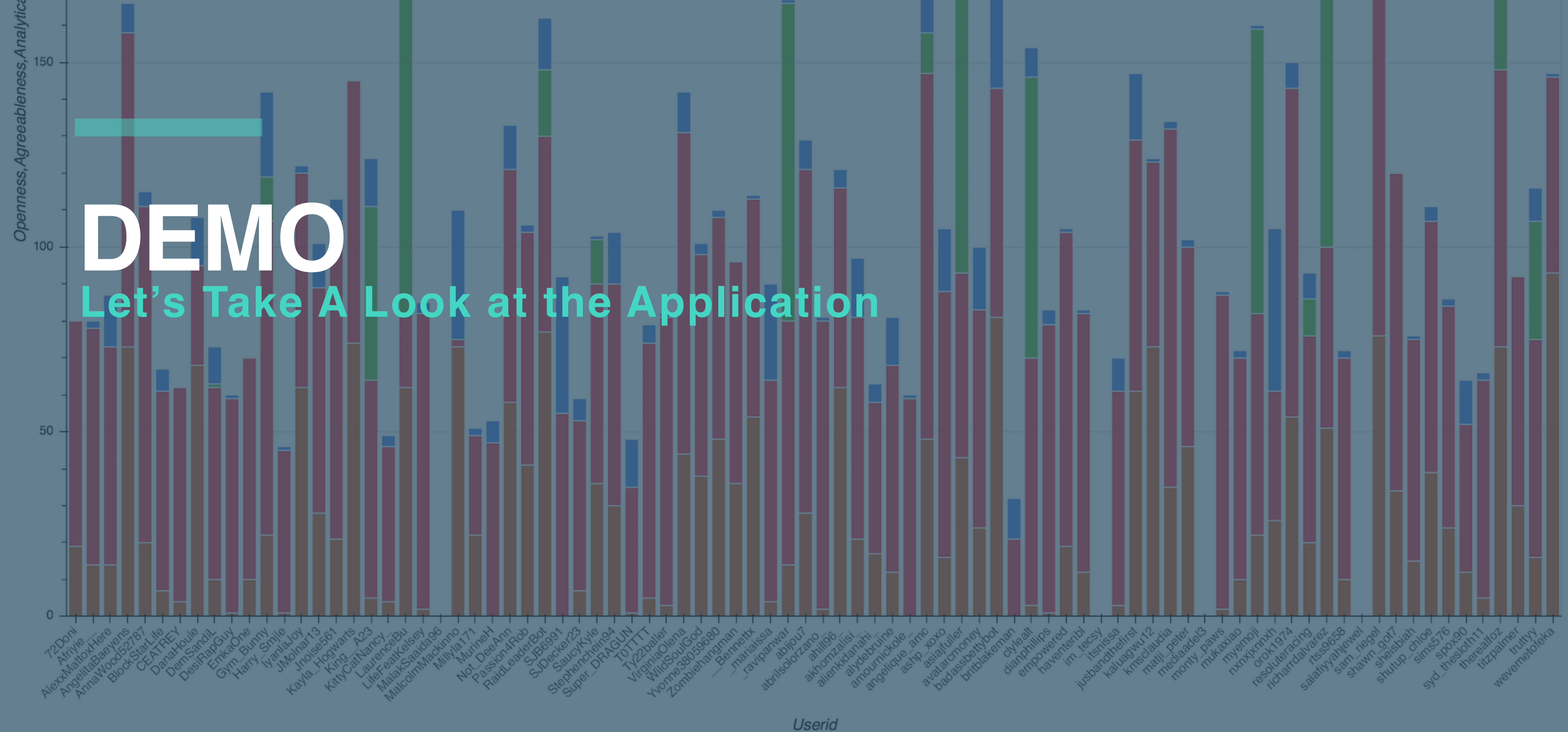
We need to acknowledge and fix our sales challenges. We cannot blame the economy for our lack of execution! We are missing critical sales opportunities. Our clients are hungry for analytical tools to improve their business outcomes. In fact, it is in times such

Results



DEMO

Let's Take A Look at the Application



Twitter Sentiment with Watson and PixieDust

```
In [6]: from operator import add
import re
tagsRDD = tweets.filter(lambda word: word.startswith("#")) \
    .map(lambda word : (word, 1)) \
    .reduceByKey(add, 10).map(lambda (a,b): (b,a)).sortByKey(False).map(lambda (a,b):(b,a))
```

<https://github.com/ibm-cds-labs/pixiedust/blob/master/notebook/Twitter%20Sentiment%20with%20Watson%20and%20Pixiedust.ipynb>

```
In [7]: from pyspark.sql.types import *
```




UPDATING THE VP

“This is great, but C-Suite executives need to be able to run the application from Notebook, select filters and see real-time charts without writing code!”

ter Feed

55 Tweets

trump,hillary

Clinton isn't w...
American people, only her
donors. Donald Trump...
us. <https://t.co/ThxR5BEgTW>

♥SheilaSpirit

@realDonaldTrump
@DebbieB69009667 that's a
first. How many will be fired
for reports of truth! She had
her hair OVER Her ear bud
ears...

