

VIS 2023

NLP4Vis: Natural Language Processing for Information Visualization *Half-day Tutorial*

Enamul Hoque, York University

<https://nlp4vis.github.io/>

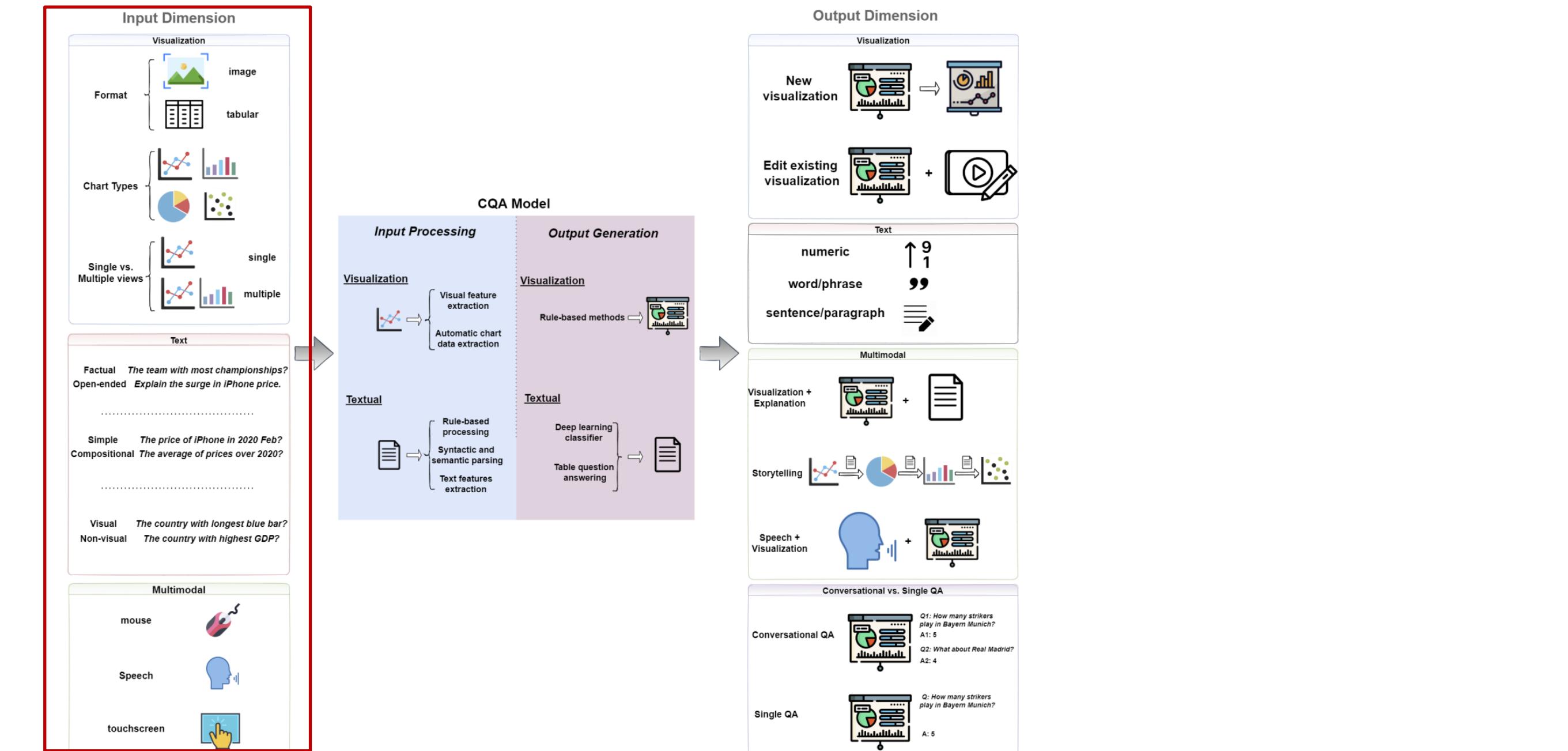
Tutorial Overview

- **Part 1:** Introduction [15 mins]
 - Why NLP + Vis?
 - An overview of NLP + Vis Research
 - An overview of the tutorial
- **Part 2:** Deep Learning for NLP [50 mins]
 - Background
 - Large language models (LLMs)
- **Part 3: NLP4Vis applications [50 mins]**
- **Part 4:** Future challenges and research opportunities [25 mins]

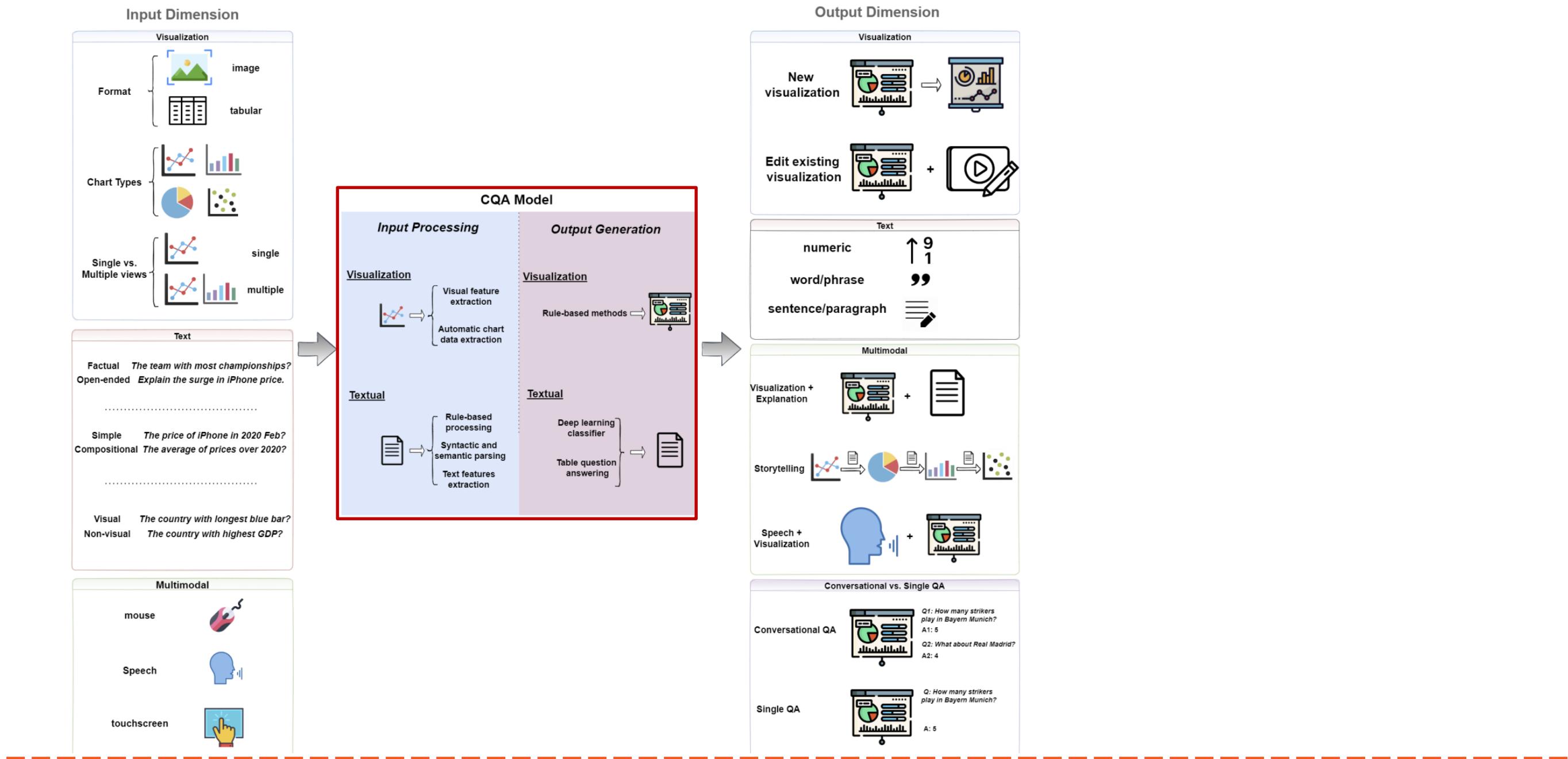
NLP4Vis Applications

- Visual text analytics
- Natural language interfaces for visualizations
- Text generations for visualizations
- Automatic visual story generation
- Visualization retrieval and recommendation etc.

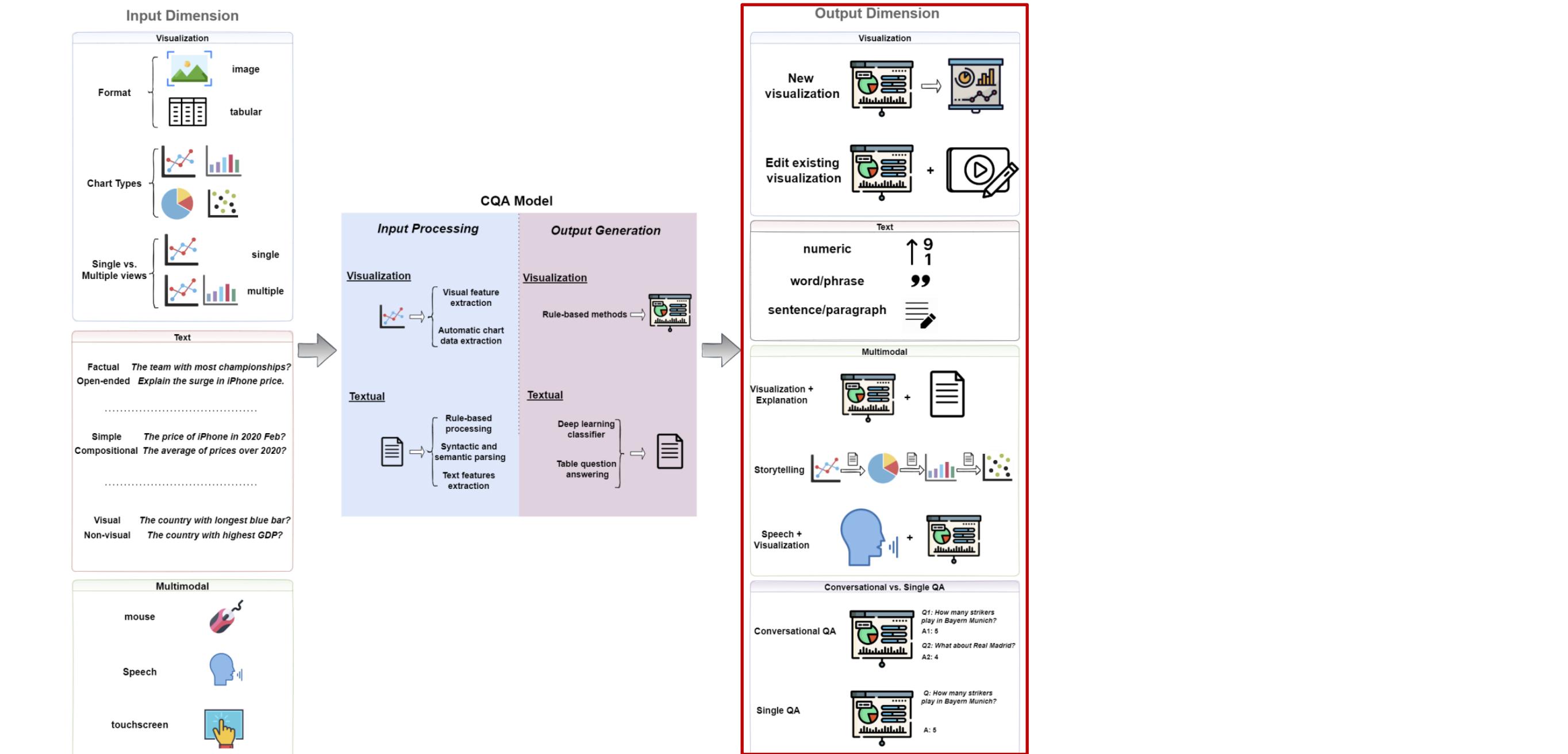
The Problem Space



The Problem Space



The Problem Space



NLP4Vis

- Natural language as *input*
- Natural language as *output*

NLP4Vis

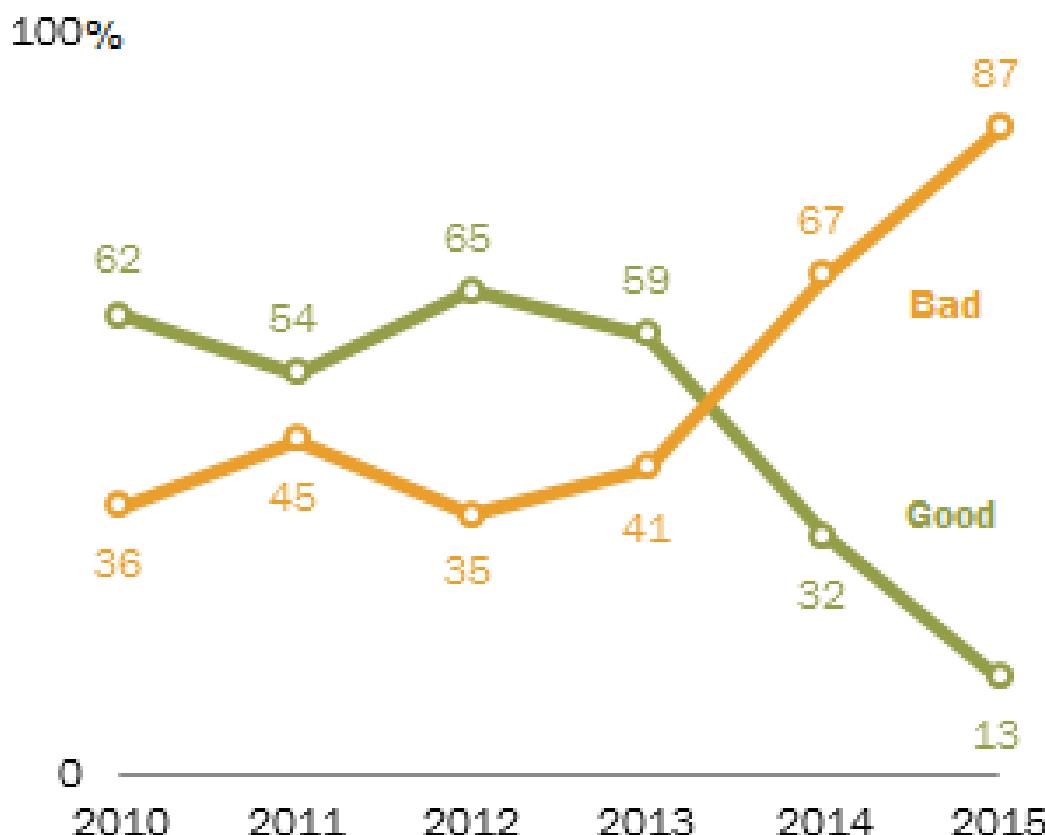
- **Natural language as *input***
- Natural language as *output*

Natural Language as Input

- Factoid Chart Question Answering: ChartQA (Masry et al., 2022)

Rapid Decline in Brazilians' Assessment of Economy

Current economic situation in Brazil is ...



Question: Which year has the most divergent opinions about Brazil's economy?

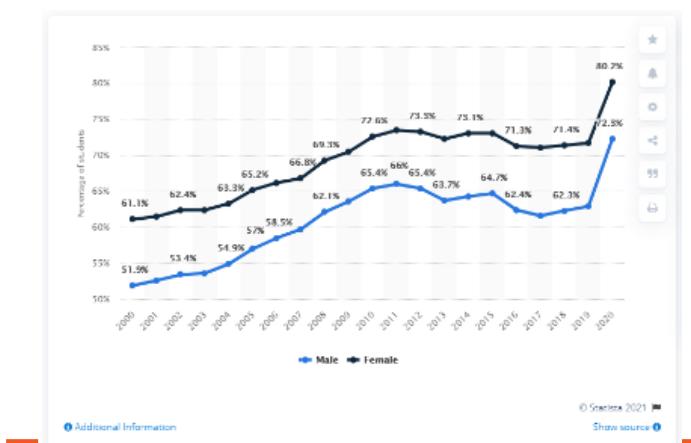
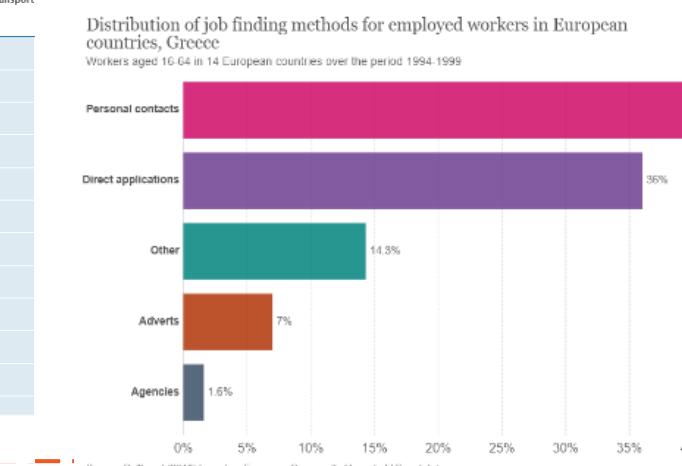
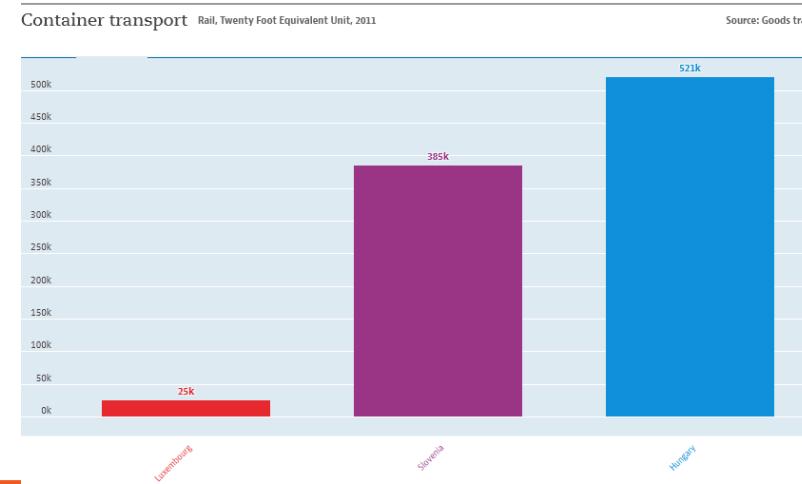
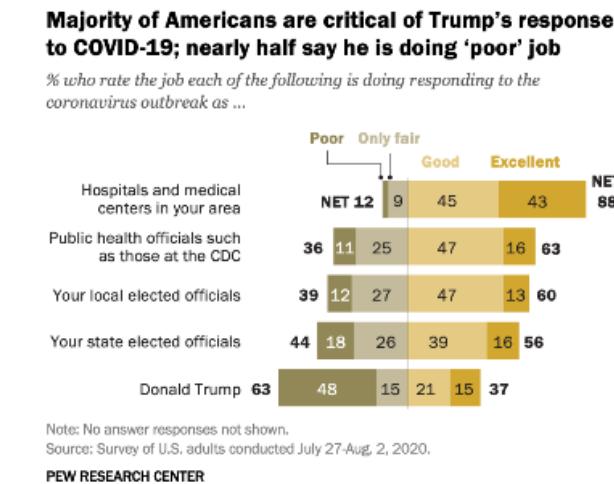
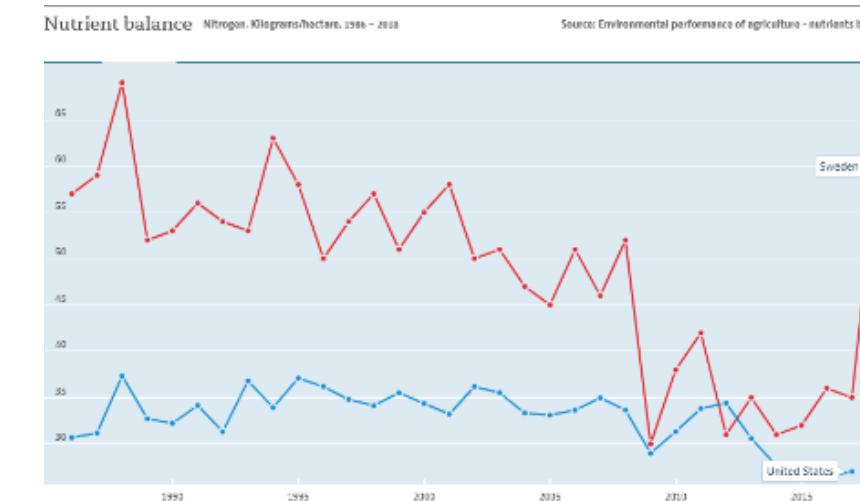
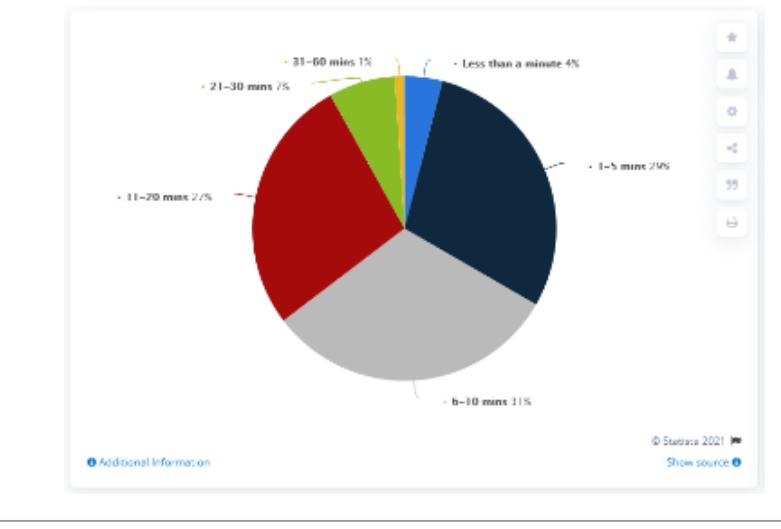
Answer: 2015

Question: What is the peak of the orange line?

