Collecting and Characterizing
Natural Language Utterances for Specifying Data Visualizations

Arjun Srinivasan

Georgia Tech

tableau

Nikhila Nyapathy

Georgia Tech

Bongshin Lee

Microsoft

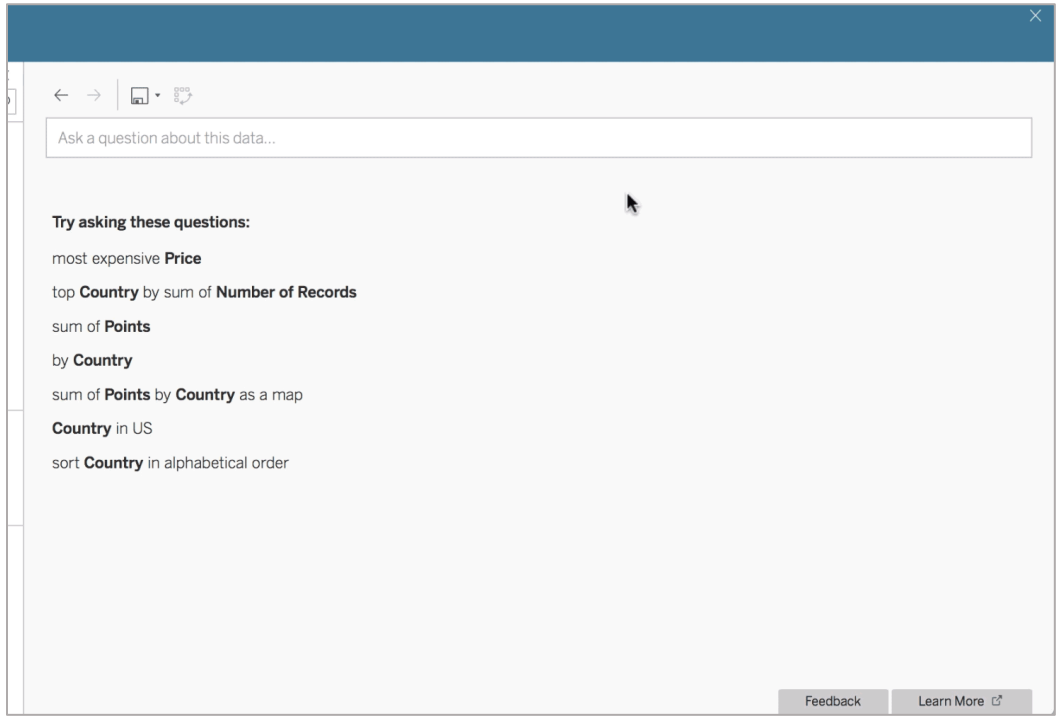
Steven M. Drucker

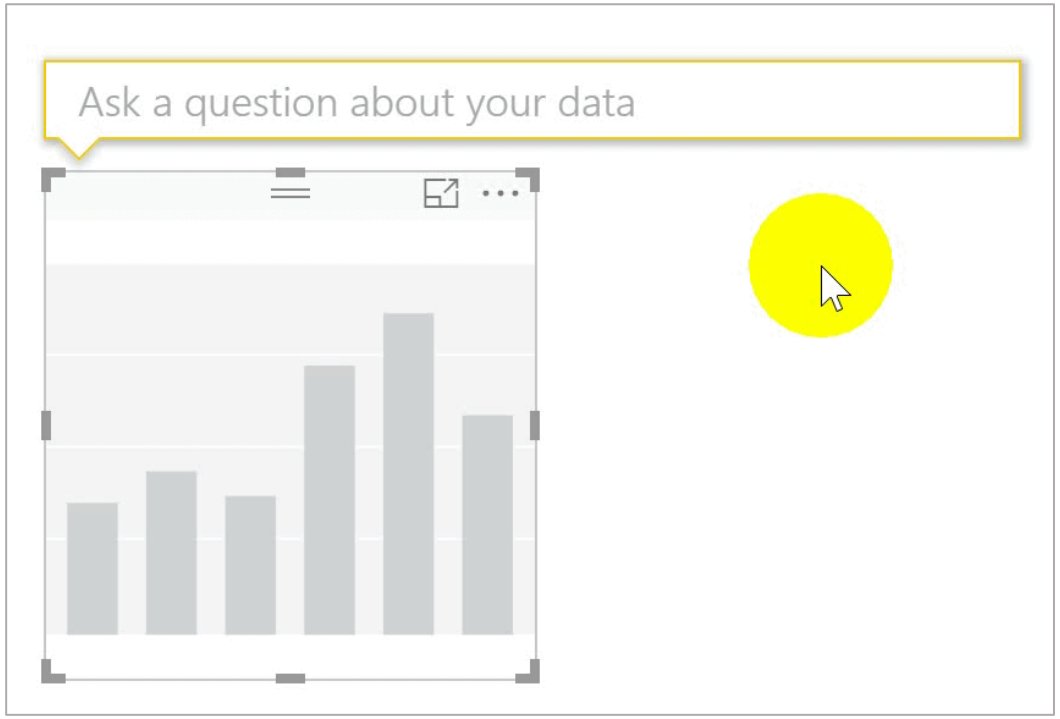
Microsoft

John Stasko

Georgia Tech

Natural Language Interfaces for Data Visualization
are becoming popular…





tableau

Ask Data

Power BI

Q&A

What types of utterances do people naturally use for specifying data visualizations?

Online Study

Online Study

- **1 dataset** per session

Cars

Model	Origin	Year	Acceleration	MPG	...
Volkswagen Dasher	Europe	1974	15.5	26	...
Honda Civic	Japan	1976	17.4	33	...
Ford Fiesta	USA	1978	14.4	36.1	...
Mercedes-Benz 240d	Europe	1980	21.8	30	...
Dodge Aspen	USA	1980	18.7	19.1	...
...					

Movies