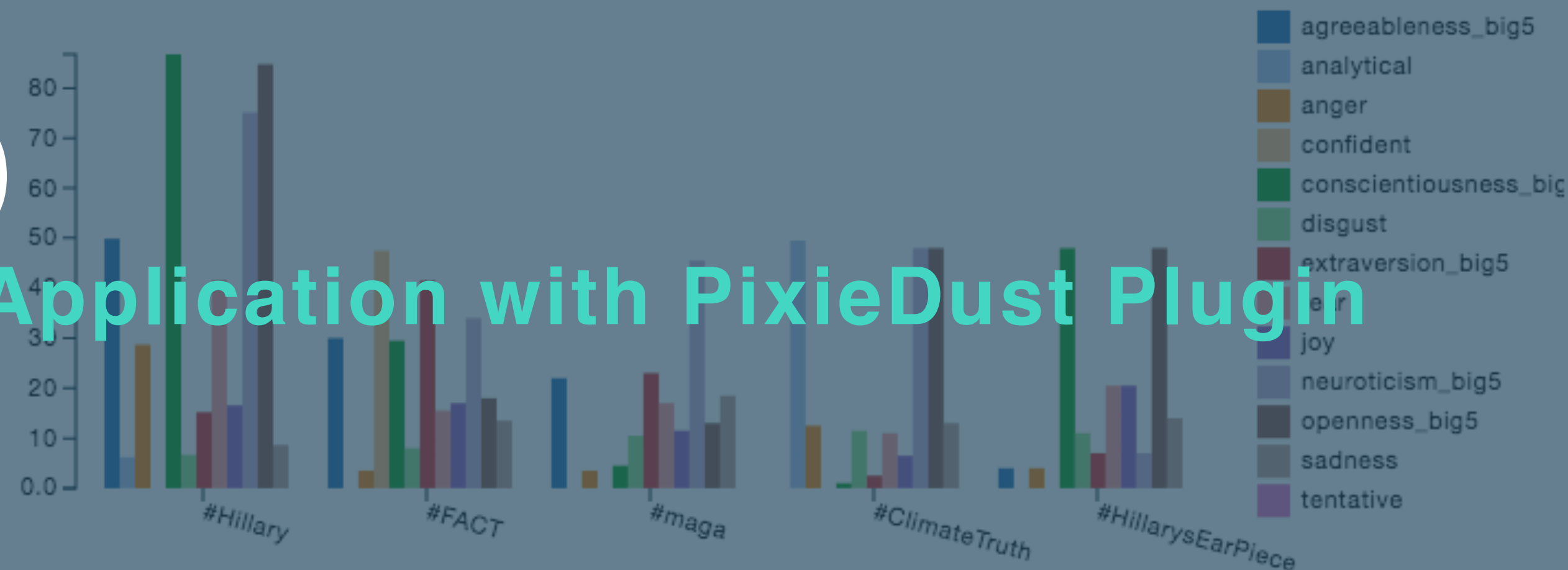
Mark

@CNNSitRoom wow Pence
agrees with Trump that Putin
is a better leader than Potus.
That's how you make
America great again

Shub

@Bioreducer @emily3183
@lwanski @realDonaldTrump

Trending Sentiments



Trending Hashtags



Streaming Logs

```
Thu Sep 08 2016 18:32:52 GMT-0400 (EDT)
Starting twitter stream
Twitter stream started
Receiver Started: TwitterReceiver-0
Batch started with 50 records
Batch completed with 50 records
Batch started with 243 records
Batch completed with 243 records
Batch started with 202 records
Batch completed with 202 records
Batch started with 203 records
Batch completed with 203 records
Batch started with 194 records
Batch completed with 194 records
Batch started with 209 records
Batch completed with 209 records
Batch completed with 206 records
Batch started with 209 records
Batch completed with 209 records
Batch started with 225 records
Batch completed with 225 records
Batch started with 210 records
Batch completed with 210 records
Batch completed with 183 records
Batch started with 212 records
Batch completed with 212 records
Batch started with 230 records
Batch completed with 230 records
Batch started with 250 records
Batch completed with 250 records
Batch started with 220 records
Batch completed with 220 records
Batch started with 206 records
Batch completed with 206 records
Batch started with 218 records
Batch completed with 218 records
Batch started with 210 records
Batch completed with 210 records
```

DEMO

Embedded Application with PixieDust Plugin

Sentiment Analysis of Twitter Hashtags with Spark

<https://github.com/ibm-cds-labs/pixiedust/blob/master/notebook/Twitter%20Sentiment%20with%20Watson%20and%20Pixiedust.ipynb>

