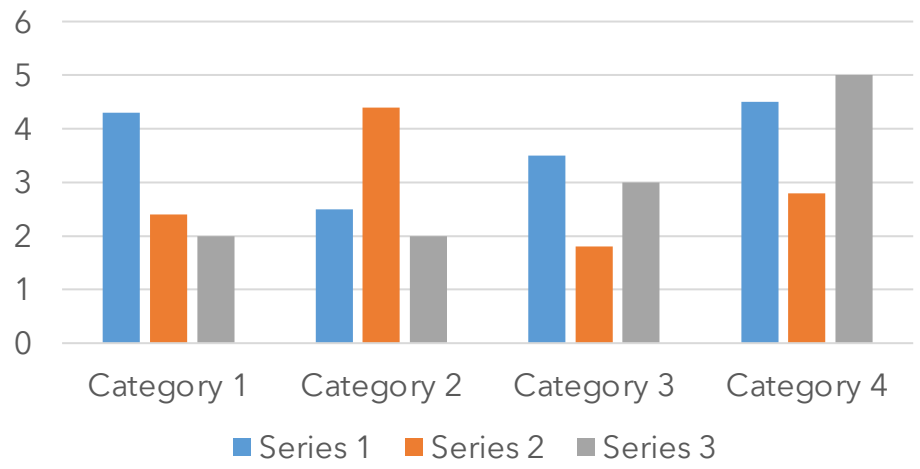
EUROVIS Leipzig 2023

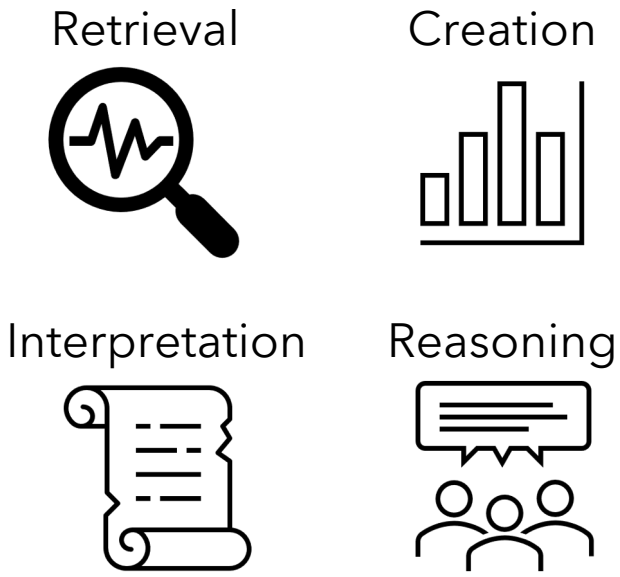
# The State of the Art in Creating Visualization Corpora for Automated Chart Analysis

Chen Chen, Zhicheng Liu

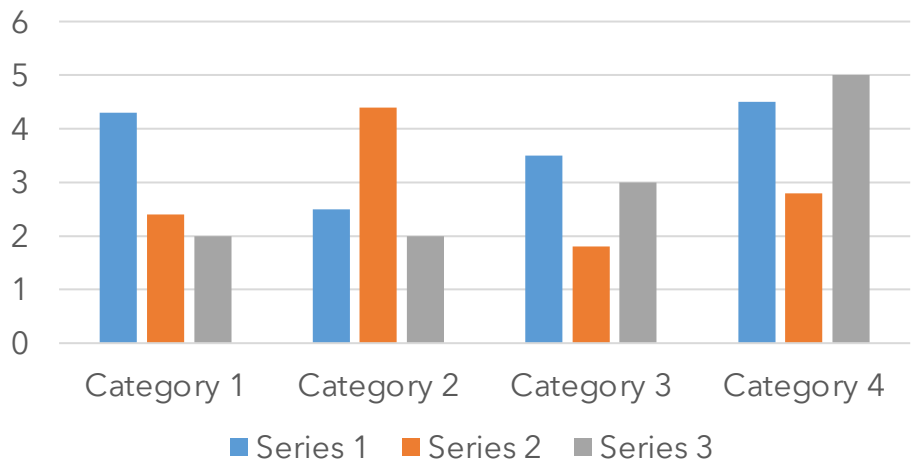
HCIL · University of Maryland · Department of Computer Science · HDI

# Background


6
5
4
3
2
1
0
Category 1
Category 2
Category 3
Category 4
Series 1
Series 2
Series 3


Retrieval
Creation
Interpretation
Reasoning

# Background

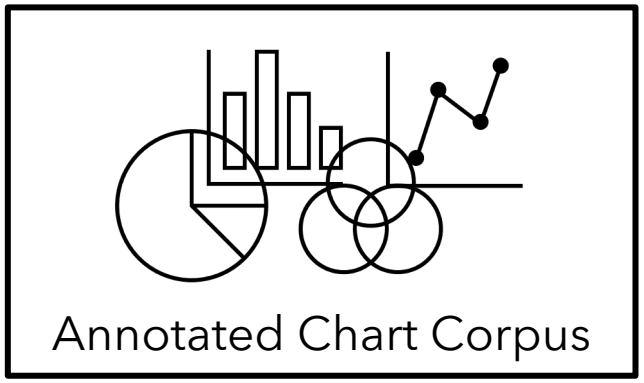



Retrieval

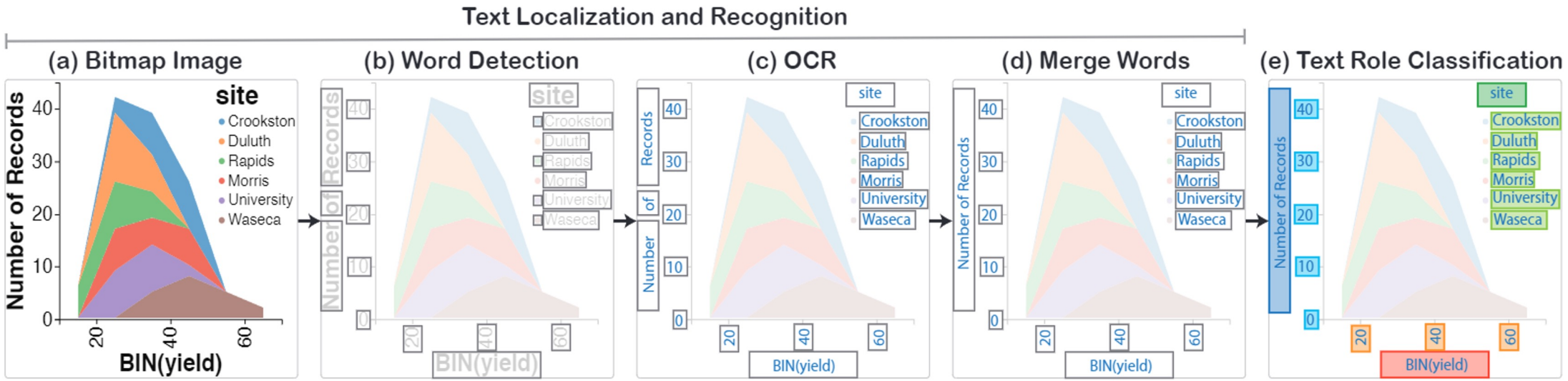
Creation

Interpretation

Reasoning



# An Example: REV




Poco and Heer. *Reverse-Engineering Visualizations: Recovering Visual Encodings from Chart Images*

collected 4,318 charts from Vega, Quartz, and academic papers;
annotated the bounding box, content, and role for each text element;
evaluated text recognition pipeline using the corpus.

# Motivation

EUROVIS Leipzig 2023

However, we have observed that many papers do **not**

- Use existing corpora but build own corpus instead
- Make the corpora with annotations publicly available
- Release their annotation tools
- Examine how different chart corpora influence model performance

# Motivation

Little research on:

- the common practices for creating the corpora
- what constitutes a good chart corpus
- the potential pitfalls and gaps in existing corpus-based research

# Motivation

Little research on:

- the common practices for creating the corpora
- what constitutes a good chart corpus
- the potential pitfalls and gaps in existing corpus-based research

*It is difficult to*
*compare chart analysis techniques*
*measure scientific progress*
*identify unsolved research problems*

# Survey Goal

A comprehensive understanding of the state of the art in creating corpora for automated chart analysis research

- summarize current patterns and practices of corpora creation
- identify research gaps and opportunities
- recommend desired properties of benchmark corpora
- discuss research ideas on tools and methods

# Outline

- Survey
  - Method
  - Task Taxonomy
    (1) Why: the goal; (2) How: the method; (3) What: the output
  - Corpus Property
    (1) Format; (2) Scope; (3) Collection Method; (4) Annotations; (5) Diversity
- Challenges and Future Directions
  - Open opportunities
  - Desire properties of benchmark corpora
  - Desired tools

# Survey Method

EUROVIS Leipzig 2023

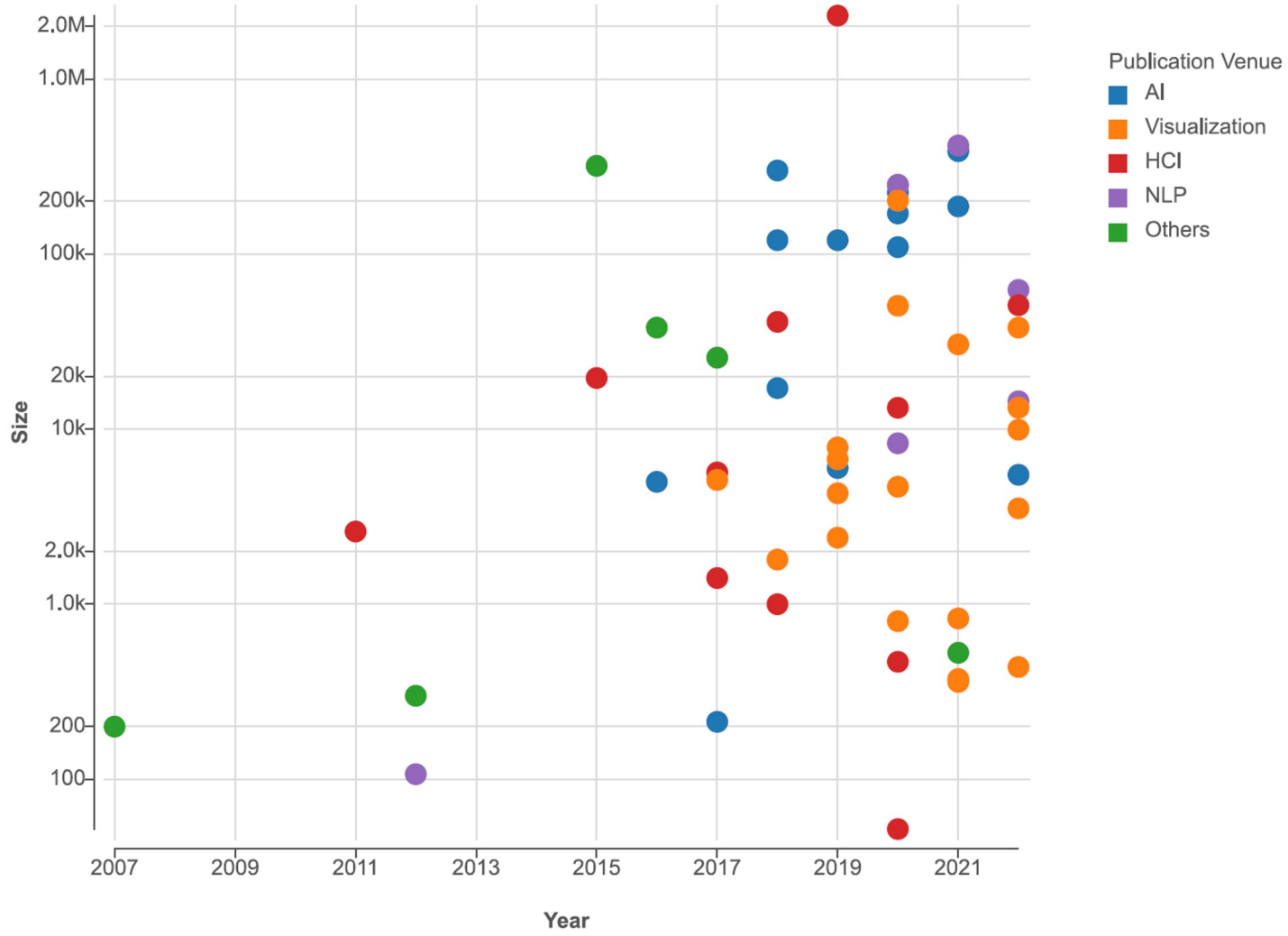
- Search 41
  - Start with AI4VIS [WWS*21] and ML4VIS [WCWQ21]
  - Three criteria: primary contribution, corpus description, and chart design

