

# Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication



**Arjun Srinivasan**



**Steven M. Drucker**



**Alex Endert**



**John Stasko**



**What** are data facts?

**How** can we integrate data facts into visualization tools?

**Why** is this integration beneficial?

**What are data facts?**

How can we integrate data facts into visualization tools?

Why is this integration beneficial?

**BRACE YOURSELVES**





**AUTO-INSIGHT  
SYSTEMS ARE COMING**



**ai** AUTOMATED  
INSIGHTS

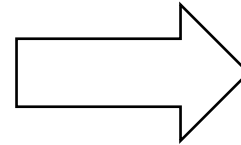
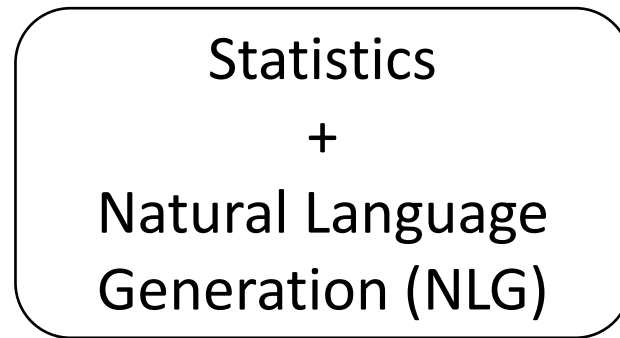
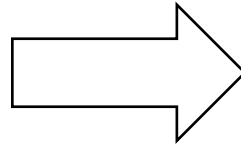
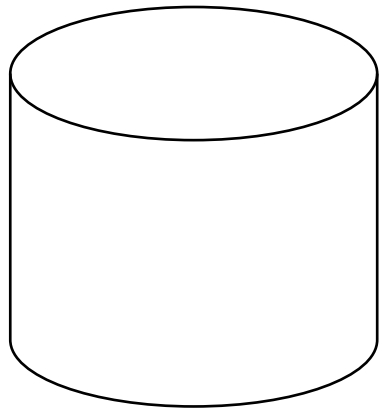
✓ **Insights are ready** ✕  
You have insights for Google Analytics internuntius.  
[View insights](#)

 Power BI

 [Explore](#) 



# Auto-insight Systems



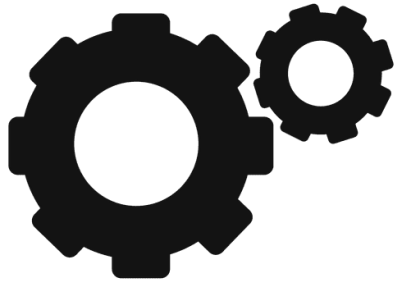
Sales for US is 3 times Europe

There is a rising trend of  
SUV's market share

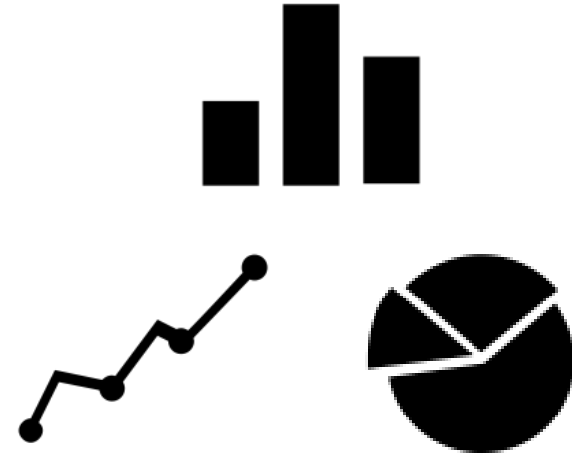
There is a falling trend of  
Sedan's market share

Acceleration and Horsepower  
exhibit a moderate correlation

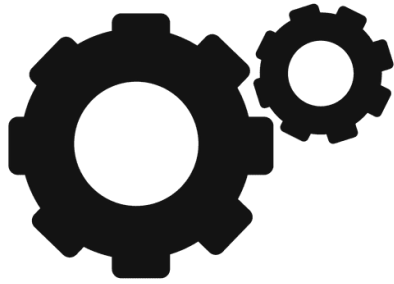
Displacement and Horsepower  
exhibit a strong correlation



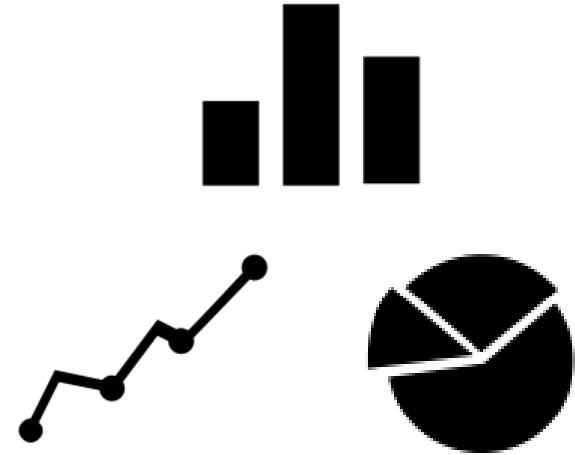
Auto-insight Systems



Visualization Systems

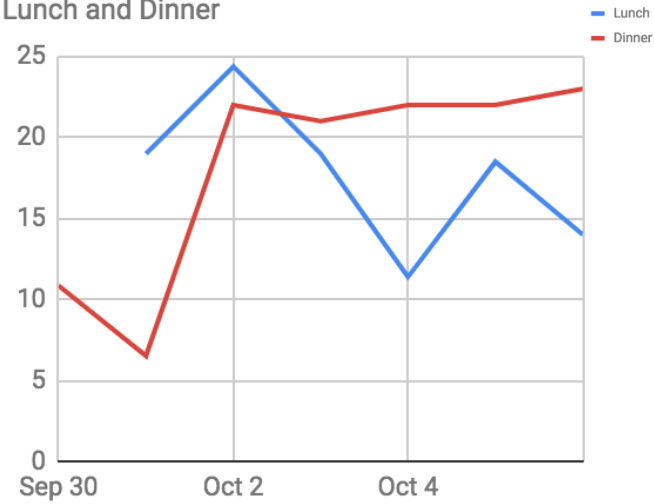


Auto-insight Systems



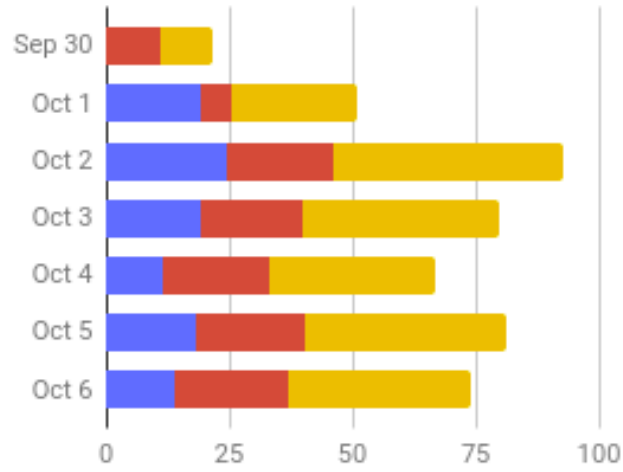
Visualization Systems

Lunch and Dinner



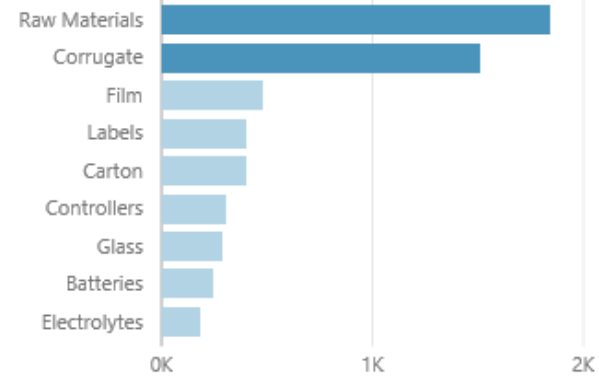
For every day, "Dinner" increases by about 2.41.

Lunch, Dinner and Food total



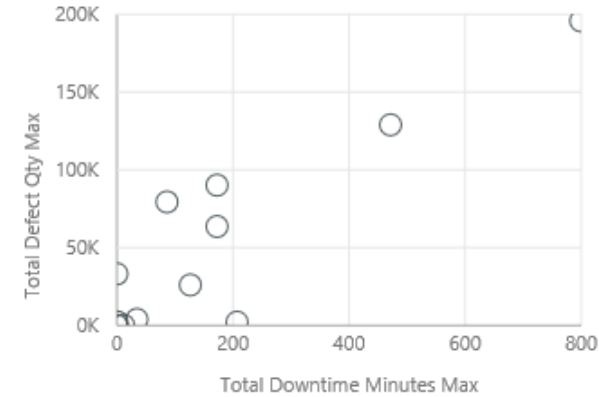
30 September has the lowest value for "Food total" (10.86) and the second-lowest value for "Dinner" (10.86).

