A Hodgepodge of Visualization Research: Provenance, User Studies, Misinformation

Alexander Lex
http://vdl.sci.utah.edu
Visualization = Human Data Interaction
The purpose of computing is insight, not numbers. [Card, Mackinlay, Shneiderman]

[Richard Wesley Hamming]
Banana  *M. acuminata*

Date  *P. dactylifera*

Cress  *Arabidopsis thaliana*

Rice  *Oryza sativa*

Sorghum  *Sorghum bicolor*

Brome  *Brachypodium distachyon*
A
B
C

Must

Must Not

Universal Set
THE BANANA CHART REDESIGNED: UPSET

Largest Intersection Includes All Sets
THE BANANA CHART REDESIGNED: UPSET

Three Leftmost Species Are Most Similar
THE BANANA CHART REDESIGNED: UPSET

Rightmost species is most different
The canonical way to show set data with > 3 sets
Second-most cited VIS paper of the last decade
Available in various languages: http://upset.app/
TECHNICAL CONTRIBUTIONS

Interaction in Notebooks

EMPIRICAL & THEORETICAL WORK

Visual Misinformation

INFRASTRUCTURE

Evaluating Complex Systems
PERSISTENT AND REUSABLE INTERACTIONS

... IN COMPUTATIONAL NOTEBOOKS

Kiran Gadhave, Zach Cutler

[EuroVis 2024]
PERSISTENT AND REUSABLE INTERACTIONS IN COMPUTATIONAL NOTEBOOKS

A Jupyter Plugin

https://vdl.sci.utah.edu/persist/

pip install persist_ext

Demo Notebook: https://tinyurl.com/5db7nynn
Avalanches are a major hazard in Utah. The Utah Avalanche Center collects data about avalanches, including where it occurred (location, elevation), how it occurred, how big it was, etc.
Supposed you’re doing data analysis in Python

What’s the pandas code…

• …to change the order of columns?
• …to drop a column?
• …to change the label of a column?

Nothing here is hard, but it’s annoying.
PERSIST MAKES THIS EASY
Have you ever plotted something and wished you could just “fix” things as you spot them?

**What is this talk about?**

- Elevation where the avalanche occurred
- How deep (big) the avalanche was
- Utah’s Elevation Range

Elevation where the avalanche occurred
PERSIST MAKES THIS EASY

```python
PR.plot.scatterplot(avalanches, "Elevation_feet\:Q", "Vertical_inches\:Q", df_name="avalanches")
```
Lots of **vis tools** support these operations

Most **data wrangling happens in code**: it’s just more powerful

**Opportunity**: bring **interactive operations to code**!

Persist works **INSIDE** your Jupyter Notebook
BRIDGING BETWEEN DATA ANALYSIS MODALITIES
What are Modalities?

1. Interactive Vis
2. Code
INTERACTIVE VISUALIZATION: BENEFITS

Intuitive
Easy to use
Uses human perceptual capabilities
**Limited Expressivity**

Some operations are difficult

  e.g., conditional queries..

**Not reusable**

need to redo analysis when data changes

**Not reproducible**
Flexible and powerful
you basically can do anything
Reusable
if your data changes, re-run
Reproducible
everything is documented

CODE: BENEFITS

```python
# Keep this cell
avv_df = pd.read_csv('./avalanches.csv')

# Remove NaN coordinates
avv_df = avv_df[avv_df['Coordinates']!=avv_df['Coordinates']]

# Split into latitude & longitude
avv_df[['lat', 'lon']] = avv_df['Coordinates'].str.split(',', expand=True)

# Remove values outside of Utah bounds
avv_df = avv_df[(avv_df['lat'].astype('float')) & (avv_df['lon'].astype('float')) < 42]]
avv_df = avv_df[(avv_df['lat'].astype('float')) & (avv_df['lon'].astype('float')) > -110]

# Keep columns we need
avv_df = avv_df[['Date', 'Region', 'Trigger', 'lat', 'lon']]

avv_df.head()
```
CODE / SCRIPTING: **DOWNSIDES**

It’s hard
requires extensive training
reading documentation
not discoverable

It’s time consuming
even simple things require effort

Some operations are difficult
e.g., labeling data points
COMPUTATIONAL NOTEBOOKS: A MIDDLE GROUND?