# Survey Method

EUROVIS Leipzig 2023

- Search 41
  - Start with AI4VIS [WWS*21] and ML4VIS [WCWQ21]
  - Three criteria: primary contribution, corpus description, and chart design
- Augment 56
  - Graph traversal over citation network
  - Excluded papers before 2007

# Overview of 56 Chart Corpora



# Bottom-up Coding
## task

EUROVIS Leipzig 2023

### Why: the goal

- Create a chart corpus
- Extract chart semantics
- Modify an existing chart
- Generate chart designs automatically
- Retrieve charts matching certain criteria
- Generate natural language descriptions

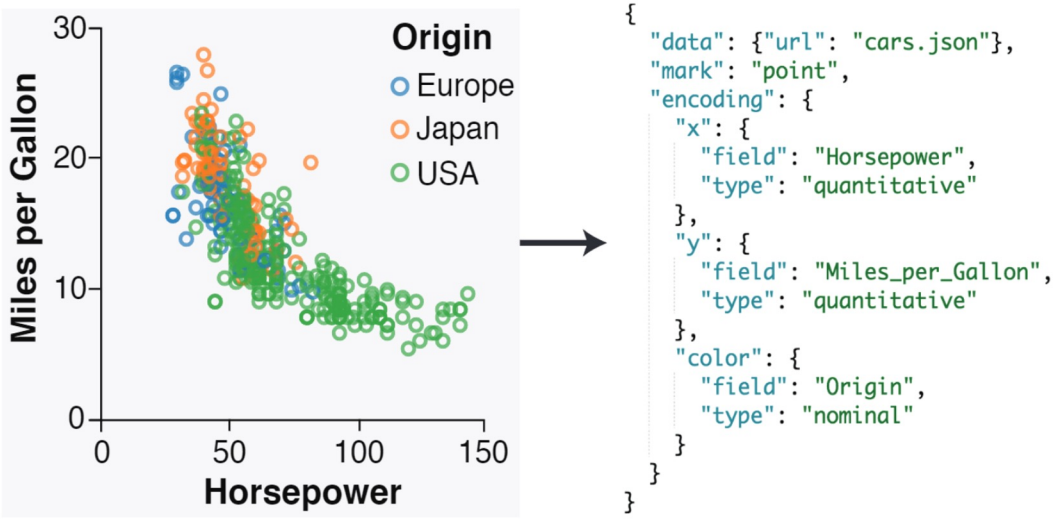

Battle et al. *Beagle: Automated Extraction and Interpretation of Visualizations from the Web*

# Bottom-up Coding
## task

EUROVIS Leipzig 2023

### Why: the goal

- Create a chart corpus
- Extract chart semantics
- Modify an existing chart
- Generate chart designs automatically
- Retrieve charts matching certain criteria
- Generate natural language descriptions




Poco and Heer. *Reverse-Engineering Visualizations: Recovering Visual Encodings from Chart Images*

# Bottom-up Coding
## task

EUROVIS Leipzig 2023

### Why: the goal

- Create a chart corpus
- Extract chart semantics
- Modify an existing chart
- Generate chart designs automatically
- Retrieve charts matching certain criteria
- Generate natural language descriptions

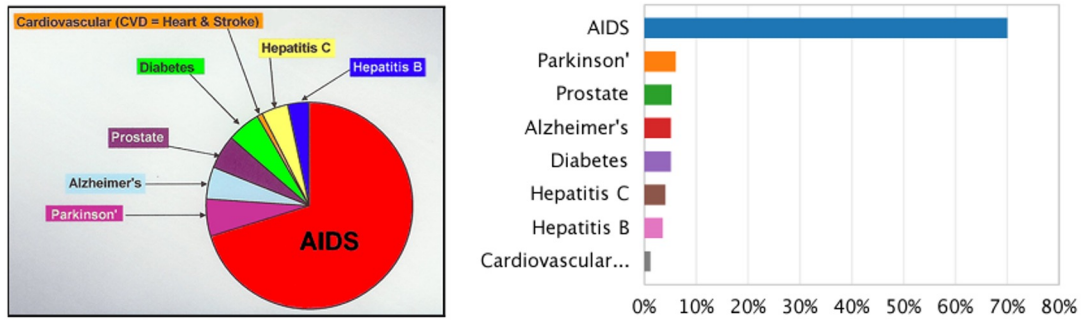



Figure 1: Chart Redesign. Left: A pie chart of NIH expenses per condition-related death. The chart suffers from random sorting, highly saturated colors, and erratic label placement. Right: Plotting the data as a sorted bar chart enables more accurate comparisons of data values [6, 20].

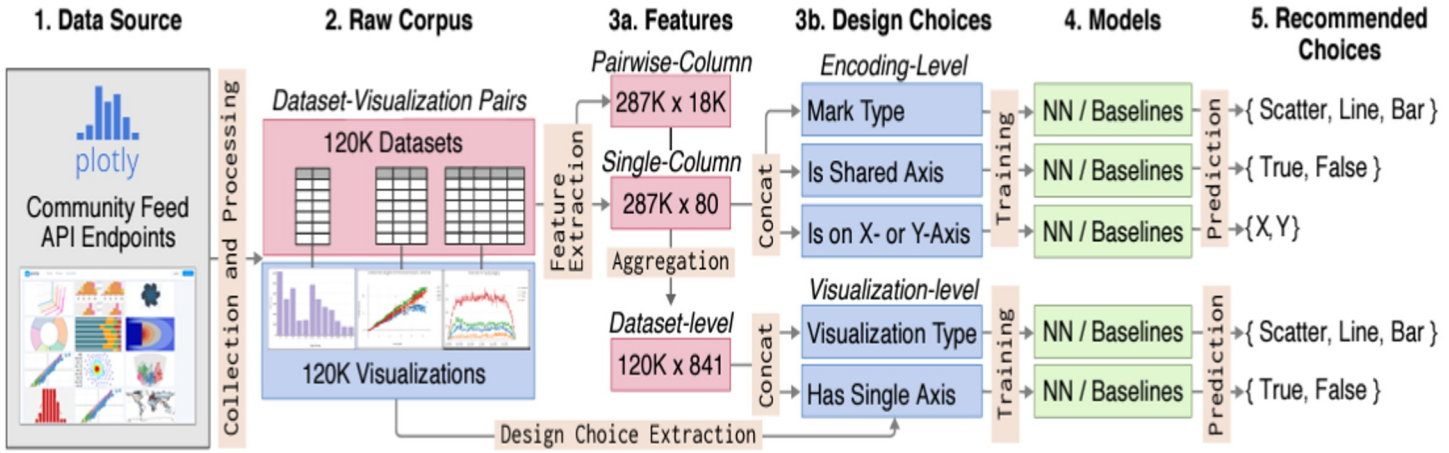
Savva et al. *ReVision: Automated Classification, Analysis and Redesign of Chart Images*

# Bottom-up Coding
task

EUROVIS Leipzig 2023

## Why: the goal

- Create a chart corpus
- Extract chart semantics
- Modify an existing chart
- Generate chart designs automatically
- Retrieve charts matching certain criteria
- Generate natural language descriptions

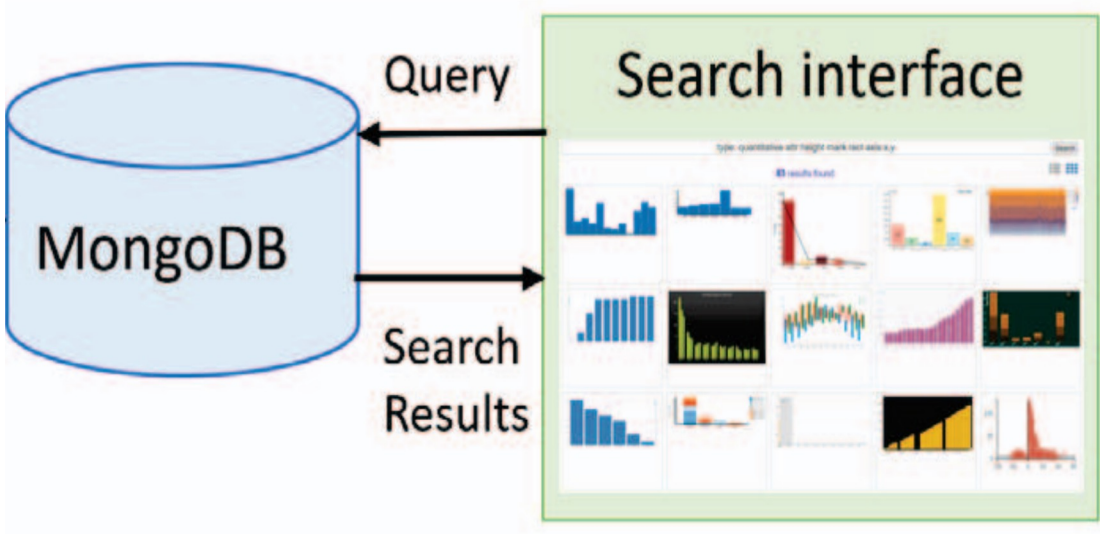



Hu et al. *VizML: A Machine Learning Approach to Visualization Recommendation*

# Bottom-up Coding
## task

Why: the goal

- Create a chart corpus
- Extract chart semantics
- Modify an existing chart
- Generate chart designs automatically
- Retrieve charts matching certain criteria
- Generate natural language descriptions



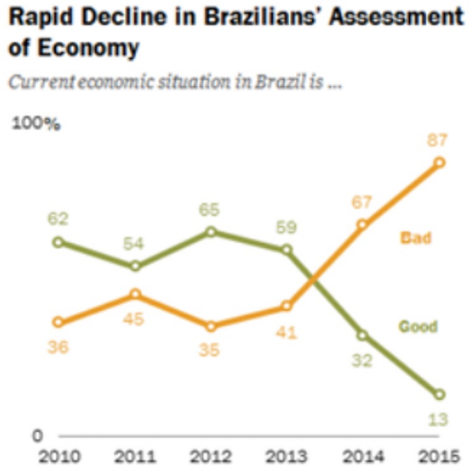

Hoque and Agrawala. *Searching the visual style and structure of d3 visualizations*

# Bottom-up Coding
## task

EUROVIS Leipzig 2023

### Why: the goal

- Create a chart corpus
- Extract chart semantics
- Modify an existing chart
- Generate chart designs automatically
- Retrieve charts matching certain criteria
- Generate natural language descriptions




**Q1:** Which year has the most divergent opinions about Brazil's economy?

**Answer:** 2015

**Q2:** What is the peak value of the orange line?

**Answer:** 87

Masry et al. *ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning*

# Bottom-up Coding
## task

EUROVIS Leipzig 2023

Why: the goal

- Create a chart corpus
- Extract chart semantics
- Modify an existing chart
- Generate chart designs automatically
- Retrieve charts matching certain criteria
- Generate natural language descriptions

How: the method

- modern neural networks
- classic machine learning
- heuristics-based algorithms

# Bottom-up Coding
## task

EUROVIS Leipzig 2023

Why: the goal

- Create a chart corpus
- Extract chart semantics
- Modify an existing chart
- Generate chart designs automatically
- Retrieve charts matching certain criteria
- Generate natural language descriptions

