



Interactive Visual Data Exploration with Spark in Databricks Cloud

Hossein Falaki
@mhfalaki

About Databricks

Founded by creators of Apache Spark

Offers Spark as a service in the cloud

Dedicated to open source Spark

- › Largest organization contributing to Apache Spark
- › Drive the roadmap

Databricks Cloud

Databricks Workspace



Databricks Platform

- › Notebooks
- › Dashboards
- › Job launcher

- › Latest version
- › Configured / Optimized

- › Start clusters in seconds
- › Dynamically scale up & down

Spark

Fast & General distributed computing engine:
batch, streaming, iterative

Capable of handling petabytes of data

Even faster by caching data in-memory

Versatile programming interfaces



Spark: Mixing SQL with Python/Scala

```
// Query an existing table and get results back as Schema RDD
rdd = hiveContext.sql("select article, text from wikipedia")

// Perform transformations
words = rdd.flatMap(lambda r: r.text.split())

// Collect sample of data in driver machine
sampled_words = words.sample(fraction = 0.001)
```

Databricks Platform

Start clusters in seconds

Zero-cost management

Dynamically scale up and down

Clusters						
Name	Memory	State	Nodes	Notebooks	Dashboard Cluster	Alter
Sales Analysis	102 GB	Running	Spark Master Worker 0 Worker 1 Worker 2	Sales Analysis Tom's Analysis	⊕ Make Dashboard Cluster	⊖ Restart ⊖ Remove
Default Cluster	254 GB	Running	Spark Master Worker 0 Worker 1 Worker 2 Worker 3 Worker 4 Worker 5 Worker 6 Worker 7 Worker 8	Sales	Current Dashboard Cluster	⊖ Restart ⊖ Remove
Marketing Pipeline	127 GB	Running	Spark Master Worker 0 Worker 1 Worker 2 Worker 3	MarketingData	⊕ Make Dashboard Cluster	⊖ Restart ⊖ Remove
⊕ Add ...						

Databricks Workspace

Notebooks

- › SQL
- › Python
- › Scala

Dashboards

Job Launcher

Notebooks

Exploring Diamonds

Command took 0.012s.



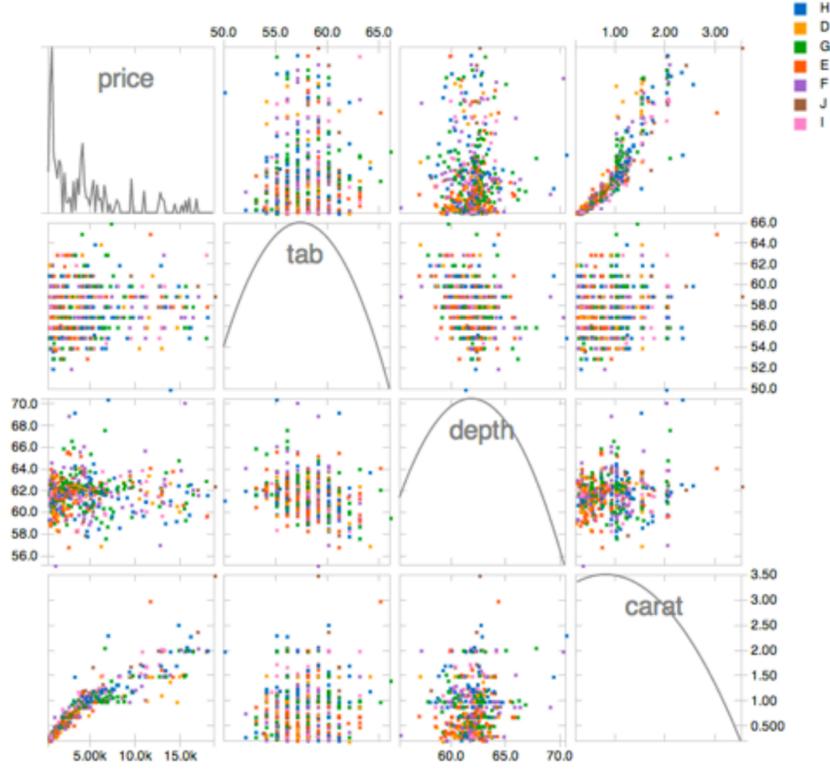
Command took 0.002s.