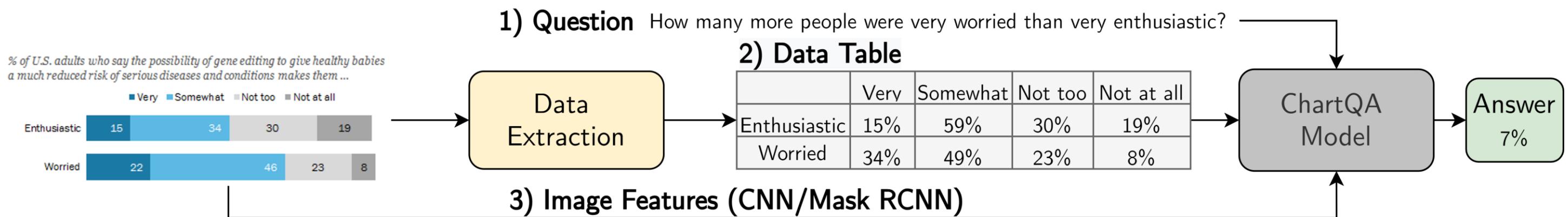
Answer: 87

ChartQA Dataset

- Real-world charts crawled from various online sources
- **9.6k** human-authored and **23.1K** Machine-generated question



ChartQA Apptoach



13

CharQA Evaluation

Models	FigureQA				DVQA (ORACLE / OCR)		PlotQA		ChartQA	
	Val1	Val2	Test1	Test2	Test-Familiar	Test-Novel	Test V1	Test V2	Val	Test
Gold Data Table Provided										
TaPas	98.10%	98.09%	-	-	53.40%	53.40%	21.56%	19.55%	49.16%	51.80%
VisionTaPas	97.59%	97.96%	-	-	99.36%	99.37%	80.18%	58.29%	59.32%	61.84%
T5	95.75%	95.75%	-	-	94.33%	81.42%	93.24%	85.99%	59.11%	59.80%
VL-T5	96.45%	96.43%	-	-	98.90%	80.18%	96.38%	84.70%	58.80%	59.12%
Gold Data Table Not Provided										
TaPas	90.32%	90.43%	89.52%	89.57%	50.28% / 48.82%	50.24% / 48.68%	15.09%	12.90%	39.68%	41.28%
VisionTaPas	91.46%	91.45%	90.68%	90.64%	95.38% / 94.43%	95.46% / 94.54%	65.30%	42.50%	42.60%	45.52%
T5	87.97%	87.83%	87.56%	87.57%	90.20% / 89.01%	77.97% / 76.89%	72.62%	56.22%	40.15%	41.04%
VL-T5	88.60%	88.49%	88.20%	88.18%	94.80% / 93.75%	77.04% / 76.14%	75.90%	56.02%	38.43%	41.56%
PReFIL	94.84%	93.26%	94.88%	93.16%	96.37% / 80.88%	96.53% / 80.04%	-	-	4.53%	4.8%
PlotQA*	-	-	-	-	— / 57.99%	— / 59.54%	53.96%	22.52%	36.15%	38.00%
STL-CQA	-	-	-	-	97.35% / —	97.51% / —	-	-	-	-

- VisionTaPas achieves state-of-the-art performance.
- Lower accuracies in our dataset compared to previous datasets
 - Mainly due to the human-written visual and logical reasoning questions

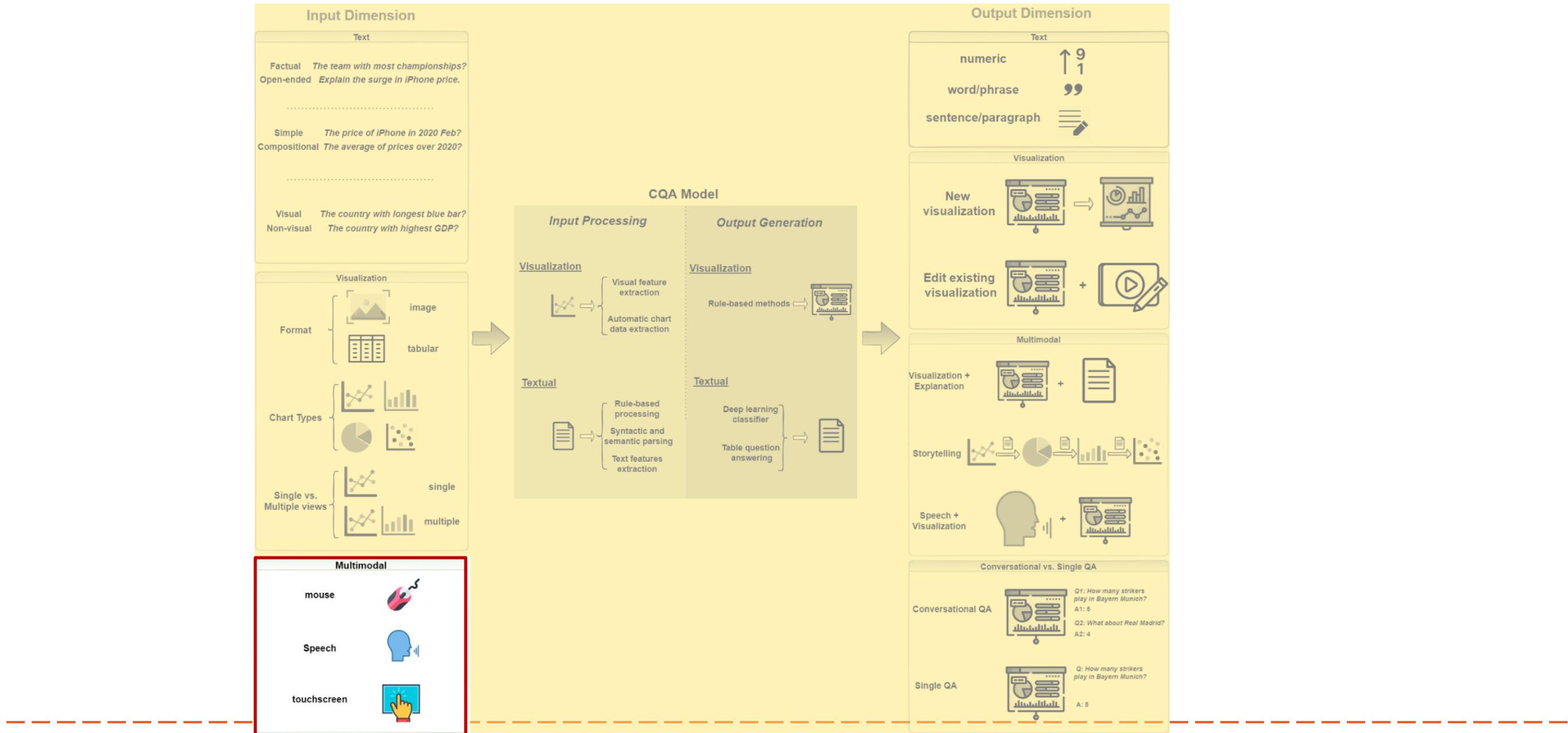
OpenAI's Study of GPT-4 on ChartQA Benchmark



Research ▾ Product ▾ Developers ▾ Safety Company ▾

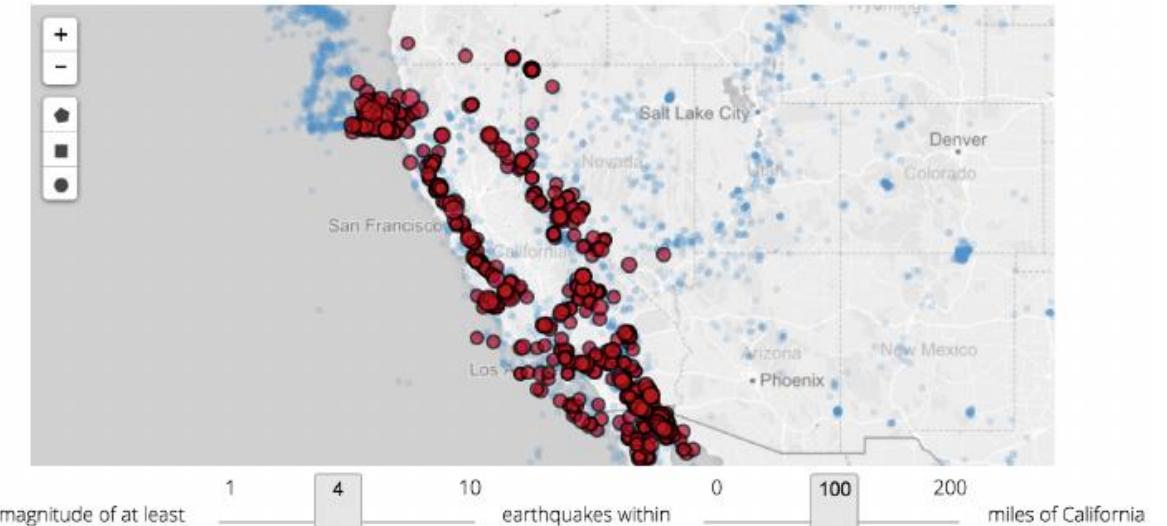
Benchmark	GPT-4 Evaluated few-shot	Few-shot SOTA	SOTA Best external model (includes benchmark-specific training)
VQAv2 VQA score (test-dev)	77.2% 0-shot	67.6% <u>Flamingo 32-shot</u>	84.3% <u>PaLI-17B</u>
TextVQA VQA score (val)	78.0% 0-shot	37.9% <u>Flamingo 32-shot</u>	71.8% <u>PaLI-17B</u>
ChartQA Relaxed accuracy (test)	78.5% ^A	-	58.6% <u>Pix2Struct Large</u>
AI2 Diagram (AI2D) Accuracy (test)	78.2% 0-shot	-	42.1% <u>Pix2Struct Large</u>
DocVQA ANLS score (test)	88.4% 0-shot (pixel-only)	-	88.4% <u>ERNIE-Layout 2.0</u>
Infographic VQA ANLS score (test)	75.1% 0-shot (pixel-only)	-	61.2% <u>Applica.ai TILT</u>
TVQA Accuracy (val)	87.3% 0-shot	-	86.5% <u>MERLOT Reserve Large</u>
LSMDC Fill-in-the-blank accuracy (test)	45.7% 0-shot	31.0% <u>MERLOT Reserve 0-shot</u>	52.9% <u>MERLOT</u>

The Problem Space

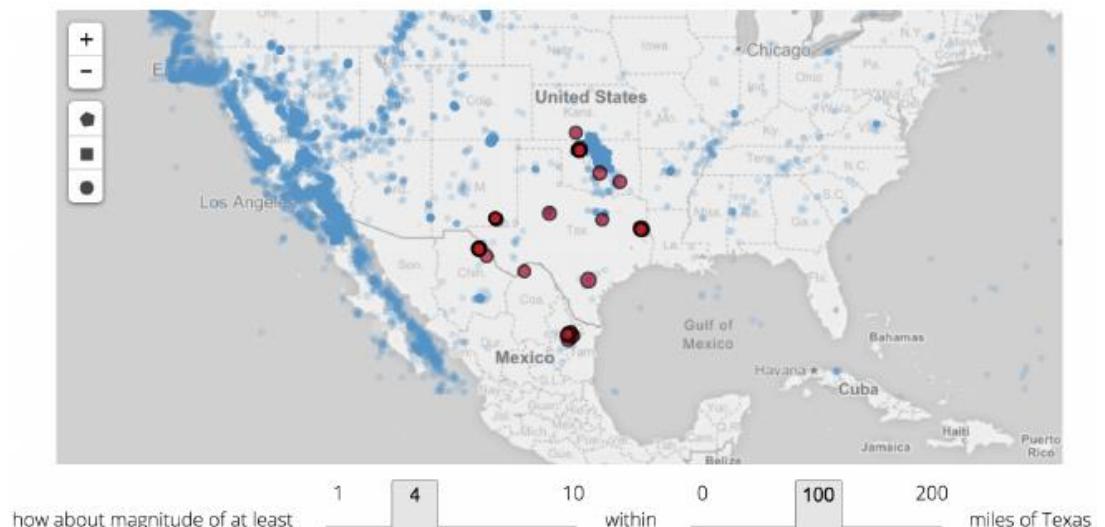


Multimodal Inputs for Visualizations

- Examples of Ambiguity Widgets
 - Eviza (Setlur et al., 2016)
- Allows users to rectify queries



(a) Previous query: “Large earthquakes near California”



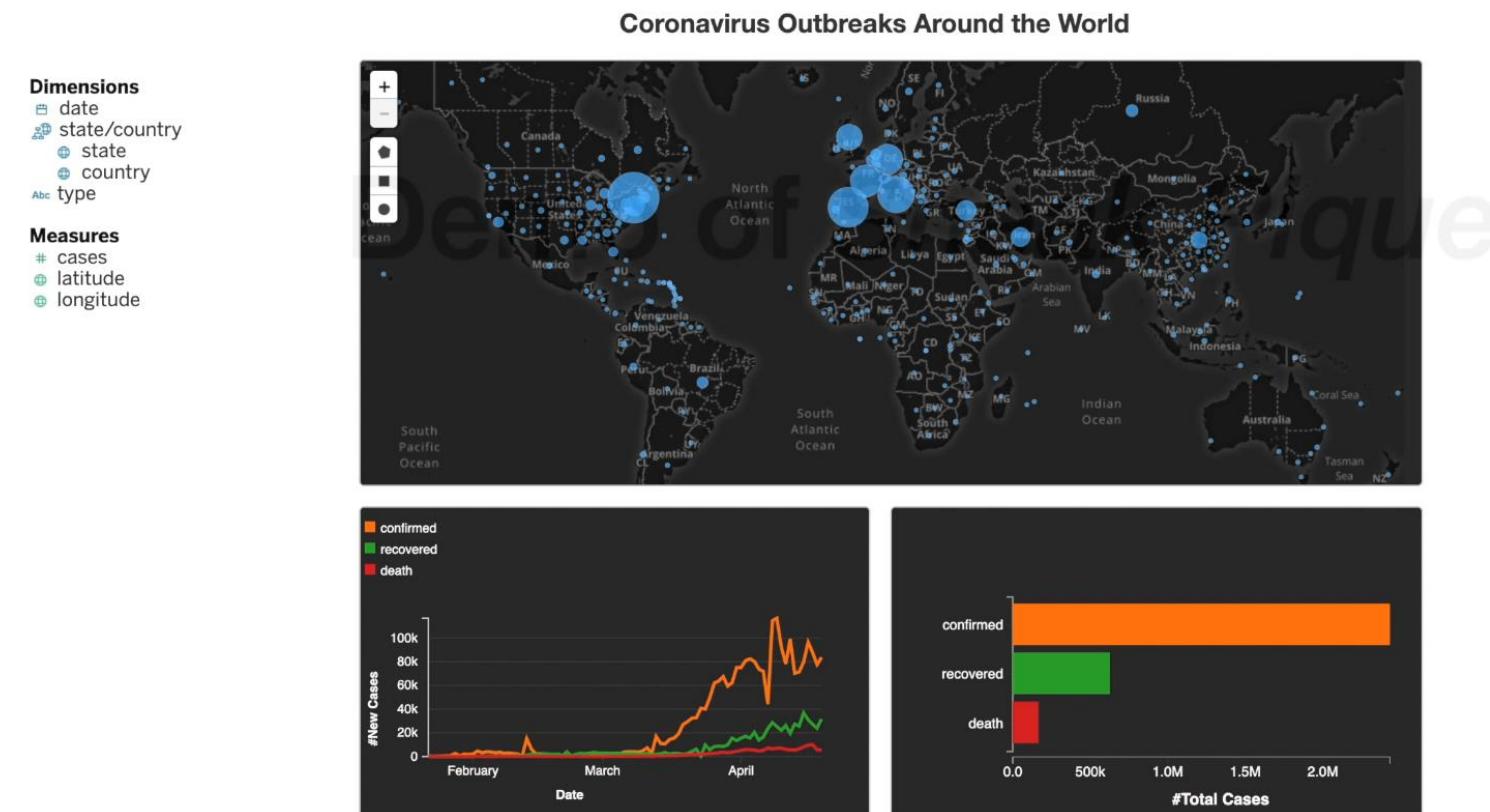
(b) Subsequent query: “how about near Texas?”

Multimodal Inputs for Visualizations

- Examples of query completion through text and interactive vis.

