

Vegas

The Missing Matplotlib for
Scala/Spark

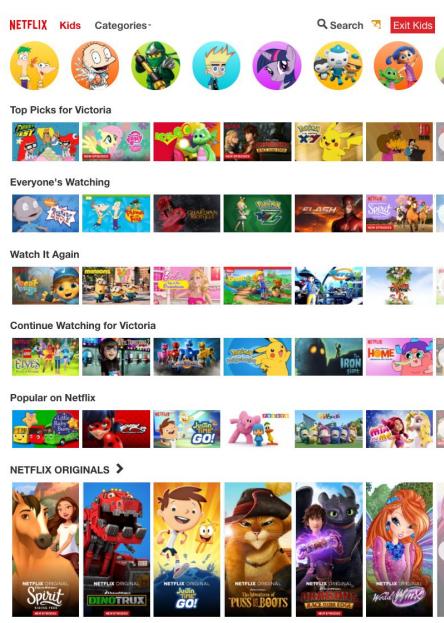
| DB Tsai
Roger Menezes

NETFLIX

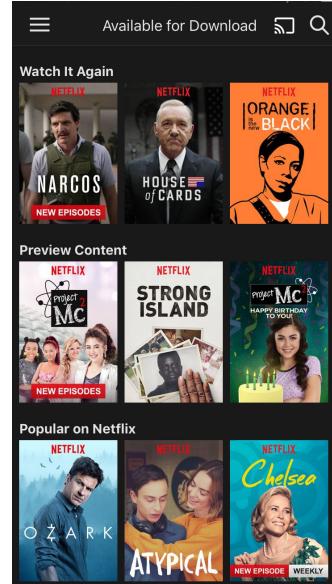
Netflix Recommendations



Homepage



Kids Page



Downloads Page

Every aspect
of the
Experience is
Machine
Learned



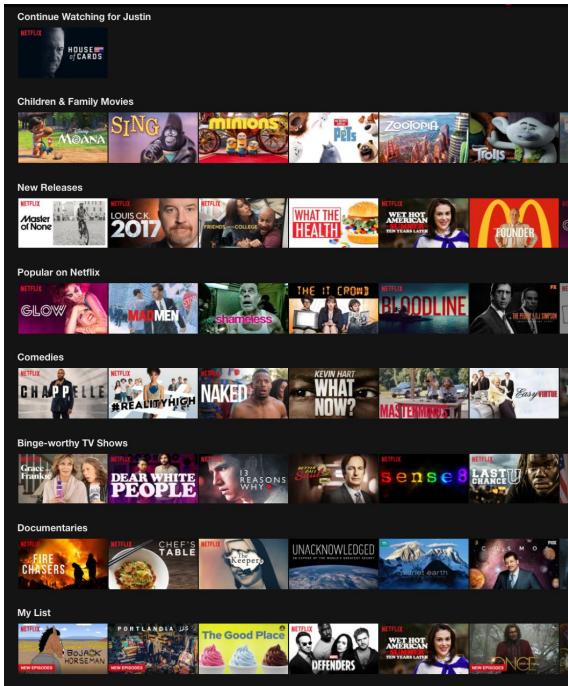
#netflixeverywhere

2017

- > 100M members
- > 190 countries



Multiple Devices



NETFLIX

NETFLIX ORIGINAL

**MARVEL
THE DEFENDERS**

Watch Now

They're not friends. But these four New Yorkers will fight as one to save the city they love.

PLAY

Continue Watching for Justin



My List



Popular on Netflix



Trending Now



Watch It Again



Search



NETFLIX ORIGINALS



Top Picks for Justin



Because you watched Maria Bamford: Old Baby ➔



Sims: 10 rows/page average

Children & Family Movies



Genres: 23 rows/page average

New Releases



Recently Added



NETFLIX

NETFLIX ORIGINAL

MARVEL

Search

**Billboard:**

Watch Now

They're not friends. But these four New Yorkers will fight as one to save the city they love.