Title	Major Genre	Release Year	Worldwide Gross	IMDB Rating	...
Titanic	Thriller	1997	1.84G	7.4	...
The Dark Knight	Action	2008	1.02G	8.9	...
Shrek 2	Adventure	2004	919M	7.5	...
Ratatouille	Comedy	2007	620M	8.1	...
I am Legend	Horror	2007	585M	8.1	...
...					

Product Sales

Order ID	Product	Order Quantity	Profit Ratio	Region	Order Date	...
CA-2016-124352	Hoover Upright Vacuum With Dirt Cup	3	868.59	Central	10/15/2016	...
CA-2016-109365	Xerox 1892	3	116.28	West	11/03/2016	...
US-2016-103674	Avaya 5410 Digital phone	5	271.96	West	12/06/2016	...
CA-2017-107727	Easy-staple paper	3	29.472	Central	10/19/2017	...
CA-2017-134404	AT&T 1080 Corded phone	2	164.388	East	12/27/2017	...
...	...	...	...	...	...	...

Online Study

- 1 dataset per session
- **10 visualizations** per dataset

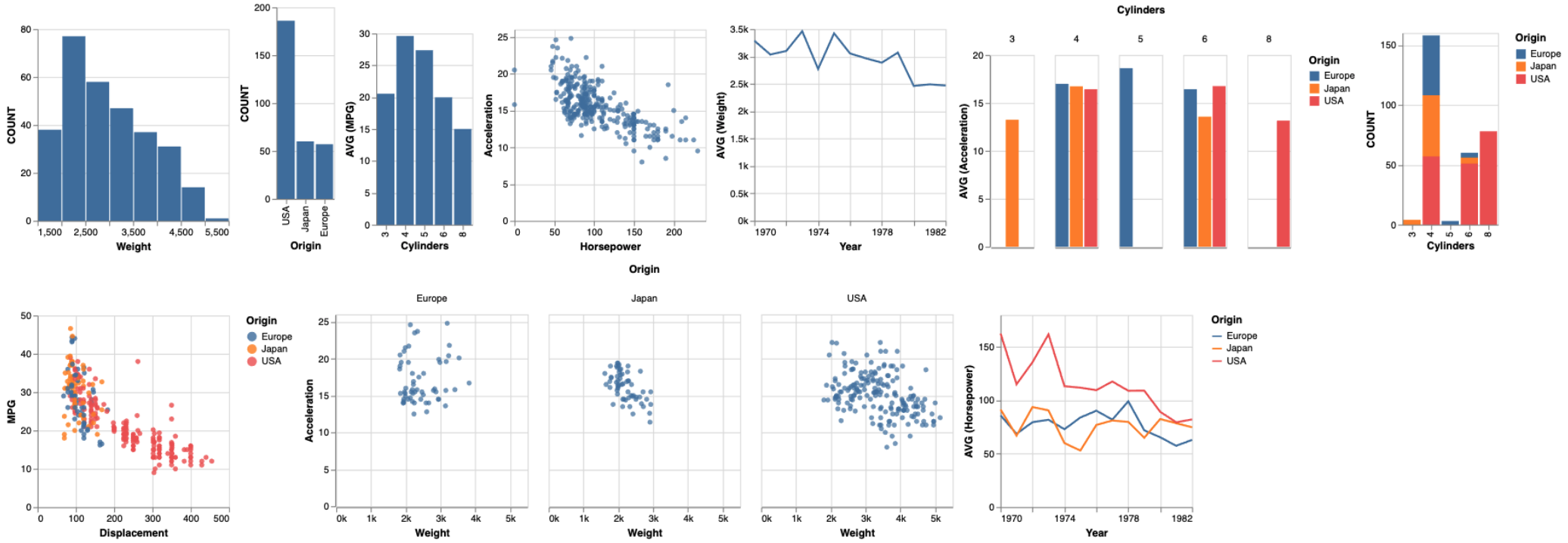
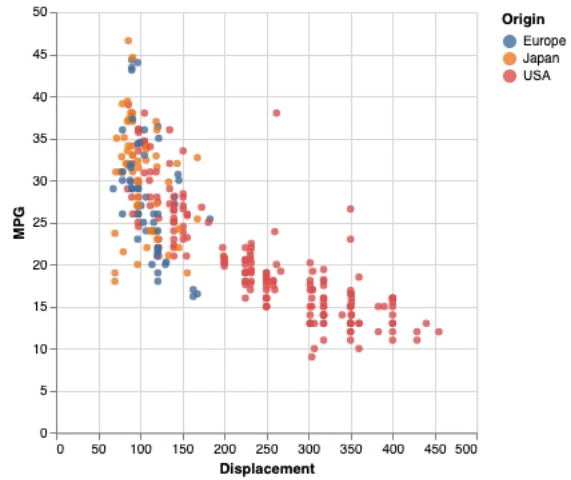


