Fast-forwarding to Desired Visualizations with zenvisage

Aditya Parameswaran
Assistant Professor
Computer Science
University of Illinois

http://data-people.cs.illinois.edu

The Emergence of Interactive Visualization Tools

Now billions of $$$ of revenue/year!
Data Visualization Tools

- Billions in revenue
- Huge audience
- Interactions not code

Data Visualization is Data Science for the 99%!

Deriving insights is laborious and time-consuming!

- Decrease in insights
- Decrease in exploration
- Increase in errors
- Increase in frustration
- Increase in wasted time
**Standard Data Visualization Recipe:**

1. **Load** dataset into data viz tool
2. **Start** with a desired hypothesis/pattern
3. **Select** viz to be generated
4. **See** if it matches desired pattern
5. **Repeat** 3-4 until you find a match
Laborious and Time-consuming!

Key Issue:
Visualizations generated by varying
• subsets of data, and
• attributes visualized

Too many visualizations to look at!
Lessons from History: Use Automation!

“Astronomers surely will not have to continue to exercise the patience which is required for computation. It is this that deters them from... working on hypotheses and from discussion of observations... For it is unworthy of excellent men to lose hours like slaves in the labor of calculation which could be safely relegated (to) machines.”

[Gottfried Leibniz, 1700s]

“...intolerable labor and fatiguing monotony of a continued repetition of similar calculations representing the lowest occupation of human intellect”

[Charles Babbage, 1800s]

Source: “The Information” by James Gleick, highly recommended!
Key Insight: Automation

We can automate that!

Desiderata for automation:

• **Expressive** – specify what you want
• **Interactive** – interact with results, don’t require programming
• **Scalable** – get interesting results quickly

Drawing from DB, DM, HCI

Enter Zenvisage:
(zen + envisage: to effortlessly visualize)
Overview: Demo!

Apologies: my demo will be on a real-estate dataset...

Hopefully highlights some of the functionalities...

I will show a demo on the DES data at the end...
Overview: Backup
Two Interoperable Modes

• **First Mode**: Interactions, drawing, drag-and-drop
  – Simple needs
  – Starting point / context

• **Second Mode**: the Zenvisage Query Language (ZQL)
  – Sophisticated needs
  – Multiple steps
  – SQL : data :: ZQL : visualizations

*Can switch back and forth, as user needs evolve*

Both modes developed after many iterations
ZQL: High Level Overview

ZQL is a visualization exploration language

Captures four key operations on viz collections

- Compose
- Filter
- Compare
- Sort

Two types of ZQL operators (pictorial representation)

Compose Op

Processing Op

User-specified, with system defaults.

X1 = year
Y1 = soldprice
Z1 = city.*
Example 1: Comparisons

Find the states where the *soldprice* trend is most similar to the *soldpricepersqft* trend.

\[
\begin{align*}
X1 &= \text{year} \\
Y1 &= \text{soldprice} \\
Z1 &= \text{state.}* \\
X2 &= \text{year} \\
Y2 &= \text{soldpricepersqft} \\
Z1 \\
X3 &= \text{year} \\
Y3 &= \text{soldprice} \\
Z2
\end{align*}
\]
Example 1: Comparisons
Example 2: Explanations/Diffs
Find visualizations on which the *states of CA* and *NY* are most different.

\[
X_1 = *
\]
\[
Y_1 = *
\]
\[
Z_1 = \text{state.}'CA'\]

\[
X_2, Y_2 = \text{ArgDissimilar}_*(X_1, Y_1)(0, 0)
\]
\[
Z_2 = \text{state.}'CA', 'NY'\]

\[
X_1 \quad \quad \quad \quad Y_1 \quad \quad \quad \quad Z_2 = \text{state.}'NY'\]
Example 2: Explanations/Diffs
Example 3: Drill-downs

Find cities in NY where the trend for soldprice is most different from the overall NY trend.
Example 3: Drill-downs
Other Examples

- **Drill-down**: find NY cities with an opposite soldprice trend to the NY state trend
- **Exceptions**: find cities that have an increasing soldprice trend with an increasing foreclosure ratio trend
- **Correlations**: find attributes that have a similar trend to soldprice for NY cities
- **Pivot**: for cities similar to NYC on soldprice trends, find typical trends for foreclosure ratios
ZQL Query Execution

Leverage a relational database as a backend

Naïve translation approach:

Repeat:
  Generate each visualization via a SQL query:
  \[
  \text{SELECT } X, F(Y) \text{ FROM } R \text{ WHERE } Z=z \text{ GROUP BY } X
  \]
  Apply processing on result

Often 1000s of SQL queries issued!
\[\rightarrow \text{wasteful, extremely high latency}\]

Solution: Sharing, Speculation, Parallelism, Pruning
\[= 100-500x \text{ Reduction}\]
Effortless Visual Exploration of Large Datasets with

Ingredients

- Drag-and-drop & sketch interactions
- Sophisticated visual expl. language, ZQL
- ZQL optimization engine: SmartFuse
- Perceptually-aware pattern matching algorithms

More demo next!
Human-in-the-loop Data Analysis Tools

Increasing sophistication of analysis

Share & Collaborate
OrpheusDB
orpheus-db.github.io

Play & View
zenvisage.github.io

Touch & Feel
dataspread.github.io

Please consider using or contributing!
http://data-people.cs.illinois.edu; adityagp@twitter
Backup Slides
ZQL: A Bird’s Eye View

Output spec and identifiers

Composition of visualizations, often using values from previous steps

Sorting, comparing, and filtering visualizations

<table>
<thead>
<tr>
<th>Name</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>Constraints</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>'quarter'</td>
<td>'soldprice'</td>
<td>'metro'. 'Peoria'</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 1: Comparisons

Find the states where the soldprice trend is most similar to the soldpricepersqft trend.

➔ Comparing a pair of y-axes for different “z”
Example 2: Explanations/Diffs

Find visualizations on which the states of CA and NY are most different.

→ Comparing across different “x”, “y” for two “z”
Example 3: Drill-downs

Find *cities in NY* where the trend for *soldprice* is most different from (or most similar to) the *overall NY trend*.

➔ *Comparing across different granularities of “z”*
SmartFuse: Intelligent Query Optimizer
[VLDB’16,’17,’18]

**Sharing**
1) SELECT X, F(Y) WHERE Z=z GRP. BY X
SELECT X, F(Y), Z GRP. BY X, Z ORDER BY Z
2) GRP. BY X1
GRP. BY X2

**Speculation**
Consider all outcomes of
e.g., find sales for products whose profit is decreasing

**Pruning**
Use to avoid generating visualizations
e.g., products whose profit curves won’t meet

Multiplicative increase in # of groups!

NP-Hard (partition)

Realistic Workload Speedup
~100x
~20x
~2-3x
~5x
User Study Takeaways (20 Participants)

Faster \[ \mu = 115s, \sigma = 51.6 \text{ vs. } \mu = 172.5s, \sigma = 50.5 \]

More accurate \[ \mu = 96.3\%, \sigma = 5.82 \text{ vs. } \mu = 69.9\%, \sigma = 13.3 \]

“In Tableau, there is no pattern searching. If I see some pattern in Tableau, such as a decreasing pattern, and I want to see if any other variable is decreasing in that month, I have to go one by one to find this trend. But here I can find this through the query table.”

“you can just [edit] and draw to find out similar patterns. You'll need to do a lot more through Matlab to do the same thing.”

“The obvious good thing is that you can do complicated queries, and you don't have to write SQL queries... I can imagine a non-cs student [doing] this.”
Example 4