▶ PLAY**✓ MY LIST**

Continue Watching for Justin

Continue Watching:

My List

My List:

PORTLANDIA

NEW EPISODES

The Good Place

NEW EPISODES

DEFENDERS

NEW EPISODES

WET HOT AMERICAN

TEN YEARS LATER

ONCE UPON A TIME

NEW EPISODES

Popular on Netflix

Popular on Netflix:

shameless

THE IT CROWD

BLOODLINE

THE PRACTICE

Trending Now

Trending Now:

TOO REAL

THE RANCH

NETFLIX

FLASH

STANDUPS

NETFLIX

Watch It Again

Watch It Again:

ORANGE IS BLACK

PORTLANDIA

THE MONEY PIT

TOUCHED BY ANGEL

WHITE GOLD

TOAST

NETFLIX ORIGINALS

**Originals Row**

NETFLIX ORIGINAL

NARCOS

NEW EPISODES

NETFLIX ORIGINAL

OZARK

NEW EPISODES

NETFLIX ORIGINAL

ATYPICAL

NEW EPISODES

NETFLIX ORIGINAL

ARRESTED DEVELOPMENT

NEW EPISODES

NETFLIX ORIGINAL

DEFENDERS

NEW EPISODES

NETFLIX ORIGINAL

Chelsea

Top Picks for Justin

Top Picks:

Because you watched Maria Bamford: Old Baby ➤

Because You Watched:

Children & Family Movies

Genres:

New Releases

New Releases:

Recently Added

Recently Added:

ML at Netflix

- Optimize the Experimentation usecase vs Productionization
- Experimentation
 - Opportunity sizing, Data Exploration
 - Tweaks to ML algos
 - Feature Selection
 - Model Evaluation

Notebooks

- Optimal for Experimentation
- Sharing reproducible research
 - Facilitates feedback loop with PMs
- End to end ML experiment.
 - Interactivity drives productivity

Python Notebooks

The image displays three distinct environments for Python notebooks:

- Top Left:** A screenshot of a web browser window titled "r_notebook_example" at "localhost:8889/notebooks/...". It shows a Jupyter notebook interface with a code cell (In [5]) and a scatter plot.
- Top Middle:** A screenshot of a desktop application window titled "IPy: Notebook spectrogram". It shows a Jupyter notebook interface with a code cell (In [1]) and a spectrogram plot.
- Bottom Right:** A screenshot of the "Jupyter nbviewer" website. It shows a Jupyter notebook interface with a code cell (In [1]) and a map of the United States with colored states representing statistical significance levels (HH, LH, LL, HL, Non-significant) for different P-values (0.100, 0.010, 0.001).

Code Examples:

```
In [5]: library(plotly)
set.seed(100)
d <- diamonds[sample(nrow(diamonds), 1000), ]
plot_ly(d, type = 'scatter', mode = 'markers',
        x = ~carat, y = ~price,
        color = ~carat, size = ~carat,
        text = ~paste("Clarity: ", clarity))
```

```
In [1]: from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

```
In [2]: fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x);
ax1.set_title('Raw audio signal')
ax2.specgram(x);
ax2.set_title('Spectrogram');
```

```
ax.set_extent(extent, crs=ccrs.PlateCarree())
ax.add_collection(polys)
ax.outline_patch.set_visible(False)

boxes, labels = maps.lisa_legend_components(lisa, p_thres=p_thres)
plt.legend(boxes, labels, loc='lower left', frameon=False)
ax.set_title('P-value = %.3f' % p_thres)