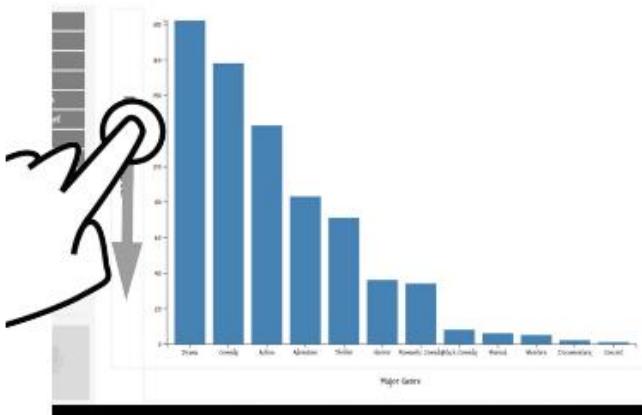
Sneak Pique (Setlur et al., 2020) 

type "Show coronavirus cases in Italy"

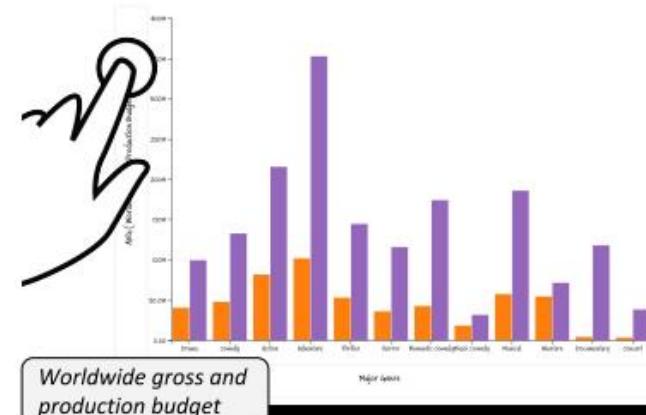


Multimodal Inputs for Visualizations

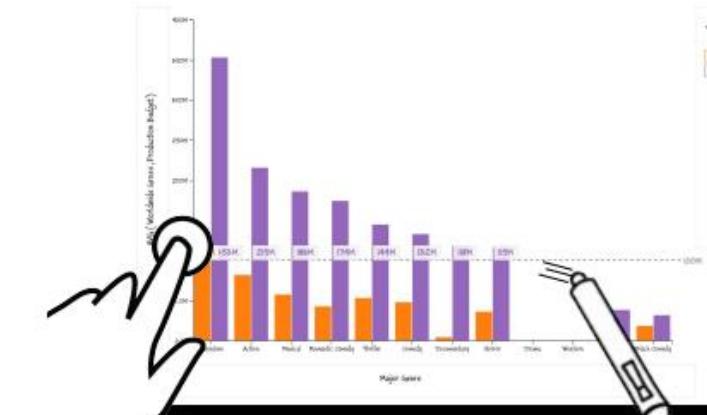
- combining speech, touch and pen : InChorus (Srinivasan et al., 2020)



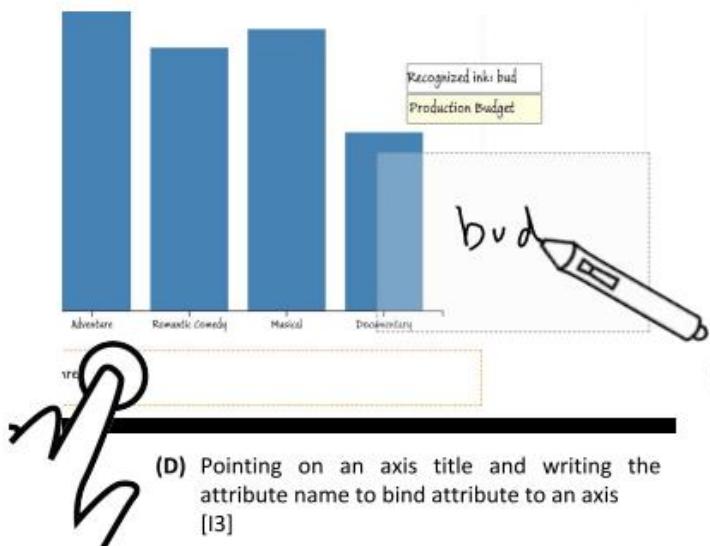
(A) Swiping along an axis to sort [I17]



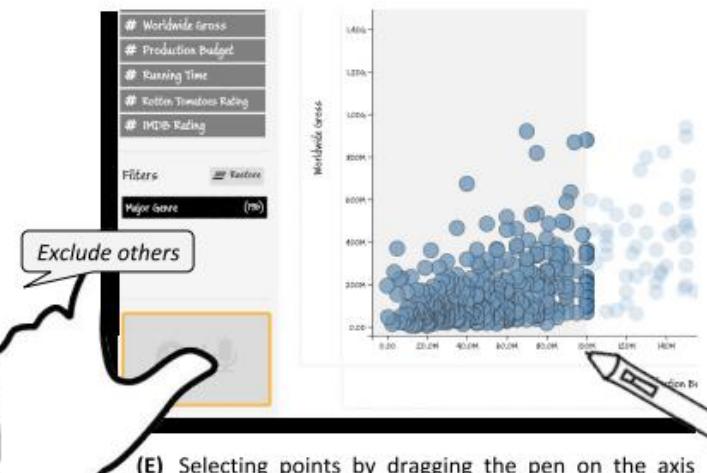
(B) Pointing on an axis title and uttering attribute names to bind attributes to an axis [I6]



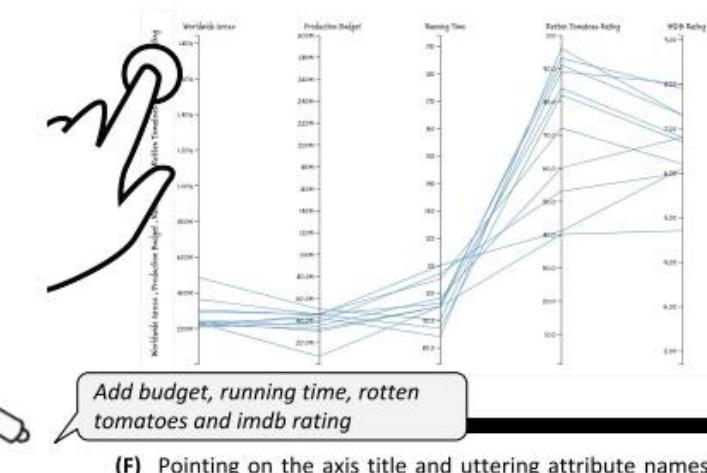
(C) Dragging a finger along the axis scale to see mark details based on value & erasing marks to filter [I25 + I20]



(D) Pointing on an axis title and writing the attribute name to bind attribute to an axis [I3]



(E) Selecting points by dragging the pen on the axis scale & deleting unselected points through speech [I21]



(F) Pointing on the axis title and uttering attribute names with the modifier "Add" to bind additional attributes to an axis [I7]

NLP4Vis

- Natural language as *input*
- **Natural language as *output***

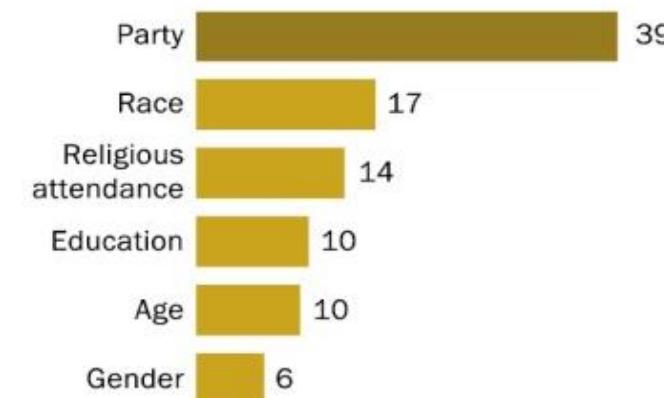
Chart-to-text

Dataset: 44.1K Chart-summary pairs Problem Setup #1

Chart:

Partisan gaps dwarf race, education, other differences in political values

Average percentage point gap across 30 political values items by ...



Notes: Indicates average gap between the share of two groups taking the same position across 30 values items. Party=difference between Rep/Lean Rep and Dem/Lean Dem. Race=white non-Hispanic vs. black non-Hispanic. Religious attendance=attend religious services weekly or more vs. attend less often. Education=college grad vs. non-college grad. Age=18-49 vs. 50+. Source: Survey of U.S. adults conducted Sept 3-15, 2019.

PEW RESEARCH CENTER

Table:

Demographic	Average Percentage Point Gap
Party	39
Race	17
Religious Attendance	14
Education	10
Age	10
Gender	6

Metadata:

- Title: Partisan gaps dwarf race, education, other differences in political values
- Chart type: Bar
- Topic: U.S. Politics & Policy

Gold Summary:

Across all 30 political values, the differences between Republicans and Democrats dwarf all other differences by demographics or other factors. The 39-point average gap is more than twice the difference between white and nonwhite adults (17 percentage points); people who regularly attend religious services and those who do not (14 points); college graduates and those who have not completed college (10 points); younger and older adults (also 10 points); and men and women (6 points).

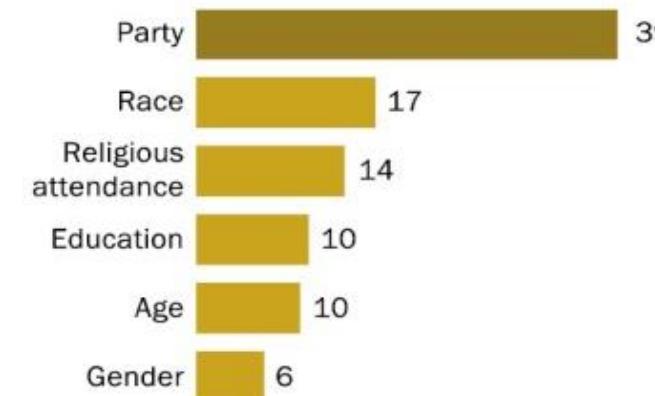
Chart-to-text

Problem Setup #2

Chart:

Partisan gaps dwarf race, education, other differences in political values

Average percentage point gap across 30 political values items by ...



Notes: Indicates average gap between the share of two groups taking the same position across 30 values items. Party=difference between Rep/Lean Rep and Dem/Lean Dem. Race=white non-Hispanic vs. black non-Hispanic. Religious attendance=attend religious services weekly or more vs. attend less often. Education=college grad vs. non-college grad. Age=18-49 vs. 50+. Source: Survey of U.S. adults conducted Sept 3-15, 2019.

PEW RESEARCH CENTER

Table:

A table showing the average percentage point gap for different demographic factors. A large red 'X' is overlaid on the table. The data is summarized in the following table:

Demographic	Average Percentage Point Gap
Party	39
Race	-
Religious Attendance	-
Education	10
Age	10
Gender	6

Metadata:

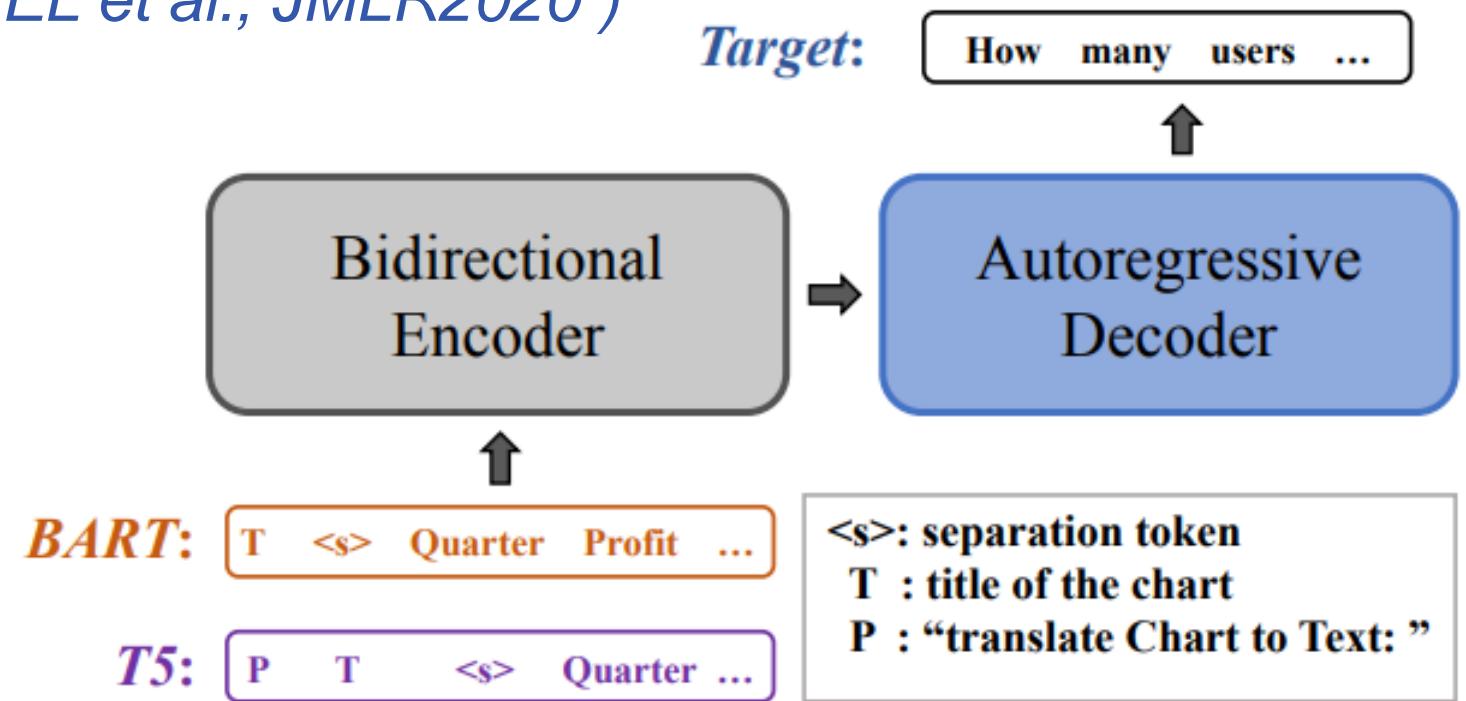
- Title: Partisan gaps dwarf race, education, other differences in political values
- Chart type: Bar
- Topic: U.S. Politics & Policy

Gold Summary:

Across all 30 political values, the differences between Republicans and Democrats dwarf all other differences by demographics or other factors. The 39-point average gap is more than twice the difference between white and nonwhite adults (17 percentage points); people who regularly attend religious services and those who do not (14 points); college graduates and those who have not completed college (10 points); younger and older adults (also 10 points); and men and women (6 points).

Chart-to-text Example Models

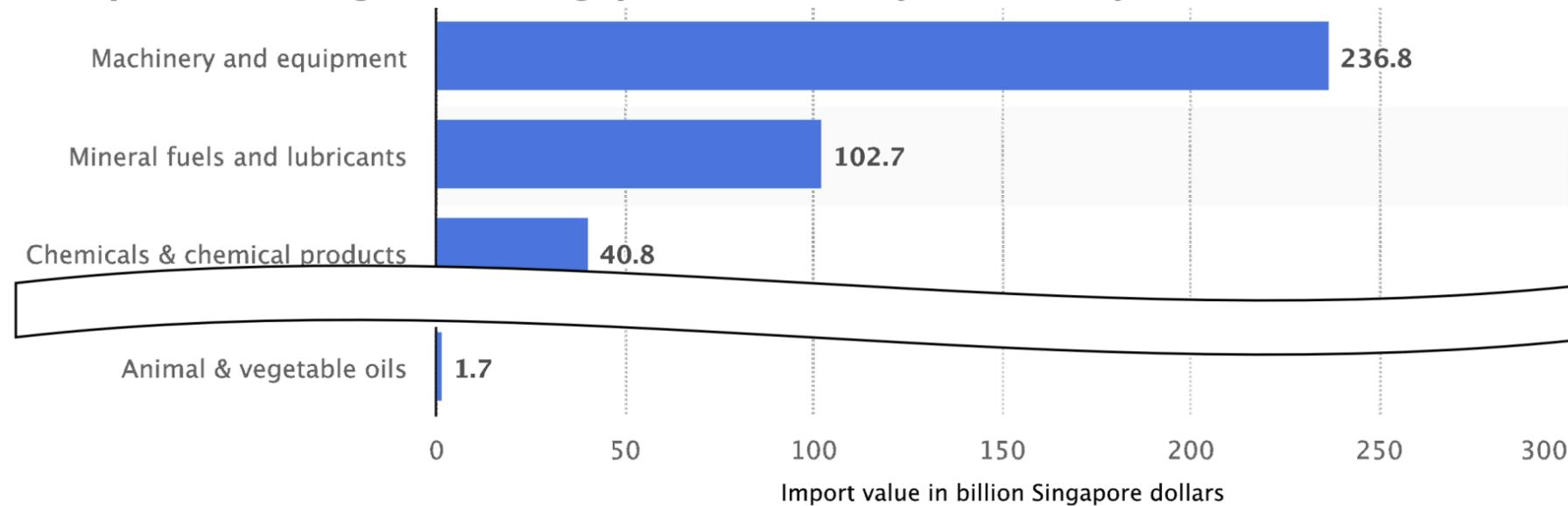
BART/T5 (*LEWIS et al., ACL2020, RAFFEL et al., JMLR2020*)



- Full fine-tuning BART/T5 on our datasets
- Setup 1: Linearizes the table as the input
- Setup 2: Send OCR text from the chart image as the input
- Prefix to T5: “translate Chart to Text: ”

Chart-to-text Sample Output

Import value of goods to Singapore for 2019, by commodity

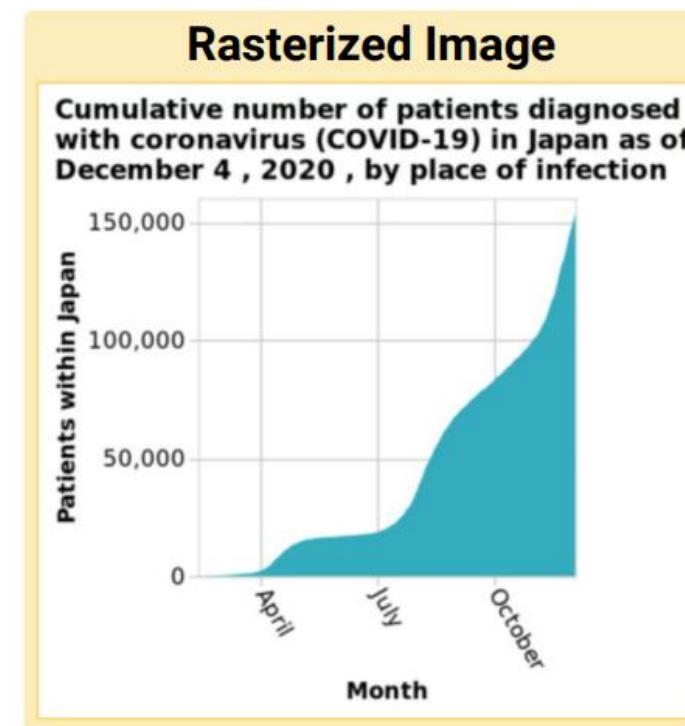


Gold: In 2019, Singapore imported approximately 236.8 billion Singapore dollars worth of machinery and equipment, making it the country's largest import commodity by value. This was followed by the import of mineral fuels and lubricants, valued at 102.7 billion Singapore dollars.

TAB-T5: Machinery and equipment was the most valuable commodity for Singapore in 2019, with an import value of 236.8 billion Singapore dollars. Mineral fuels and lubricants were the second most valuable commodity for Singapore, with an import value of 102.7 billion Singapore dollars.

VisText

- 12.4K Charts with Generated+crowdsourced caption
- Scene graph a hierarchical representation of a chart's visual elements



Data Table

Cumulative number of patients diagnosed with coronavirus (COVID-19) in Japan as of December 4, 2020, by place of infection

Month	Patients within Japan
Feb 11, 2020	16
...	...

Scene Graph

```
{title: "Cumulative number ...", x: -76, y: -50,},  
axes: [{x-axis: "Month", x: 100, y: 55.6,},  
       {y-axis: "Patients within Japan", x: ...},  
       {x-tick: [{x: 33, val: "April"}, ...]}],  
marks: [...],  
...}
```

Generated L1 Caption

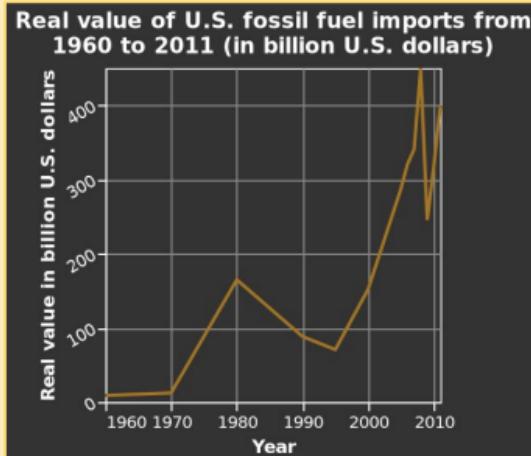
Here is a area chart is labeled Cumulative number of patients diagnosed with coronavirus (COVID-19) in Japan as of December 4, 2020, by place of infection. On the x-axis, Month is measured with a categorical scale starting with April and ending with October. There is a linear scale with a minimum of 0 and a maximum of 150,000 along the y-axis, labeled Patients within Japan.