How: the method

- modern neural networks
- classic machine learning
- heuristics-based algorithms

What: the output

- chart components
  - marks, encodings, layouts…
- synthesized descriptions
  - caption, summarization…
- derived properties
  - quality, similarity…

# Coding and Analysis
## corpus property

EUROVIS Leipzig 2023

**Format**

**Scope**

- chart type
- design variation

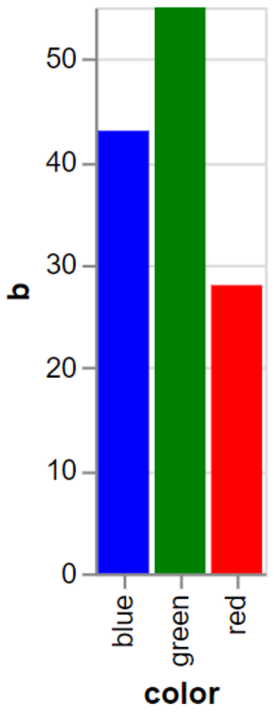
**Collection Method**

**Annotations**

- annotation type
- annotation method

**Diversity**

## Bitmap



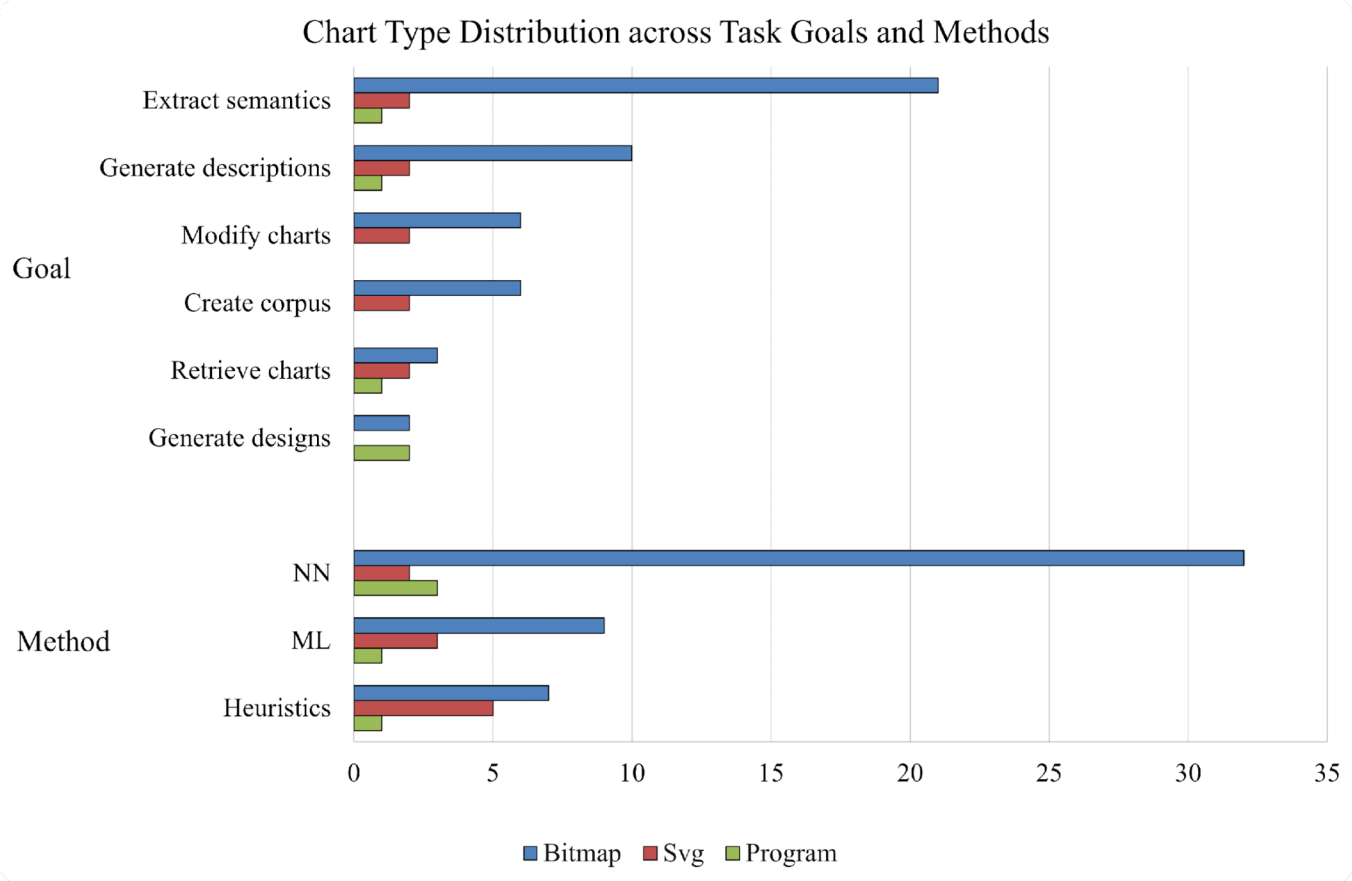

## SVG

```
<svg xmlns="http://www.w3.org/2000/svg" xmlns:xlink="http://www.w3.org/1999/xlink" version="1.1" class="marks"
 width="104" height="258" viewBox="0 0 104 258" style="background-color: white;">
 <g fill="none" stroke-miterlimit="10" transform="translate(39,5)">
  <g class="mark-group role-frame root" role="graphics-object" aria-roledescription="group mark container">
   <g transform="translate(0,0)">
    <path class="background" aria-hidden="true" d="M0.5,0.5h60v200h-60Z" fill="transparent" stroke="#ddd">
    </path>
    <g>
     <g class="mark-group role-axis" aria-hidden="true">…</g>
     <g class="mark-group role-axis" role="graphics-symbol" aria-roledescription="axis" aria-label="X-axis ti
     tled 'color' for a discrete scale with 3 values: blue, green, red">…</g>
     <g class="mark-group role-axis" role="graphics-symbol" aria-roledescription="axis" aria-label="Y-axis ti
     tled 'b' for a linear scale with values from 0 to 55">…</g>
     <g class="mark-rect role-mark marks" role="graphics-object" aria-roledescription="rect mark container">
      <path aria-label="color: red; b: 28" role="graphics-symbol" aria-roledescription="bar" d="M41,98.18181
      818181819h18v101.81818181818181h-18Z" fill="red"></path>
      <path aria-label="color: green; b: 55" role="graphics-symbol" aria-roledescription="bar" d="M21,0h18v2
      00h-18Z" fill="green"></path>
      <path aria-label="color: blue; b: 43" role="graphics-symbol" aria-roledescription="bar" d="M1,43.63636
      3636363626h18v156.36363636363637h-18Z" fill="blue"></path>
     </g>
    </g>
    <path class="foreground" aria-hidden="true" d display="none"></path>
   </g>
  </g>
 </g>
</svg>
```

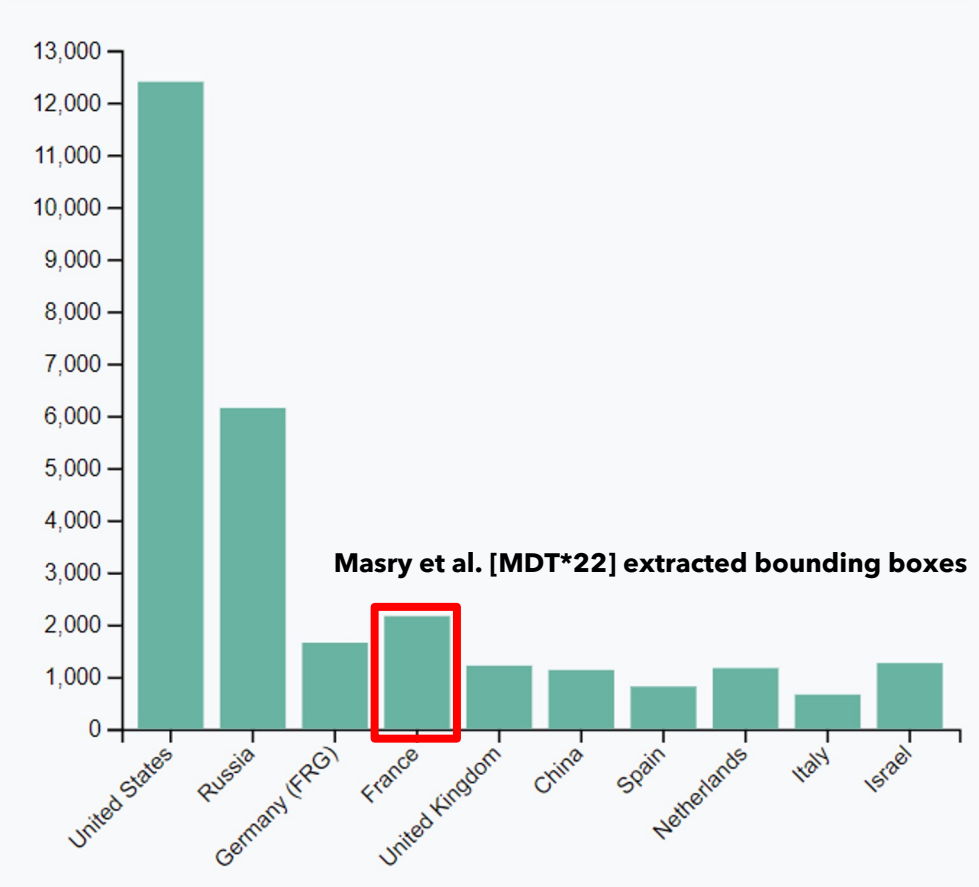
## Program

```
{
  "$schema": "https://vega.github.io/schema/vega-lite/v5.json",
  "description": "A bar chart that directly encodes color names in the data.",
  "data": {
    "values": [
      {
        "color": "red",
        "b": 28
      },
      {
        "color": "green",
        "b": 55
      },
      {
        "color": "blue",
        "b": 43
      }
    ]
  },
  "mark": "bar",
  "encoding": {
    "x": {
      "field": "color",
      "type": "nominal"
    },
    "y": {
      "field": "b",
      "type": "quantitative"
    },
    "color": {
      "field": "color",
      "type": "nominal",
      "scale": null
    }
  }
}
```

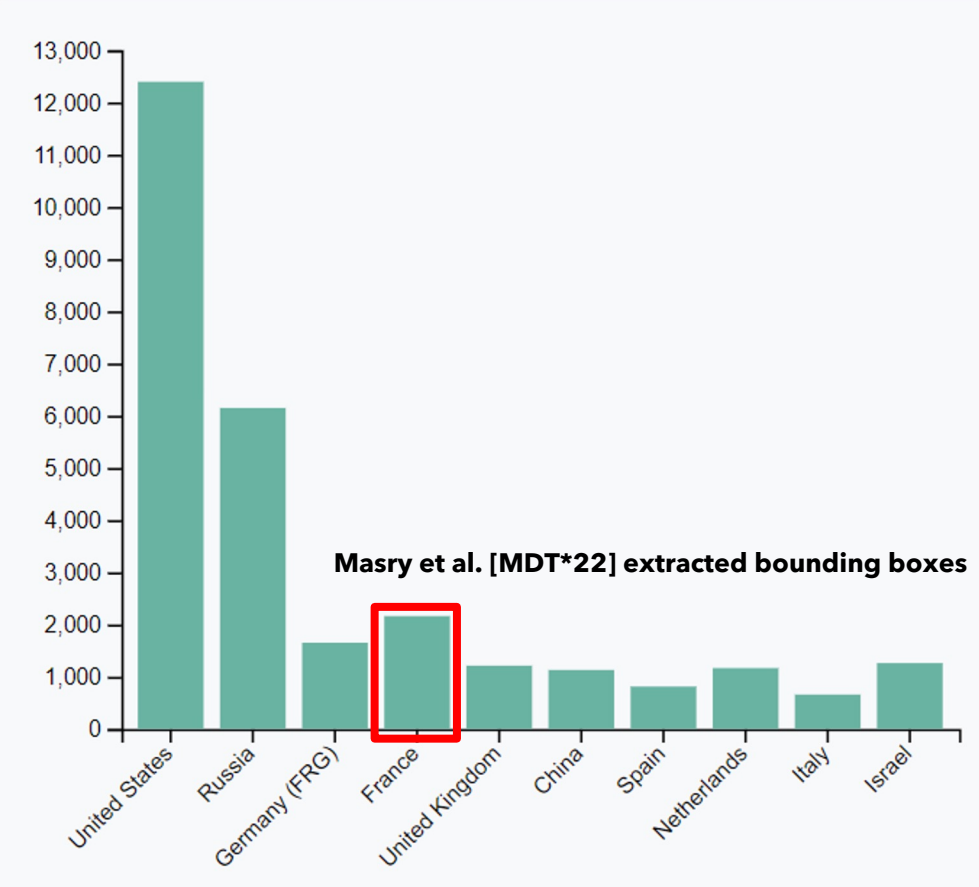
*https://vega.github.io/vega-lite/examples/bar_color_disabled_scale.html*