plt.show()
```

Python Notebooks

- Seamless Experience - ML experimentation
- Well known Scientific computing libraries
- Huge catalog of Visualization plotting libraries
 - Matplotlib, Seaborn, Bokeh, BQPlot, Lightning, etc.

Scala Notebooks

- Zeppelin, Jupyter, Databricks, Spark-Notebooks, ...
- Computing library gap filling up
- Lack of Visualization Libraries
 - Main friction point in adoption
 - End to End ML use case not convincing

Introducing Vegas

- Visualization Library in Scala
- Mainly built for the notebook use case
- Scala wrapper around Vega-Lite
- Missing Matplotlib for the Scala and Spark world.

VegaLite

- Statistical Visualization
- Design considerations for vega-lite
 - Imperative vs Declarative API

NETFLIX

DECLARATIVE

STATISTICAL

VISUALIZATION
GRAMMAR

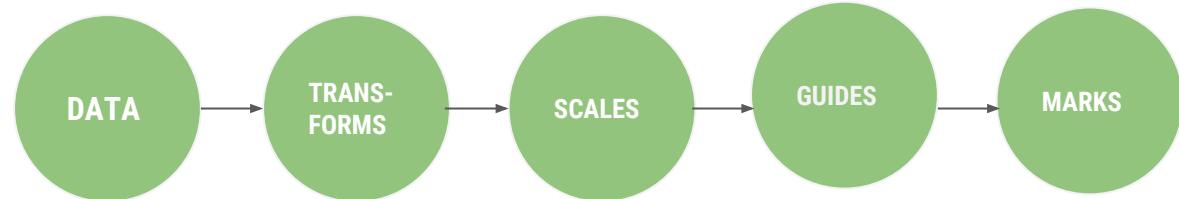
IN SCALA

VEGAS

You tell it **WHAT** should be done with the data, and it knows **HOW** to do it!

Operations such as *filtering, aggregation, faceting* are built into the visualization, rather than putting the burden on the user to massage the data into shape.

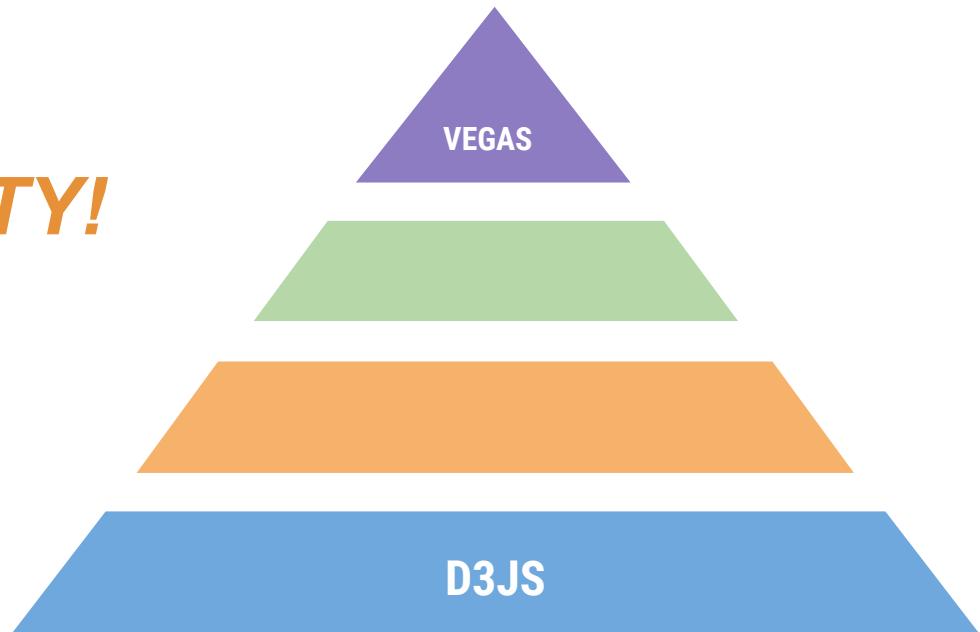
Complex visualizations can be built with a few high level abstractions:



cf : Altair Talk by Brian Granger in PyData 2016 <https://youtu.be/v5mrwq7yJc4>

Added Bonus of Declarative Visualizations:

INTERACTIVITY!



VEGAS CODE EXPANDS OUT TO D3JS CODE!

Anatomy of a plot: Channels