Crowdsourced L2/L3 Caption

By December 4th 2020, approximately 160,000 people in Japan had been diagnosed with COVID-19. The first person diagnosed with COVID-19 in Japan was diagnosed in March 2020. The greatest increase in cumulative number of patients in Japan diagnosed with COVID-19 occurred between November and December 2020.

VisText Sample Output

Rasterized Image



Scene Graph

```
title Real value of U.S. fossil fuel imports from  
1960 to 2011 (in billion U.S. dollars) x 100 y  
-33.84 x-axis x 100 y 21 Year y-axis x -29.088 y  
100 Real value in billion U.S. dollars xtick x 0  
val 1960 x 39 val 1970 x 78 val 1980 x 118 val  
1990 x 157 val 2000 x 196 val 2010 ytick y 200  
val 0 y 156 val 100 y 111 val 200 y 67 val 300 y  
22 val 400 marks line XY 0 196.04 desc XY 39.221  
194.516 desc XY 78.43 126.778 desc XY 117.651  
160.747 desc XY 137.256 168.604 desc XY 156.861  
132.467 desc XY 176 ...
```

Model Generated L1 Caption

Here a line graph is called Real value of U.S. fossil fuel imports from 1960 to 2011 (in billion U.S. dollars). The y-axis measures Real value in billion U.S. dollars while the x-axis measures Year.

Model Generated L2/L3 Caption

The real value of US fossil fuel imports has increased from 1960 to 2011. The real value of US fossil fuel imports has increased from 1960 to 2011. The real value of US fossil fuel imports has increased from 1960 to 2011.

- Correctly identifies upward trends.
- But Repeats this claim twice

NLP4Vis

- Natural language as *input*
- Natural language as *output*

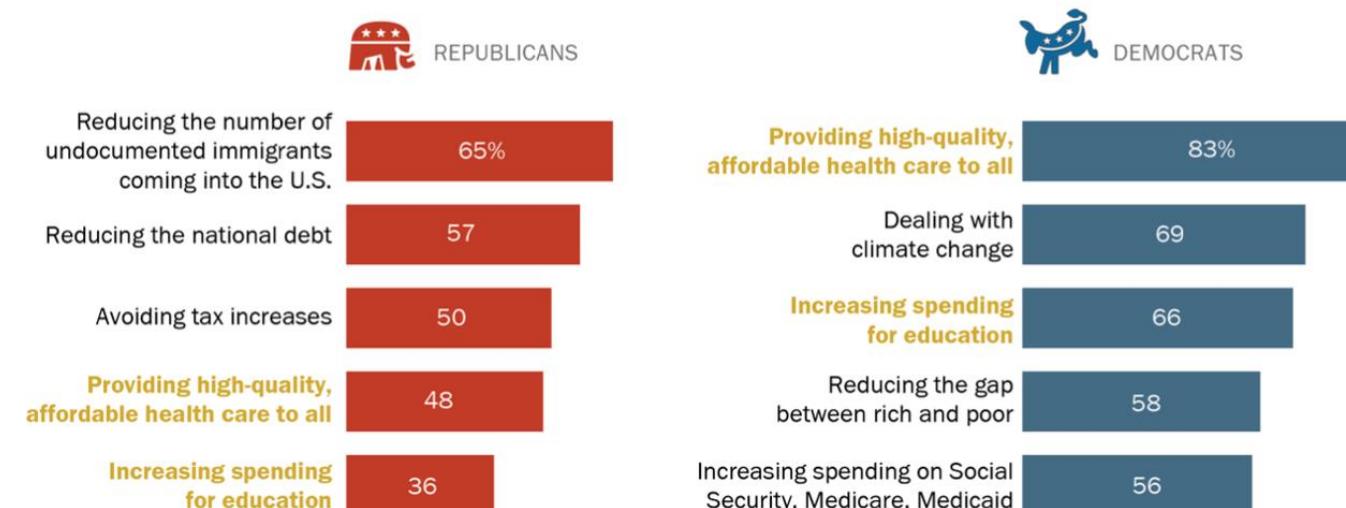
Open-ended Question Answering with Charts

Question

Compare the Democrats and Republicans views about providing health care to the population?

Republicans and Democrats have different ideas about what government should do to improve the lives of future generations of Americans

% of **Republicans/Democrats** saying each of the following should be a **top priority** in order for the federal government to improve the quality of life for future generations



Notes: These categories represent the top five responses (based on the share saying each is a top priority within each party group).

Source: Survey of U.S. adults conducted Dec. 11-23, 2018.

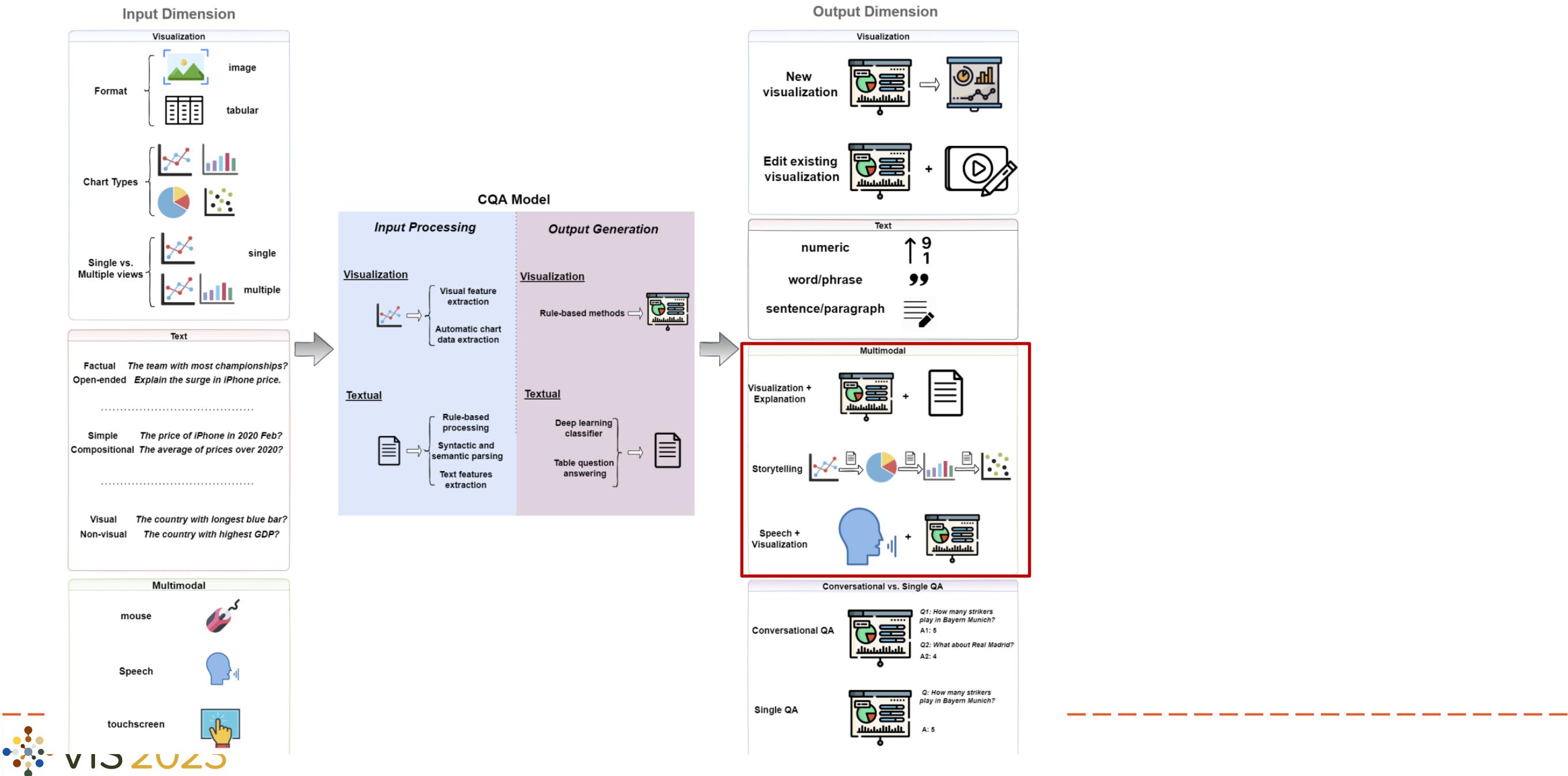
"Looking to the Future, Public Sees an America in Decline on Many Fronts"

PEW RESEARCH CENTER

Answer

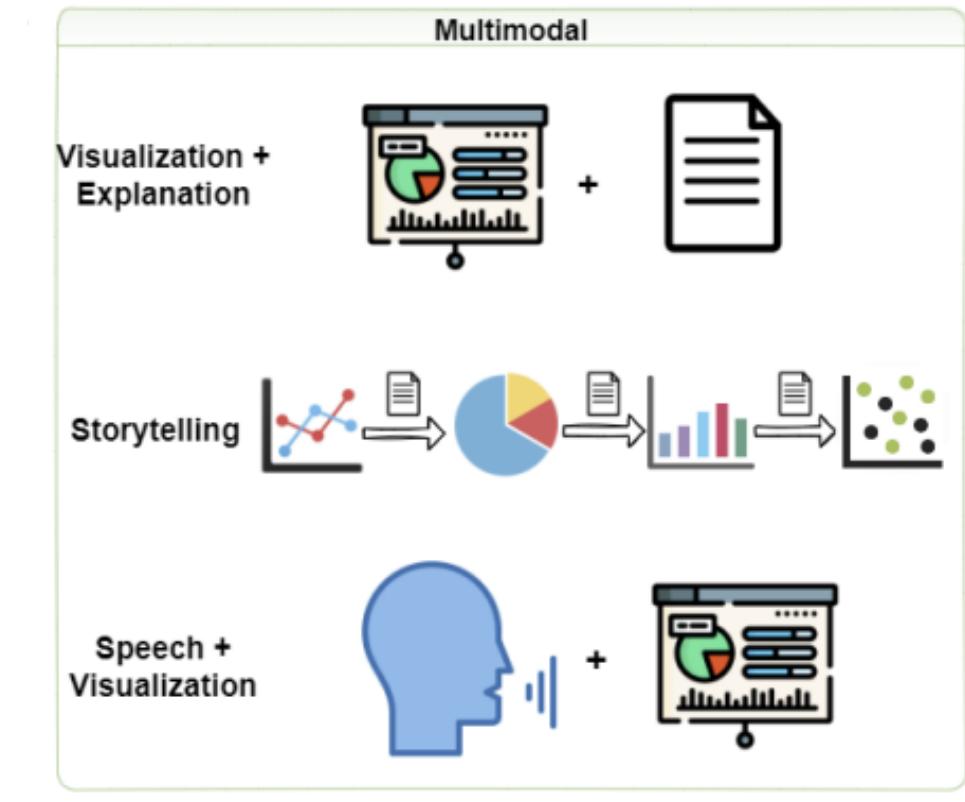
While 83% of Democrats say providing high-quality, affordable health care for all should be a top priority, a much smaller share of Republicans (48%) agree .

The Problem Space



Combining Language and Visualizations as Output

- Roles of natural language
 - generating explanatory answer
 - explaining the answer



Combining Language and Visualizations as Output

- An example of combining text and vis as a multimodal output

Age 38.69

The value of the hikers' average age is 38.69.

The average Age

A bar chart titled "The distribution of COUNT" showing the count of hikers across different age ranges. The x-axis categories are 31-40, 41-60, 20-30, over 60, and under 20. The y-axis represents the count from 0 to 40. The bars show approximately 40 for 31-40, 32 for 41-60, 25 for 20-30, 3 for over 60, and 1 for under 20.

age_amount	count
31-40	40
41-60	32
20-30	25
over 60	3
under 20	1

A donut chart titled "The proportion of Female" showing the percentage of female hikers. The chart is divided into two segments: blue (56 percent) and yellow (44 percent). A callout box indicates "Female 56".

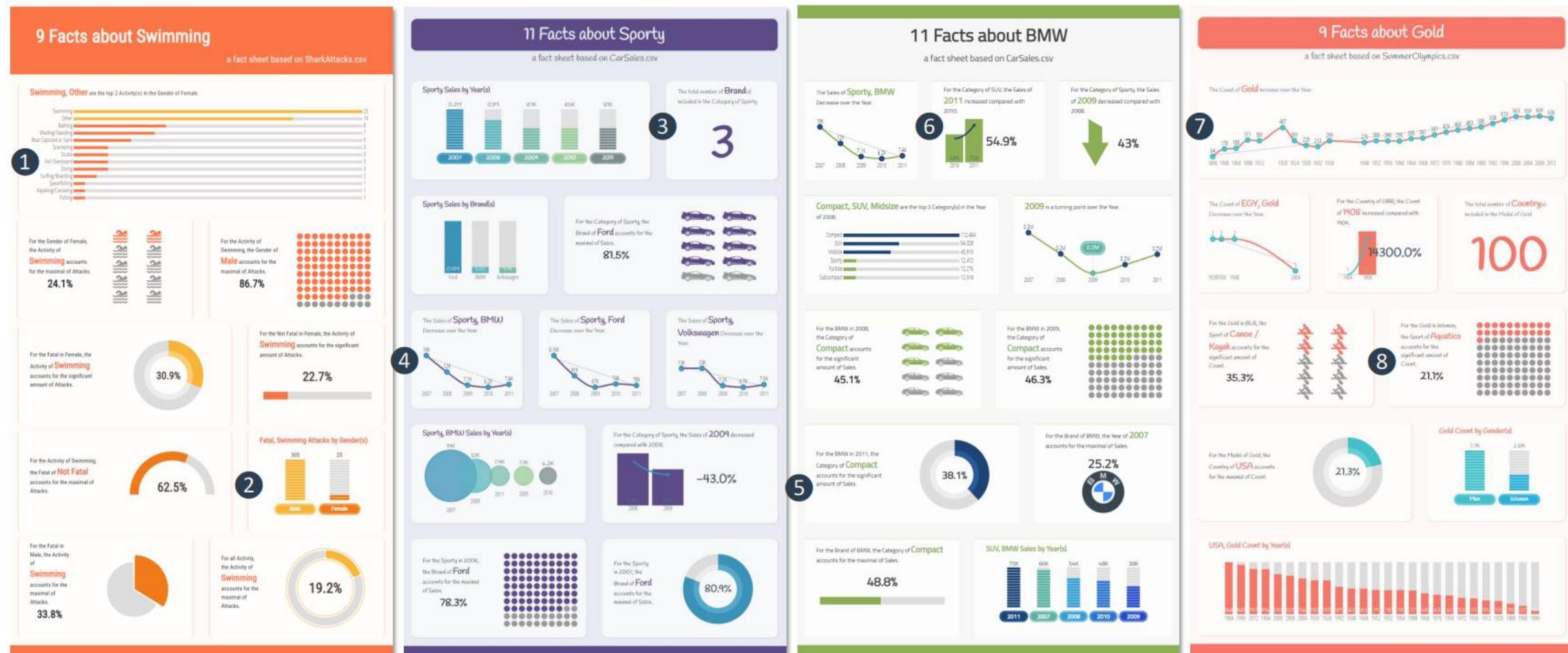
Gender	Percentage
Female	56%
Male	44%

A dot matrix visualization titled "The proportion of Female" where the density of dots represents the proportion of females. The matrix is mostly yellow with some grey dots at the bottom, corresponding to the 56% female proportion shown in the donut chart.

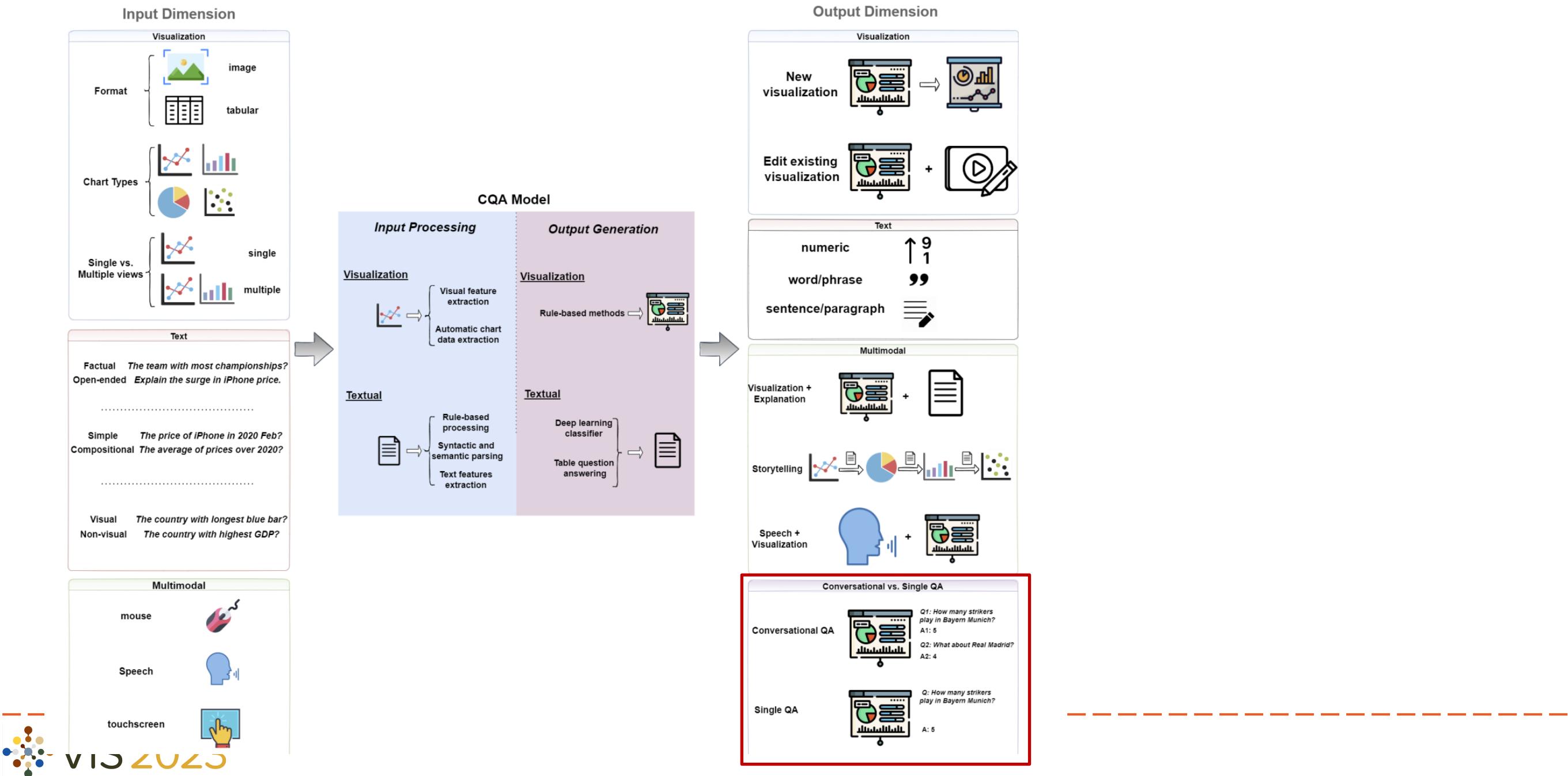
● value ● distribution ● proportion ● rank