Observable

R Markdown

Jupyter Notebooks
Yes!
Afford both scripting and interactive visualization
But visualizations are a dead end
can’t “use” interaction in code
e.g., changing a label, or filtering a value
GAPS BETWEEN CODE AND INTERACTIVE OUTPUTS*

**Semantic Gap**
- Information only flows from code to visualization

**Temporal Gap**
- Changes made to code are preserved
- Changes made to vis are lost

**Layout Gap**
- Changes in code are messy

*Wu, Hellerstein, Satyanarayan, UIST 2020*
THESIS: **BRIDGING BETWEEN CODE AND INTERACTIVE VIS** IS **USEFUL**

Easy handoffs are important!
THE PERSIT APPROACH
PRINCIPLE

Track events in interactive visualizations

Map them to data frame operations

Operations then applied to data frame
HOW IT WORKS

Code to create chart

```python
import pandas as pd

df = pd.read_csv('cars.csv')
PersistChart(scatterplot(df))
```

Provenance

- df1: Select 3 points
- df2: Remove selections
- df3: Select 2 points
- df4: Assign category v4 to selection
- df5: Select 3 points

Interactive Visualization

Dataframe updates

```
In code
```

<table>
<thead>
<tr>
<th>Name</th>
<th>MPG</th>
<th>Cylinders</th>
<th>origin</th>
<th>Engine</th>
<th>is_selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>ford</td>
<td>18</td>
<td>8</td>
<td>USA</td>
<td>NA</td>
<td>FALSE</td>
</tr>
<tr>
<td>dodge</td>
<td>15</td>
<td>8</td>
<td>USA</td>
<td>NA</td>
<td>FALSE</td>
</tr>
<tr>
<td>volkswagen</td>
<td>22</td>
<td>4</td>
<td>Europe</td>
<td>V4</td>
<td>FALSE</td>
</tr>
<tr>
<td>amc</td>
<td>16</td>
<td>8</td>
<td>USA</td>
<td>NA</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

27 rows x 8 columns
OPERATIONS

- Selection
- Edit column names, edit cells
- Sort rows/columns
- Drop columns
- Filter (in/out) items
- Label items
- Categorize items
- Change data types
PERSIST WORKFLOW
import altair as alt
from vega_datasets import data
import persist_ext as PR

cars_df = data.cars()  # Load cars dataset

brush = alt.selection_interval()  # Create a 2d brush

# create scatterplot and link to brush
scatterplot = alt.Chart(cars_df).mark_point().encode(
    x='Miles_per_Gallon:Q',
    y='Weight_in_lbs:Q',
    color=alt.condition(brush, alt.value('steelblue'), alt.value('gray'))
).add_params(
    brush
)

PR.PersistChart(scatterplot)
import altair as alt
from vega_datasets import data
import persistence as PR

cars_df = data.cars()  # Load cars dataset
brush = alt.selection_interval()  # Create a 2d brush

# create scatterplot and link to brush
scatterplot = alt.Chart(cars_df).mark_point().encode(
    x='Miles_per_Gallon:Q',
    y='Weight_in_lbs:Q',
    color=alt.condition(brush, alt.value("steelblue"), alt.value("gray"))
).add_params(brush)

PR.PersistChart(scatterplot)
import altair as alt
from vega_datasets import data
import persist_ext as PR

cars_df = data.cars()  # Load cars dataset
brush = alt.selection_interval()  # Create a 2d brush

# create scatterplot and link to brush
scatterplot = alt.Chart(cars_df).mark_point().encode(
    x="Miles_per_Gallon",
    y="Weight_in_lbs",
    color=alt.condition(brush, alt.value("steelblue"), alt.value("gray")))
.add_params(
    brush
)
PR.PersistChart(scatterplot)
Branches and **Choosing a State** in provenance support **non-linear analysis**, addressing the **layout gap**.

```python
import altair as alt
from vega_datasets import data
import persist_ext as PR

cars_df = data.cars() # Load cars dataset

brush = alt.selection_interval() # Create a 2d brush

# create scatterplot and link to brush
scatterplot = alt.Chart(cars_df).mark_point().encode(
    x='Miles_per_Gallon:Q',
    y='Weight_in_lbs:Q',
    color=alt.condition(brush, alt.value('steelblue'), alt.value('gray'))
).add_params(brush)

PR.PersistChart(scatterplot)
```
Branches and Choosing a State in provenance support non-linear analysis, addressing the layout gap.
Persist re-runs the interactions in the output, addressing the temporal gap.
Persist applies interactions to data frames that can be accessed in code, addressing the **semantic gap**

- **Manipulated data frame accessed in code**
  - count: 2392→79
VISUALIZATION OPTIONS

Arbitrary Vega-Altair Charts

An Interactive Data Table
Persist works with most Vega-Altair charts

“Listens” to native operations (selections)

Updates Vega charts:

Use original chart spec when possible (e.g., filters)

Update spec when necessary (categories, labels)
EXAMPLE CHARTS
DEMO
EVALUATION
IN-LAB STUDY

STUDY DESIGN
We recruited ELEVEN participants for the study. Participants all had prior experience with Python and Pandas.