SHAPE CHANNEL

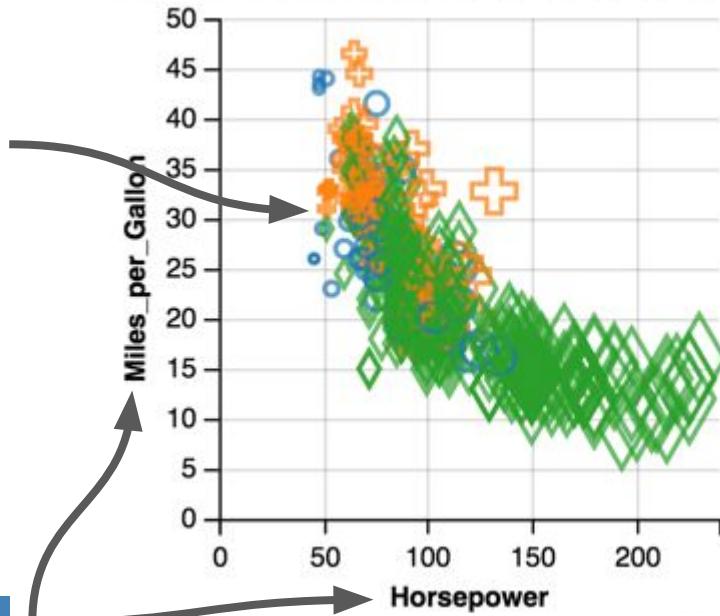
COLOR CHANNEL

X/Y CHANNEL

Origin
Europe
Japan
USA

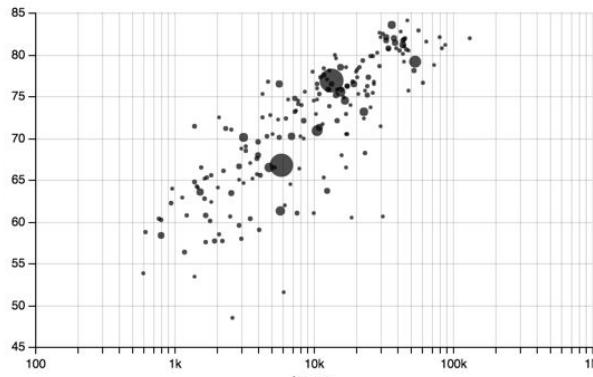
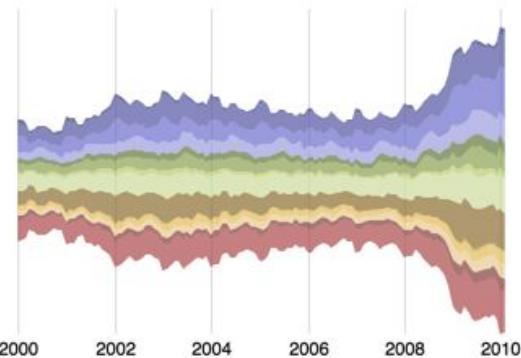
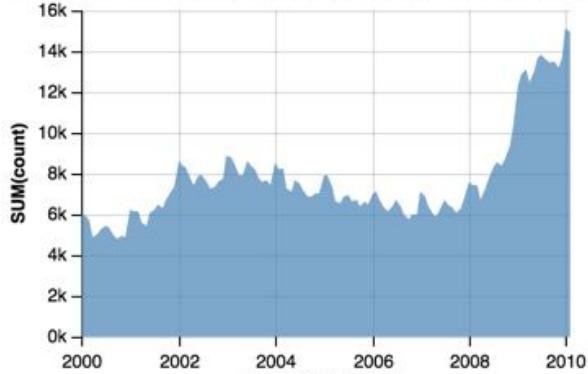
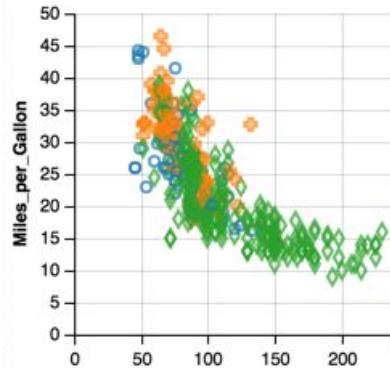
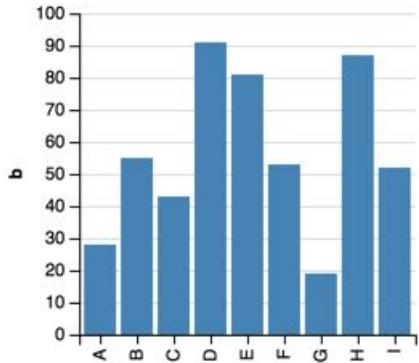
Horsepower
50
100
150
200

Origin
Europe
Japan
USA

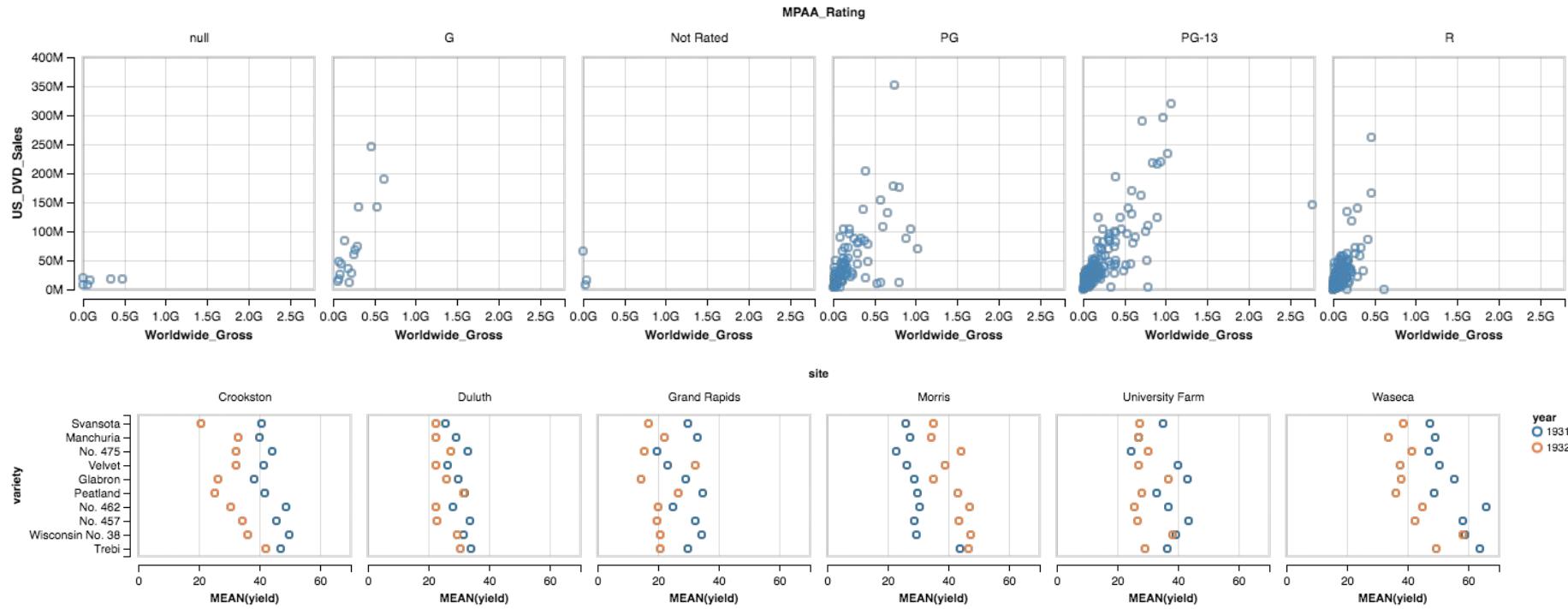


Features...

1. *Supports most plot types*



2. Trellis plots

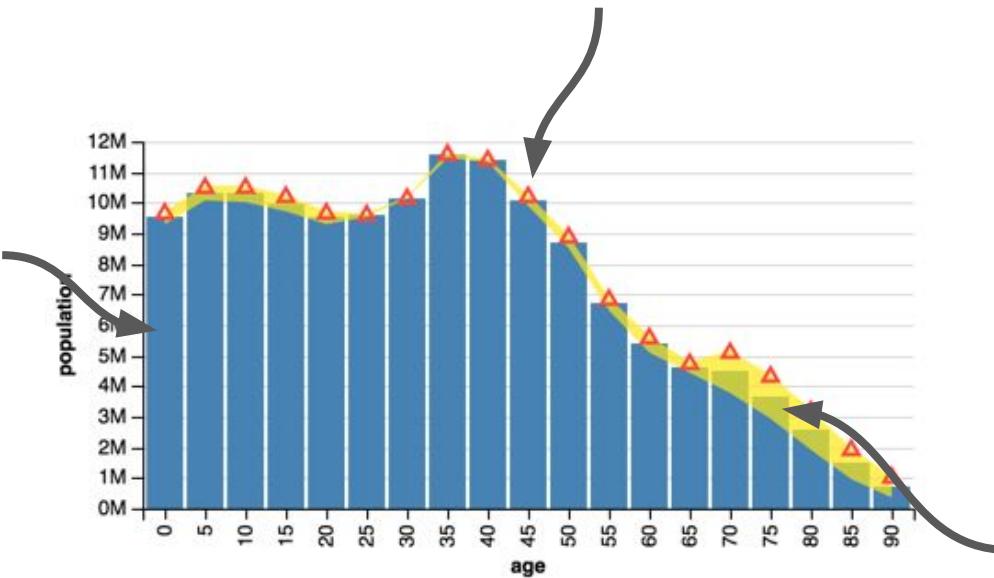


3. Layers

LAYER 1.

LAYER 2.

LAYER 3.



4. Notebook and Consoles

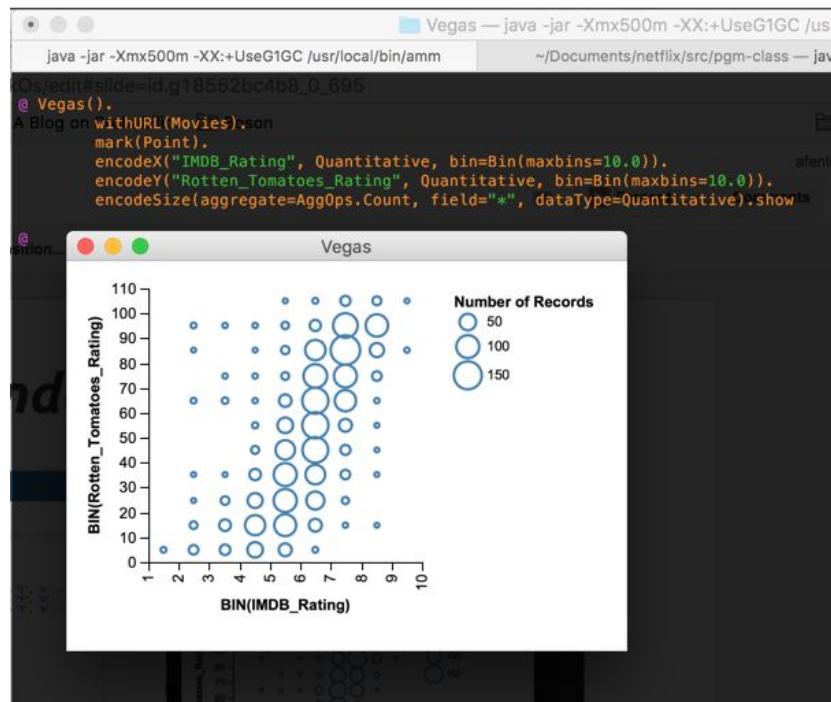
Zepplin Notebook ▾

Vegas Examples

A simple bar chart with embedded data.