Combining Language and Visualizations as Output

- DataShot (Yun et al., 2019):

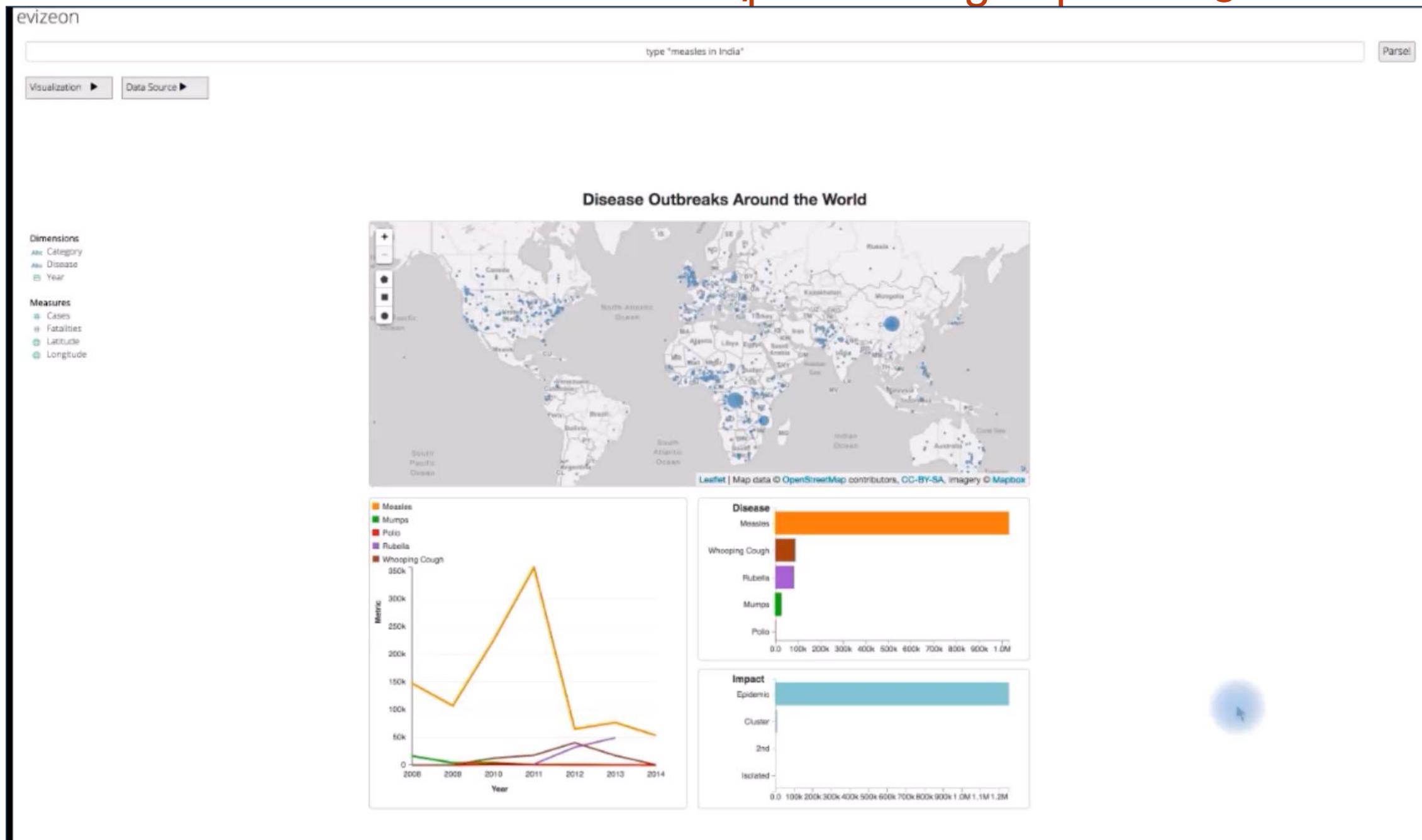


The Problem Space



Conversational QA with visualization

Evizeon (Hoque et al., TVCG 2017) “Measuring Disease Outbreaks with a WikiCongo”



Applications of LLMs for Visualizations

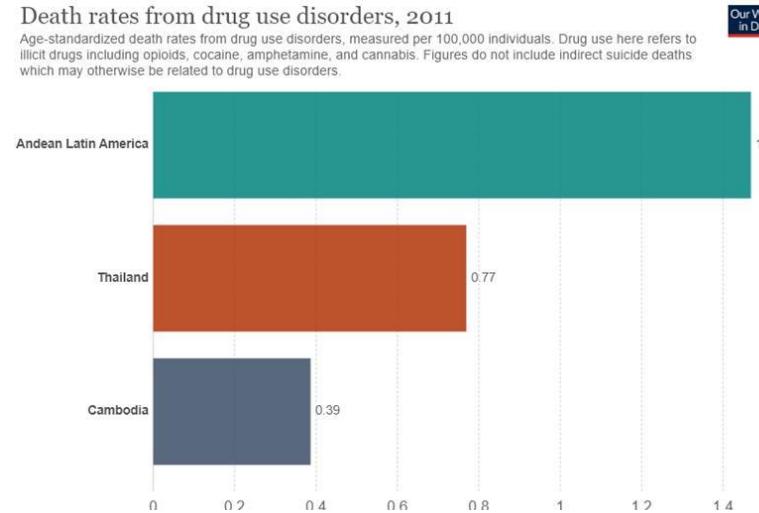
- **Background**
 - Introduction to NLP
 - Language modeling
- **Model architectures**
 - Transformer architecture
 - Encoder, decoder, encoder-decoder
 - Pre-training and fine-tuning
- **Large language models (LLMs)**
 - Scaling LMs to LLMs
 - Prompt engineering
 - In context learning
 - Instruction tuning

Do LLMs Work for Charts?

Can we achieve good performance without fine-tuning the model?

- Does **chain-of-thought** visual Reasoning help?
- Do **few-shot prompts** work for charts?

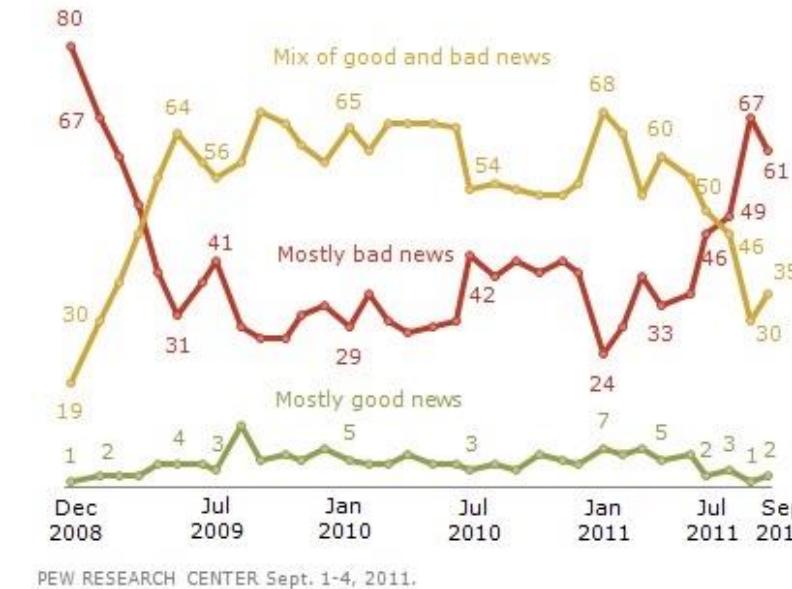
Does Chain-of-thought Visual Reasoning help?



Question: Is the average value of Andean Latin America and Cambodia more than the value of Thailand?

CCR: "The value of Andean Latin America is 1.47 and the value of Cambodia is 0.77. So the average value of Andean Latin America and Cambodia is $(1.47 + 0.77) / 2 = 1.12$. The value of Thailand is 0.39. Since $1.12 > 0.39$, the average value of Andean Latin America and Cambodia is more than the value of Thailand. The answer is Yes"

Recent Economic News is...



Question: What's the total sum of peak points of all three lines?

The peak point of red line is 80, the peak point of yellow line is 68, the peak point of green line is 7. The total sum of peak points is $80 + 68 + 7 = 155$. The answer is 155.

In-context Learning for Charts

- Do few-shot prompts work for charts?
 - Identify small number of demonstrations for each chart-related tasks.
 - Carefully analyze the benchmark to choose **diverse** sample demonstrations
 - Experiment with different few-shot prompting setups to verify the necessity and effectiveness of each demonstration.
- Evaluation:
 - Few-shot prompts with LLMs are as effective (or better) than fine-tuning.

Applications of LLMs for Visualizations

- **Background**
 - Introduction to NLP
 - Language modeling
- **Model architectures**
 - Transformer architecture
 - Encoder, decoder, encoder-decoder
 - Pre-training and fine-tuning
- **Large language models (LLMs)**
 - Scaling LMs to LLMs
 - Prompt engineering
 - In context learning
 - Instruction tuning

Vision-Language Pretraining Approach for Charts

- Existing pretrained models
 - Not optimal for chart-specific tasks
 - Trained on large text/image corpus without specific focus on charts.
 - Charts differ from natural images
 - Visually communicate the data using graphical marks (e.g., bars, lines) and text (e.g., titles, labels, legends)

Chart Pretraining Corpus

What we need:

- A very large corpus of diverse real-world charts
- Associated texts with rich semantics

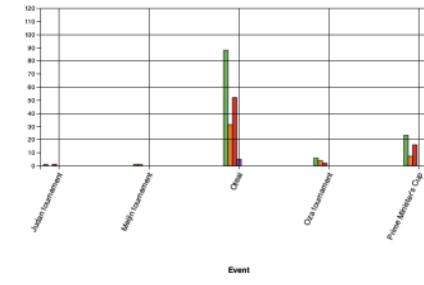
Solution?

- Data augmentation

Data Augmentation

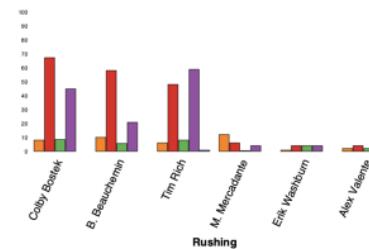
- Chart generation from Data Table
 - Web Data Commons
- Chart Generation:
 - Using D3 and Vega-Lite to generate charts
- Diversifying visual styles:
 - Colors and shapes
 - Position and distance
 - Guides
- **Corpus size:**
 - 612K charts (189K through augmentation)

GoBase.org - Kodama Kunio - Detailed Statistics

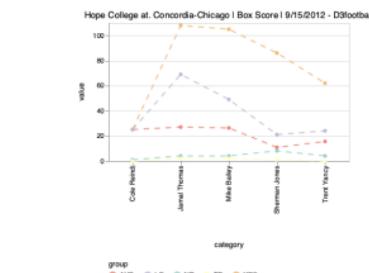


(a)

UMass Dartmouth at: Westfield State | Box Score | 9/21/2013 - D3football

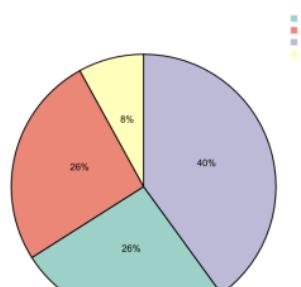


(c)

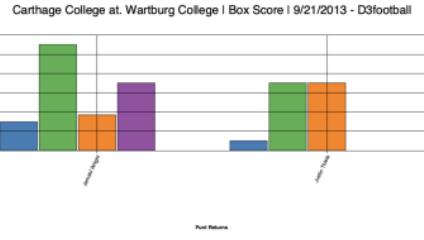


(e)

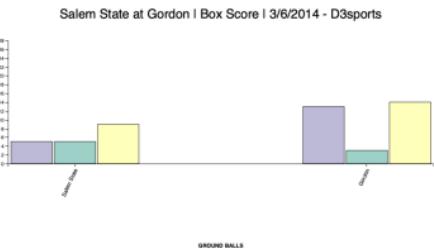
Authentic Collection Tech Fleece



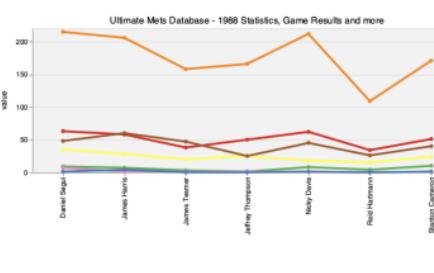
(g)



(b)

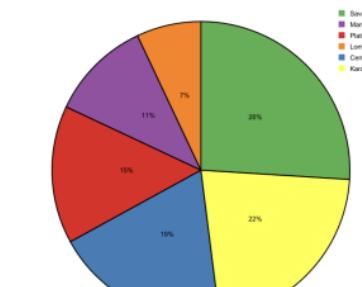


(d)



(f)

How to call Togo from Malawi | NobelCom

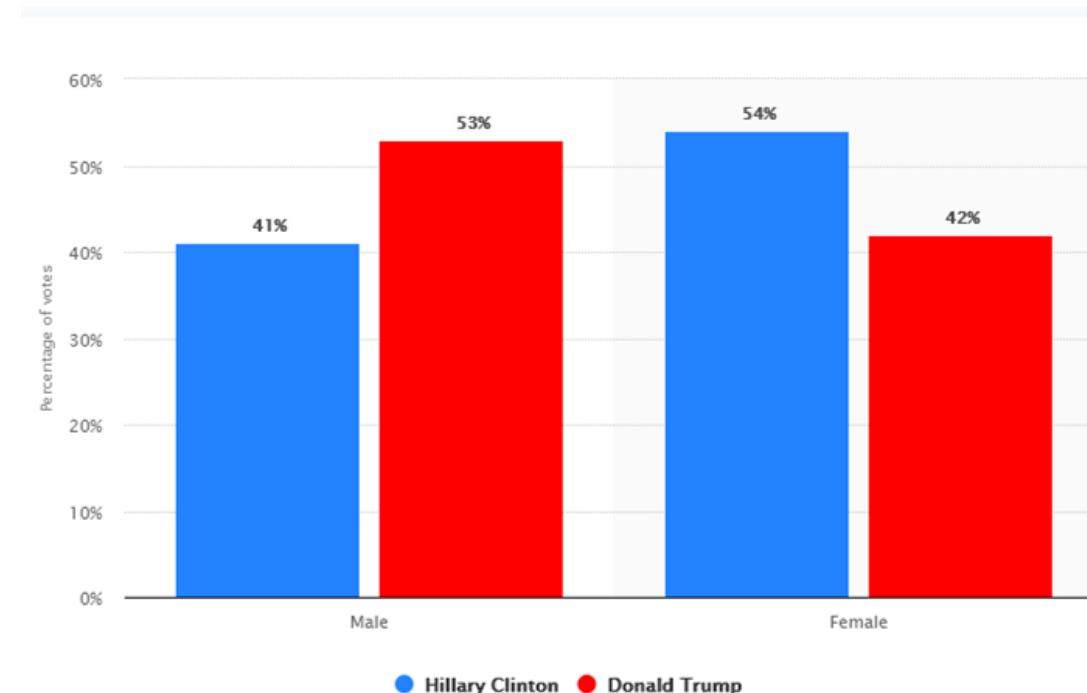


(h)

Text Augmentation by Knowledge Distillation

- Most charts in pretraining corpus have either no associated summaries or the summaries **lack quality** (e.g., too short and not very informative)

This graph shows the percentage of votes of the 2016 presidential elections in the United States on November 9, 2016, by gender. According to the exit polls, about 54 percent of female voters nationwide voted for Hillary Clinton.



- Human raters recently:

Summaries generated by **LLMs** >> Reference summaries in various text summarization datasets

Text Augmentation by Knowledge Distillation

- Approach:
 1. curate a small dataset (3700 charts)
 - Prompt **InstructGPT** the **underlying data table** as input and one exemplar (i.e., 1-shot in-context learning)
 2. finetune **Flan-T5 XL** on the small dataset
 3. Using **finetuned Flan-T5** model to generate summaries for all the charts with underlying data tables (**454K**)

Augmentation by Knowledge Distillation for Text Generation Tasks

Analyze the following bar chart in one paragraph and round the numbers. The numbers represent Rating of statistical capacity (0-100).

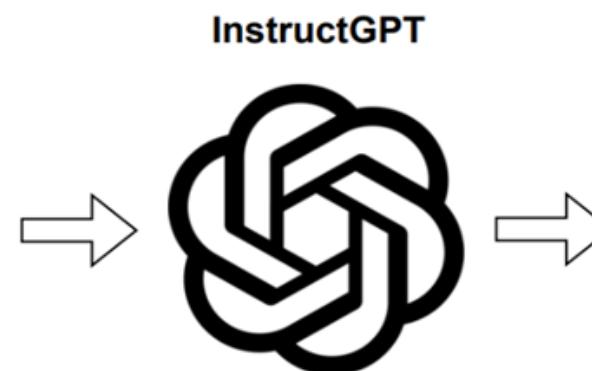
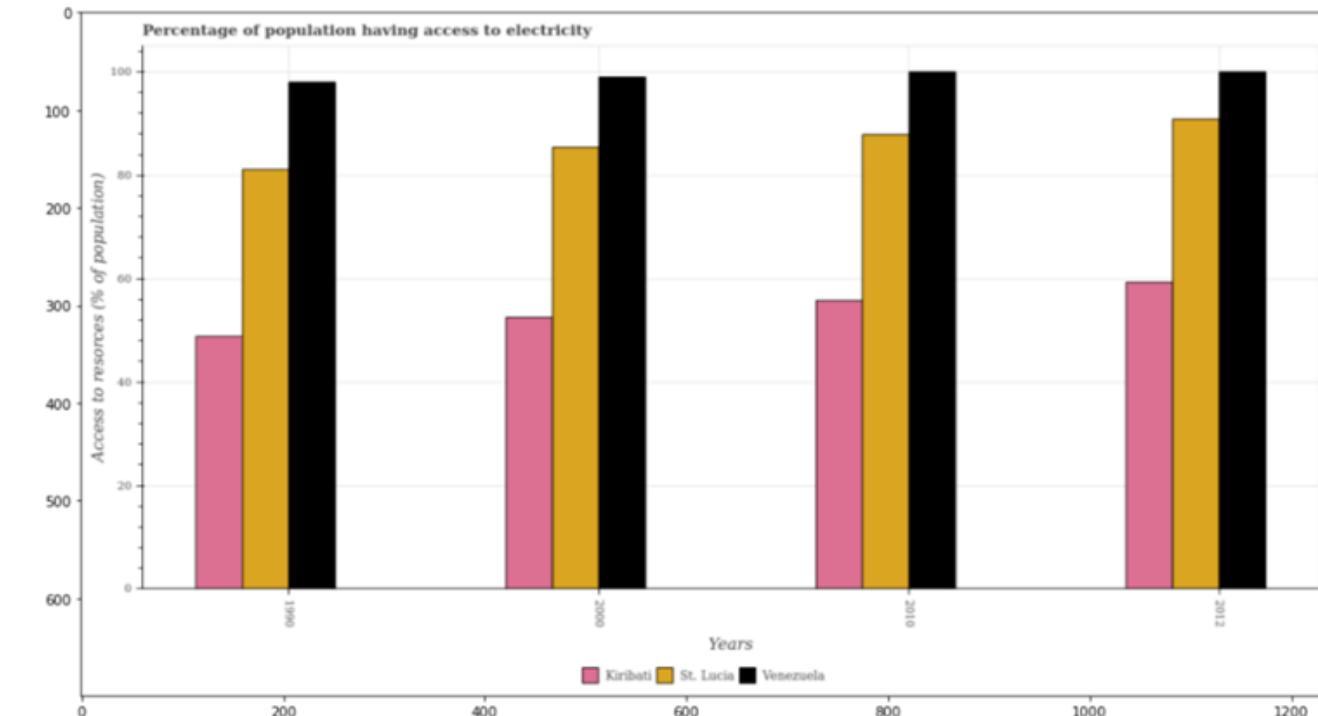
Chart: Years | Botswana | Samoa & 2010 | 30 | 50.0 & 2011 | 30 | 50.0 & 2012 | 40 | 60.0 & 2013 | 30 | 60.0

Summary: The bar chart shows the rating of statistical capacity for Botswana and Samoa from 2010 to 2013. In 2010, both countries had a statistical capacity rating of 30, but by 2012, Samoa's rating improved to 60 while Botswana's rating only increased to 40. In 2013, both countries had a rating of 60 for Samoa and 30 for Botswana. It can be concluded that Samoa has shown a better improvement in its statistical capacity compared to Botswana over the four-year period.