Total Defect Reports  
BY MATERIAL TYPE



**CATEGORY OUTLIERS**  
'Raw Materials' and 'Corrugate' have noticeably greater 'Total Defect Reports'.

Total Downtime Minutes Max and Total Defect...  
BY DATE

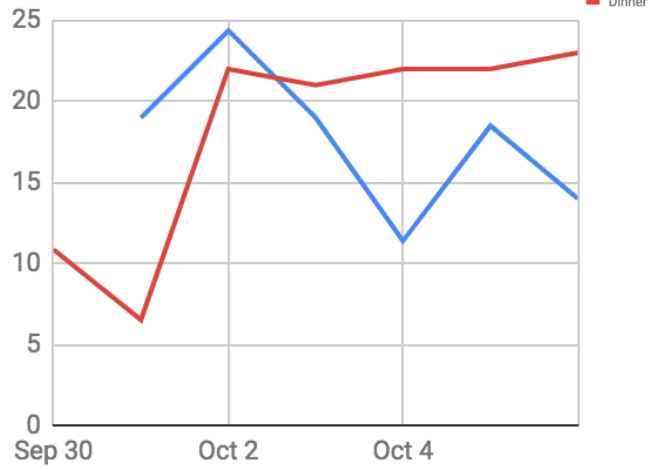


**CORRELATION**  
There is a correlation between 'Total Downtime Minutes Max' and 'Total Defect Qty Max' for '6'.



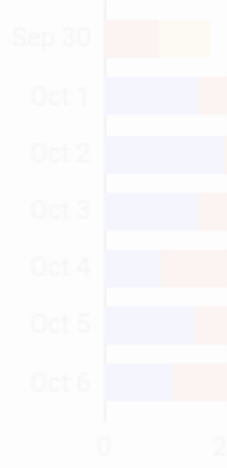


Lunch and Dinner



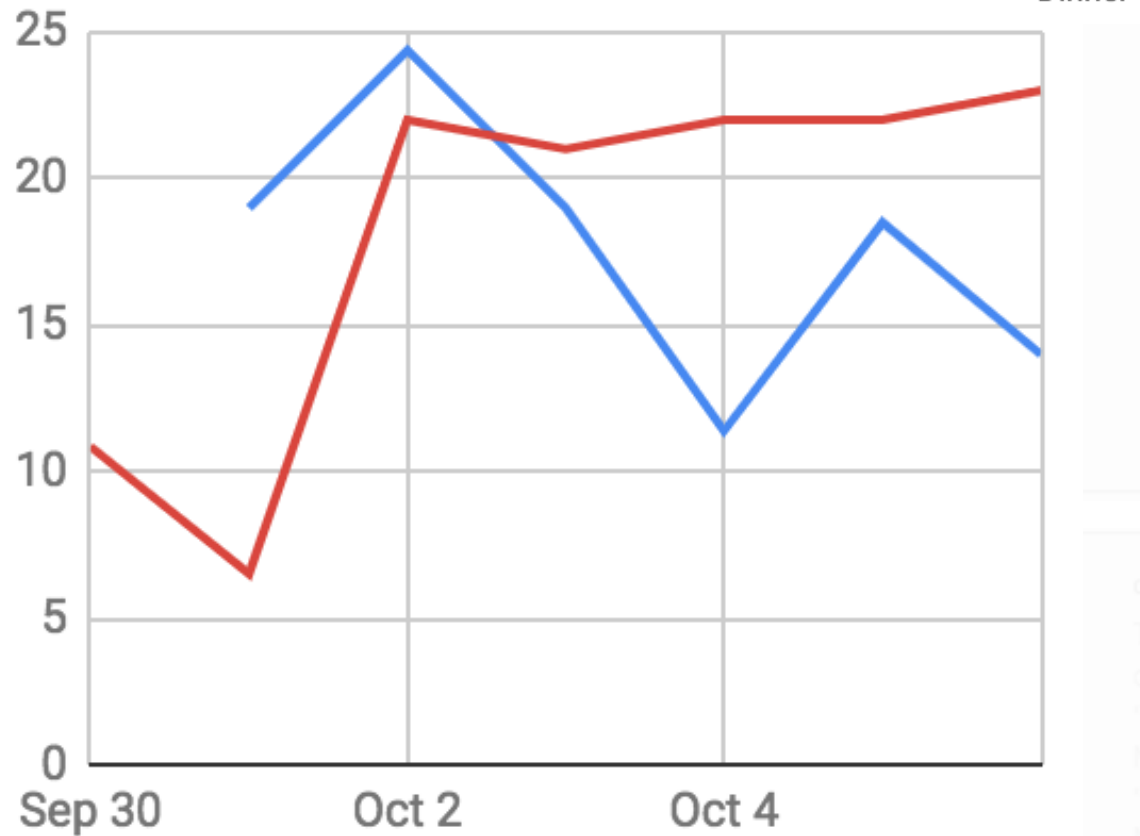
For every day, "Dinner" increases by about 2.41.

Lunch, Dinner



30 September total" (10.86) as for "Dinner" (10.86)

Lunch and Dinner



For every day, "Dinner" increases by about 2.41.



Google Sheets

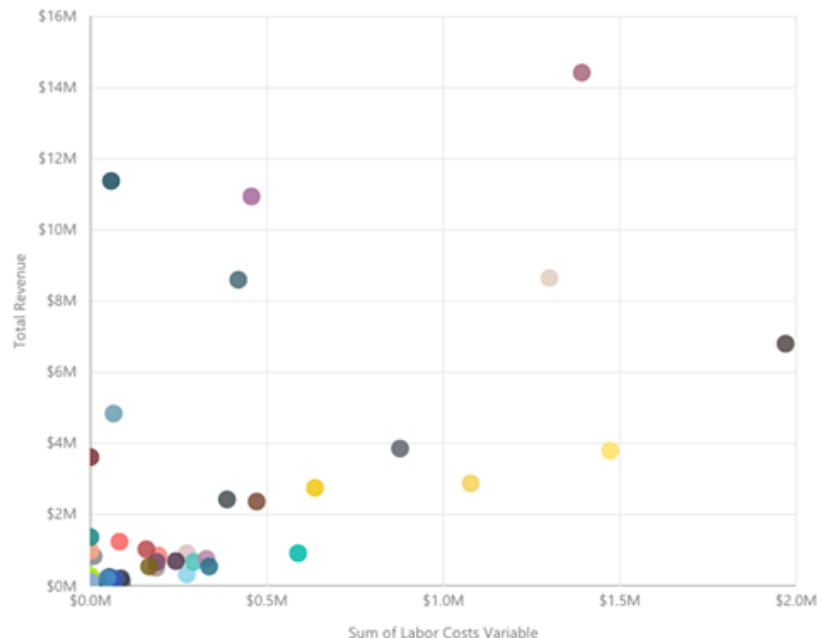
VIS '17 expenses

REGORY OUTLIERS  
 'Low Materials' and  
 'Irregular' have  
 noticeably greater  
 'Total Defect  
 Reports'.

CORRELATION  
 There is a  
 correlation between  
 'Total Downtime  
 Minutes Max' and  
 'Total Defect Qty  
 Max' for '6'.

Total Downtime Minutes Max

Sum of Labor Costs Variable and Total Revenue by City



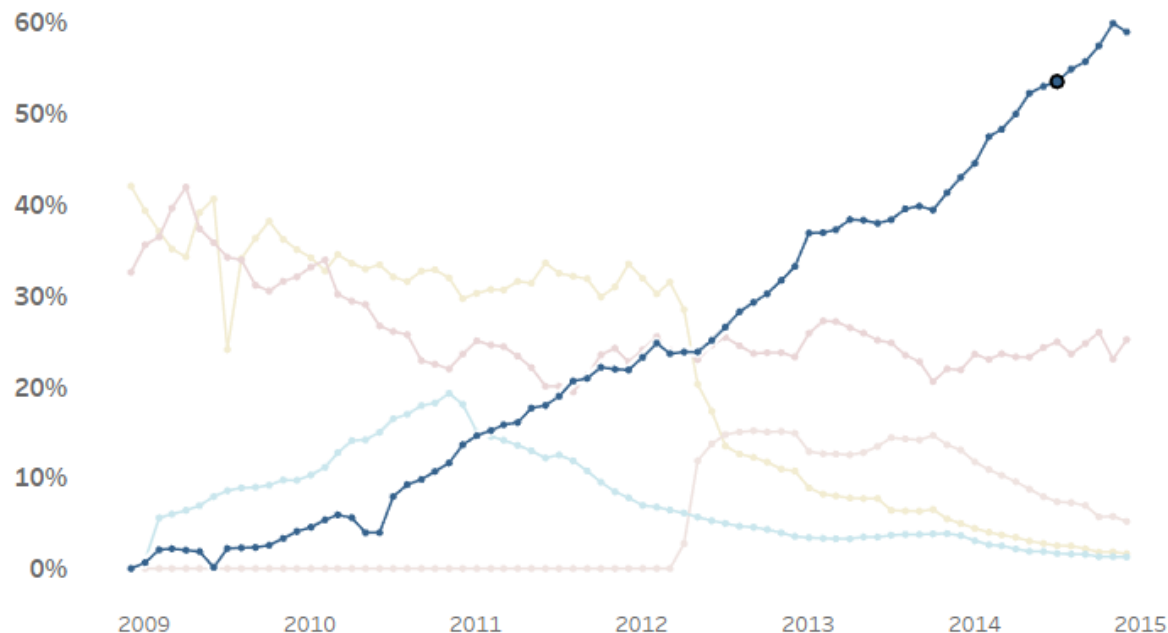
The analysis measures Total Revenue and Sum of Labor Costs Variable across 50 cities.

- As Sum of Labor Costs Variable increased, Total Revenue increased based on the data provided. Specifically, when Sum of Labor Costs Variable increased by \$1, Total Revenue increased \$4. There may be other factors contributing to Total Revenue, but there is evidence of a very strong relationship.
- When organized into groups of similar Sum of Labor Costs Variable and Total Revenue values, one distinct group stands out. There were 44 cities that had values of Sum of Labor Costs Variable between \$0 and \$1.5 million and Total Revenue between \$0 and \$4.8 million.
- The distribution of Sum of Labor Costs Variable ranges from \$0 to \$2 million. The average Sum of Labor Costs Variable per city is \$287,794 and the median is \$84,637.
- The minimum value for Total Revenue is \$0 and the maximum value is \$14.4 million. The average Total Revenue per city is \$2.1 million and the median is \$687,266.

powered by Narrative Science

### Percentage of Market Share per Mobile OS

December 2008 to December 2014



### Automated Insights

Natural Language Generation

In July 2014, 54 percent global mobile operating system market share was owned by Android. Goodness gracious! That's more than half!

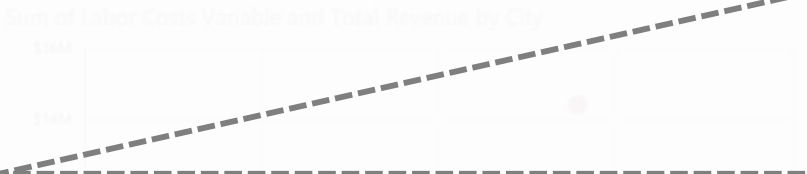
Between December 2008 and December 2014, their peak month was November 2014 at 60 percent.

The leader in usage at the time was Android.

ai AUTOMATED INSIGHTS

narrative science

ai AUTOMATED INSIGHTS



Percentage of Market Share per Mobile OS  
December 2008 to December 2014

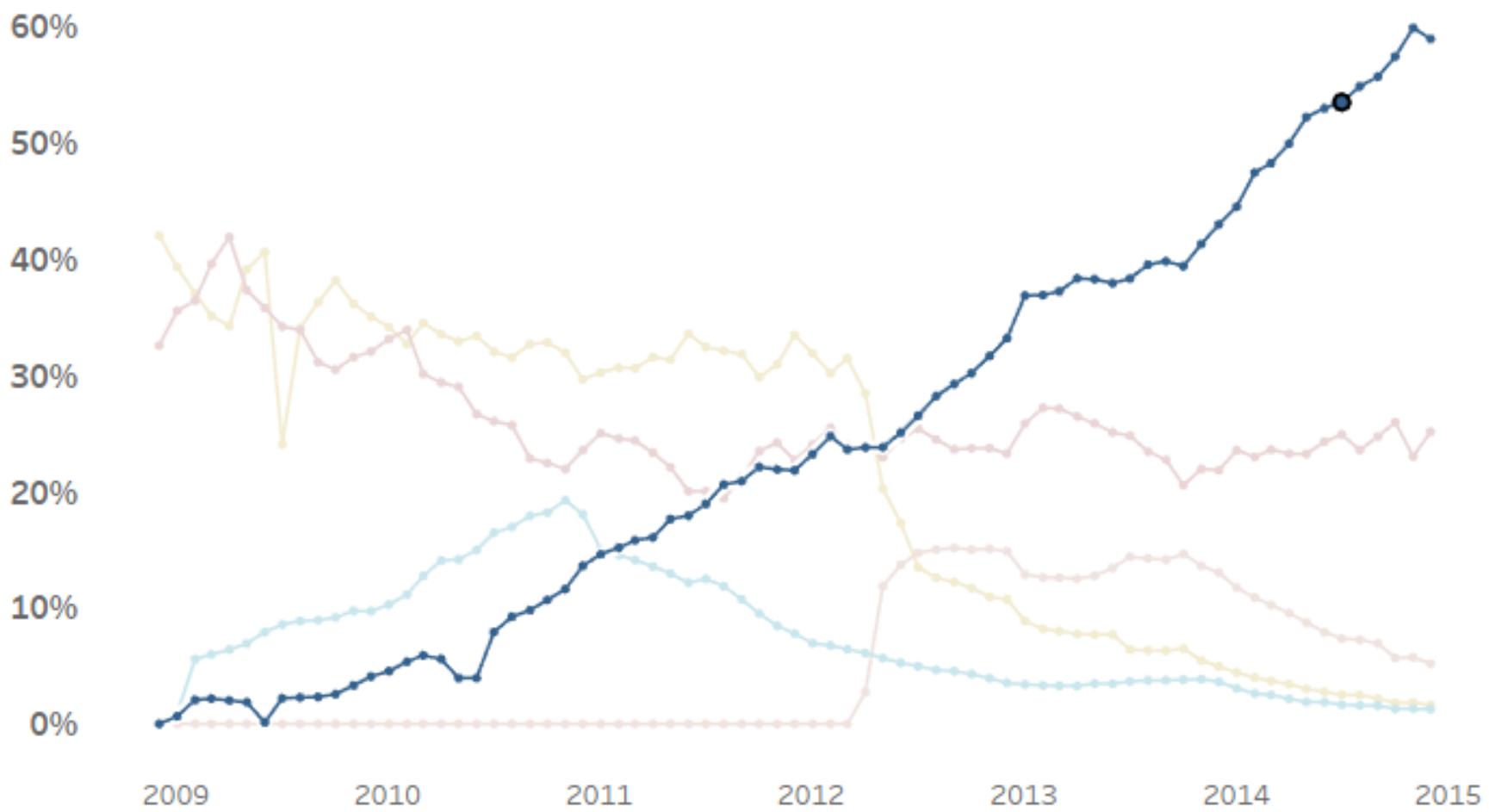
Automated Insights  
Natural Language Generation

60%

In July 2014, 54 percent global

Percentage of Market Share per Mobile OS  
December 2008 to December 2014

Automated Insights  
Natural Language Generation



In July 2014, 54 percent global mobile operating system market share was owned by Android. Goodness gracious! That's more than half!