```
Vegas("A simple bar chart with embedded data.").  
withData(SeqC  
  Map("a" -> "A", "b" -> 28), Map("a" -> "B", "b" -> 55), Map("a" -> "C", "b" -> 43),  
  Map("a" -> "D", "b" -> 91), Map("a" -> "E", "b" -> 81), Map("a" -> "F", "b" -> 53),  
  Map("a" -> "G", "b" -> 19), Map("a" -> "H", "b" -> 87), Map("a" -> "I", "b" -> 52)  
)).  
encodeX("a", Ordinal).  
encodeY("b", Quantitative).  
mark(Bar).  
show
```

Category	Value
A	28
B	55
C	43
D	91
E	81
F	53
G	19
H	87
I	52



5. *Built-in spark support*

Vegas

```
.withDataFrame(myDataFrame)  
.encodeX("population")  
.encodeY("age")
```

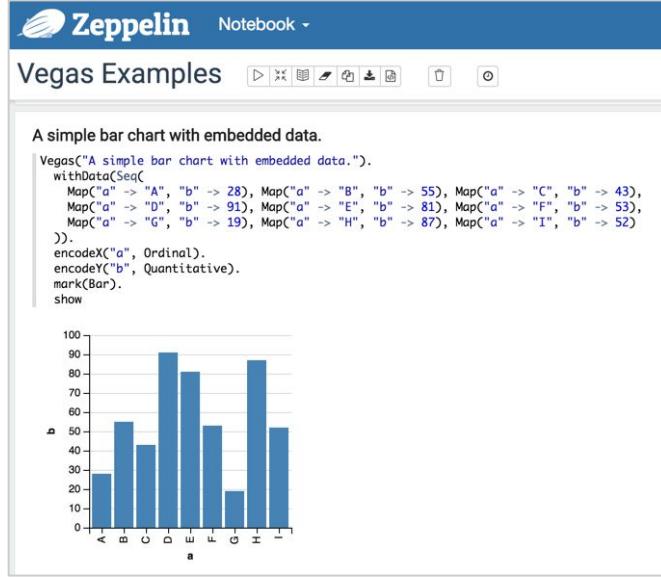
PASS IN DF.

MAPPED COLUMNS

6. *Visual statistics*

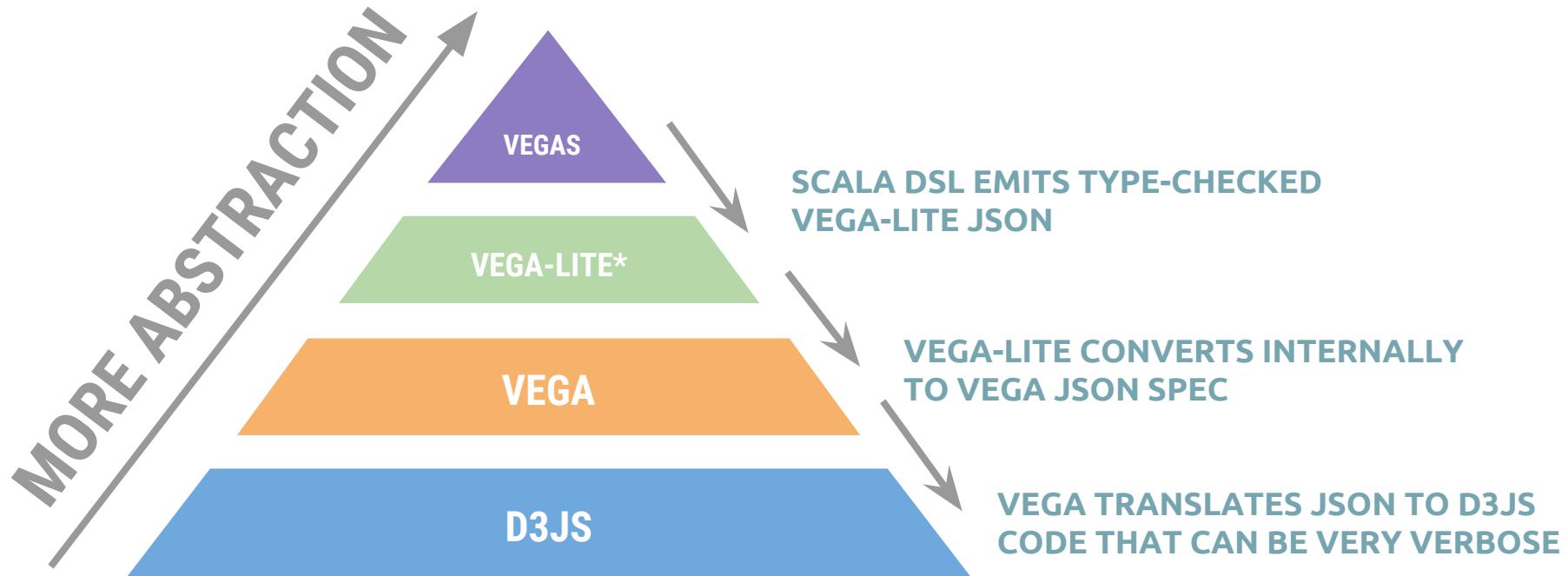
- Advanced Binning
- Sorting
- Scaling
- Custom Transforms
- Time Series
- Aggregation
- Filtering
- Math functions (log, etc)
- Missing data support
- Descriptive Statistics

How It Works !



1. Specify in Scala
2. Embed HTML
(iFrame)
3. Render within
iFrame using JS

A SCALA DSL FOR VEGA-LITE



* Vega-Lite

Example 3

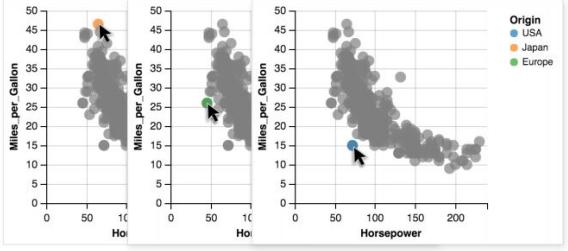
Other Channels + Transforms

What's coming

1. Interactive selections

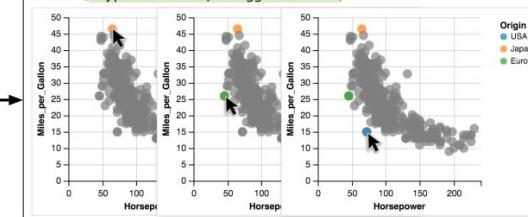
(a) Highlight a single point on click