- 48 bitmap
- 10 SVG
- 5 program



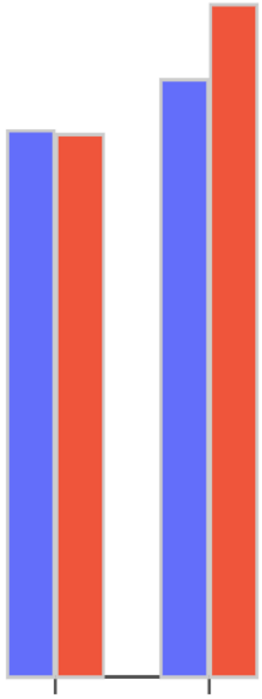

```
▼<g class="tick" opacity="1" transform="translate(0,23.576923076923038)">
    <line stroke="#000" x2="-6"></line>
    <text fill="#000" x="-9" dy="0.32em">12,000</text>
  </g>
▼<g class="tick" opacity="1" transform="translate(0,0.5)">
    <line stroke="#000" x2="-6"></line>
    <text fill="#000" x="-9" dy="0.32em">13,000</text>
  </g>
</g>
<rect x="7.254901960784309" y="13.984615384615381" width="29.019607843137255" height="286.0153846153846" fill="#69b3a2"></rect>
<rect x="43.52941176470588" y="158.12307692307692" width="29.019607843137255" height="141.87692307692308" fill="#69b3a2"></rect>
<rect x="79.80392156862744" y="261.8538461538462" width="29.019607843137255" height="38.14615384615382" fill="#69b3a2"></rect>
<rect x="116.078431372549" y="250.1076923076923" width="29.019607843137255" height="49.892307692307696" fill="#69b3a2"></rect>
```




**Masry et al. [MDT*22] extracted bounding boxes**

**Poco and Heer [PH17] parsed text elements**

```
<g class="tick" opacity="1" transform="translate(0,23.576923076923038)">
  <line stroke="#000" x2="-6"></line>
  <text fill="#000" x="-9" dy="0.32em">12,000</text>
</g>
<g class="tick" opacity="1" transform="translate(0,0.5)">
  <line stroke="#000" x2="-6"></line>
  <text fill="#000" x="-9" dy="0.32em">13,000</text>
</g>
<rect x="7.254901960784309" y="13.984615384615381" width="29.019607843137255" height="286.0153846153846" fill="#69b3a2"></rect>
<rect x="43.52941176470588" y="158.12307692307692" width="29.019607843137255" height="141.87692307692308" fill="#69b3a2"></rect>
<rect x="79.80392156862744" y="261.8538461538462" width="29.019607843137255" height="38.14615384615382" fill="#69b3a2"></rect>
<rect x="116.078431372549" y="250.1076923076923" width="29.019607843137255" height="49.892307692307696" fill="#69b3a2"></rect>
```



```
<g id="collection17" class="collection17">
  <g id="collection16" class="collection16">
    <rect id="rect24" class="rect24" x="100" y="137.50867194686887" width="16" height="162.49132805313113" style="fill: rgb(99, 110, 250); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
    <rect id="rect25" class="rect24" x="117" y="138.65125984428636" width="16" height="161.34874015571364" style="fill: rgb(239, 85, 59); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
  </g>
  <g id="collection18" class="collection16">
    <rect id="rect26" class="rect24" x="153" y="122.37662949583793" width="16" height="177.62337050416207" style="fill: rgb(99, 110, 250); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
    <rect id="rect27" class="rect24" x="170" y="100" width="16" height="200" style="fill: rgb(239, 85, 59); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
  </g>
</g>
```

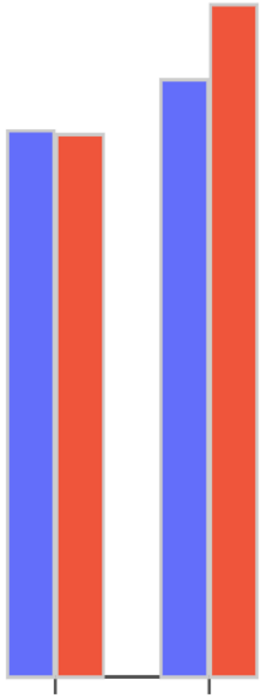
**(a) Mascot**

```
<g class="barlayer mlayer">
  <g class="trace bars" style="opacity: 1;">
    <path d="M36.55,260V59.32H182.75V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(99, 110, 250); fill-opacity: 1;"/>
    <path d="M402.05,260V40.64H548.25V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(99, 110, 250); fill-opacity: 1;"/>
  </g>
  <g class="trace bars" style="opacity: 1;">
    <path d="M182.75,260V60.73H328.95V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(239, 85, 59); fill-opacity: 1;"/>
    <path d="M548.25,260V13H694.45V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(239, 85, 59); fill-opacity: 1;"/>
  </g>
</g>
```

**(b) Plotly**

```
<g class="mark-rect role-mark marks" role="graphics-object" aria-roledescription="rect mark container">
  <path aria-label="sex: Female; value: 18.10519; smoker: no" role="graphics-symbol" aria-roledescription="bar" d="M10.000000000000007,49.123416666666664h20v150.87658333333334h-20Z" fill="#4c78a8"/>
  <path aria-label="sex: Female; value: 17.97788; smoker: yes" role="graphics-symbol" aria-roledescription="bar" d="M30.000000000000007,50.18433333333334h20v149.81566666666666h-20Z" fill="#f58518"/>
  <path aria-label="sex: Male; value: 19.79124; smoker: no" role="graphics-symbol" aria-roledescription="bar" d="M60.00000000000001,35.07300000000002h20v164.92699999999996h-20Z" fill="#4c78a8"/>
  <path aria-label="sex: Male; value: 22.2845; smoker: yes" role="graphics-symbol" aria-roledescription="bar" d="M80,14.29583333333333h20v185.70416666666668h-20Z" fill="#f58518"/>
</g>
```

**(c) Vega-Lite**



```
<rect id="rect24" class="rect24" x="100" y="137.50867194686887" width="16" height="162.49132805313113" style="fill: rgb(99, 110, 250); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
```

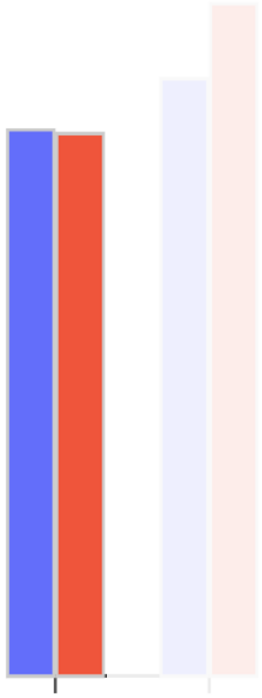
(a) Mascot

```
<path d="M36.55,260V59.32H182.75V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(99, 110, 250); fill-opacity: 1;"/>
```

(b) Plotly

```
<path aria-label="sex: Female; value: 17.97788; smoker: yes" role="graphics-symbol" aria-roledescription="bar" d="M30.000000000000007,50.18433333333334h20v149.81566666666666h-20Z" fill="#f58518"/>
```

(c) Vega-Lite



```
<g id="collection16" class="collection16">
  <rect id="rect24" class="rect24" x="100" y="137.50867194686887" width="16" height="162.49132805313113" style="fill: rgb(99, 110, 250); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
  <rect id="rect25" class="rect24" x="117" y="138.65125984428636" width="16" height="161.34874015571364" style="fill: rgb(239, 85, 59); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
</g>
```

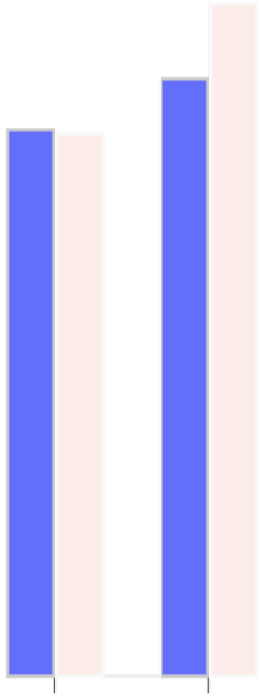
(a) Mascot

```
<g class="barlayer mlayer">
  <g class="trace bars" style="opacity: 1;">
    <path d="M36.55,260V59.32H182.75V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(99, 110, 250); fill-opacity: 1;"/>
    <path d="M402.05,260V40.64H548.25V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(99, 110, 250); fill-opacity: 1;"/>
  </g>
  <g class="trace bars" style="opacity: 1;">
    <path d="M182.75,260V60.73H328.95V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(239, 85, 59); fill-opacity: 1;"/>
    <path d="M548.25,260V13H694.45V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(239, 85, 59); fill-opacity: 1;"/>
  </g>
</g>
```

(b) Plotly

```
<g class="mark-rect role-mark marks" role="graphics-object" aria-roledescription="rect mark container">
  <path aria-label="sex: Female; value: 18.10519; smoker: no" role="graphics-symbol" aria-roledescription="bar" d="M10.000000000000007,49.123416666666664h20v150.87658333333334h-20Z" fill="#4c78a8"/>
  <path aria-label="sex: Female; value: 17.97788; smoker: yes" role="graphics-symbol" aria-roledescription="bar" d="M30.000000000000007,50.18433333333334h20v149.81566666666666h-20Z" fill="#f58518"/>
  <path aria-label="sex: Male; value: 19.79124; smoker: no" role="graphics-symbol" aria-roledescription="bar" d="M60.00000000000001,35.07300000000002h20v164.92699999999996h-20Z" fill="#4c78a8"/>
  <path aria-label="sex: Male; value: 22.2845; smoker: yes" role="graphics-symbol" aria-roledescription="bar" d="M80,14.295833333333333h20v185.70416666666668h-20Z" fill="#f58518"/>
</g>
```

(c) Vega-Lite



```
▼<g id="collection17" class="collection17">
  ▼<g id="collection16" class="collection16">
     <rect id="rect24" class="rect24" x="100" y="137.50867194686887" width="16" height="162.49132805313113" style="fill: rgb(99, 110, 250); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
     <rect id="rect25" class="rect24" x="117" y="138.65125984428636" width="16" height="161.34874015571364" style="fill: rgb(239, 85, 59); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
   </g>
  ▼<g id="collection18" class="collection16">
     <rect id="rect26" class="rect24" x="153" y="122.37662949583793" width="16" height="177.62337050416207" style="fill: rgb(99, 110, 250); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
     <rect id="rect27" class="rect24" x="170" y="100" width="16" height="200" style="fill: rgb(239, 85, 59); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
   </g>
 </g>
```