Between December 2008 and December 2014, their peak month was November 2014 at 60 percent.

The leader in usage at the time was Android.



- For every day, “Dinner” increases by about 2.41
- There is a correlation between Acceleration and Horsepower
- Between December 2008 and December 2014, Android’s peak month was November 2014 at 60%
- Five cylinder cars have highest average Acceleration

“*Insights*”?



# Defining “*Insights*”

- Toward measuring visualization insight.  
North (2006)
- Defining insight for visual analytics.  
Chang et al. (2009)

- For every day, “Dinner” increases by about 2.41
- There is a correlation between Acceleration and Horsepower
- Between December 2008 and December 2014, Android’s peak month was November 2014 at 60%
- Five cylinder cars have highest average Acceleration

# ~~Insights~~ **Data Facts**

# Data Fact:

“a textual description of the result of one or more statistical functions applied to the data used to create a visualization.”

“a textual description of the result of one or more statistical functions applied to the data used to create a visualization.”

**Can we do more with data facts?**



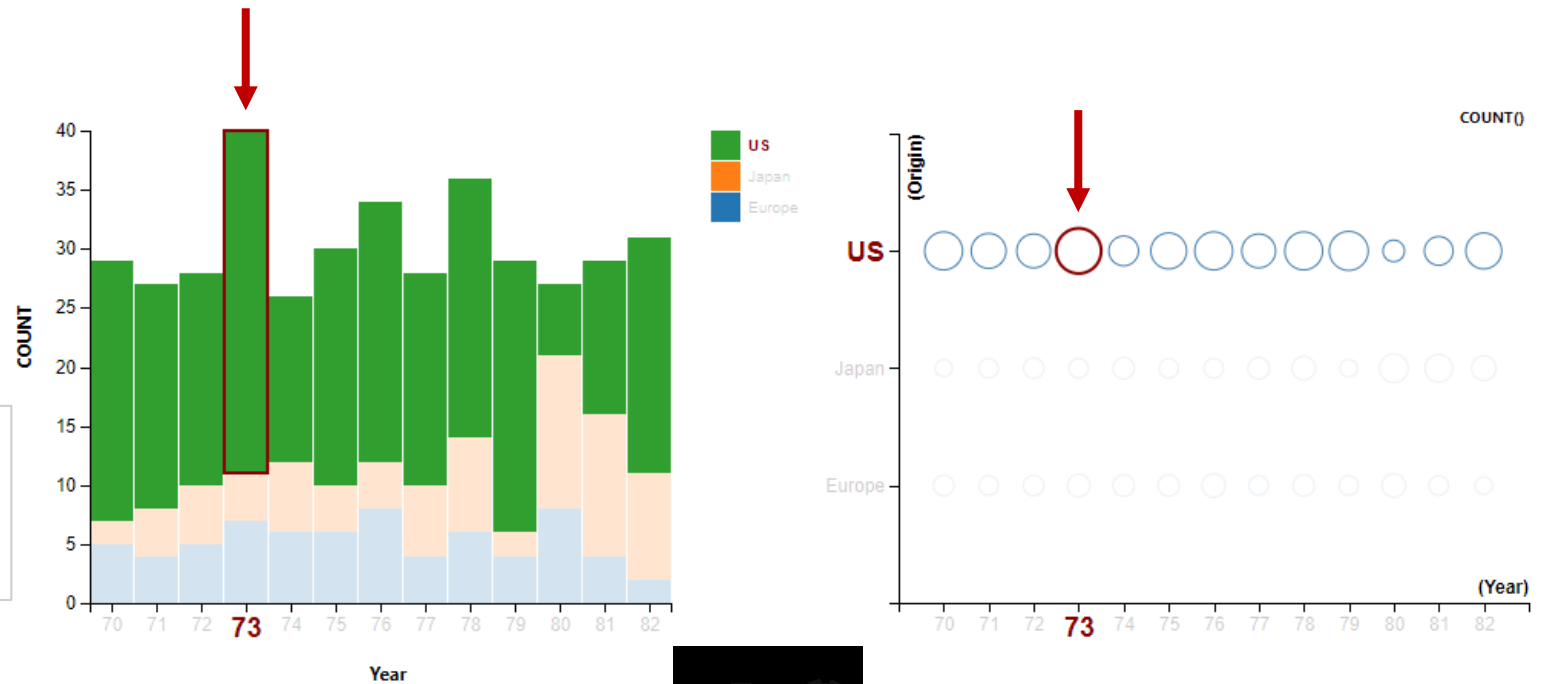
# Can data facts present communication-oriented alternatives?

US manufactures highest number of cars. The highest number of US cars were manufactured in 1973.

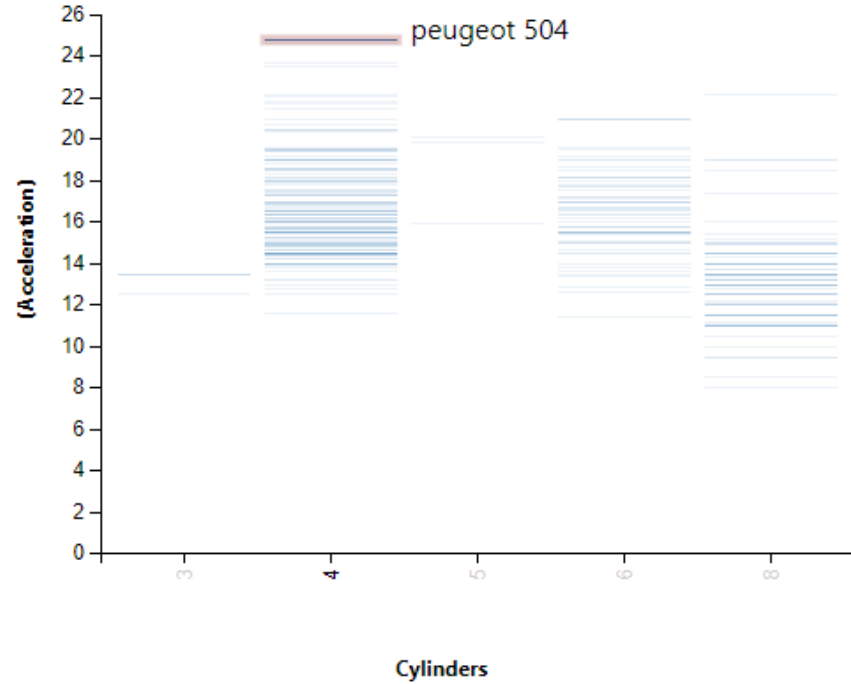


# Can data facts present communication-oriented alternatives?

US manufactures highest number of cars. The highest number of US cars were manufactured in 1973.



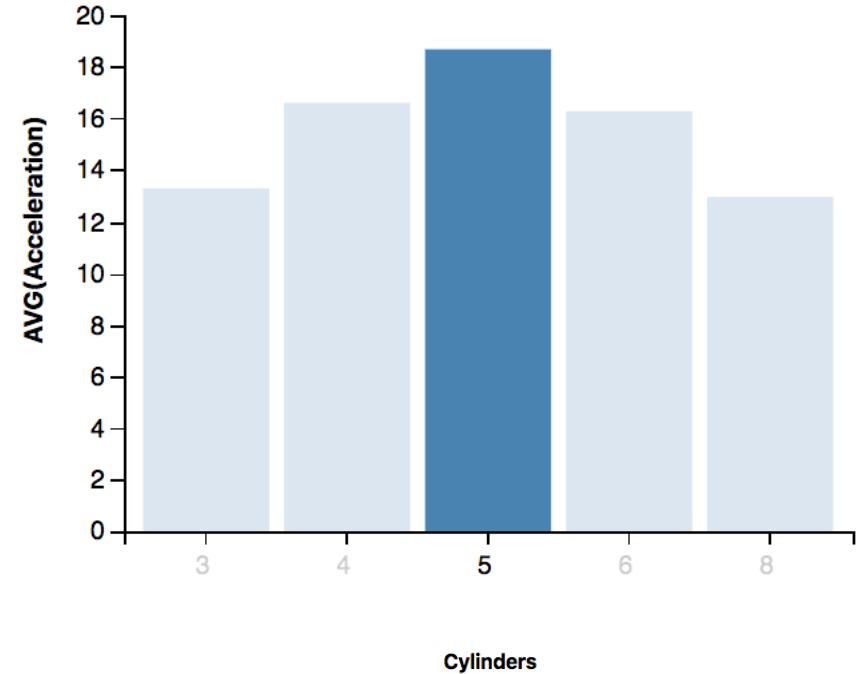
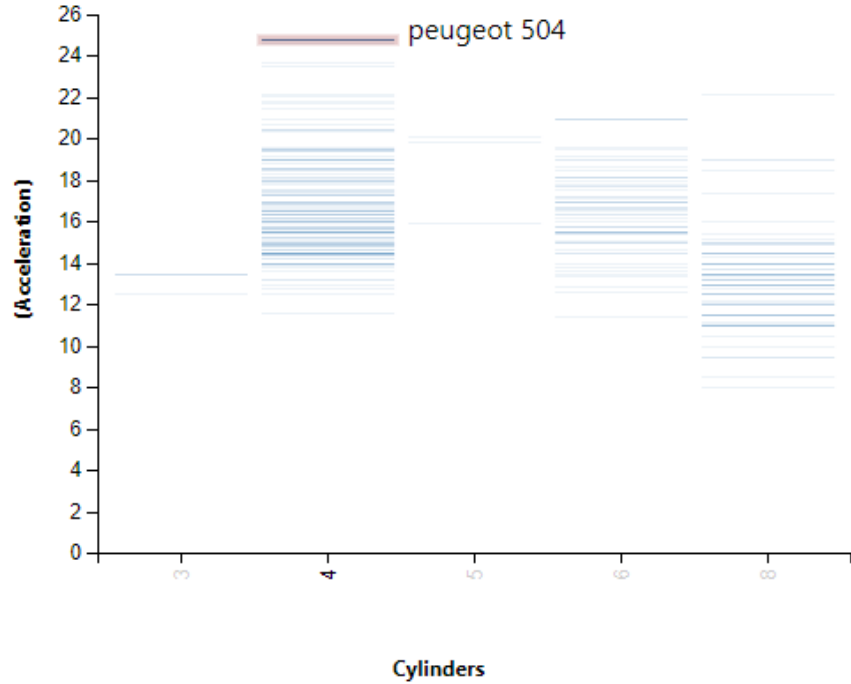
# Can data facts aid exploratory data analysis?