Analyze the following bar chart in one paragraph and round the numbers. The chart title is Percentage of population having access to electricity and the numbers represent Access to resources (% of population).

Chart: Years | Kiribati | St. Lucia | Venezuela & 1990.0 | 48.86 | 81.14 | 98.0 & 2000.0 | 52.53 | 85.41 | 99.0 & 2010.0 | 55.8 | 87.87 | 100.0 & 2012.0 | 59.33 | 90.88 | 100.0

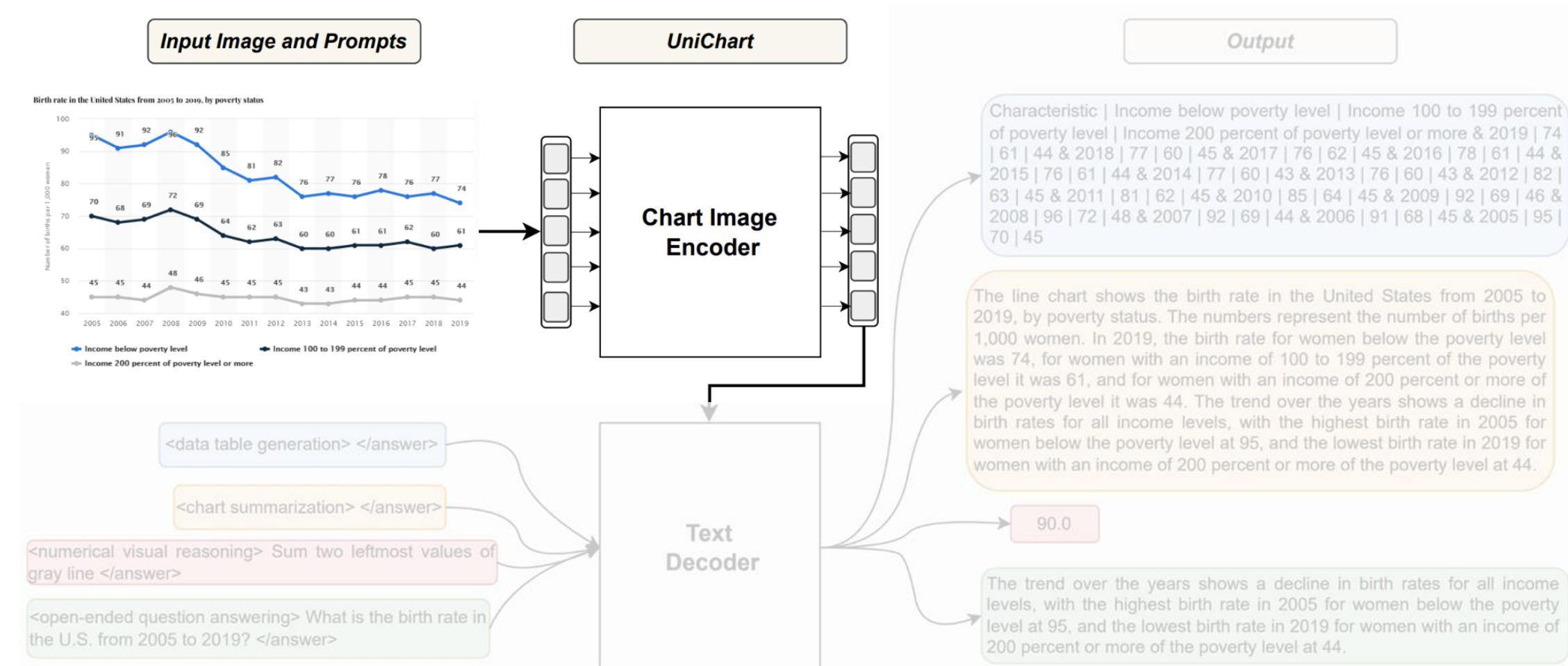
Summary:



The bar chart shows the percentage of population having access to electricity in Kiribati, St. Lucia, and Venezuela from 1990 to 2012. In 1990, 48.86%, 81.14%, and 98% of the population had access to electricity in Kiribati, St. Lucia, and Venezuela respectively, while in 2012, the percentages increased to 59.33%, 90.88%, and 100% respectively. It is evident that all three countries have experienced a significant increase in access to electricity over the 22-year period, with Venezuela having the highest access rate of 100% in 2012.

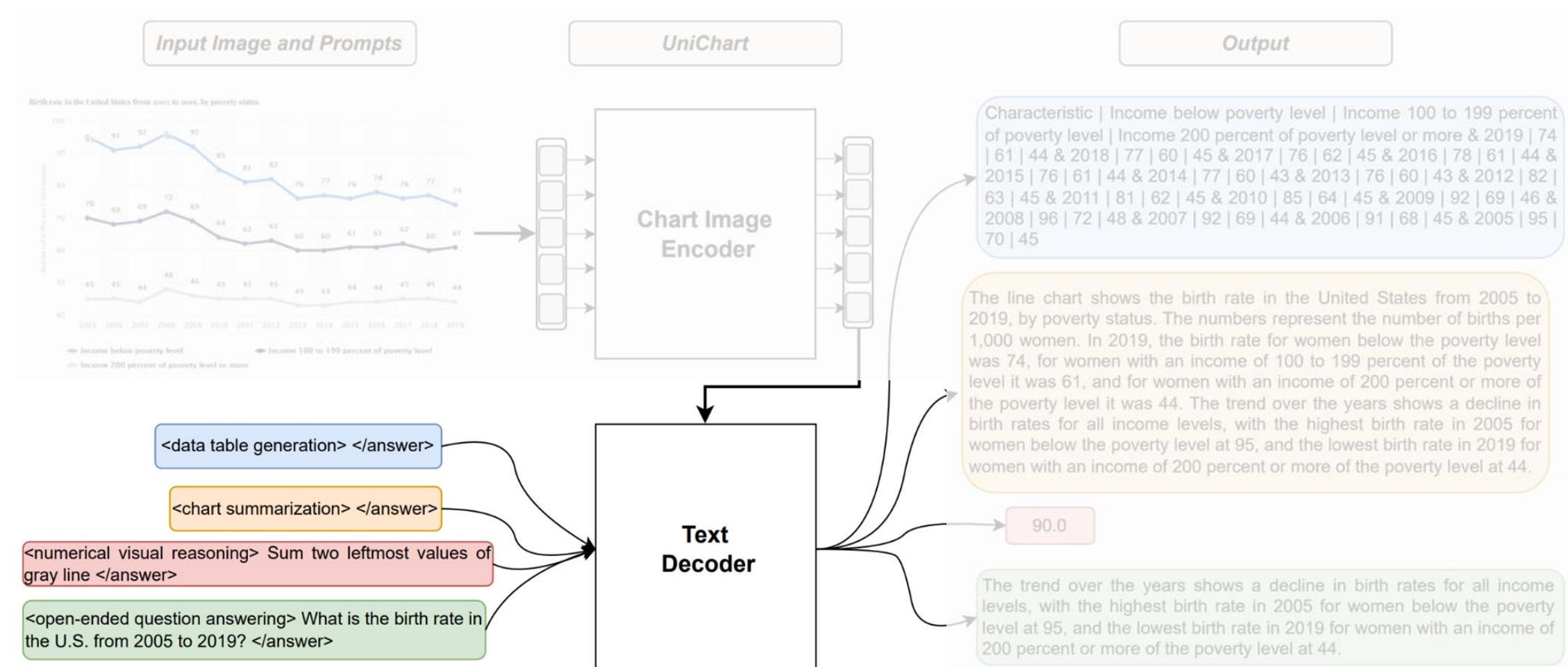
UniChart Model Architecture

- End-to-end **encoder-decoder** architecture
- Adopted **Donut**: a pretraining approach for document understanding
- Image Encoder
 - **Swin Transformer** architecture

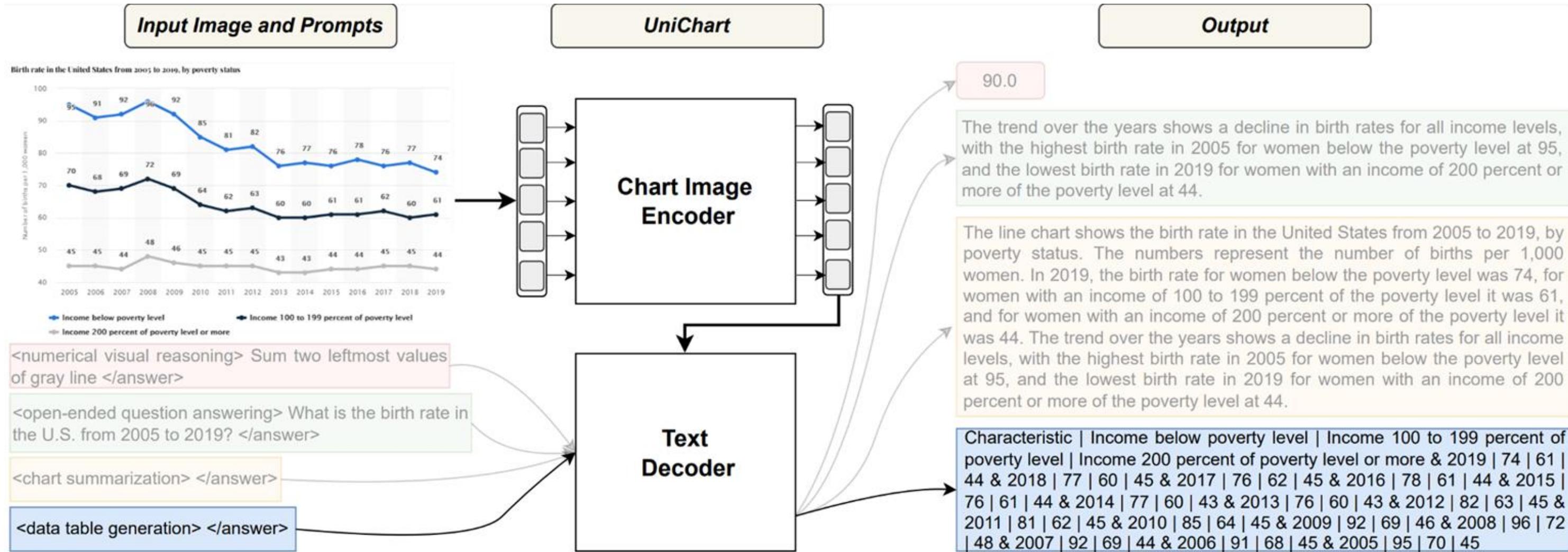


UniChart Model Architecture

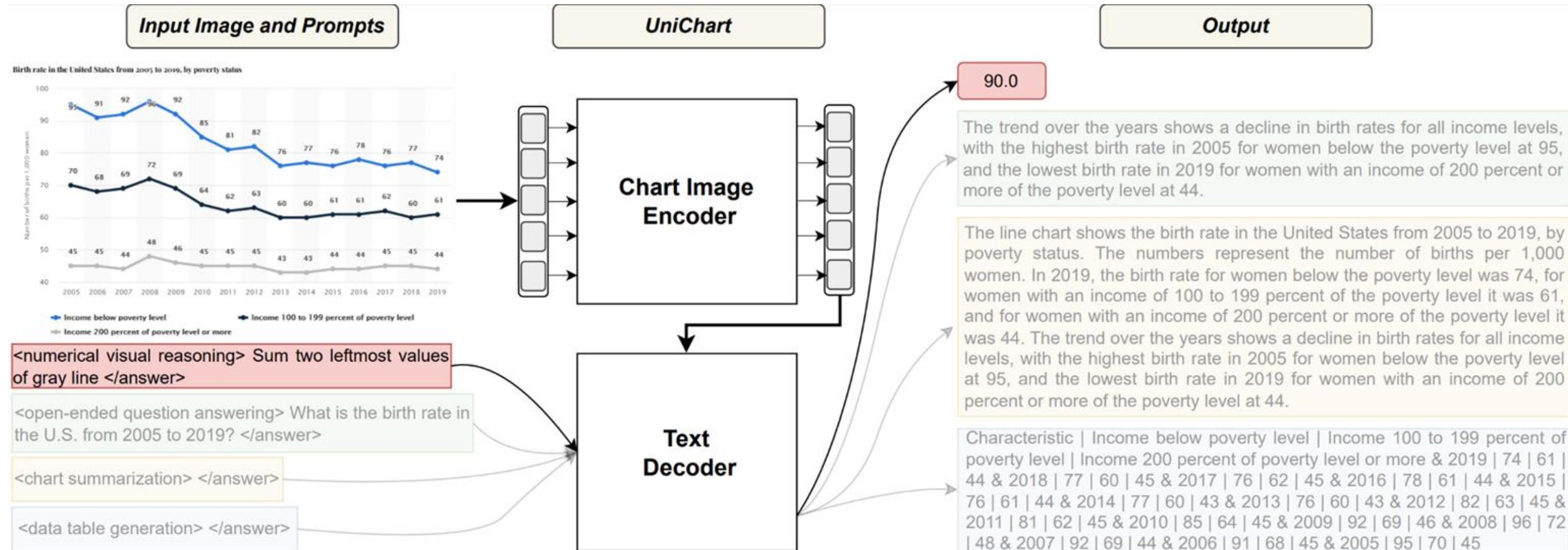
- Text Decoder
 - **BART** decoder (the same used in Donut)



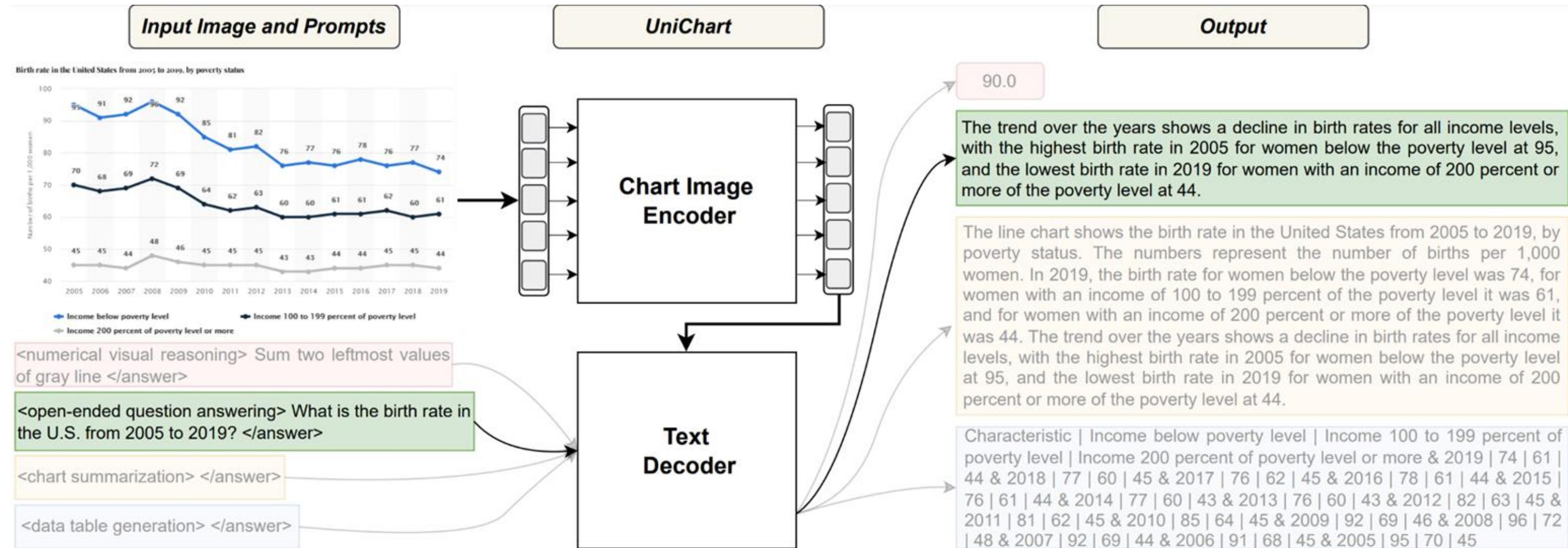
Pretraining Objectives



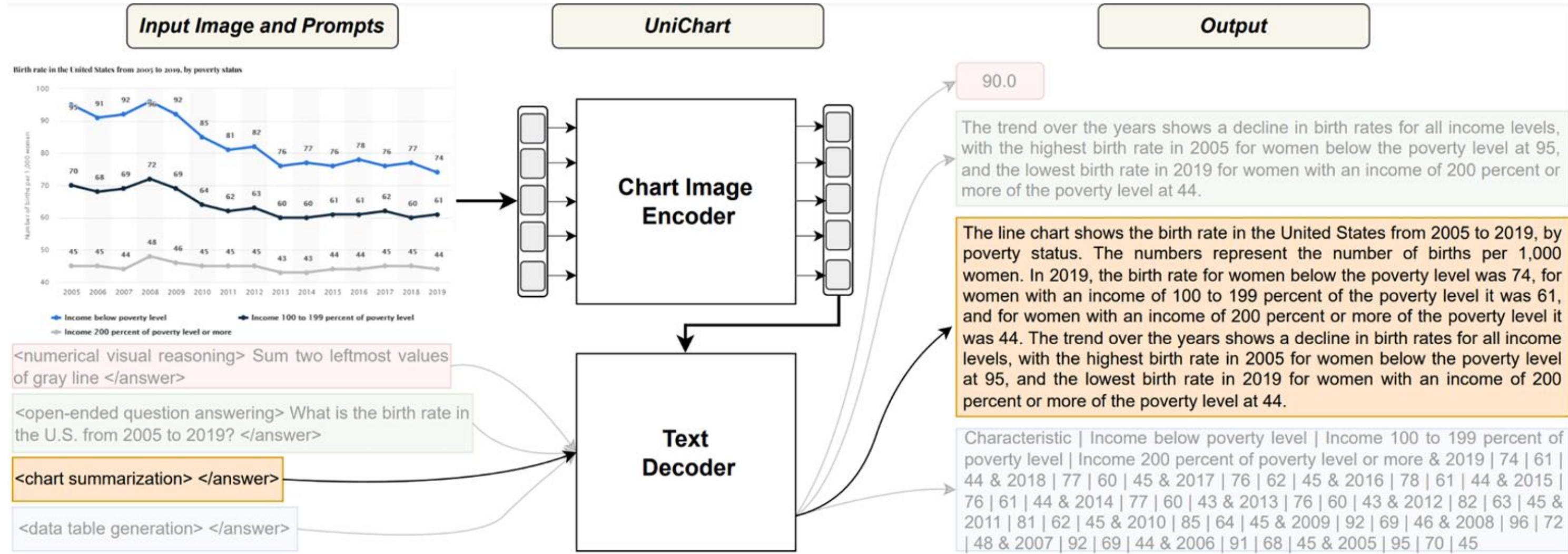
Pretraining Objectives



Pretraining Objectives



Pretraining Objectives



Main Results

Model	ChartQA (RA)			OpenCQA (BLEU)	Chart-to-Text (BLEU)		Chart-to-Table (RNSS RMS _{F1})	
	aug.	human	avg.		Pew	Statista	ChartQA	WebCharts
VisionTaPas [76]	61.44	29.60	45.52	-	-	-	-	-
T5 [76]	56.96	25.12	41.04	9.28	10.49	35.29	-	-
VL-T5 [76]	56.88	26.24	41.56	14.73	-	-	-	-
Pix2Struct [52]	81.6	30.5	56.0	-	10.3	38.0	-	-
MatCha [63]	90.2	38.2	64.2	-	12.2	39.4	85.21 83.49	44.37 17.94
UniChart	88.56	43.92	66.24	14.88	12.48	38.21	94.01 91.10	60.73 43.21

Result Analysis

- Noticeable performance gap between UniChart and the baselines reported by Masry et al., 2022.
- UniChart outperforms SoTA pretrained models, **Pix2Struct** and **MatCha** on **ChartQA**.