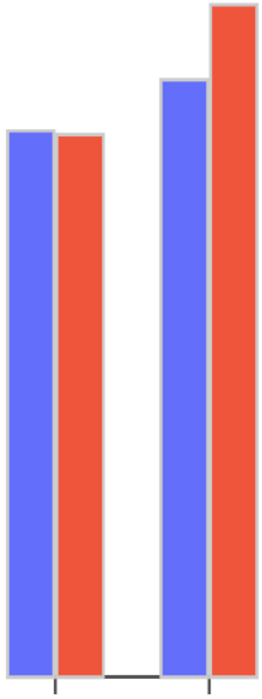
(a) Mascot

```
▼<g class="barlayer mlayer">
 ▼<g class="trace bars" style="opacity: 1;">
    <path d="M36.55,260V59.32H182.75V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(99, 110, 250); fill-opacity: 1;"/>
    <path d="M402.05,260V40.64H548.25V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(99, 110, 250); fill-opacity: 1;"/>
  </g>
 ▼<g class="trace bars" style="opacity: 1;">
    <path d="M182.75,260V60.73H328.95V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(239, 85, 59); fill-opacity: 1;"/>
    <path d="M548.25,260V13H694.45V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(239, 85, 59); fill-opacity: 1;"/>
   </g>
 </g>
```

(b) Plotly

```
▼<g class="mark-rect role-mark marks" role="graphics-object" aria-roledescription="rect mark container">
   <path aria-label="sex: Female; value: 18.10519; smoker: no" role="graphics-symbol" aria-roledescription="bar" d="M10.000000000000007,49.123416666666664h20v150.87658333333334h-20Z" fill="#4c78a8"/>
   <path aria-label="sex: Female; value: 17.97788; smoker: yes" role="graphics-symbol" aria-roledescription="bar" d="M30.000000000000007,50.18433333333334h20v149.81566666666666h-20Z" fill="#f58518"/>
   <path aria-label="sex: Male; value: 19.79124; smoker: no" role="graphics-symbol" aria-roledescription="bar" d="M60.00000000000001,35.07300000000002h20v164.92699999999996h-20Z" fill="#4c78a8"/>
   <path aria-label="sex: Male; value: 22.2845; smoker: yes" role="graphics-symbol" aria-roledescription="bar" d="M80,14.295833333333333h20v185.70416666666668h-20Z" fill="#f58518"/>
 </g>
```

(c) Vega-Lite



```
▼<g id="collection17" class="collection17">
  ▼<g id="collection16" class="collection16">
     <rect id="rect24" class="rect24" x="100" y="137.50867194686887" width="16" height="162.49132805313113" style="fill: rgb(99, 110, 250); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
     <rect id="rect25" class="rect24" x="117" y="138.65125984428636" width="16" height="161.34874015571364" style="fill: rgb(239, 85, 59); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
   </g>
  ▼<g id="collection18" class="collection16">
     <rect id="rect26" class="rect24" x="153" y="122.37662949583793" width="16" height="177.62337050416207" style="fill: rgb(99, 110, 250); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
     <rect id="rect27" class="rect24" x="170" y="100" width="16" height="200" style="fill: rgb(239, 85, 59); stroke: rgb(204, 204, 204); stroke-width: 1; stroke-dasharray: none;"/>
   </g>
 </g>
```

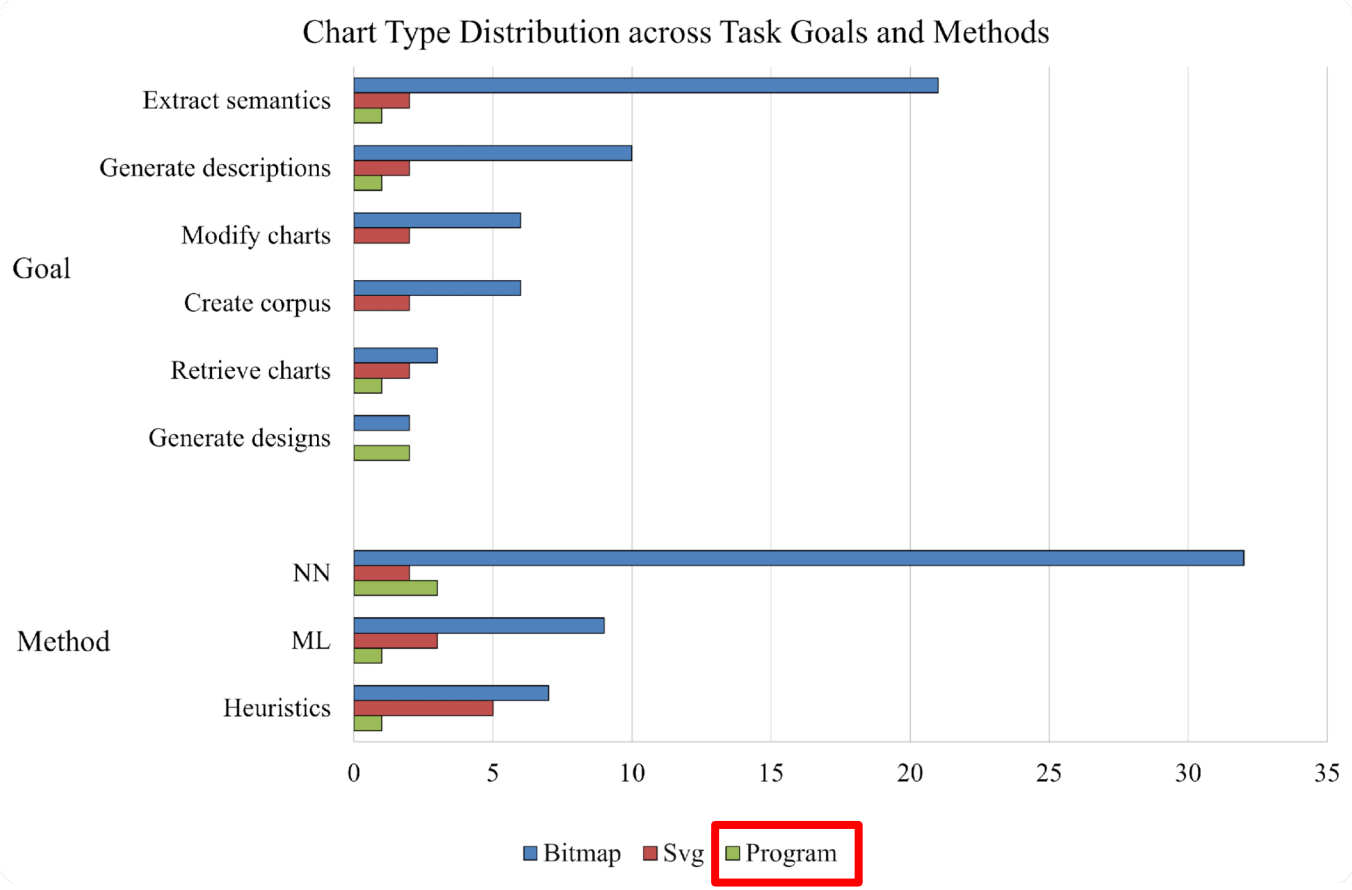
(a) Mascot

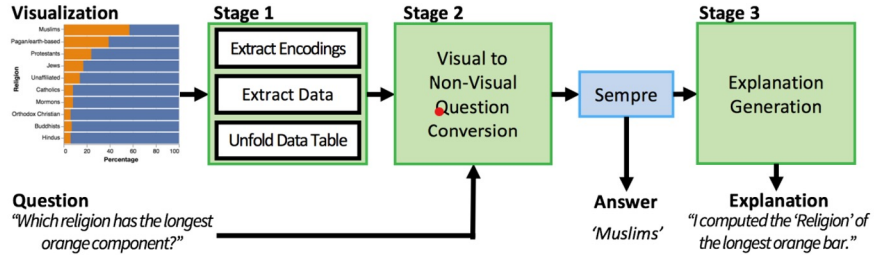
```
▼<g class="barlayer mlayer">
  ▼<g class="trace bars" style="opacity: 1;">
     <path d="M36.55,260V59.32H182.75V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(99, 110, 250); fill-opacity: 1;"/>
     <path d="M402.05,260V40.64H548.25V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(99, 110, 250); fill-opacity: 1;"/>
   </g>
  ▼<g class="trace bars" style="opacity: 1;">
     <path d="M182.75,260V60.73H328.95V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(239, 85, 59); fill-opacity: 1;"/>
     <path d="M548.25,260V13H694.45V260Z" style="vector-effect: non-scaling-stroke; opacity: 1; stroke-width: 0px; fill: rgb(239, 85, 59); fill-opacity: 1;"/>
   </g>
 </g>
```

(b) Plotly

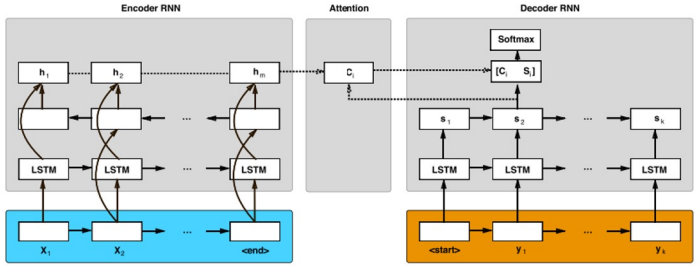
```
▼<g class="mark-rect role-mark marks" role="graphics-object" aria-roledescription="rect mark container">
   <path aria-label="sex: Female; value: 18.10519; smoker: no" role="graphics-symbol" aria-roledescription="bar" d="M10.000000000000007,49.123416666666664h20v150.87658333333334h-20Z" fill="#4c78a8"/>
   <path aria-label="sex: Female; value: 17.97788; smoker: yes" role="graphics-symbol" aria-roledescription="bar" d="M30.000000000000007,50.18433333333334h20v149.81566666666666h-20Z" fill="#f58518"/>
   <path aria-label="sex: Male; value: 19.79124; smoker: no" role="graphics-symbol" aria-roledescription="bar" d="M60.00000000000001,35.07300000000002h20v164.92699999999996h-20Z" fill="#4c78a8"/>
   <path aria-label="sex: Male; value: 22.2845; smoker: yes" role="graphics-symbol" aria-roledescription="bar" d="M80,14.29583333333333h20v185.70416666666668h-20Z" fill="#f58518"/>
 </g>
```

(c) Vega-Lite

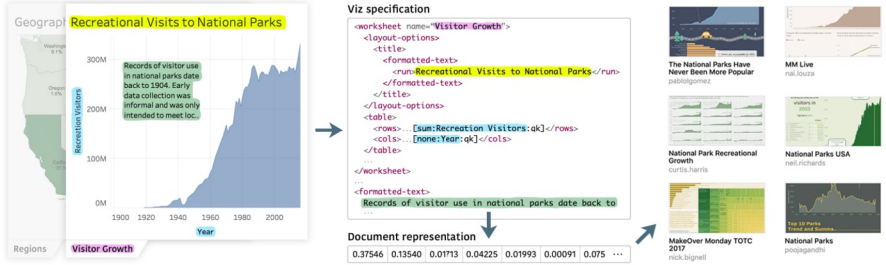



Kim et al. [KHA20]

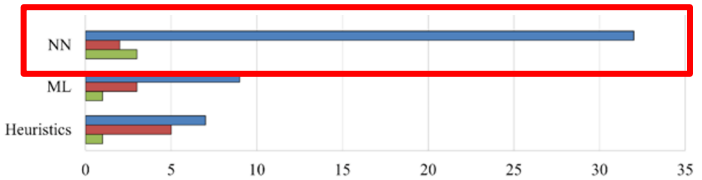



Dibia and Demiralp [DD19]