Cylinders: 4 has item (peugeot 504)  
with highest value for Acceleration



# Can data facts aid exploratory data analysis?



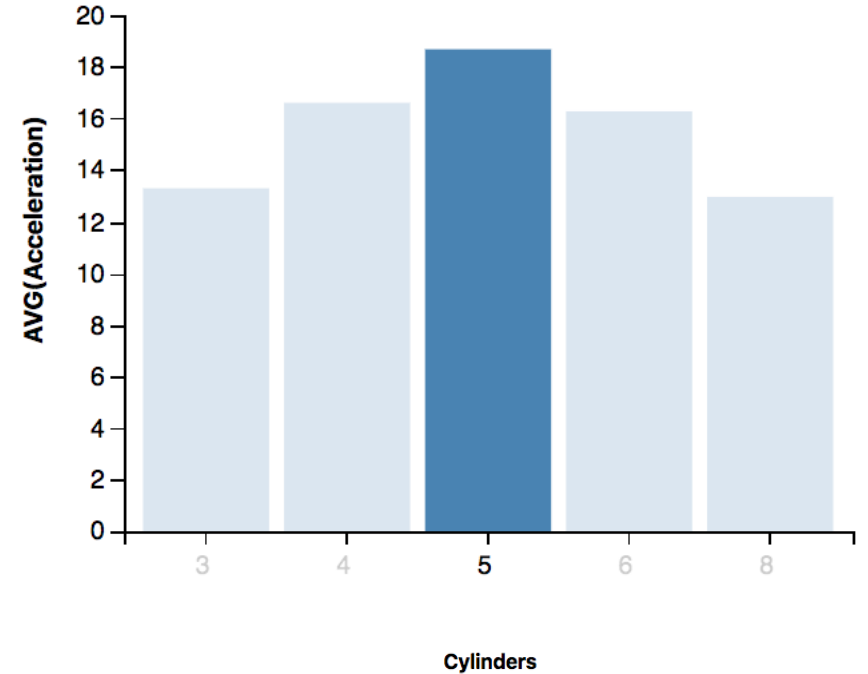
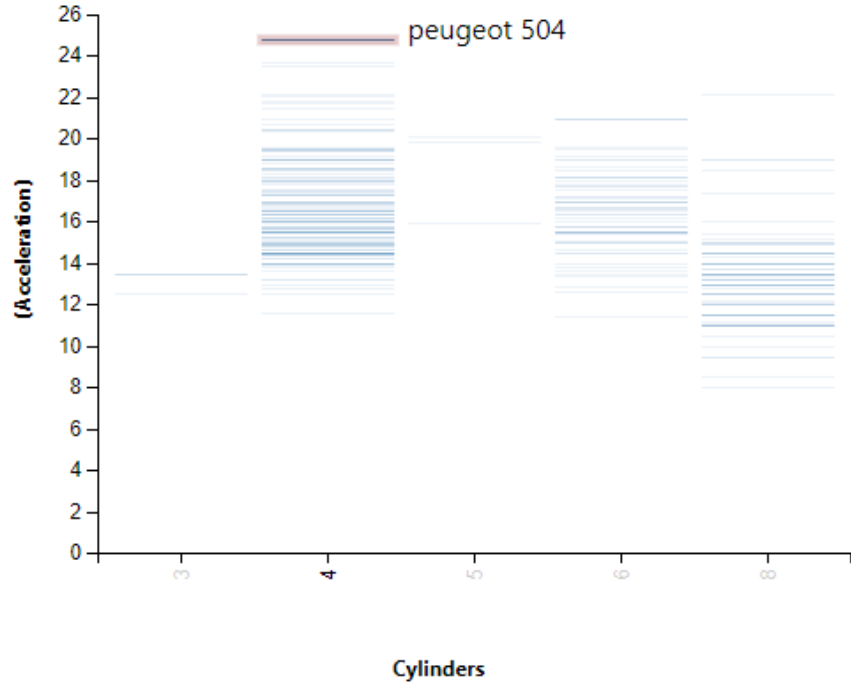
Cylinders: 4 has item (peugeot 504) with highest value for Acceleration



Cylinders: 5 has highest average Acceleration



# Can data facts aid exploratory data analysis?



Cylinders: 4 has item (peugeot 504) with highest value for Acceleration



Cylinders: 5 has highest average Acceleration



What are data facts?

**How** can we integrate data facts into visualization tools?

Why is this integration beneficial?

X:

Y:

Color:

Size:

Mark:

Show Possible Visualizations

Search for a fact in the dataset

# Voder

**Manual View Specification**



**Interactive Data Facts**

*Disc-shaped voice-box translation devices that enabled the Phylosians to converse in audible language with humanoid visitors*





X:

Y:

Color:

Size:

Mark:

Show Possible Visualizations

Data fact tier: 1

Search for a fact in the dataset







X:

Y:

Color:

Size:

Mark:

Show Possible Visualizations

---

Data fact tier: 1

**View Specification Panel**

Search for a fact in the dataset

+



X:

Y:

Color:

Size:

Mark:

Show Possible Visualizations

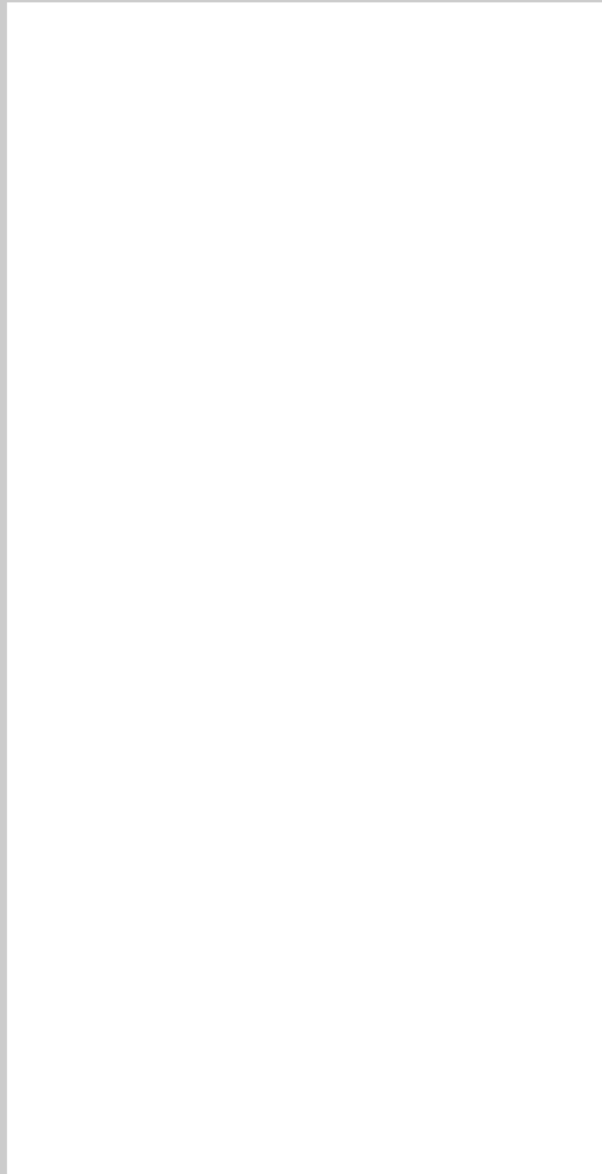
Data fact tier:



**Specified Visualization**

+

Search for a fact in the dataset





X:

Y:

Color:

Size:

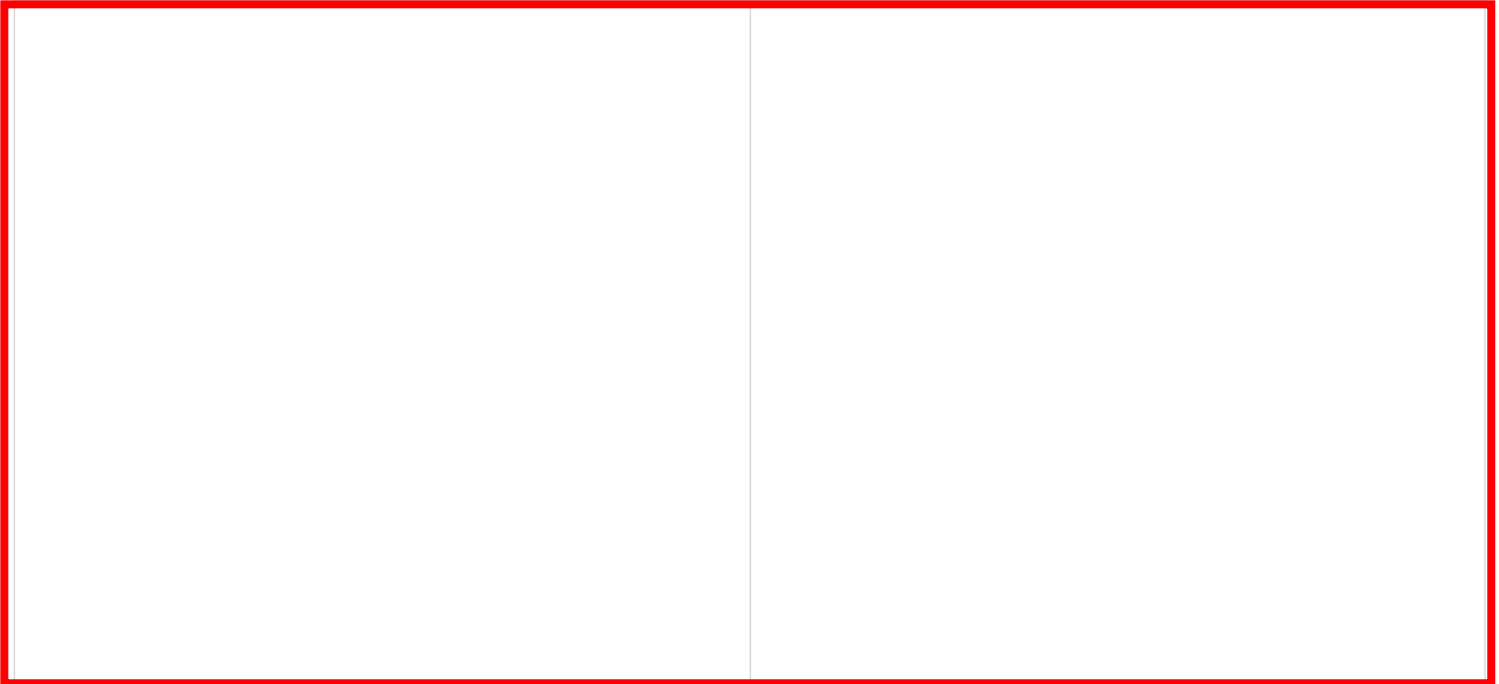
Mark:

Show Possible Visualizations

Data fact tier: 1

Search for a fact in the dataset

**Data Facts**





X:

Y:

Color:

Size:

Mark:

Show Possible Visualizations

Data fact tier:

Search for a fact in the dataset

+

**Bookmarked Data Facts**



X:

Y:

Color:

Size:

Mark:

Show Possible Visualizations

Data fact tier:

**Data Fact Search Panel**

+

Search for a fact in the dataset

Search results area



X:

Y:

Color:

Size:

Mark:

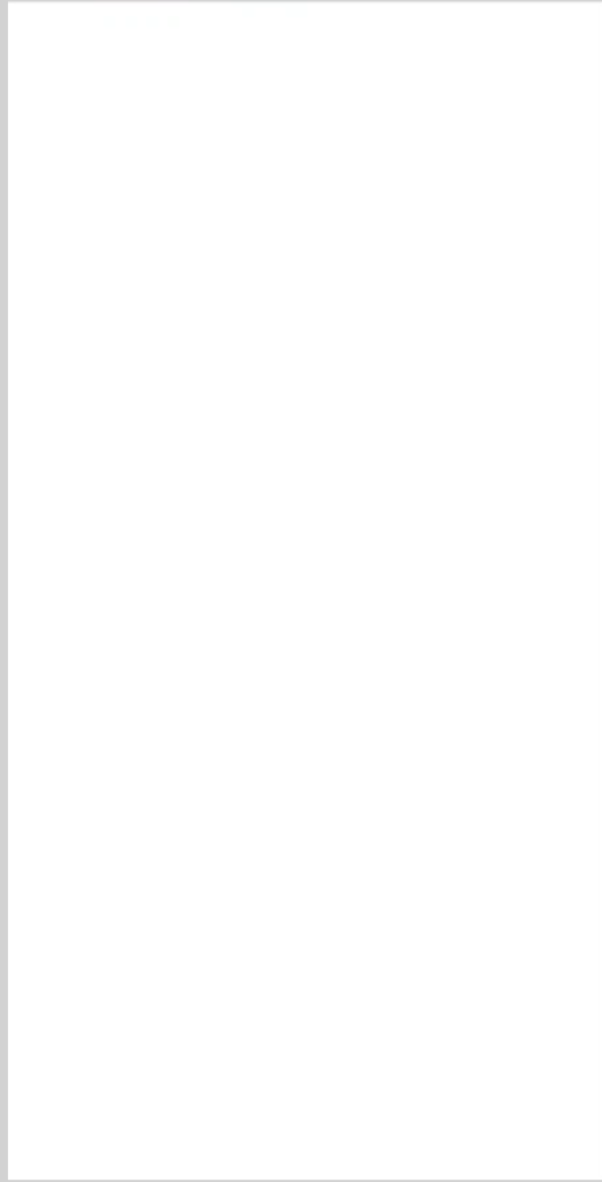
Show Possible Visualizations

Data fact tier:

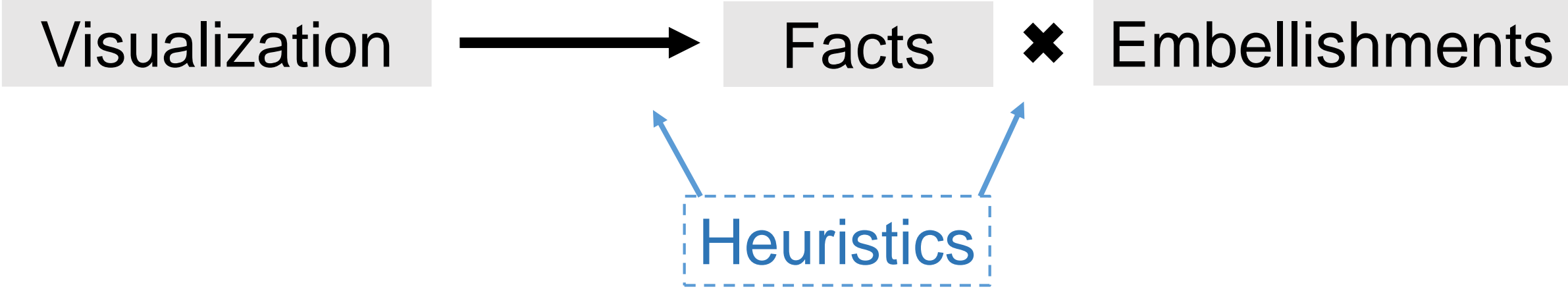


+

Search for a fact in the dataset



Demo

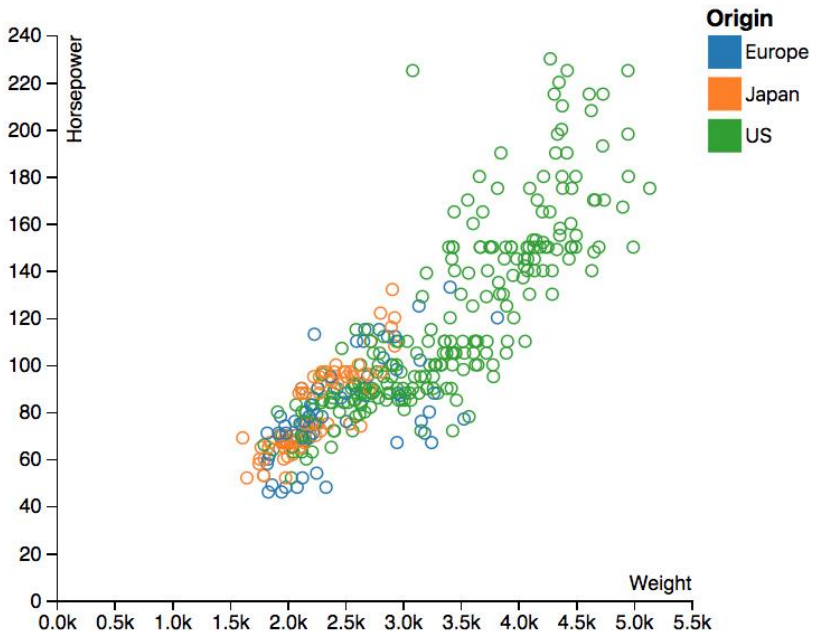






Q x Q x C

Horsepower x  
Weight x  
Origin



Visualization



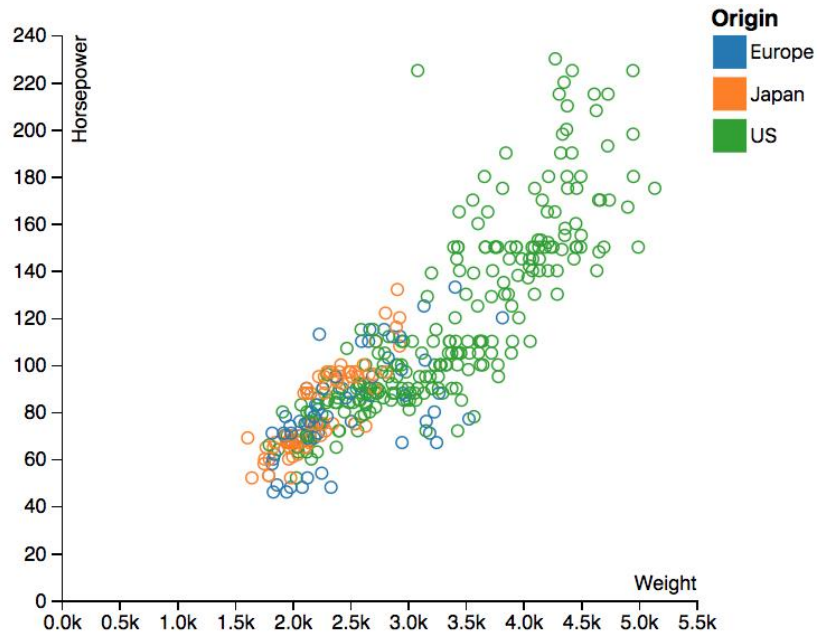
Facts



Embellishments

Q x Q x C

Horsepower x  
Weight x  
Origin



**Correlation**

Pearson's  $r < -.75$  or  $> .75$

**Distribution**

(#points in BIN > 75%)



**Visualization**



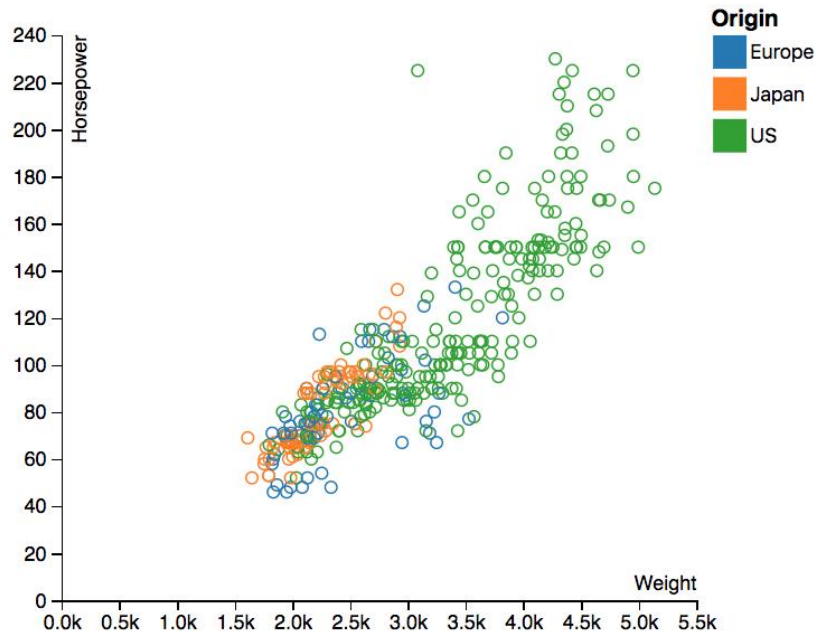
Facts



Embellishments

Q x Q x C

Horsepower x  
Weight x  
Origin



**Correlation**

Pearson's  $r < -.75$  or  $> .75$

**Distribution**

(#points in BIN > 75%)



- Overall, Horsepower and Weight have a strong correlation
- Items with Origin: Japan exhibit a strong correlation between Horsepower and Weight
- Items with Origin: US exhibit a strong correlation between Horsepower and Weight
- Most items with Origin: Japan have low Horsepower and low Weight

Visualization



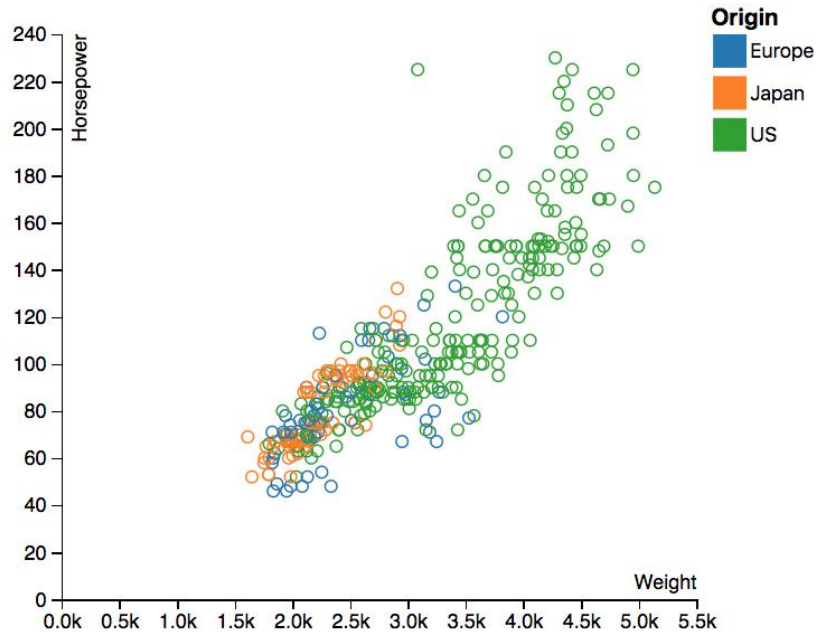
Facts



Embellishments

Q x Q x C

Horsepower x  
Weight x  
Origin



- Overall, Horsepower and Weight have a strong correlation
- Items with Origin: Japan exhibit a strong correlation between Horsepower and Weight
- Items with Origin: US exhibit a strong correlation between Horsepower and Weight
- Most items with Origin: Japan have low Horsepower and low Weight