Main Results

Model	ChartQA (RA)			OpenCQA	Chart-to-Text (BLEU)		Chart-to-Table (RNSS RMS _{F1})	
	aug.	human	avg.	OpenCQA	Pew	Statista	ChartQA	WebCharts
VisionTaPas [76]	61.44	29.60	45.52	-	-	-	-	-
T5 [76]	56.96	25.12	41.04	9.28	10.49	35.29	-	-
VL-T5 [76]	56.88	26.24	41.56	14.73	-	-	-	-
Pix2Struct [52]	81.6	30.5	56.0	-	10.3	38.0	-	-
MatCha [63]	90.2	38.2	64.2	-	12.2	39.4	85.21 83.49	44.37 17.94
UniChart	88.56	43.92	66.24	14.88	12.48	38.21	94.01 91.10	60.73 43.21

Result Analysis

- UniChart outperforms SoTA pretrained models, **Pix2Struct** and **MatCha** on **ChartQA**.
- UniChart achieves a **higher BLUE score** compared to the SoTA VL-T5 model on **OpenCQA**.

Main Results

Model	ChartQA (RA)			OpenCQA (BLEU)	Chart-to-Text (BLEU)		Chart-to-Table (RNSS RMS _{F1})	
	aug.	human	avg.		Pew	Statista	ChartQA	WebCharts
VisionTaPas [76]	61.44	29.60	45.52	-	-	-	-	-
T5 [76]	56.96	25.12	41.04	9.28	10.49	35.29	-	-
VL-T5 [76]	56.88	26.24	41.56	14.73	-	-	-	-
Pix2Struct [52]	81.6	30.5	56.0	-	10.3	38.0	-	-
MatCha [63]	90.2	38.2	64.2	-	12.2	39.4	85.21 83.49	44.37 17.94
UniChart	88.56	43.92	66.24	14.88	12.48	38.21	94.01 91.10	60.73 43.21

Result Analysis

- UniChart outperforms SoTA pretrained models, **Pix2Struct and MatCha** on **ChartQA**.
- UniChart achieves a **higher BLUE score** compared to the SoTA VL-T5 model on **OpenCQA**.
- Better performance of UniChart on WebCharts demonstrates its **generalizability** across diverse visual styles, even in a **zero-shot setup on unseen charts**.

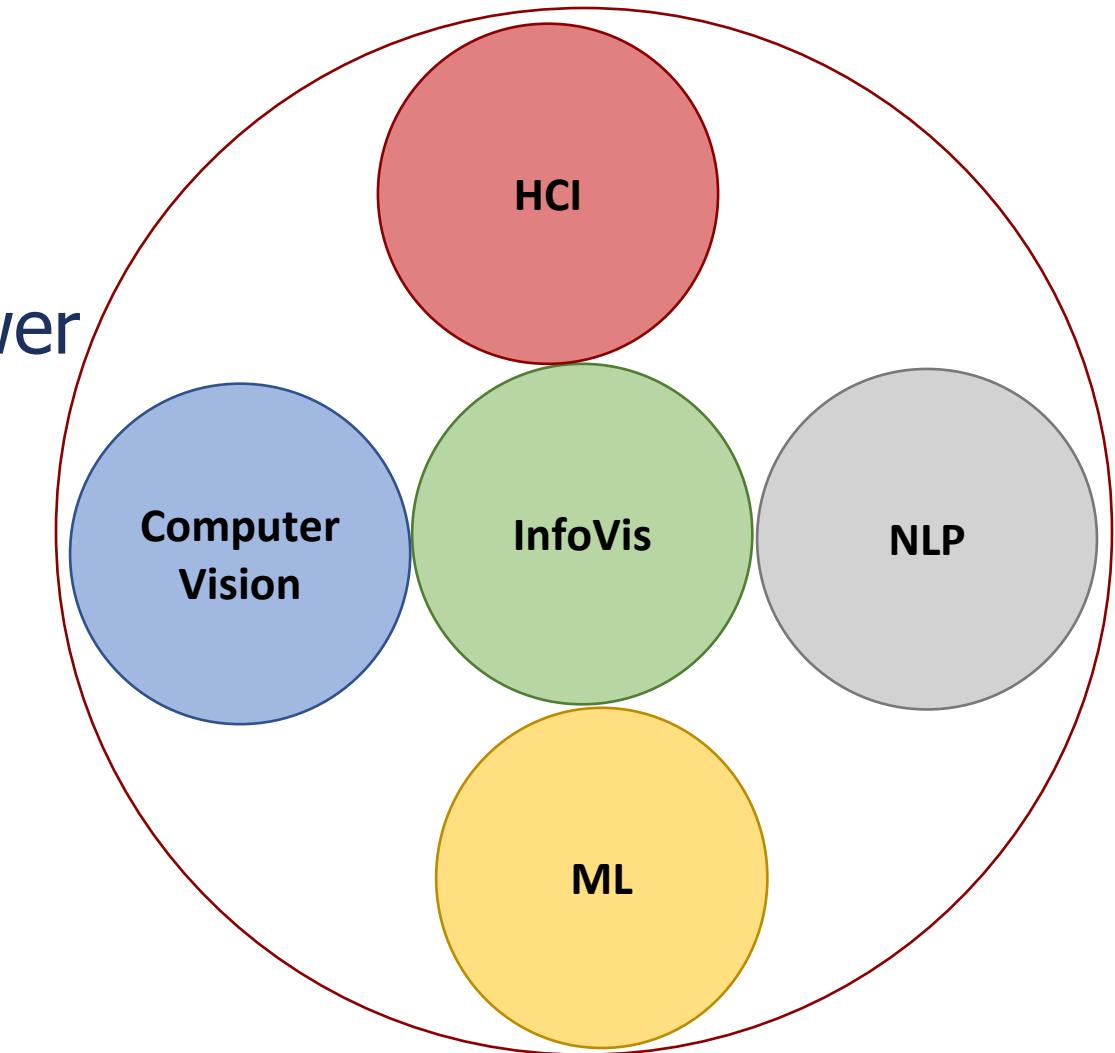
Tutorial Overview

- **Part 1:** Introduction [15 mins]
 - Why NLP + Vis?
 - An overview of NLP + Vis Research
 - An overview of the tutorial
 - **Part 2:** Deep Learning for NLP [50 mins]
 - Background
 - Large language models (LLMs)
 - **Part 3:** NLP4Vis applications [50 mins]
- ➡ **Part 4:** Future challenges and research opportunities [25 mins]

Open Challenges and Research Opportunities

Natural language interfaces

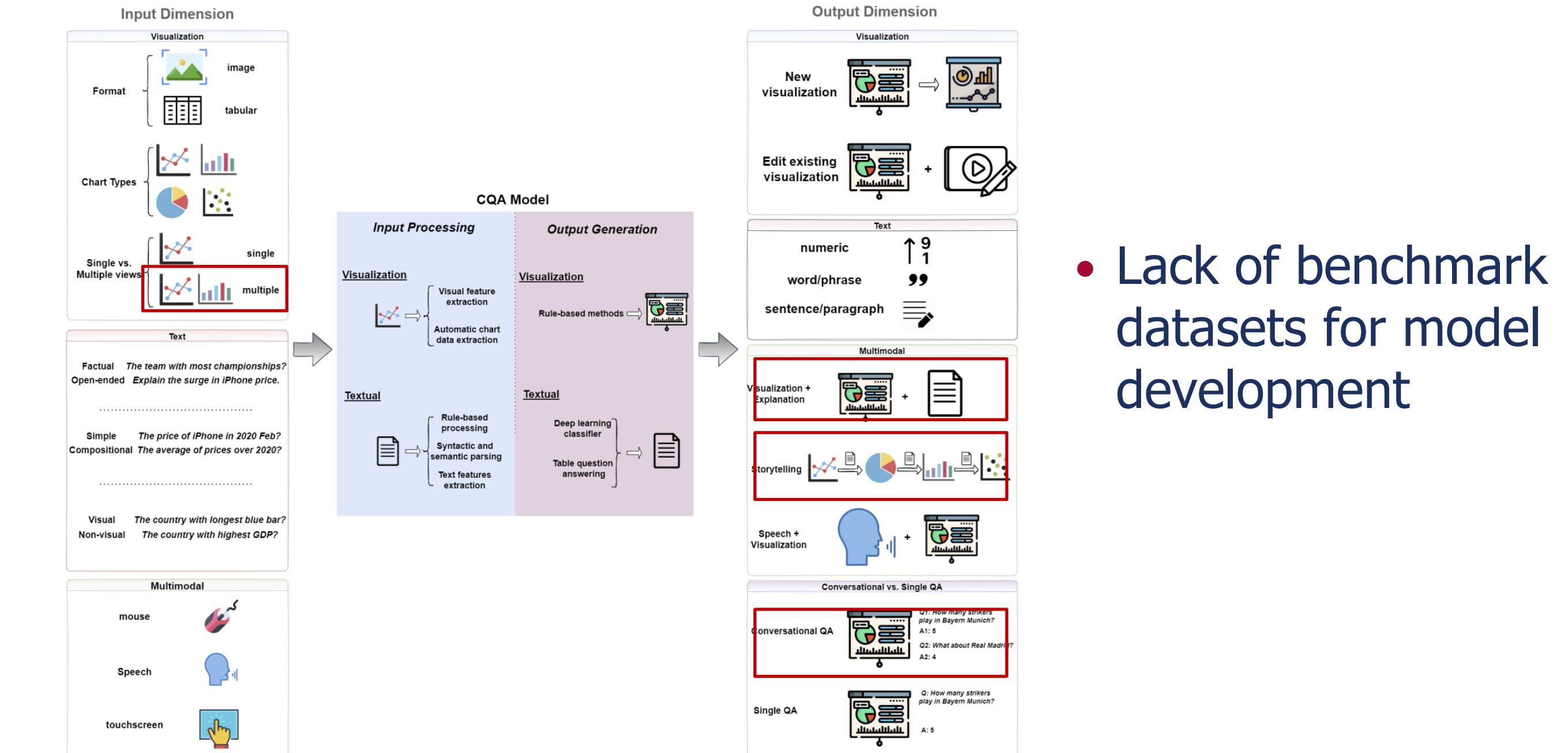
- Richness and ambiguities of natural language
- Complex reasoning required to predict the answer
- Computer vision challenges for automatic understanding of image charts
- The research is quite interdisciplinary



Open Challenges and Research Opportunities

- Need for large-scale real-world benchmark datasets
 - Most existing datasets lack realism
 - For many problem setups, there is no benchmark
 - More efforts are needed to create large-scale benchmark

Open Challenges and Research Opportunities

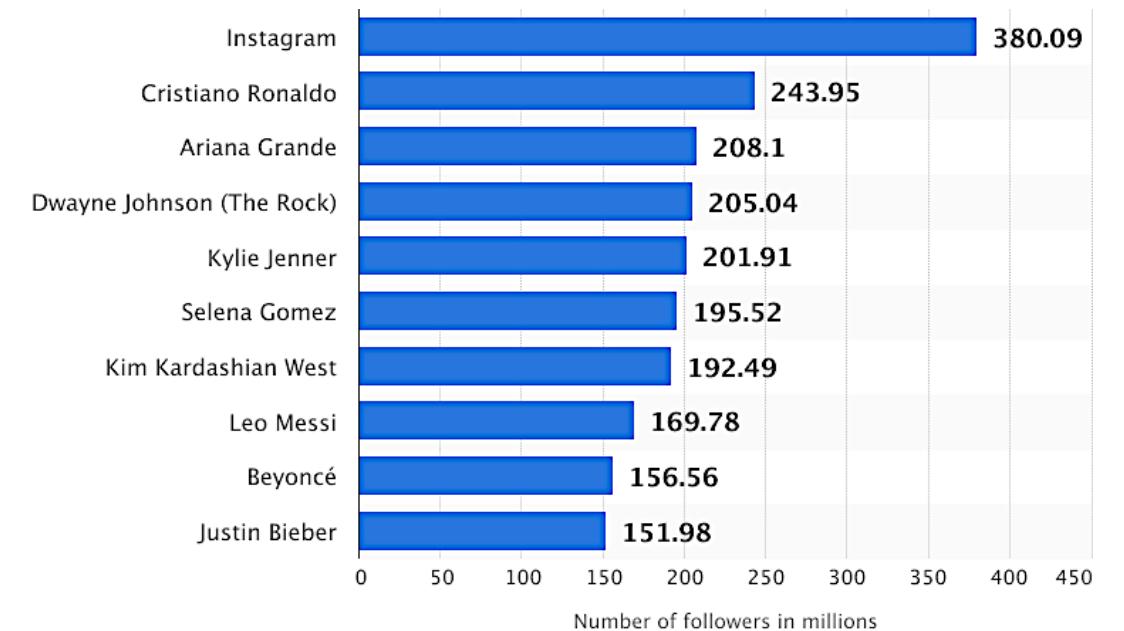


Open Challenges and Research Opportunities

Challenges with natural language generation:

- Hallucinations
- Factual errors
- Perceptual and reasoning aspects
- Computer Vision Challenges

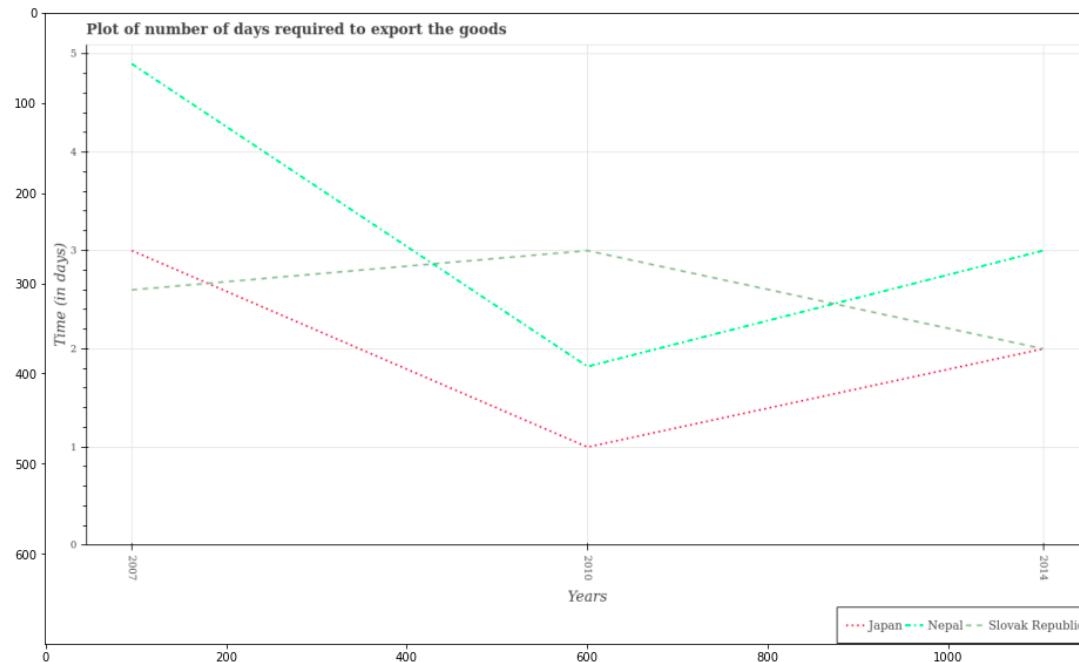
Instagram accounts with the most followers worldwide



OCR-T5: As of December 2020, **Cristiano Ronaldo** was the Instagram account with the most followers worldwide, with **380.09** million followers. The **Brazilian singer had** **243.95** million followers on the photo sharing app.

Challenges for Language Generations

ChatGPT Output



Summary: The line chart shows the number of days required to export goods for three countries – Japan, Nepal, and Slovak Republic – in 2007, 2010, and 2014. In 2007, Japan had the **lowest** number of days required to export goods at 3 days, followed by the Slovak Republic at 2.6 days, and Nepal at 4.9 days. By 2010, all three countries had reduced the number of days required to export goods, with Japan having the lowest number at 1 day, followed by the Slovak Republic at 3 days, and Nepal at 1.82 days. In 2014, Japan and the Slovak Republic both required 2 days to export goods, while Nepal took 3 days. The line chart indicates that all three countries made progress in reducing the time required to export goods between 2007 and 2014, with Japan and Nepal showing the most improvement. Overall, the chart highlights the efforts made by these countries to improve their export procedures and reduce the time required to export goods.

Most Americans believe it is acceptable to monitor others, except U.S. citizens

% of U.S. adults who say it is acceptable or unacceptable for the American government to monitor communications from ...

	Acceptable	Unacceptable
American citizens	40%	57%
Citizens of other countries	54	44
American leaders	60	38
Leaders of other countries	60	37
Terrorism suspects	82	15

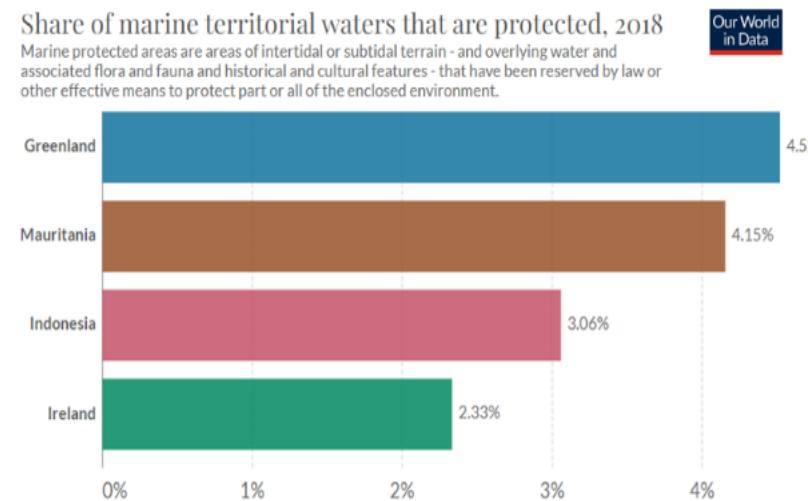
Source: Survey conducted Nov. 26, 2014-Jan. 3, 2015.

PEW RESEARCH CENTER

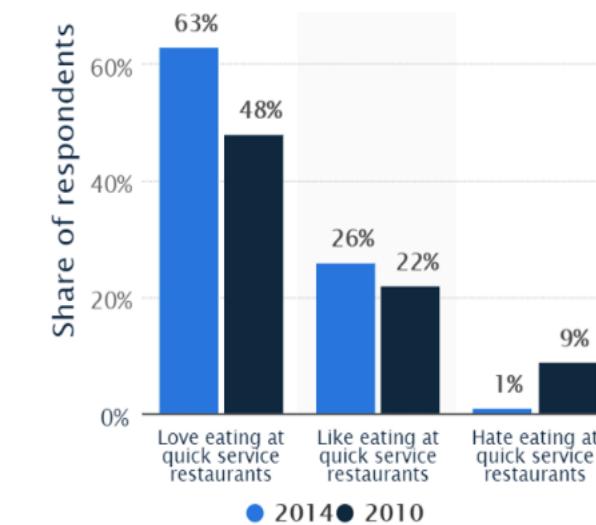
Summary: The chart shows the percentage of U.S. adults who believe that it is acceptable or unacceptable for the American government to monitor communications from different groups of people. The majority of Americans believe that it is acceptable to monitor citizens of other countries, American leaders, and leaders of other countries, with 54%, 60%, and 60% respectively saying it is acceptable. However, only 40% of Americans believe it is acceptable to monitor communications of American citizens. The chart also shows that the vast majority of Americans, 82%, believe it is acceptable to monitor communications of terrorism suspects. The survey was conducted by the Pew Research Center from November 26, 2014, to January 3, 2015.