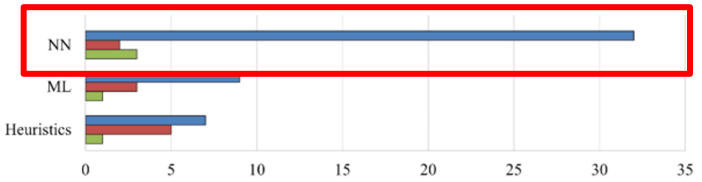


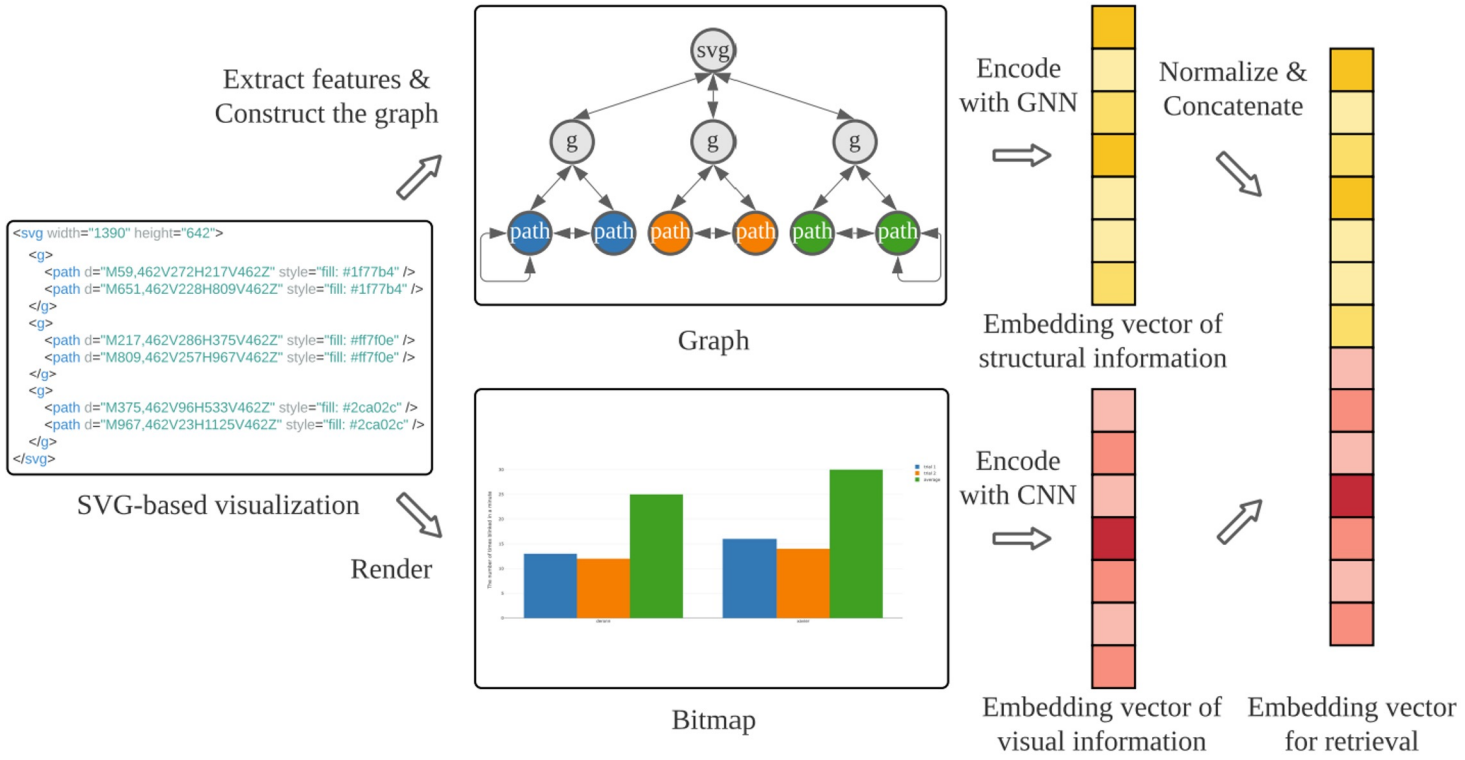

Oppermann et al. [OKM20]

Bitmap charts are naturally compatible with CNNs.



# Bitmap charts are naturally compatible with CNNs. SVGs could be better analyzed with GNNs.




Li et al. *Structure-aware Visualization Retrieval*

Assumptions or inclusion criteria during the chart selection process. They are usually specified to constrain the research problem space to achieve feasible solutions.

Assumptions or inclusion criteria during the chart selection process. They are usually specified to constrain the research problem space to achieve feasible solutions.

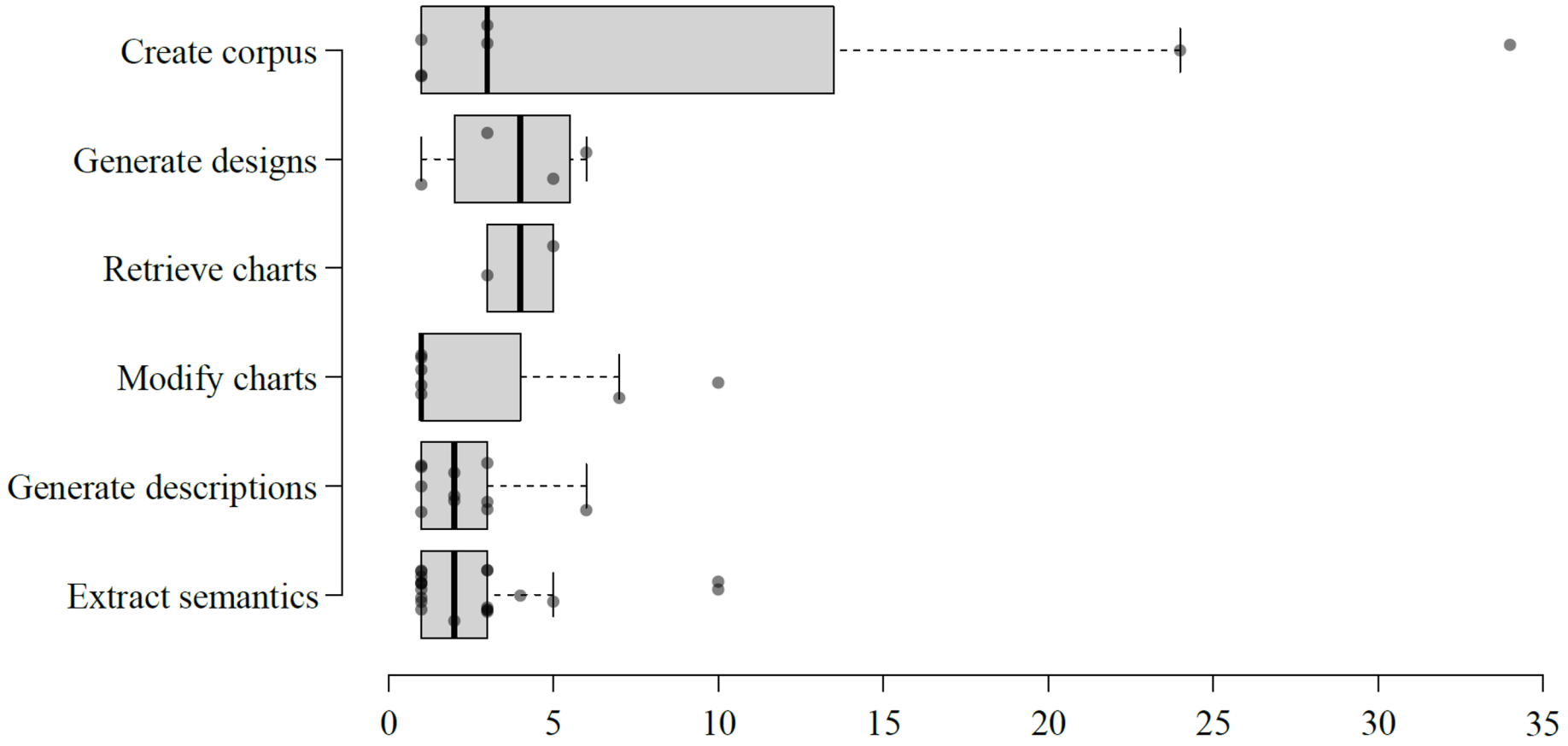
We identified two kinds of scopes:

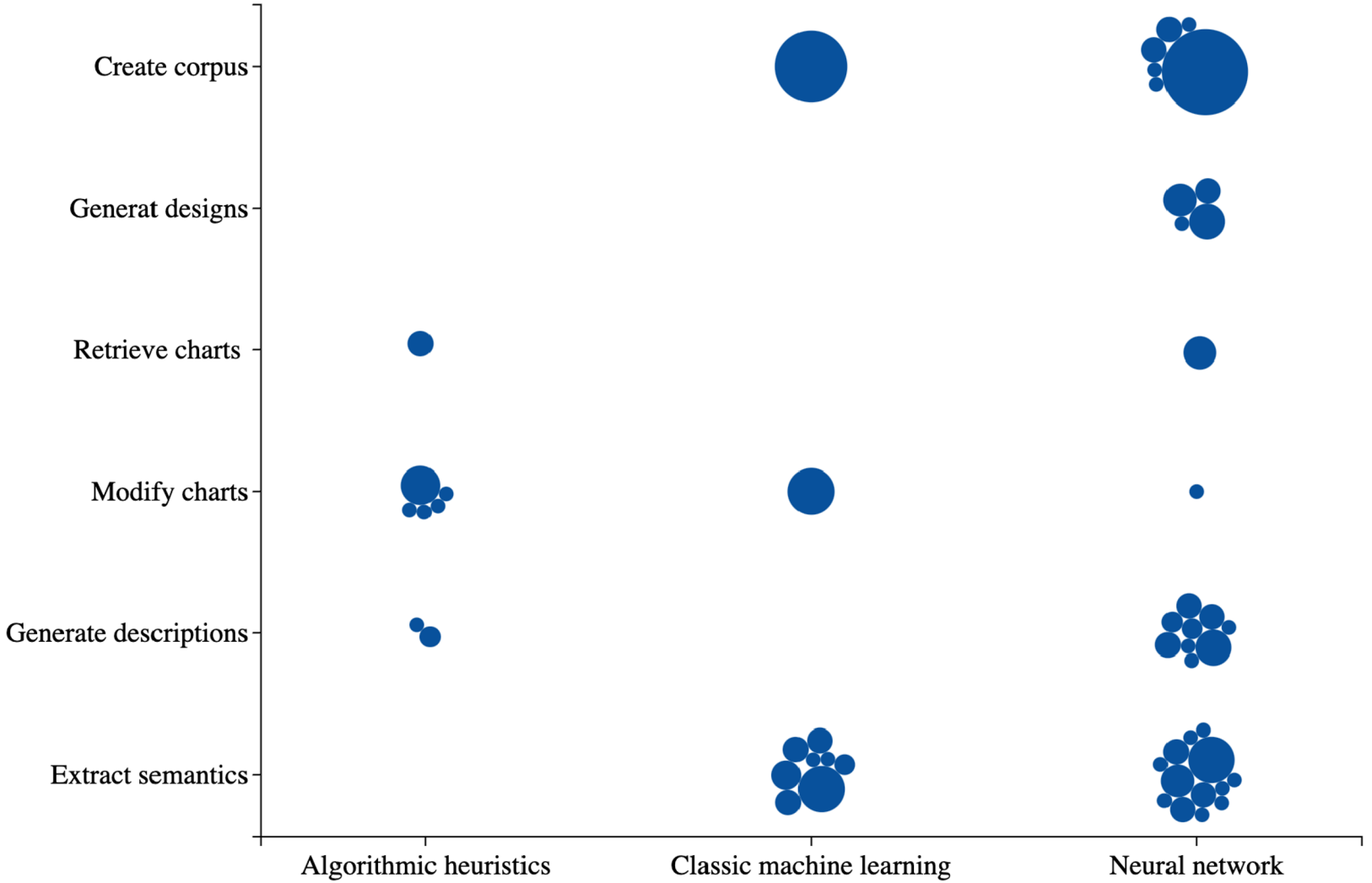
- Chart type
- Design variation

| Chart Type | Frequency | Percentage |
|---|---|---|
| Bar | 38 | 73.07% |
| Line | 31 | 59.62% |
| Pie | 18 | 34.62% |
| Scatterplot | 16 | 30.77% |
| Infographics | 9 | 17.31% |
| Area | 9 | 17.31% |
| Map | 8 | 15.38% |
| Treemap | 4 | 7.69% |
| Boxplot | 4 | 7.69% |
| Heatmap | 3 | 5.77% |
| Table | 3 | 5.77% |
| Venn | 3 | 5.77% |
| Parallel Coordinate | 3 | 5.77% |
| Sunburst | 3 | 5.77% |
| Donut | 3 | 5.77% |
| Node-link Diagram | 3 | 5.77% |
| Radar | 2 | 3.85% |
| Matrix | 2 | 3.85% |
| Tick | 2 | 3.85% |
| Pareto | 2 | 3.85% |