Visualization



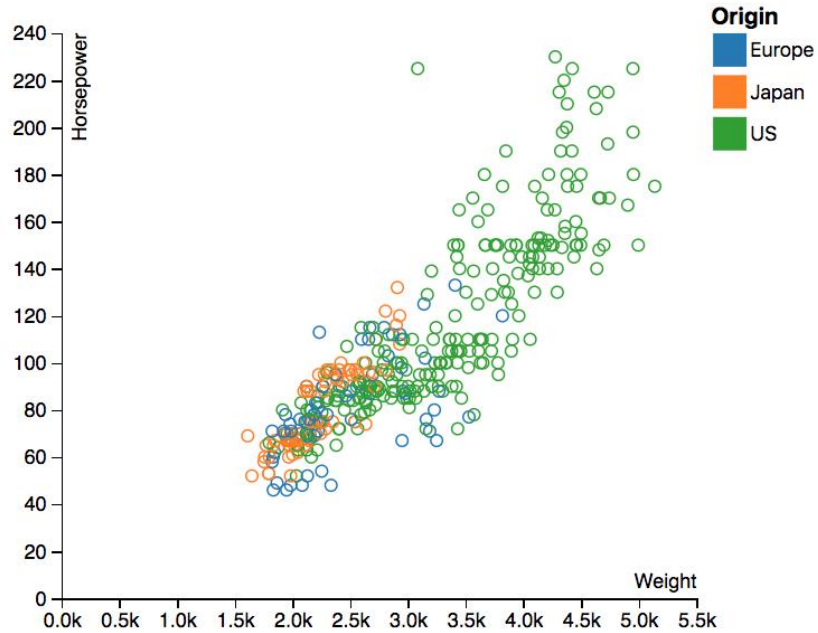
Facts



Embellishments

Q x Q x C

Horsepower x  
Weight x  
Origin



- Overall, Horsepower and Weight have a strong correlation
- Items with Origin: Japan exhibit a strong correlation between Horsepower and Weight
- Items with Origin: US exhibit a strong correlation between Horsepower and Weight
- Most items with Origin: Japan have low Horsepower and low Weight

Stroke	Opacity	Convex Hull	Quadrant Lines	Regression Line	Item Label	Text Highlight
				●		
●	●	●		●		
●	●	●		●		
●	●	●	●			

Visualization



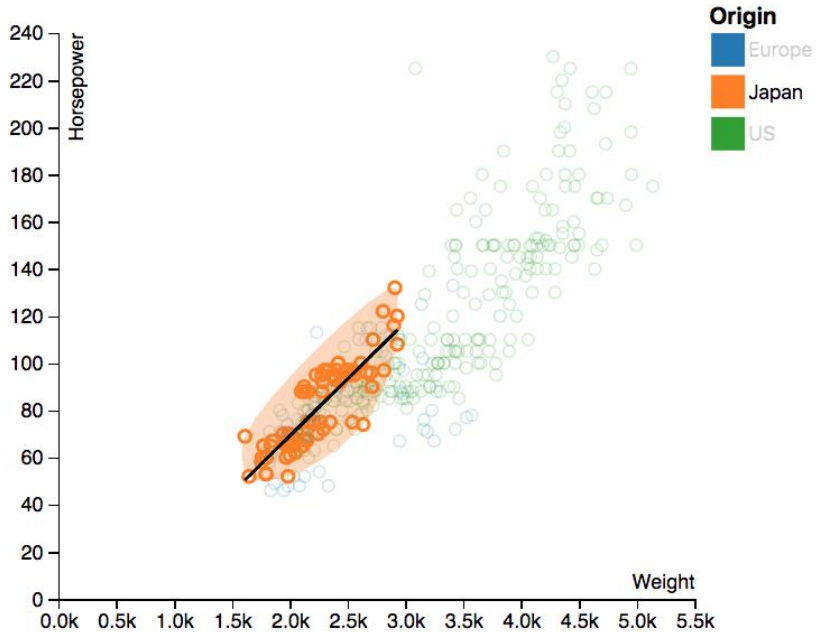
Facts



Embellishments

Q x Q x C

Horsepower x  
Weight x  
Origin



• Overall, Horsepower and Weight have a strong correlation

• Items with Origin: Japan exhibit a strong correlation between Horsepower and Weight

• Items with Origin: US exhibit a strong correlation between Horsepower and Weight

• Most items with Origin: Japan have low Horsepower and low Weight

Stroke

Opacity

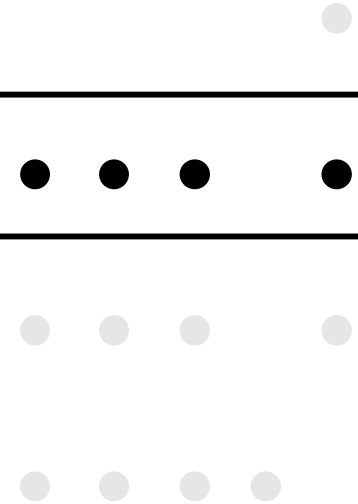
Convex Hull

Quadrant Lines

Regression Line

Item Label

Text Highlight



What are data facts?

How can we integrate data facts into visualization tools?

**Why** is this integration beneficial?

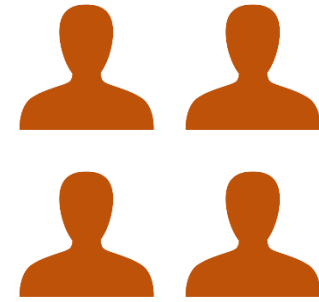


# User Study

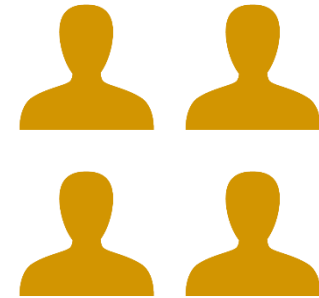


(Colleges) 18 Attributes  
1300 Data Items

**Task:** Explore the data and present your findings.



(Experts)

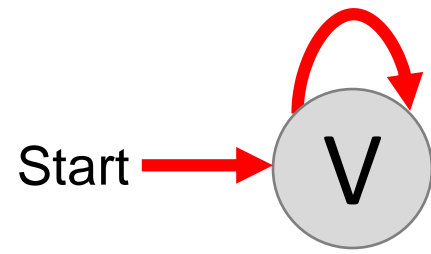



(Intermediate)



(Novices)

# Iterating between Visualizations and Facts during Exploration





**Voder** Visualizations (0) Facts (0)  ○ Explore ● Present

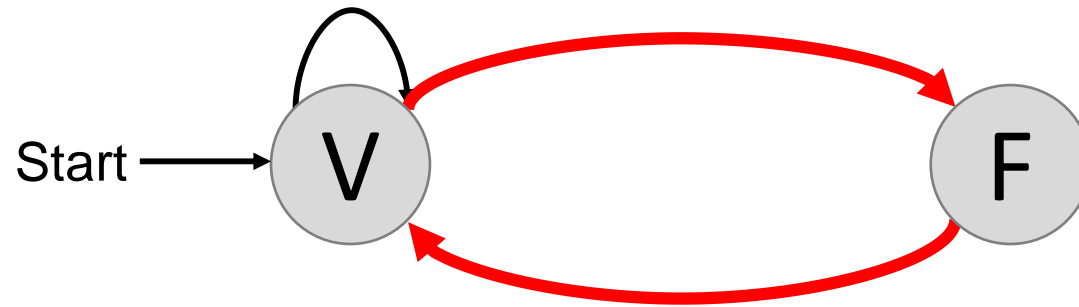
X:  ▼  
Y:  ▼  ▼  
Color:  ▼  
Size:  ▼  ▼  
Mark:  ▼

Show Possible Visualizations

Data fact tier:  ▼

Search for a fact in the dataset  





**Voder** Visualizations (1) Facts (1) Explore Present

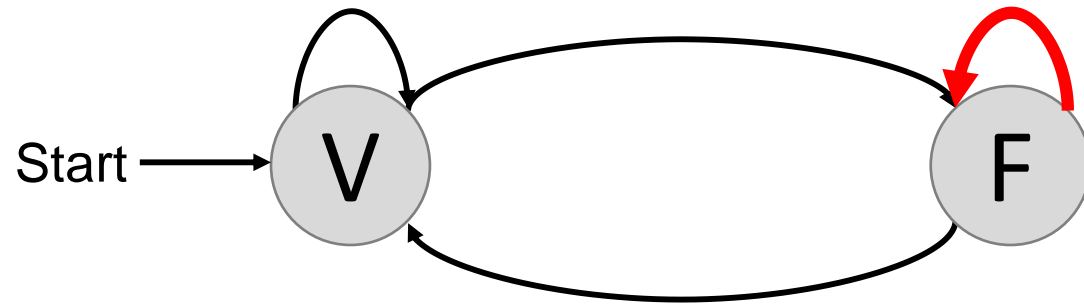
X: Region Y: Admission Rate Color: Size: Mark: Tick

Show Possible Visualizations

Data fact tier: 1

Search for a fact in the dataset

- Southeast has item (Cleveland State Community Colleges) with highest value for Admission Rate
- Far West has item (Stanford University) with lowest value for Admission Rate
- Southeast has highest total Admission Rate
- New England has lowest average Admission Rate
- Outlying Areas has highest average Admission Rate
- Outlying Areas has lowest total Admission Rate
- SUM(Admission Rate) of Southeast is 226.18 times Outlying Areas



**Voder** Visualizations (1) Facts (1) ⬇️ ○ Explore ● Present

X:  Y: Median Debt Color:  Size:  Mark: Tick

Show Possible Visualizations

Data fact tier: 1

Search for a fact in the dataset

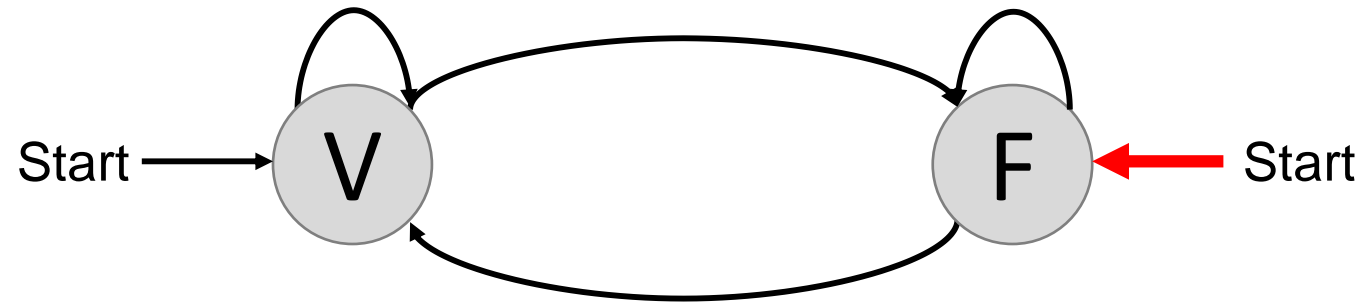
Southern California Institute of Architecture appears to be an outlier 👁️ ★

Southern California Institute of Architecture has highest value for Median Debt ★

Motlow State Community College appears to be an outlier 👁️ ★

Motlow State Community College has lowest value for Median Debt ★

Most values for Median Debt are in the range 14.6 k - 20.8 k 👁️ ★



**Voder** Visualizations (0) Facts (0) ○ Explore ● Present

X:

Y:

Color:

Size:

Mark:

Show Possible Visualizations

Data fact tier:

Search for a fact in the dataset

# Participant Feedback

- Varied preferences for using facts for interpretation.