Scatter plot matrix

We can use the scatter plot matrix to quickly explore correlation between all features

Command took 0.003s.

```
» select * from diamonds where rand() < 0.01
```



Type comment here!

admin @ 6/16 7:16 PM
As expected - the primary finding is that there is significant correlation between price and carat, and a lack of correlation between depth and price
[Show Less](#)

admin @ 6/16 7:16 PM
Additionally, tab values seem to be quantized

Plot Options...

Command took 4.207s.

Supports Python, Scala, SQL

Interactive commands and plots

On-line collaboration

Dashboards

Diamonds Dashboard

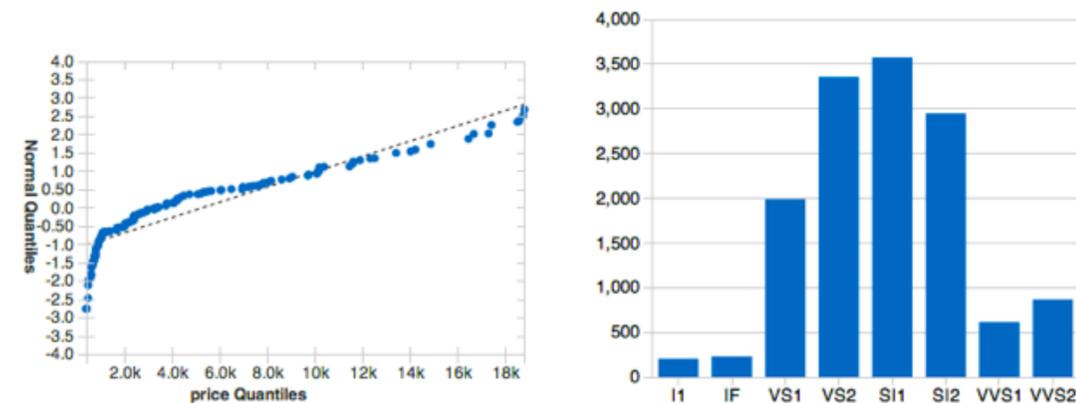
Provide a quick overview of some pertinent elements of the data that we've seen so far.

Note - Data refreshed each night at midnight with the latest sales data.

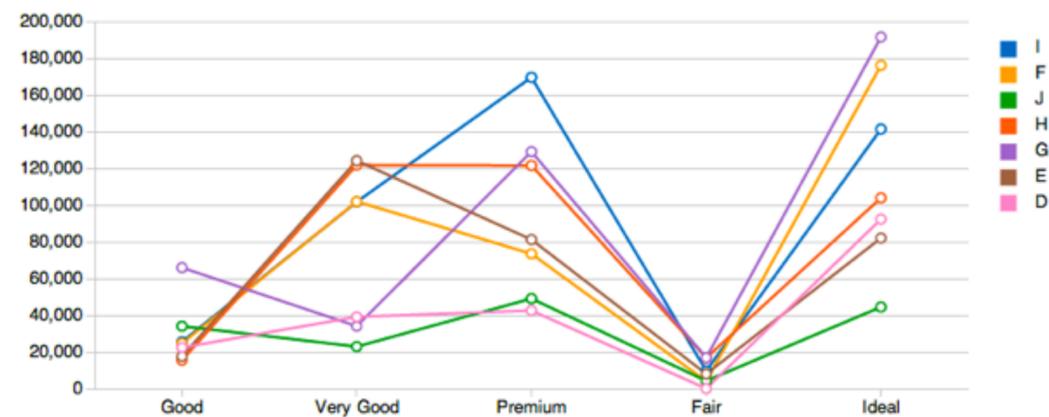


Showing data results filtered by cut

cut:



Effect of cut on price controlled for color



WYSIWYG Builder

Interactive jobs

On-click publishing

Exporting from notebooks

Job Launcher

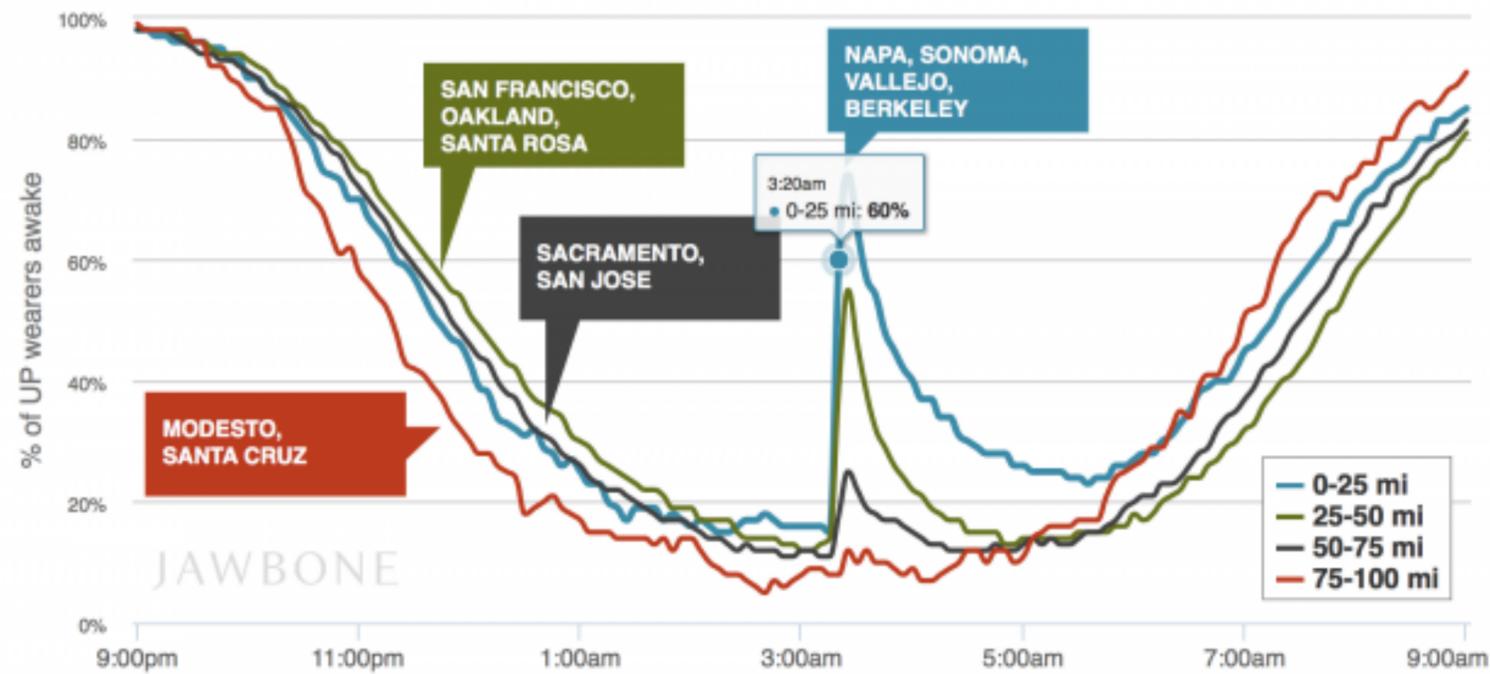
Elastic Jobs [New Job](#)

Job Name ▼	Recent Runs	Active
Market Basket Analysis	05/26/2014 06/02/2014 06/09/2014 ...	Today at 8:57 PM Cluster: Default Cluster Minimize Remove Actions + databricks/analysis/transform.jar : Triggers + Every Week: Monday
Sales Dashboard ETL	Today at 5:00 PM Today at 6:00 PM Today at 7:00 PM Today at 8:00 PM Today at 9:00 PM Today at 10:00 PM Today at 11:00 PM ...	Tomorrow at 12:00 AM Cluster: Default Cluster Edit Remove
Fraud Model Training	06/09/2014 Last Tuesday at 1:00 AM Last Wednesday at 1:00 AM Last Thursday at 1:00 AM Last Friday at 1:00 AM Last Saturday at 1:00 AM Yesterday at 1:00 AM ...	Actions + databricks/ml/training.jar : Triggers + Daily: 1am

