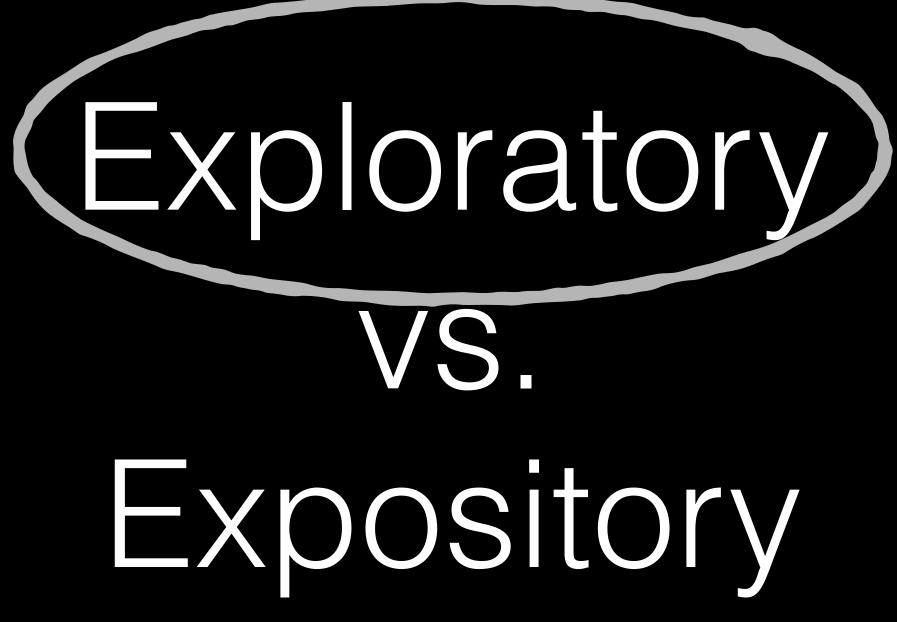
# Interactive Visual Data Exploration with Spark



Hossein Falaki @mhfalaki Databricks Inc.

### otohriolz UdldldKS



## Large data

### "Visualization is critical to data analysis." William S. Cleveland

### But we often skip it for large data

### 1. Interactivity

### interactivity with large data is challenging

### 2. Visual medium

More data points than we have pixels

## Challenges

### Use Apache Spark

### Summarize, Sample, Model

5



## Apache Spark

- over distributed data
- Capabale of handling peta bytes of data
- Enables caching distribted data
- Versatile programming interfaces

### General computing engine for batch, streaming, and iterative jobs



# Versatile programming interface

- SQL, Scala, Python, Java and (experimental) R API
- Libraries for distributed statistics and learning
- Exploratory data visualization is very much like programming
  - Point and click doesn't really cut it
  - Requires an API (grammar): ggplot, matplotlib, bokeh, etc. 0

## Interactively & Iterative jobs

- Cache data in memory to reduce latency
- Control data partitioning and parallelization to reduce latency
- Powerful API for data manipulation
  - Mix SQL with other languages
  - Create Hive tables from data in any language

## More data points than pixels

### Short answer: no

- Long answer: Summarize & visuzliae

Can we visualize 200GB of multidimensional data?

 Sample & visualize Model & visualize

## Summarize & visualize

## Sample & visualize

11

## Model & visualize

## Iterative analysis