Open Challenges and Research Opportunities

- Improving Logical and Visual Reasoning



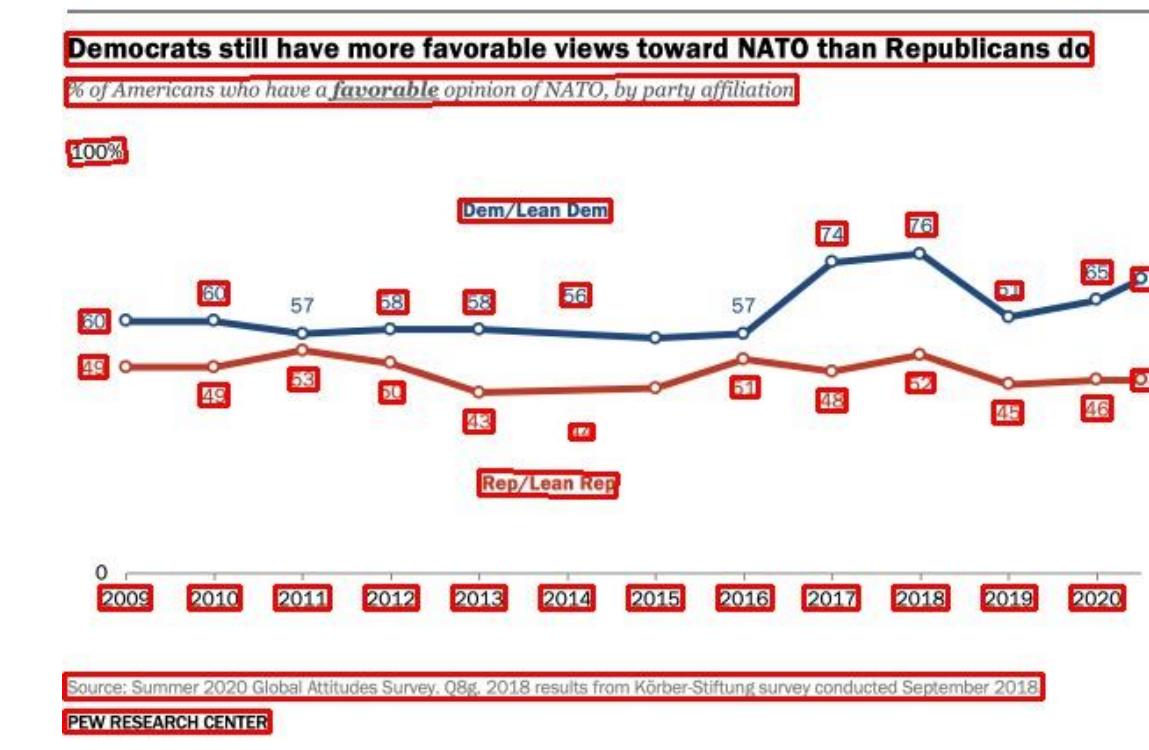
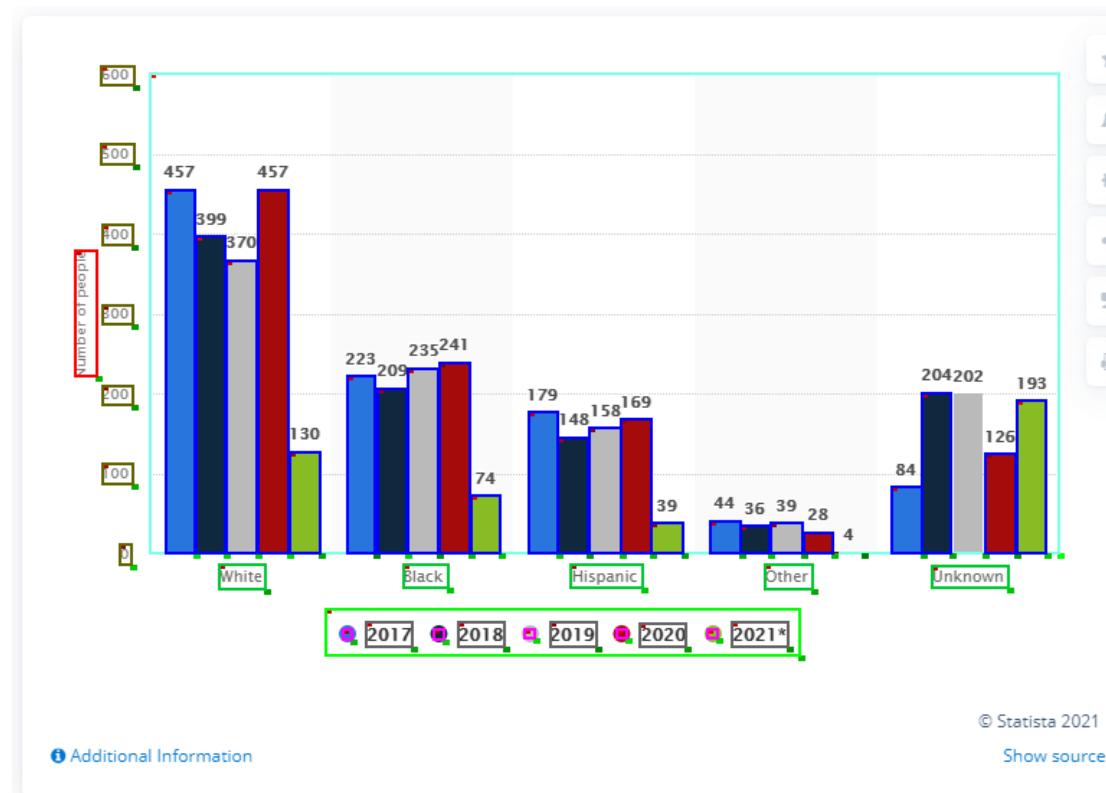
Q1: What is the difference between the sum shares of Indonesia and Ireland and share of Mauritania?
A: 1.06 **Output:** 0.547



Q2: What is the least difference between light blue bar and dark blue bar?
A: 4 **Output:** 39.32

Open Challenges and Research Opportunities

- Addressing computer vision challenges
 - chart data extraction challenges



Open Challenges and Research Opportunities

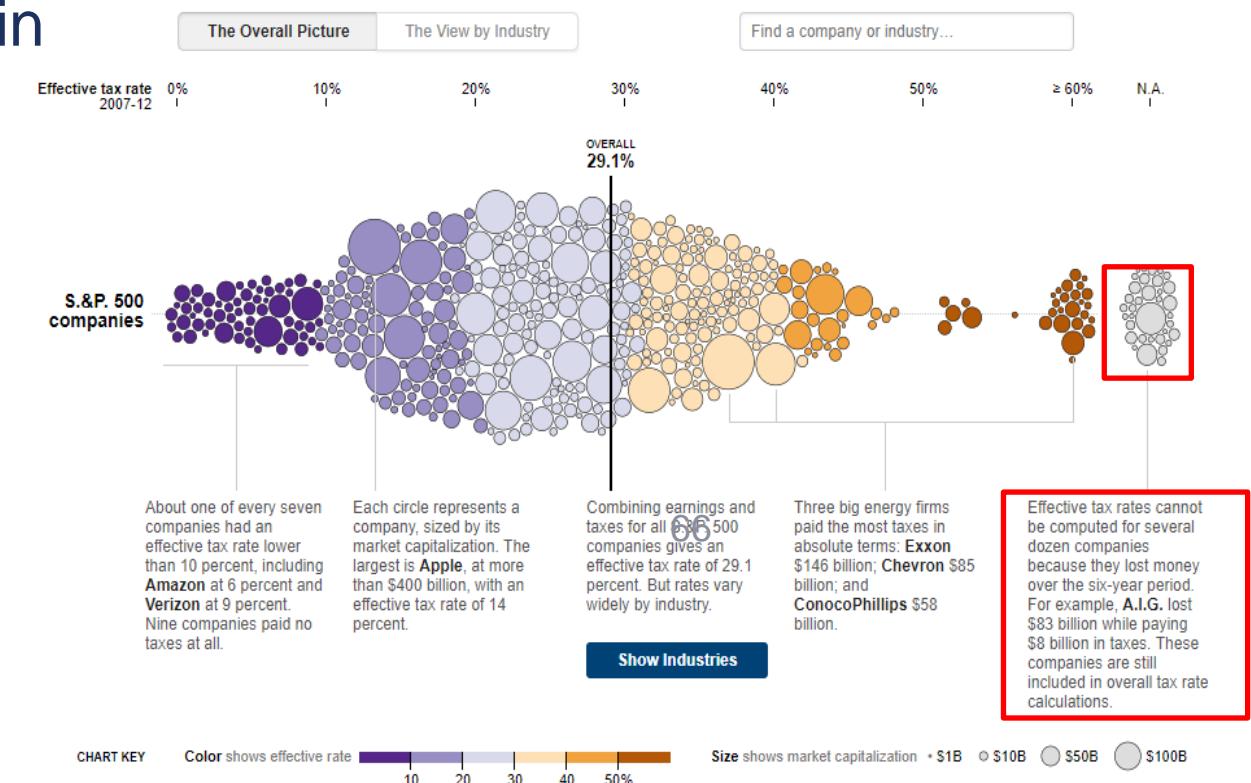
*How can we effectively **combine** text and visualization in the data stories?*

Support the author

- Suggest interesting patterns, trends, and outliers in the visualization
- Suggest explanatory text

Semi-automatic data storytelling tools

- What to show?
 - Generate visualizations and textual content
- How to show?
 - Determine the order of content
- Incorporate human in the loop



Open Challenges and Research Opportunities

- NLP4Vis to promote accessibility, inclusion, and diversity

The image shows a vertical sequence of four rounded rectangular speech bubbles, each containing a question or response from a user and an AI system. The user's questions are preceded by a microphone icon, and the AI's responses are preceded by a megaphone icon.

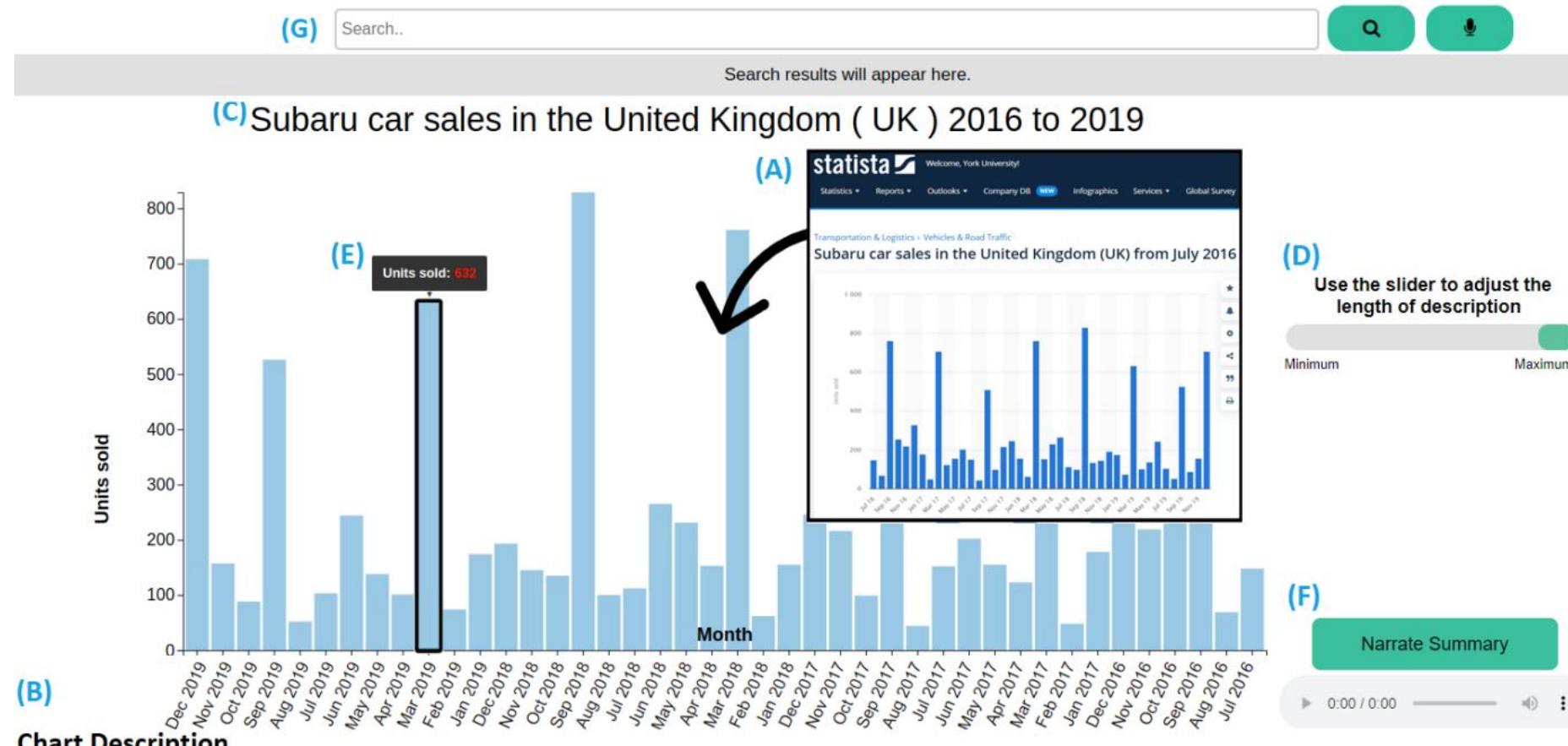
- What's the maximum data point in the graph?
- I understood you're looking for maximum. Maximum value of Total Crime for Crime Neighborhood is 48,900 belonging to Downtown Commercial.
- What about the average and variance?
- I heard you asking about the average, and variance. Average of Total Crime for Crime Neighborhood is 8,870. Variance of Total Crime for Crime Neighborhood is 78,800,000.

How people who are blind may ask questions about charts?

Sharif A, Wang OH, Muongchan AT, Reinecke K, Wobbrock JO. VoxLens: Making Online Data Visualizations Accessible with an Interactive JavaScript Plug-In. InCHI Conference on Human Factors in Computing Systems 2022 Apr 29 (pp. 1-19).

Open Challenges and Research Opportunities

- NLP4Vis to promote accessibility, inclusion, and diversity



Interested in NLP4Vis Research?



Overview of NLP Tools

- **Stanford Core NLP**: robust, has various wrappers
- **NLTK**: A popular Python library for basic NLP tasks
- **TextBlob**: An intuitive interface for NLTK
- **SpaCy**: Fast, library for information extraction and NLU tasks
- **IBM Watson**: AI platform for businesses
- **Google Cloud NLP API**: Google technology applied to NLP
- **GenSim**: topic modeling, word2vec, document indexing and similarity

Deep Learning Tools and Libraries

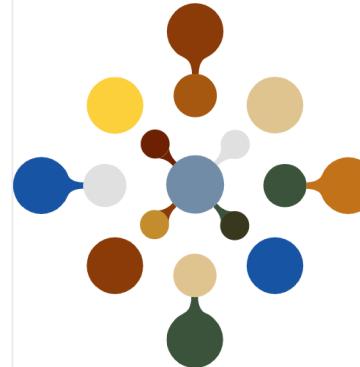
- **PyTorch-NLP:** extends PyTorch to provide basic text processing functions.
- **Huggingface:** pre-trained models, datasets and metrics, for NLP
- **TensorFlow:** written in C++ and CUDA and provides an interface to languages like Python, Java
- **Theano:** Python library to evaluate mathematical operations including multi-dimensional arrays
- **Keras:** a high-level neural network API that is capable of running on top of TensorFlow or Theano.

Some Suggested Readings

- Survey on text visualization and visual text analytics
 - Text visualization browser: Kucher et al., <http://textvis.lnu.se/>
 - Shixia Liu, Xiting Wang, Christopher Collins, Wenwen Dou, Fangxin Ouyang, Mennatallah El-Assady, Liu Jiang, Daniel A Keim, Bridging text visualization and mining: A task-driven survey, IEEE TVCG, 2018
- Survey on Natural language interfaces
 - Enamul Hoque, Parsa Kavehzadeh and Ahmed Masry, Chart question answering: State of the art and future directions, Journal of Computer Graphics Forum (Proc. EuroVis), 2022
 - Shen et al., Towards Natural Language Interfaces for Data Visualization: A Survey, IEEE TVCG,
- Textbooks on NLP and Deep Learning
 - Deep Learning by Goodfellow, Bengio, and Courville [free online](#)
 - Machine Learning — A Probabilistic Perspective by Kevin Murphy [online](#)
 - Natural Language Processing by Jacob Eisenstein [free online](#)
 - Speech and Language Processing by Dan Jurafsky and James H. Martin [\(3rd ed. draft\)](#)

Some Suggested Readings on LLM

- Courses:
 - [Stanford Course on LLM](#)
 - [Princeton Course on Understanding LLMs](#)
- Surveys:
 - [Survey of Large Language Models](#)
 - [Survey on of Prompting Methods in NLP](#)
- A few key papers:
 - [Attention Is All You Need](#)
 - [Language Models are Few-Shot Learners](#)
 - [The Power of Scale for Parameter-Efficient Prompt Tuning](#)
 - [Emergent abilities of large language models](#)
 - [Chain-of-Thought Prompting Elicits Reasoning in Large Language Models](#)
 - [Multitask Prompted Training Enables Zero-shot Task Generalization](#)



VIS 2023



<https://nlp4vis.github.io>

NLP4Vis: Natural Language Processing for Information Visualization
Half-day tutorial at IEEE Vis Conference 2023.

Overview

This tutorial will provide an introduction to natural language processing (NLP) to interested researchers in the visualization (Vis) community. It will first motivate why NLP4Vis is an important area of research and provide an overview of research topics on combining NLP and Vis techniques. Then an overview of deep learning models for NLP will be covered. A particular focus will be provided on highlighting the recent progress on large language models such as ChatGPT and how such models can be leveraged to solve various NLP tasks for visualizations. In the final part, we will focus on various application tasks at the intersection of NLP and Vis. We will conclude with an interactive discussion of future challenges for NLP+Vis applications. The audience will include researchers interested in applying NLP for visualizations as well as others who focus more generally at the intersection of AI and visualization.

Materials

Tutorial Overview

Introduction [15 mins]

- What is NLP?
- What is Vis?
- Why NLP + Vis?
- An overview of research topics on combining NLP and Vis techniques
- An overview of the tutorial

Coffee Break

Deep Learning for NLP [60 mins]

- Introduction to NLP
- Language modeling
- Model Architectures
 - Transformer Architecture
 - Encoder, decoder, encoder-decoder
 - Pre-training and fine-tuning
- Large language models (LLMs)
 - Scaling LMs to LLMs
 - Prompt Engineering
 - In context Learning
 - Instruction Tuning

NLP + Vis Applications [50 mins]

- Visual interfaces
- Natural language interfaces for visualizations
- ChartNL (e.g., Chart question answering, Text2Chart)
- Natural language generation for visualization (e.g., Chart-to-text)
- Automated data-driven storytelling
- NLP for chart accessibility and inclusions
- Live demos

Future Challenges [25 mins]

- Building benchmarks for training and evaluation
- Guidelines for fairness and bias

Questions?

**Interested in NLP+Vis?
We are looking for PostDocs and PhD students.
Reach out to me: enamulh@yorku.ca**