<https://medium.com/ibm-watson-data-lab/real-time-sentiment-analysis-of-twitter-hashtags-with-spark-7ee6ca5c1585>

Start Streaming

Back to Notebo



MEETING WITH THE VP

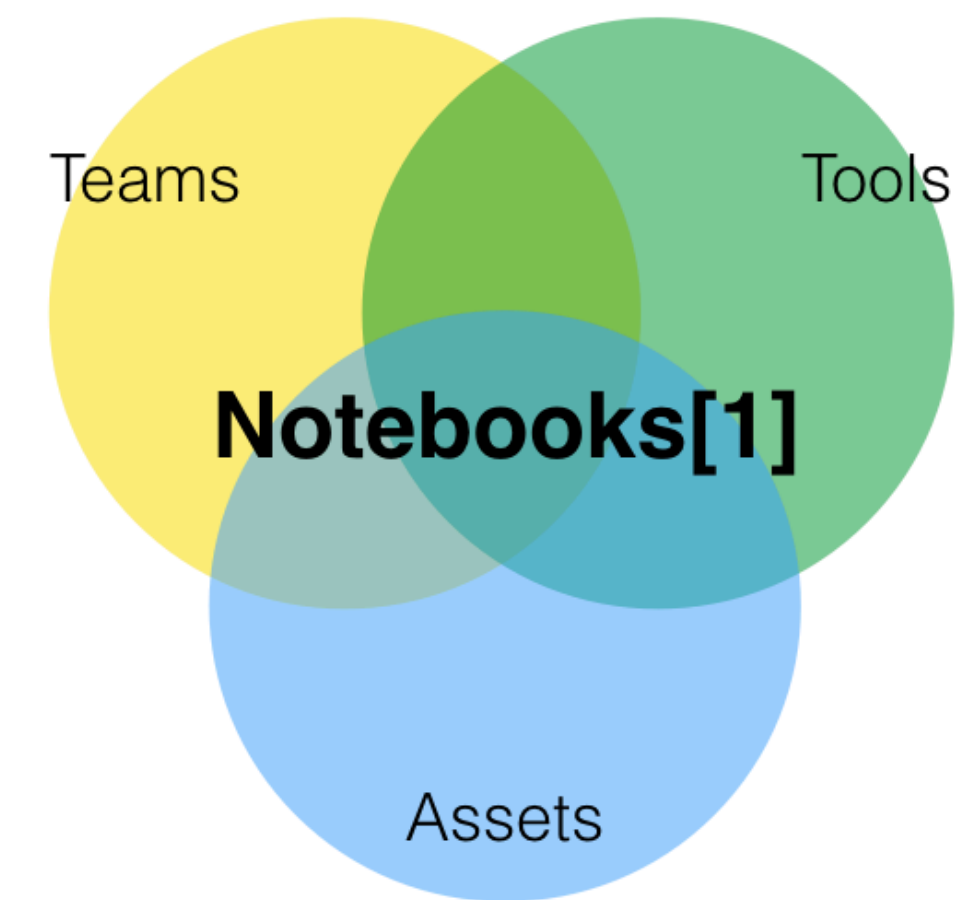
“SUCCESS!!”

CONCLUSION

- Solving the Data problems of tomorrow cannot be done by data scientists alone.
- Notebooks, considered by most to be the domain of data scientists, can help break down traditional silos and help team of all types who are working on data problems

Try it for yourself today:

- IBM Data Science Experience
<http://datascience.ibm.com/>
- Locally using PixieDust automated installer
<https://ibm-cds-labs.github.io/pixiedust/install.html>
- IBM Booth # 301





[1] Not just for data scientists



RESOURCES



- <https://github.com/ibm-cds-labs/pixiedust>
 - <https://ibm-cds-labs.github.io/pixiedust>
 - <https://medium.com/ibm-watson-data-lab/i-am-not-a-data-scientist-e7ca6ceba2>
 - <https://spark.apache.org>
 - <https://www.ibm.com/us-en/marketplace/spark-as-a-service>
 - <http://datascience.ibm.com>
 - <https://www.ibm.com/watson/developercloud/tone-analyzer.html>
 - <https://medium.com/ibm-watson-data-lab/real-time-sentiment-analysis-of-twitter-hashtags-with-spark-7ee6ca5c1585>
 - <https://gist.github.com/vabarbosa/76d08b1cc6f80d5fc80856a1f3f32014>
 - <https://gist.github.com/vabarbosa/dca176c3a68f0c101cbe475571e56bf7>
 - <https://ibm.biz/pixiedustvis>
 - <https://ibm.biz/pixiedustlab>
- 
- 

ONE LAST MESSAGE

“THANK YOU”

— NATASHA



— BEN