(P11 - novice)

*“It’s almost like this **tool is training me by showing facts based on a visualization.** Now I can use this the other way around like **if I wanted to show a fact, I know which visualization I need to check.**”*



(P4 - expert)

*“The facts shown were **useful given that I didn’t know anything about the dataset.** But if this was a type of dataset that I use on a regular basis, **I’d want the system to tell me facts specific to the domain of the dataset.**”*

# Participant Feedback

- Some experts (2/4) felt suggestion of alternative visualizations was unnecessary.



(P3 - expert)

*“I feel the system should be smart enough to select the best visualization for a statement automatically. In fact, when we’re working on building some sort of report, I tell my team not to worry about the visualization and always add the default one suggested by the system.”*



# Participant Feedback

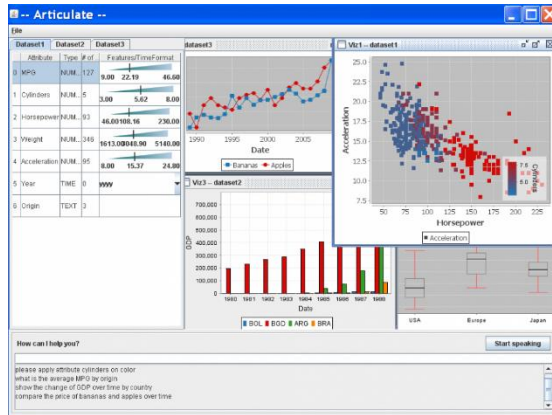
- All participants liked suggestion of alternative embellishments and highlighting via data facts.



(P6 - intermediate)

*“it was nice to have the system consistently show facts in the visualization by fading things out. **Since getting to possible styling options was easy, I could simply go in and format a chart further when I wanted to.**”*

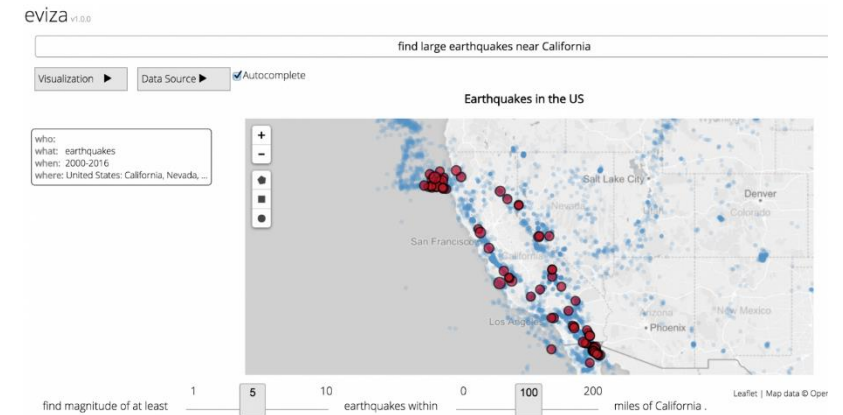
# Research Opportunity: Integrating Natural Language Understanding (NLU) and Generation (NLG)



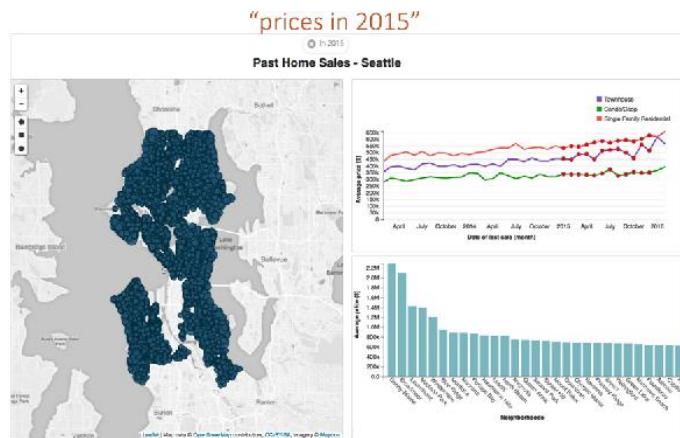
**Articulate**, Sun et al. 2011



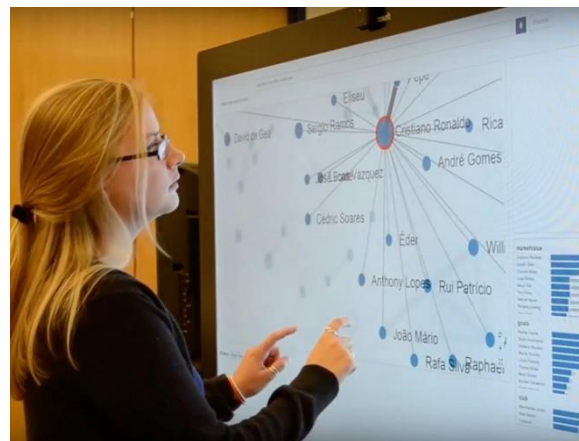
**DataTone**, Gao et al. 2015



**Eviza**, Setlur et al. 2016

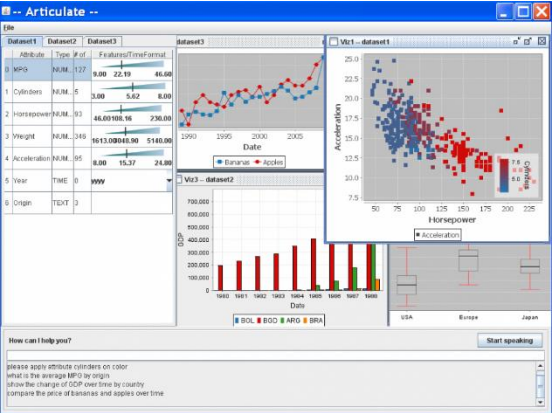


**Evizeon**, Hoque et al. 2017



**Orko**, Srinivasan & Stasko 2017

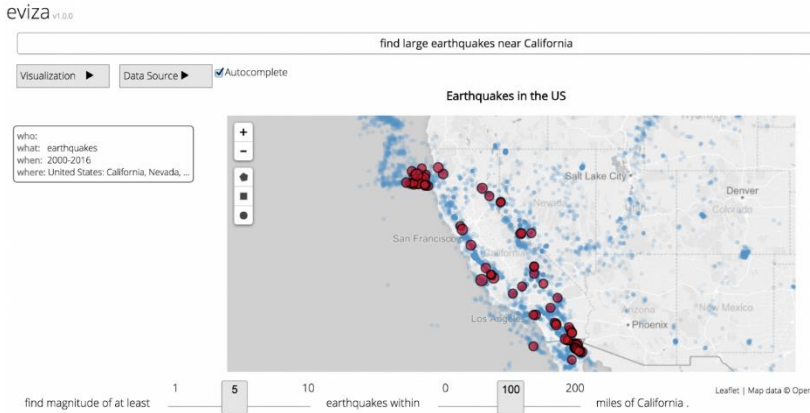
# Research Opportunity: Integrating Natural Language Understanding (NLU) and Generation (NLG)