| Chart Type | Frequency | Percentage |
|---|---|---|
| Bar | 38 | 73.07% |
| Line | 31 | 59.62% |
| Pie | 18 | 34.62% |
| Scatterplot | 16 | 30.77% |
| Infographics | 9 | 17.31% |
| Area | 9 | 17.31% |
| Map | 8 | 15.38% |
| Treemap | 4 | 7.69% |
| Boxplot | 4 | 7.69% |
| Heatmap | 3 | 5.77% |
| Table | 3 | 5.77% |
| Venn | 3 | 5.77% |
| Parallel Coordinate | 3 | 5.77% |
| Sunburst | 3 | 5.77% |
| Donut | 3 | 5.77% |
| Node-link Diagram | 3 | 5.77% |
| Radar | 2 | 3.85% |
| Matrix | 2 | 3.85% |
| Tick | 2 | 3.85% |
| Pareto | 2 | 3.85% |

| Chart Type | Frequency | Percentage |
|---|---|---|
| Bar | 38 | 73.07% |
| Line | 31 | 59.62% |
| Pie | 18 | 34.62% |
| Scatterplot | 16 | 30.77% |
| Infographics | 9 | 17.31% |
| Area | 9 | 17.31% |
| Map | 8 | 15.38% |
| Treemap | 4 | 7.69% |
| Boxplot | 4 | 7.69% |
| Heatmap | 3 | 5.77% |
| Table | 3 | 5.77% |
| Venn | 3 | 5.77% |
| Parallel Coordinate | 3 | 5.77% |
| Sunburst | 3 | 5.77% |
| Donut | 3 | 5.77% |
| Node-link Diagram | 3 | 5.77% |
| Radar | 2 | 3.85% |
| Matrix | 2 | 3.85% |
| Tick | 2 | 3.85% |
| Pareto | 2 | 3.85% |


Create corpus
Generate designs
Retrieve charts
Modify charts
Generate descriptions
Extract semantics
0
5
10
15
20
25
30
35


Create corpus
Generat designs
Retrieve charts
Modify charts
Generate descriptions
Extract semantics
Algorithmic heuristics
Classic machine learning
Neural network

| Design Variation Type | Assumption | Relevant Corpora |
|---|---|---|
| composite arrangement | only multiple-view charts | [CZL*20] |
| | no multiple-view charts | [CAM*18, WTD*20, PH17, HGH21, LWL*20] |
| | no layered charts | [JKS*17, PH17, CJP*19] |
| mark and glyph | no abstract icons or symbols | [JKS*17] |
| | only proportion-related charts | [QSC*20, CZW*19] |
| | only timeline-related infographics | [CWW*19] |
| | no handmade sketches | [JKS*17, CAM*18, SDHL15] |
| | no 3D effects | [SKC*11, DWS*22, CJP*19] |
| chart component | chart must have a legend | [PMH17, MKT22] |
| | axes being at the left and bottom | [SKC*11] |
| coordinate space | in Cartesian coordinate space | [WTD*20, PH17] |

Collection method describes how charts in a corpus were collected. We have observed four kinds of collection methods.

- reusing and transforming existing corpus
- web crawling
- manual curation
- computer-aided generation

Out of the 56 corpora:

- 17 are publicly available
- 9 were generated by modifying existing corpora
- only 4 corpora (FigureQA [KAM*18], VIF [LWL*20], SciCap [HGH21], REV [PH17]) were reused in subsequent works

Out of the 56 corpora:

- 17 are publicly available
- 9 were generated by modifying existing corpora
- only 4 corpora (FigureQA [KAM*18], VIF [LWL*20], SciCap [HGH21], REV [PH17]) were reused in subsequent works

Transformations in those 9 corpora:

- Adding new charts to increase size or diversity
- Adding new annotations to support new tasks

Gather charts matching certain criteria from targeted sources automatically.

Gather charts matching certain criteria from targeted sources automatically.

Web crawling sources:

- Search engines e.g., Google Image Search
- Galleries of online charting tools, e.g., Tableau, Vega-Lite
- Public documented materials, e.g., online Excel sheets
- Public scholarly repositories, e.g., DBLP, Semantic Scholar
- Public data sharing platforms, e.g., the Pew research, Our World In Data

When the quality and variation of chart design matters more than the size of a corpus, some works such as Cui et al. [CZW*19] decided to collect charts manually.

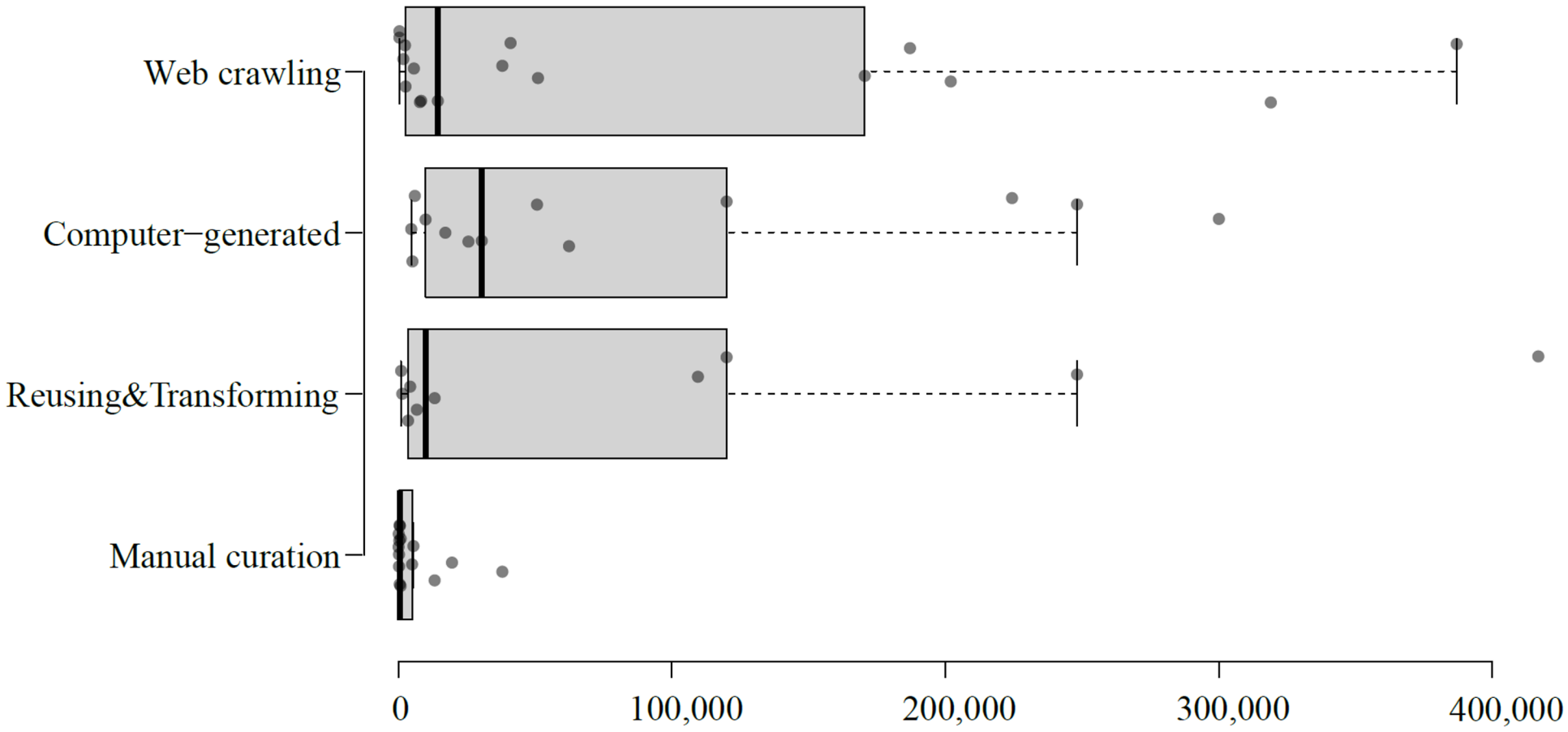
When the quality and variation of chart design matters more than the size of a corpus, some works such as Cui et al. [CZW*19] decided to collect charts manually.

Advantages:

- they could inspect each chart candidate and decide if they would like to include it in the corpus
- The sources for web crawling are still useful

To use computer charting tools to generate a corpus.

- Underlying datasets
  - Synthetic ones by varying types and distributions
  - Real ones online, e.g., World Development Indicators
- Charting tools
  - Matplotlib
  - Vega-Lite
  - Bokeh
  - GeoPandas
  - Timeline Storyteller


Web crawling
Computer−generated
Reusing&Transforming
Manual curation
0
100,000
200,000
300,000
400,000

Annotations are labels associated with charts in a corpus

- serving as ground truth for chart analysis tasks
- the sources where the charts are collected usually do not provide needed labels
- as reported in Battle et al. [BDM*18], there is a lack of consistent metadata across different sources

| Annotation Type | Relevant Corpora |
|---|---|
| bounding box | for mark or glyph [LLJ*20, LLWL21, LWL21, CRMY17, CSG*20, QSC*20, HWWL21], for legend [LWL21, CSG*20, MGKK20, HWWL21], for axes [MGKK20, CRMY17, HWWL21], for text [LWL21, ZZC*21, PH17, CRMY17, SGCV19, CSG*20, MGKK20, SKC*11, HWWL21], for main chart area [LLWL21, DWS*22, HWWL21], for chart sub-views [CZL*20] |
| chart type | [BDM*18, JKS*17, TLL*16, CAM*18, KM18, SKC*11, CWG16, GZB12, DWS*22, CJP*19] |
| question-answer pair | [KHA20, MDT*22, KPCK18, KAM*18, MBT*22, CSG*20, CPL*22, MGKK20] |
| question-caption pair | [CZK*19, MKT22] |
| text role | [ZZC*21, CWG16, PH17] |
| infographics element type | [LWL*20, QSC*20] |
| pairwise style similarity | [SDHL15, MTW*18] |
| saliency map | [BKO*17] |
| aesthetics ranking | [FWD*19] |

In-house labeling: in-person process where a small group of people gathers to annotate charts manually; usually two considerations:

- user interface for annotation
- training procedure

In-house labeling: in-person process where a small group of people gathers to annotate charts manually; usually two considerations:

- user interface for annotation
- training procedure

Crowdsourcing: online process where workers from platforms such as Amazon's MTurk are recruited to annotate charts

Template-based generation: question-answer and chart-caption annotations can be generated based on pre-defined templates

- Compared to crowdsourcing, avoids high expenses, but lacks rich linguistic variation

Template-based generation: question-answer and chart-caption annotations can be generated based on pre-defined templates