Chart 3/10

Model	Origin	Year	Acceleration	MPG	...
Volkswagen Dasher	Europe	1974	15.5	26	...
Honda Civic	Japan	1976	17.4	33	...
Ford Fiesta	USA	1978	14.4	36.1	...
Mercedes-Benz 240d	Europe	1980	21.8	30	...
Dodge Aspen	USA	1980	18.7	19.1	...
...	...	...	...	...	...

Note: The above table only shows a portion of the dataset. The complete dataset contains 303 rows and 9 columns.



Provide one or more natural language statements/queries/commands/questions you would enter in a system to specify the visualization on the right based on the given dataset.

1*: Enter text here.

+ Add More

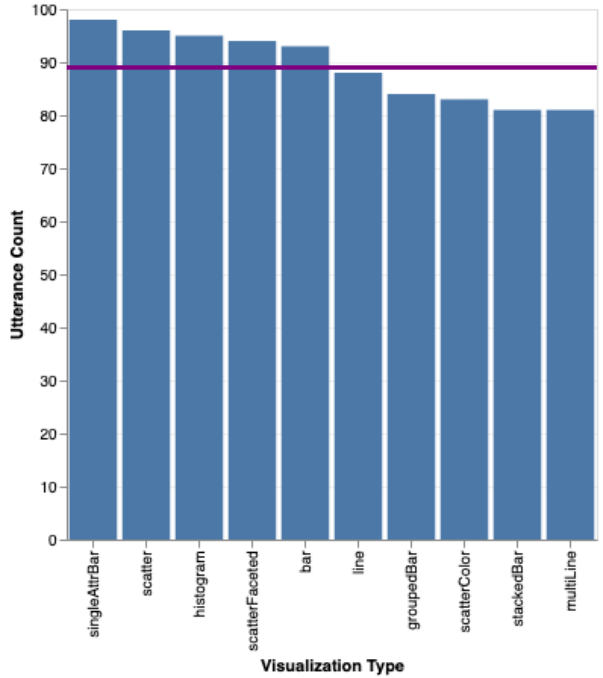
Next

(Fill out the *required fields to proceed)