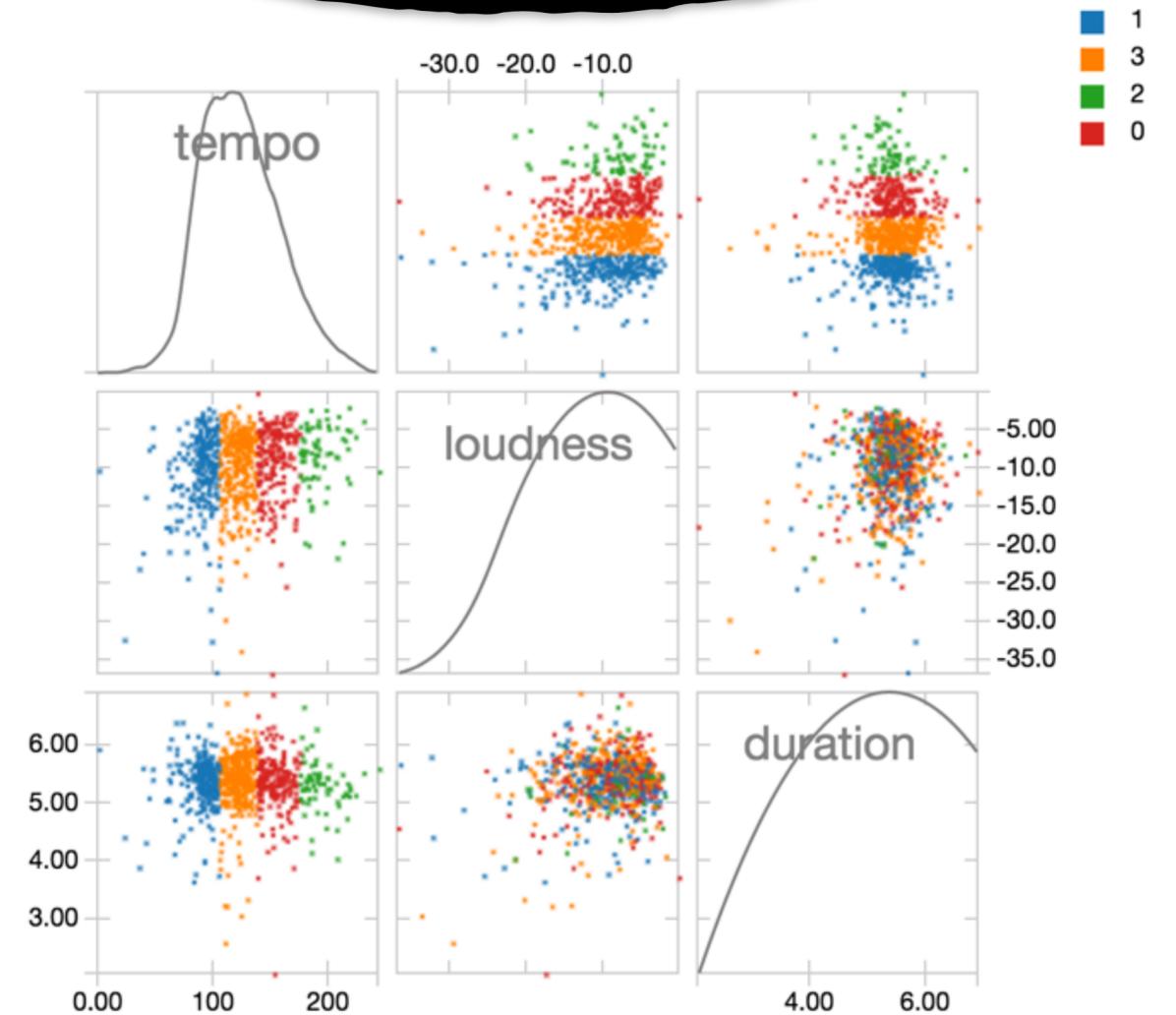
Runs arbitrary Spark jobs programmatically

Expository vs.

Exploratory



We wish all the people in the Bay Area who were affected by the earthquake a speedy recovery and a good night's sleep.



Large data

“Visualization is critical to data analysis.”

William S. Cleveland

But we often skip exploratory visualization with large data

Challenges

1. Interactivity

with large data is challenging

2. Visual medium

cannot accommodate as many pixels as data points

Solutions

1. Interactivity

In-memory computation

High parallelism

Reducing interaction latency with Spark

1. In-memory computation

- › Significantly reduces latency

2. High parallelism

- › Get more executors with Mesos or Yarn: a challenge in itself
- › Click a button to increase cluster size in Databricks Cloud

Versatile programming interface

Data visualization is very much like programming.

- › Point and click doesn't really cut it
- › Requires an API (grammar): ggplot, matplotlib, bokeh, etc.

Spark has SQL, Scala, Python, Java and (experimental) R API

Libraries for distributed statistics and machine learning

Solutions

1. Interactivity

In-memory computation

High parallelism

2. Visual medium

In-browser collaborative notebooks

Summarizing, Sampling and Modeling

More data points than pixels

Can we visualize 200GB of multidimensional data?