```
{  
  "data": {"url": "data/cars.json"},  
  "mark": "circle",  
  "select": {  
    "id": {"type": "point"}  
  },  
  "encoding": {  
    "x": {"field": "Horsepower", "type": "Q"},  
    "y": {"field": "MPG", "type": "Q"},  
    "color": [  
      {"if": {"id": "id", "field": "Origin", "type": "N"},  
      {"value": "grey"}  
    ],  
    "size": {"value": 100}  
  }  
}
```



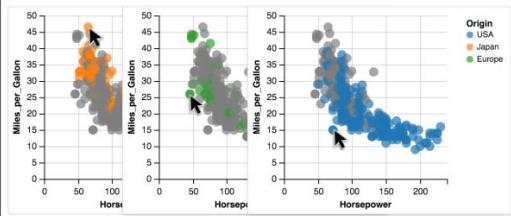
(b) Highlight a list of individual points

```
"id": {"type": "list", "toggle": true}
```



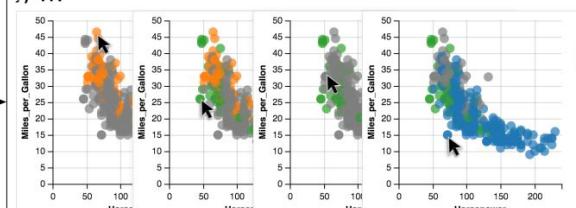
(d) Highlight a single Origin