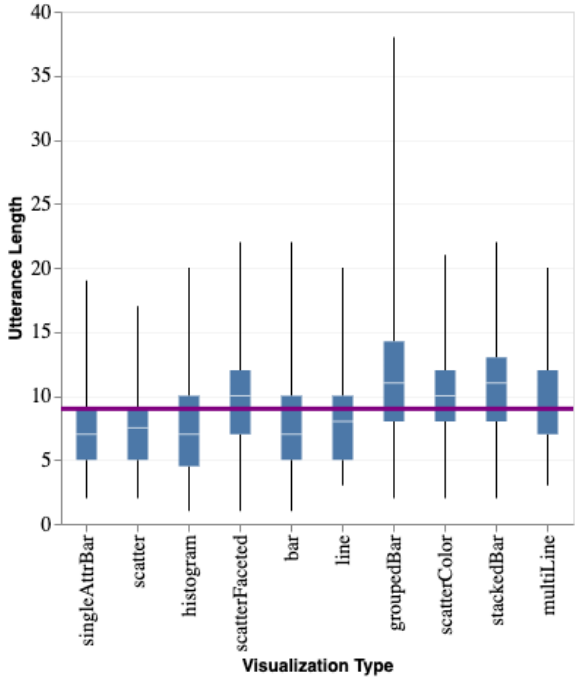
Data Collected

- 102 participants (76 full sessions; 26 partial)

- 893 utterances (Cars: 332, Movies: 290, Superstore: 271)



89 (avg.)
utterances per chart type



9 (avg.)
words per utterance

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

WHAT information do utterances contain?

HOW are utterances phrased?

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

WHAT information do utterances contain?

- Attributes

HOW are utterances phrased?

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

Explicit

WHAT information do utterances contain?

- Attributes

HOW are utterances phrased?

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

Explicit

Semantic

WHAT information do utterances contain?

- Attributes

HOW are utterances phrased?

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

Explicit

Semantic

Value-based

WHAT information do utterances contain?

- Attributes

HOW are utterances phrased?

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

WHAT information do utterances contain?

- Attributes
- Chart Type & Encodings

HOW are utterances phrased?

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

WHAT information do utterances contain?

- Attributes
- Chart Type & Encodings
- Aggregations

HOW are utterances phrased?

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

WHAT information do utterances contain?

- Attributes
- Chart Type & Encodings
- Aggregations
- Design

HOW are utterances phrased?

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

WHAT information do utterances contain?

- Attributes
- Chart Type & Encodings
- Aggregations
- Design

HOW are utterances phrased?

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

WHAT information do utterances contain?

- Attributes
- Chart Type & Encodings
- Aggregations
- Design

HOW are utterances phrased?

- Commands

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

WHAT information do utterances contain?

- Attributes
- Chart Type & Encodings
- Aggregations
- Design

HOW are utterances phrased?

- Commands
- Queries

Utterance Characterization

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart

- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

WHAT information do utterances contain?

- Attributes
- Chart Type & Encodings
- Aggregations
- Design

HOW are utterances phrased?

- Commands
- Queries
- Questions

Utterance Characterization

Details in the paper

- draw a line chart of daily sales forecasts
- Show me a bar graph of the profit for each region > Make the bars stacked with the ship status
- Please show me a histogram of weights with 500 intervals.
- Cylinders average mpg
- Count by origin
- mpg vs displacement > as scatter chart
- How much do various cars weigh?
- What is our profit based on shipping mode by customer segment?
- How does displacement relate to fuel economy for cars from Europe v. USA?

WHAT information do utterances contain?

- Attributes
- Chart Type & Encodings
- Aggregations
- Design

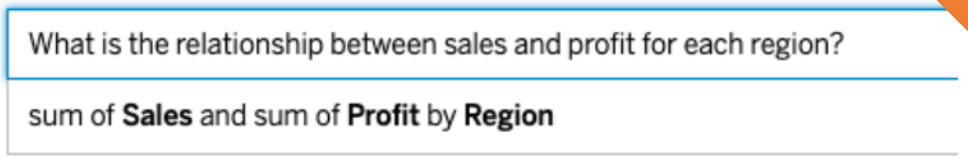
HOW are utterances phrased?

- Commands
- Queries
- Questions

Implications for System Design

Details in the paper

- Accommodating natural phrasings as part of user input in visualization tools



- Inferring different types of attribute references

Miles per gallon vs. mpg vs. fuel economy

- Balancing automated and manual view specification.

Show MPG and displacement by Origin

Show MPG and displacement and split cars by Origin

Using the Corpus

Details in the paper

nlvcorpus.github.io

Example applications:

- Benchmarking NL-based visualization tools like NL4DV
- Developing new models for NL-driven data visualization

Thank You

nlvcorpus.github.io

We present:

- A corpus of natural language utterances for specifying data visualizations.

- A characterization of these utterances along with implications for future system design.

Arjun Srinivasan

Georgia Tech

tableau

Nikhila Nyapathy

Georgia Tech

Bongshin Lee

Microsoft

Steven M. Drucker

Microsoft

John Stasko

Georgia Tech