Short answer: no

Long answer:

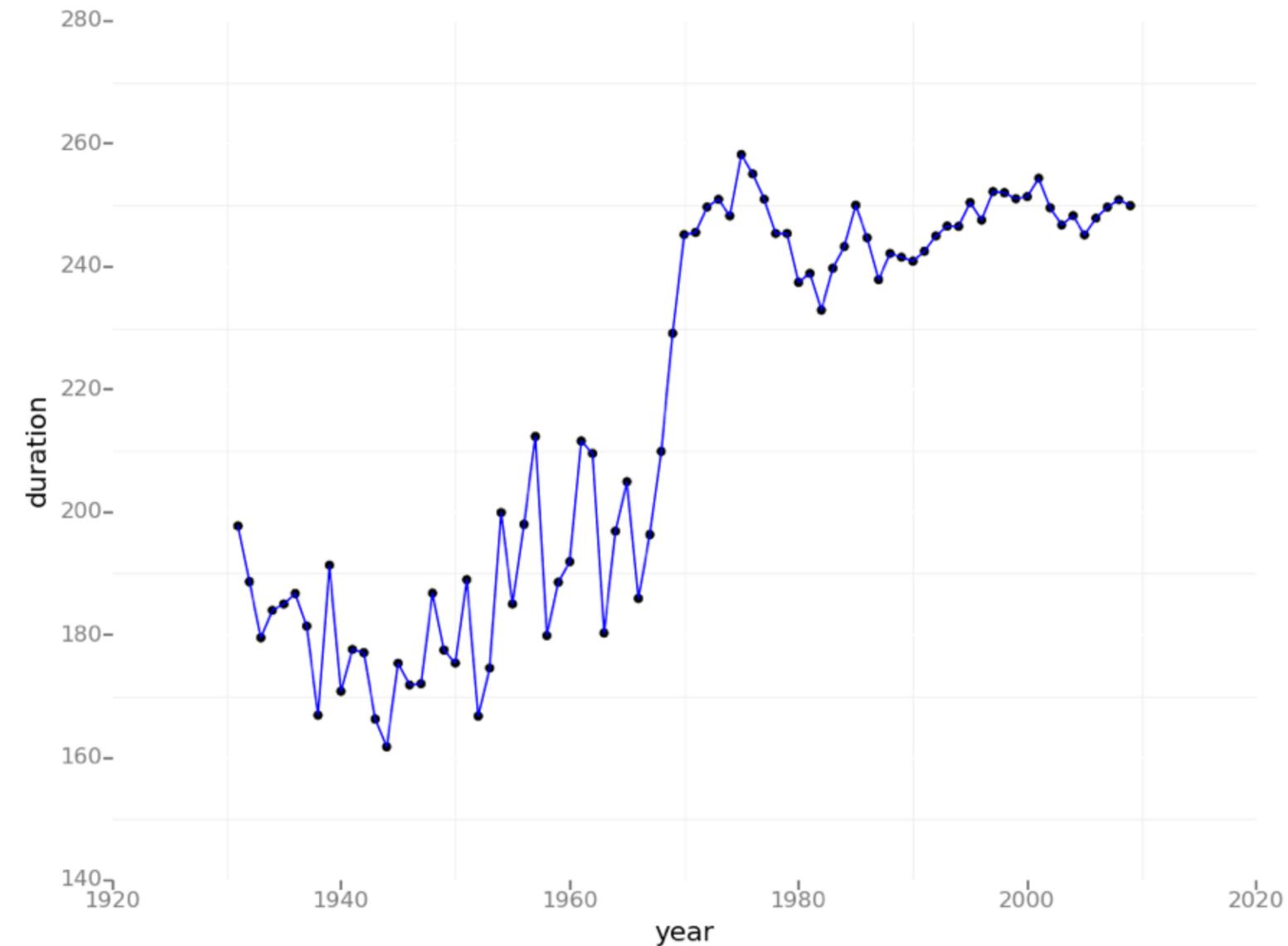
- › Summarize & visualize
- › Sample & visualize
- › Model & visualize