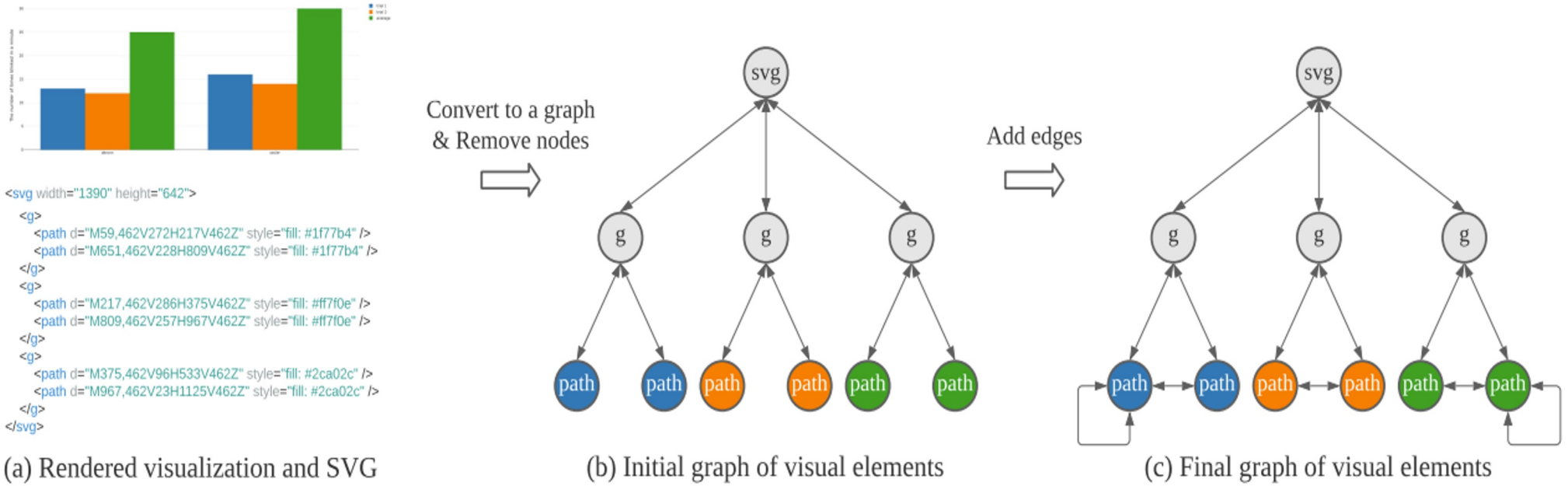
- Compared to crowdsourcing, avoids high expenses, but lacks rich linguistic variation

Automatic extraction: Application-specific API (e.g., bounding boxes in Matplotlib, data values in Excel)

Diversity measures how much the charts differ from one another.

Why it is important?

- diversity is an under-explored property that significantly influence the scalability, generalizability, and robustness of developed techniques




Li et al. *Structure-aware Visualization Retrieval*

- source websites: to collect charts from multiple sources
- chart topics: to collect charts on various topics
- chart creators: to sample from online providers whose charts are created by a larger community of content creators
- scholarly document repositories: to enrich publication venues and increasing the year range

- underlying datasets: to generate a variety of diverse synthetic datasets for plotting
- style parameters: to enumerate the style parameters in the code
- visual questions and captions: to alleviate poor linguistic variations from templates
  - Design more diverse templates [CZK*19, SS20]
  - Combine crowdsourcing, templates, and in-house labeling [MGKK20]
  - Adopt large language models [MDT*20]

## Feature tags can be a better way

**beyond chart types**

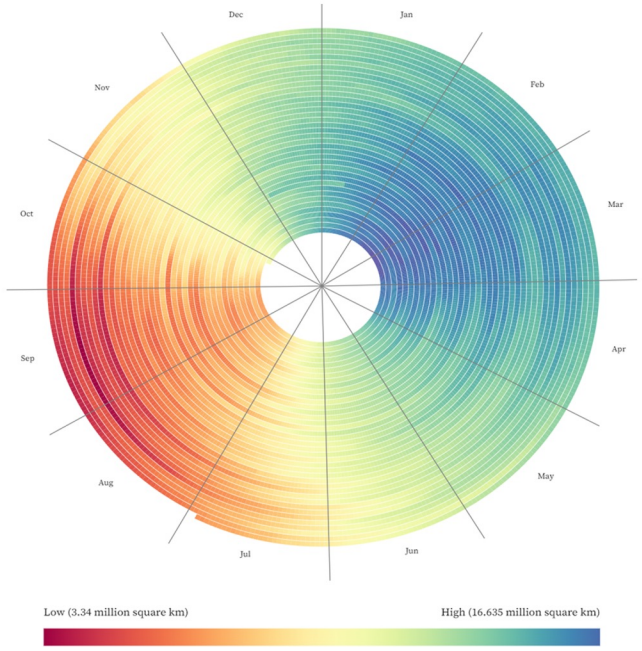
beyond chart similarity

tool/source-agnostic chart analysis

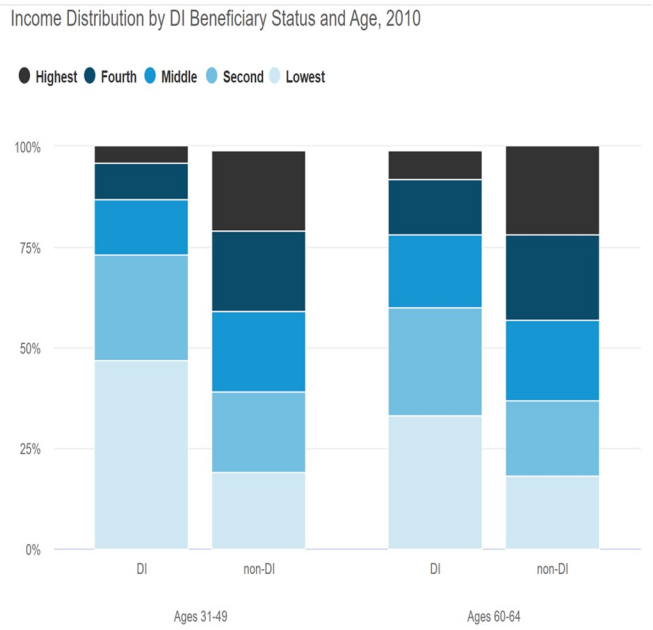
intelligent design generation

diversity quantification

understanding interaction&animation




(a) Spiral Heatmap




(b) Grouped Stacked Bar Chart

beyond chart types

**beyond chart similarity**

tool/source-agnostic chart analysis

intelligent design generation

diversity quantification

understanding interaction&animation

Chart quality (aesthetics, effectiveness, memorability, ...) is under-explored

*"when visualization creators are seeking design ideas, similarity may not be their primary desired criterion; instead, they prefer alternative or bespoke designs to broaden the scope of consideration"*

Bako et al. *Understanding how Designers Find and Use Data Visualization Examples*

beyond chart types

beyond chart similarity

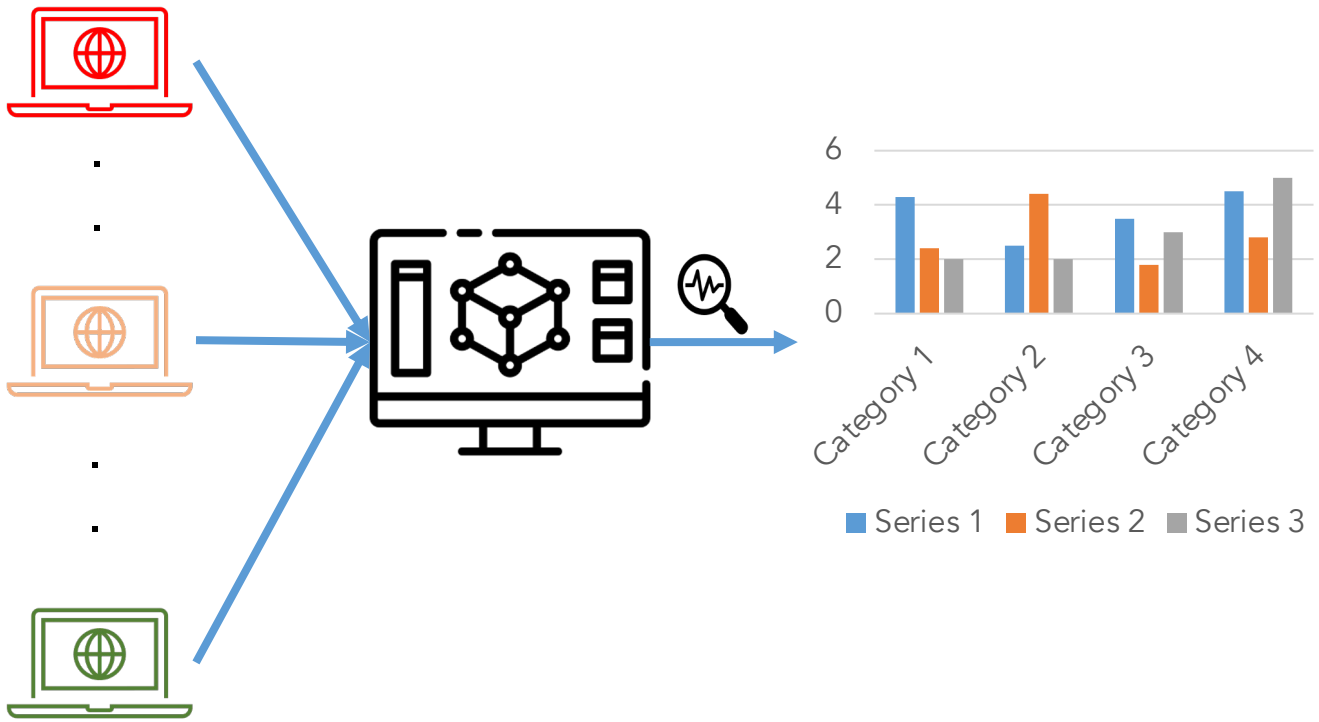
**tool/source-agnostic chart analysis**

intelligent design generation

diversity quantification

understanding interaction&animation

## Enhanced generalizability



beyond chart types

beyond chart similarity

tool/source-agnostic chart analysis

**intelligent design generation**

diversity quantification

understanding interaction&animation

Current research on automatic generation of chart design mostly relies on a **program** chart corpus from a **single charting tool**.

It is worth thinking that what techniques are required to automate design ideas from various sources if we would like to include bitmaps and SVGs.

beyond chart types

beyond chart similarity

tool/source-agnostic chart analysis

intelligent design generation

**diversity quantification**

understanding interaction&animation

No metric to measure diversity

- There is a clear need for systematic methods to evaluate chart diversity
- Such methods can guide chart selection processes

beyond chart types

beyond chart similarity

tool/source-agnostic chart analysis

intelligent design generation

diversity quantification

**understanding interaction&animation**

## Interaction & Animation

- the logic for interactions and animations is usually hidden
- How to automatically capture them remains open

## Enhanced Chart Diversity within a Corpus

- little effort in enhancing diversity in terms of chart format

## Enhanced Chart Diversity within a Corpus

- little effort in enhancing diversity in terms of chart format

## Multi-level Fine-Grained Annotations

- the lack of fine-grained annotations makes it difficult to reuse existing corpora for new tasks

## Enhanced Chart Diversity within a Corpus

- little effort in enhancing diversity in terms of chart format

## Multi-level Fine-Grained Annotations

- the lack of fine-grained annotations makes it difficult to reuse existing corpora for new tasks

## Interactivity and Animation Understanding

- semantic abstract for describing them & methods for capturing and understanding them

## Smart Web Crawler

- The implementation handles composition arrangements in HTMLs which can be random

## Smart Web Crawler

- The implementation handles composition arrangements in HTMLs which can be random

## Chart Pre-Processor

- Collected charts shall be processed, e.g., resizing for bitmaps and unifying semantics for SVGs

## Smart Web Crawler

- The implementation handles composition arrangements in HTMLs which can be random

## Chart Pre-Processor

- Collected charts shall be processed, e.g., resizing for bitmaps and unifying semantics for SVGs

## Mix-Initiative Annotating System

- Dedicated research in human-AI collaboration for the annotation process is necessary

# Thanks for listening!

EUROVIS Leipzig 2023

## Corpus Properties

format, scope, collection method, annotations, diversity

## Opportunities

- Beyond chart type
- Beyond chart similarity
- Tool/Source-agnostic analysis
- Design generation
- Diversity quantification
- Interaction & Animation

## Towards Benchmark Corpora

- Properties
  - Diversity
  - Multi-level Fine-Grained Annotations
  - Interactivity and Animation Understanding
- Needed tools
  - Smart web crawler
  - Smart image processor
  - Mix-initiative annotating system

**Contact: cchen24@umd.edu**