```
"id": {"type": "point", "project": {"fields": ["Origin"]}}
```



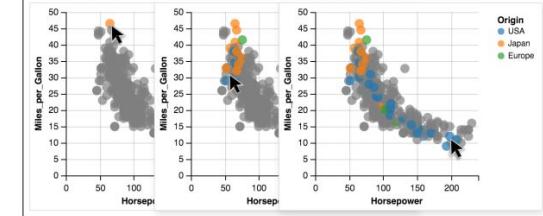
(e) Highlight a list of Origins

```
"select": {  
  "id": {"type": "list", "toggle": true, "project": {"fields": ["Origin"]}}  
}, ...
```



(c) "Paintbrush": highlight multiple points on hover

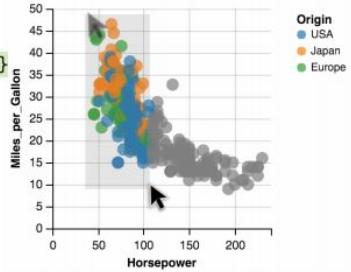
```
"id": {"type": "list", "on": "mouseover", "toggle": true}
```



2. Selections transforms

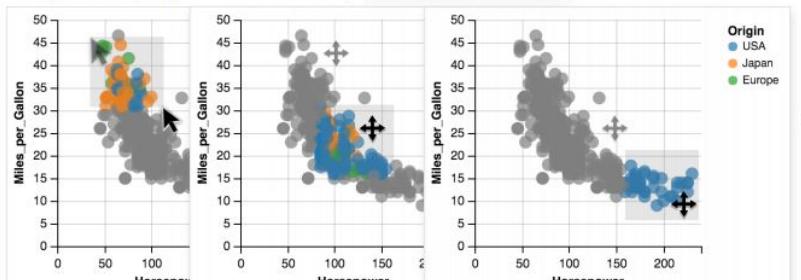
(a) Rectangular brush

```
"select": {  
  "region": {"type": "interval"}  
},  
...  
  "color": [  
    {"if": "region", ...}  
  ]  
...
```



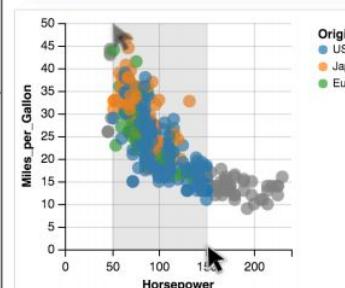
(b) Moving the brush

```
"region": {"type": "interval", "translate": true}
```

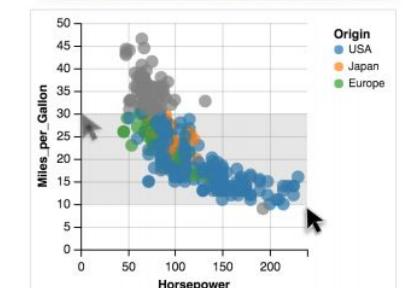


(c) Single-dimension brush

```
"region": {"type": "interval",  
  "project": {"channels": ["x"]}}
```



```
"region": {"type": "interval",  
  "project": {"channels": ["y"]}}
```



Contributors



NETFLIX

Thank you.

NETFLIX



The missing Matplotlib for Scala/Spark

<http://vegas-viz.org>

@NetflixResearch
@rogermenezes @dbtsai

NETFLIX