Summarize and visualize

Extensively used by BI tools

- › Aggregation
- › Pivoting

Most data scientists' nightly jobs summarize data



Sample and visualize

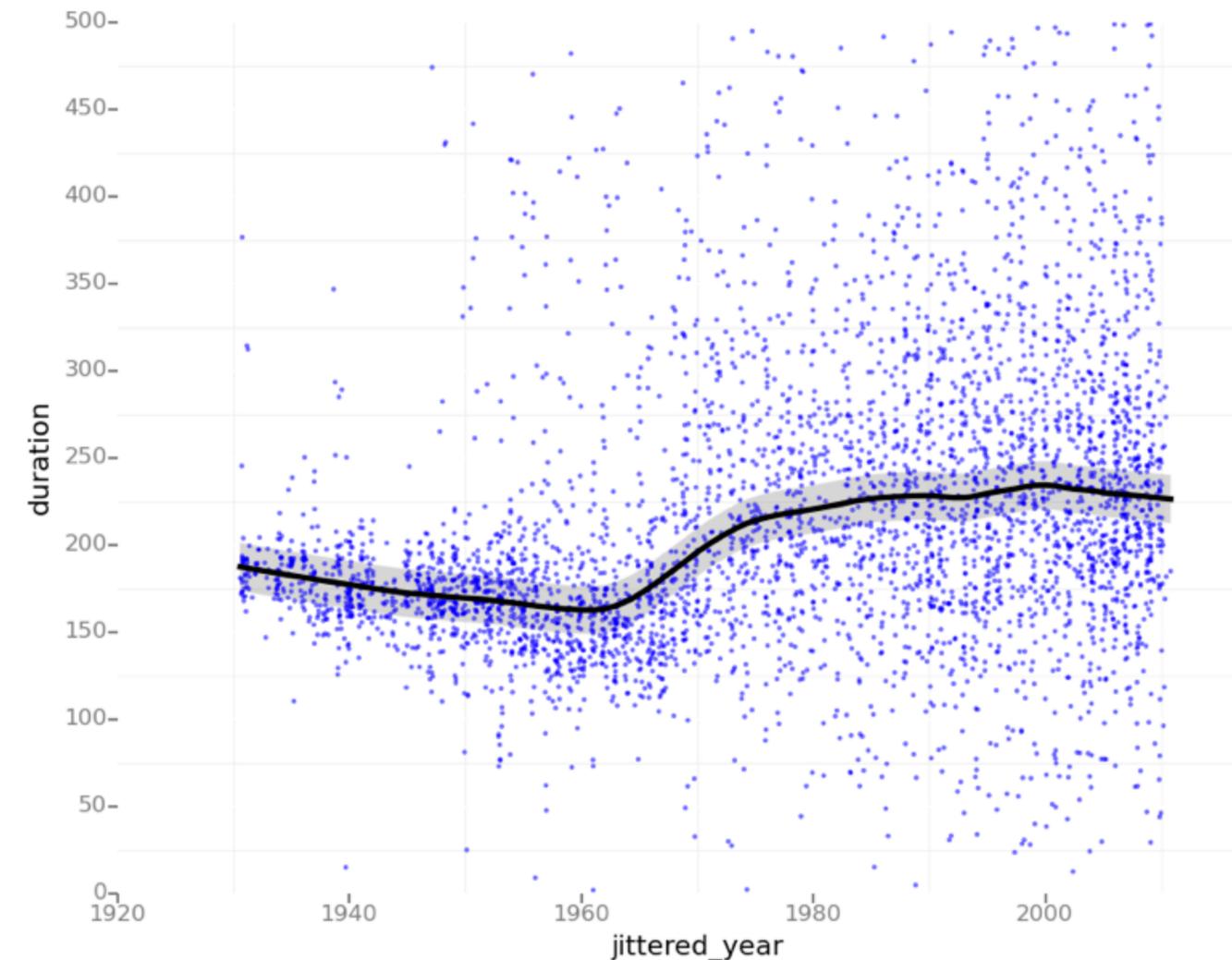
Sometimes we need to visualize (feel) individual data points

Sampling is extensively used in statistics

Spark offers native support for:

- › Approximate and exact sampling
- › Approximate and exact stratified sampling

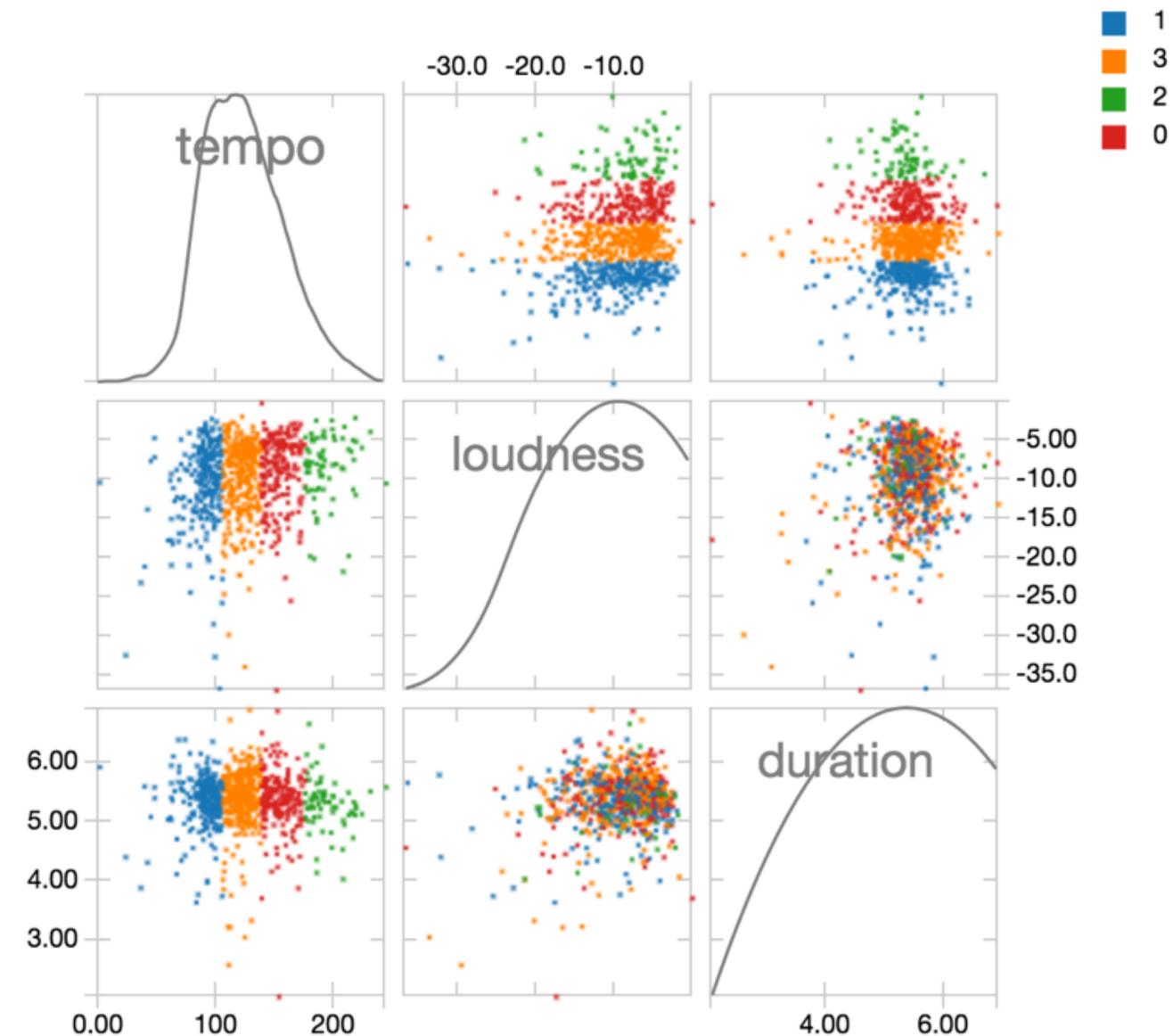
Approximate sampling is faster and is good enough in most cases



Model and visualize

MMLib supports a large (and growing) set of distributed algorithms

- › Clustering: k-means
- › Classification and regression: LM, DT, NB
- › Dimensionality reduction: SVD, PCA
- › Collaborative filtering: ALS
- › Correlation, hypothesis testing



Demo

Summary

With new big data tools we can resume interactive visual exploration of data

Using Spark we can manipulate large data in seconds

- › Cache data in memory
- › Increase parallelism

To visualize millions of data points we can

- › Summarize
- › Sample
- › Models

Databricks Cloud

databricks.com

Apache Spark

spark.apache.org

Matplotlib

matplotlib.org

Python ggplot

ggplot.yhathq.com

D3

d3js.org