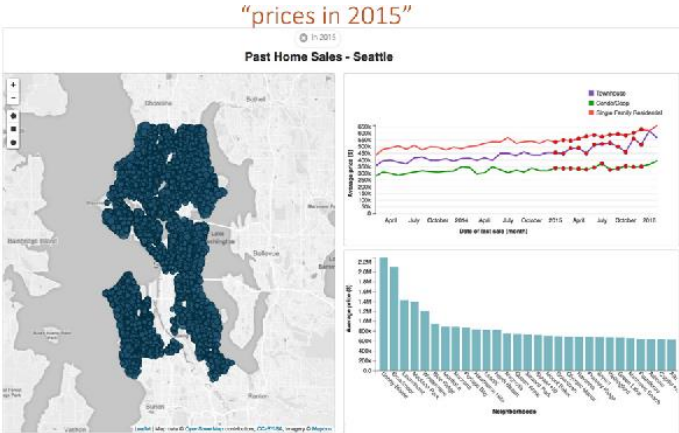
Articulate, Sun et al. 2011



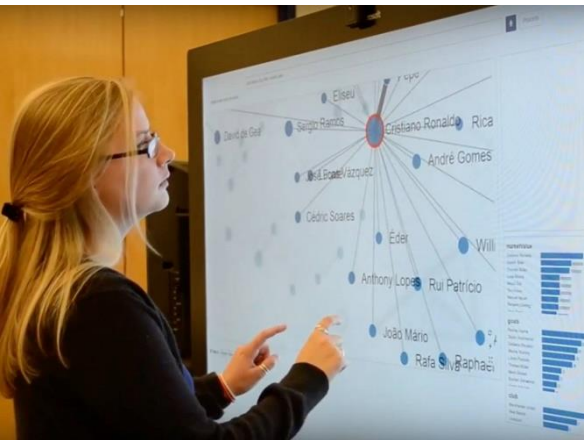
DataTone, Gao et al. 2015



Eviza, Setlur et al. 2016



Evizeon, Hoque et al. 2017

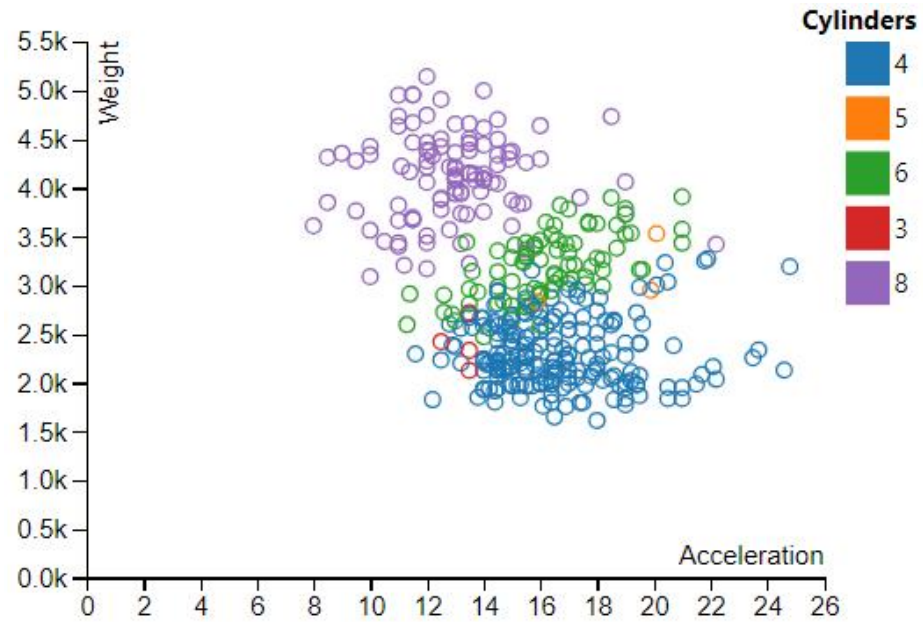


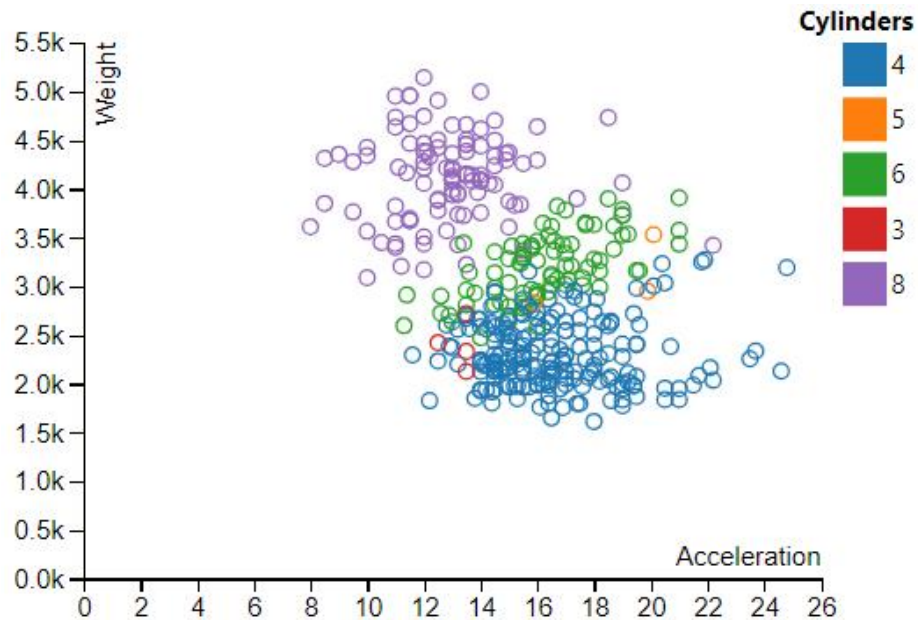
Orko, Srinivasan & Stasko 2017





Compare weight and acceleration for cars with different cylinder counts



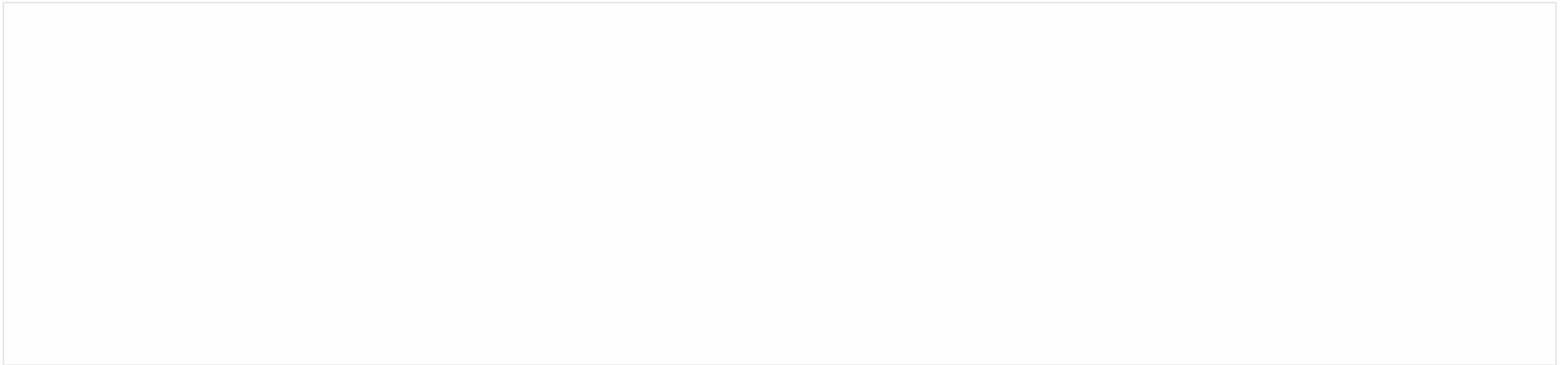


**Here are some commands you might want to try:**

- Is there a group of cars that exhibits a strong correlation between acceleration and weight?
- Highlight heavy but fast cars.



# Summary

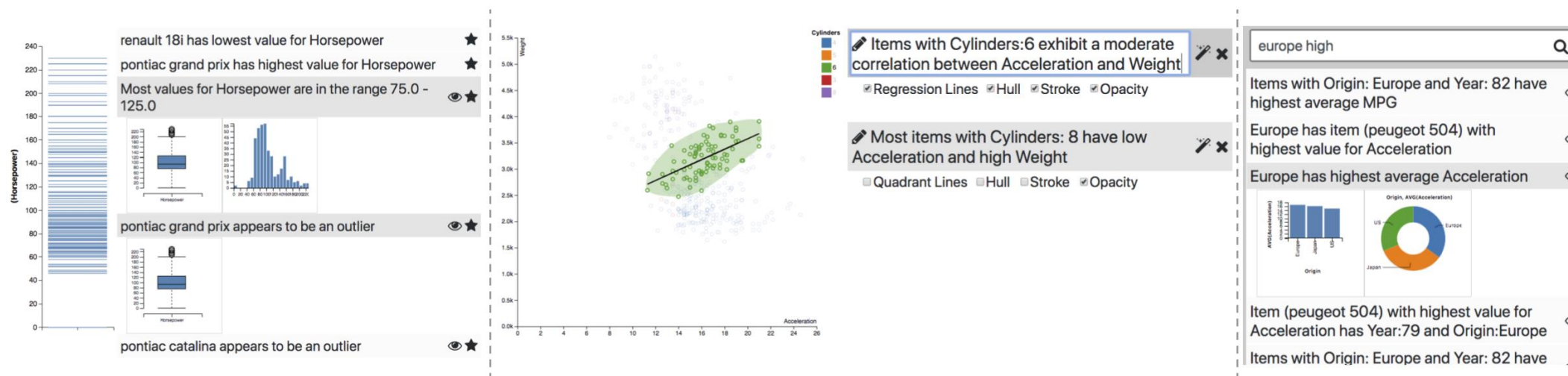


Augmenting visualizations with interactive data facts helps:

- Interpret visualizations and explore alternatives for communication
- Facilitate flexible exploration strategies allowing users to switch between top-down and bottom-up exploration



# Thank You



Augmenting visualizations with interactive data facts helps:

- Interpret visualizations and explore alternatives for communication
- Facilitate flexible exploration strategies allowing users to switch between top-down and bottom-up exploration

Arjun Srinivasan (@10\_arjun)

Steven M. Drucker

Alex Endert

John Stasko

**Backup slides...**



# Potential Risks

- Participants skipped visualizations when the system did not show facts (**Trust**)
- Ability to search for facts and select visualizations and embellishments to show those can be misused (**Deception**)

# Usage Summary

Visualizations created	86 (4-12 per session)
Corresponding data facts saved	119 (4-17 data facts per session)
System generated	102
Manually entered	17
Search queries executed	31 (9 sessions, 1-9 per session)

Overall, Horsepower and Weight have a strong correlation

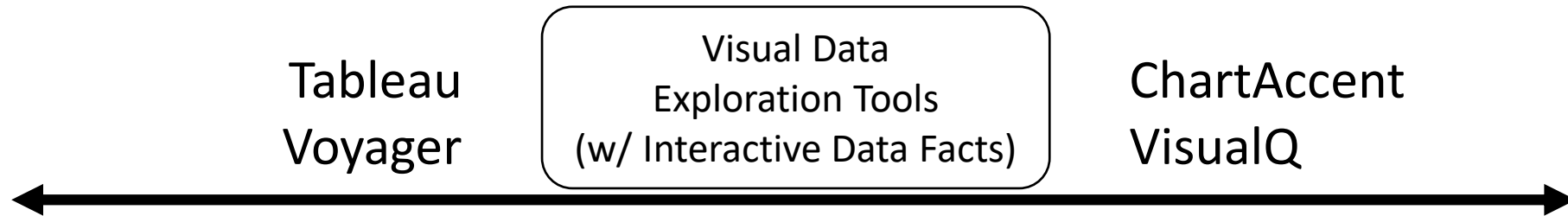
**Attributes:** Horsepower, Weight  
**Tasks:** Correlate

Most items with Origin: Japan have low Horsepower and low Weight

**Attributes:** Horsepower, Weight  
**Tasks:** Characterize Distribution,  
Filter  
**Value:** Japan

# Future Work

- Integrating NLU and NLG
- Integration with partial view specification-based tools
- Recommending exploratory facts and visualizations based on user interest



Focus: Visual Data  
Exploration

Focus: Data-driven  
Communication

Voyager: Wongsuphasawat et al. (2016)

ChartAccent : Ren et al. (2017)  
VisualQ: Kong et al. (2017)

# Static Data Facts: Pros and Cons

+ Help detect salient facts or confirm inferences

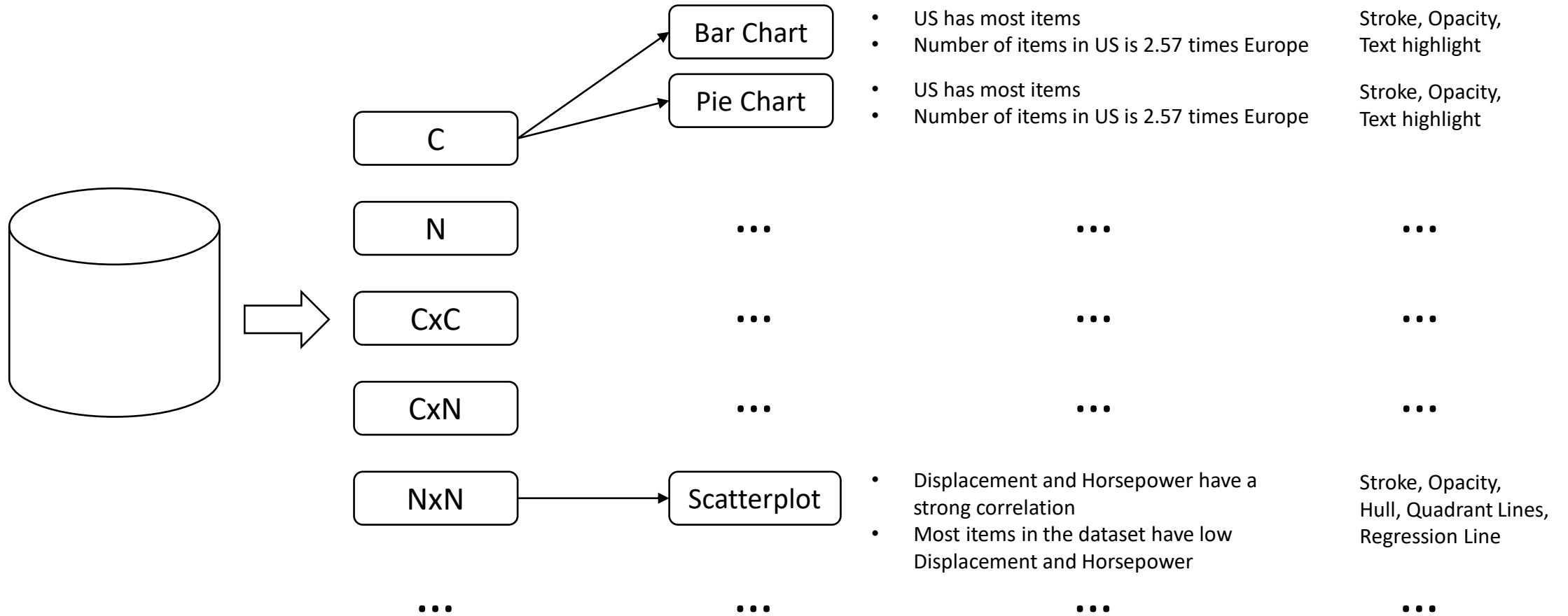
+ Provide richer information about user intent

+ Aid communication or sharing of findings

- Difficult to read facts and mentally map to visualization

- Alternative ways to show facts remain unexplored

# How it works?



# Behavior highlights

- Experts:
  - More **V->V** than non-experts
- Novices:
  - Used search most
  - Almost no V->V
- Search feature:
  - 3 participants started with search (one in each group).
  - 2 used it more after they tried it once.
  - Results were clicked 18/31 times (58%)