- FULL FACTORIAL DESIGN
- 2 DATASETS X 2 CONDITIONS X

THE ORDER OF CONDITIONS WAS RANDOMLY ASSIGNED.
For each condition, datasets were randomly assigned. Participants never saw the same dataset twice.

STUDY METRICS
- SUBJECTIVE PERFORMANCE
- TIME
- ERROR
- REPRODUCIBILITY

CONDITIONS

PANDAS CODING
1. import pandas as pd
df = pd.read_csv(...)  
2. df = df.drop("AGE")
df = df.rename(
columns={"JOB": "PROFESSION"})
df.iloc[7,0] = "PRINCIPAL"

PERSIST EXTENSION
1. import persist_ext as pr
import pandas as pd
df = pd.read_csv(...)  
2. SELECT 2 VALUES
REMOVE 2 VALUES
SELECT 2 VALUES

DATASETS
- VIDEO GAMES
- AVALANCHEs
IN-LAB STUDY

TASKS

PARTICIPANTS MADE THE FOLLOWING CHANGES TO A DATASET

REMOVE COLUMNS

RENAME COLUMNS

CHANGE DATA TYPE

EDIT VALUES

FILTER DATA

ADD CATEGORICAL COLUMN
RESULTS

11/11 notebooks using Persist were reproducible only 7/11 using pandas were

3x times faster with Persist

97% tasks correctly using Persist, compared to 85% for Pandas
“so much easier than manually coding.”
- M4

“easier as compared to the code and everything was visible [...] and it didn’t take much time.”
- M2

“Changing the category type, or adding new categories or removing anomalies from data, they were very much easier in [Persist] than coding.”
- M7

“The thing I really liked about is version control, which shows the history of all operations [...] and also saves the changes [...] into a data frame.”
- M14
DISCUSSION
Provenance better for most cases
   No code clutter
   Undo/redo
   Consistent semantics
But code generation might be more robust
   works outside of notebooks
   works w/o the library
**GENERATE CODE ON DEMAND!**

[Beta]
TRY OUT PERSIST!

Persist is **available now!**

https://vdl.sci.utah.edu/persist/

Documentation & examples

Feedback / bug reports appreciated!
EMPIRICAL & THEORETICAL WORK

HOW PEOPLE ACTUALLY LIE WITH CHARTS . . .

. . . AND WHAT WE CAN DO ABOUT IT

Max Lisnic, Marina Kogan

[CHI 2023 & CHI 2024]
HOW WE THINK PPL LIE WITH CHARTS: VIOLATIONS OF VISUALIZATION DESIGN GUIDELINES

Gun deaths in Florida

Number of murders committed using firearms

2005 Florida enacted its 'Stand Your Ground' law

873

1,000 1990s 2000s 2010s

Source: Florida Department of Law Enforcement

WELFARE VS. FULL TIME JOBS

108.6M

PEOPLE ON WELFARE

PEOPLE WITH A FULL TIME JOB

101.7M

Source: Census Bureau, 2011

Inverted y axis

Truncated y axis
BUT IS THAT REALLY HOW PEOPLE LIE WITH CHAR'S?
ANALYZED 10K COVID-19 VISUALIZATIONS SHARED ON TWITTER
Cherry-picking

New daily Covid cases this time last year (pre-vaccination) compared to this year.

Arbitrary threshold

The official mortality in the UK is as high as it was in 2009.

Causal inference

As of 10th January, when they started vaccinations there had only been 10 cases.

Issues with data validity

The effect of Pax-mutacit in Mexico needs no explanation.

Incorrect reading

More proof that masks work...

Misrepresentation of scientific results

Hydroxychloroquine probably better than vaccines...

Statistical nuance

According to official data, the vaccinated are the super-spreaders.
The effect of #ivermectin in #Mexico needs no explanation.

#ivermectinworks
89% of charts with reasoning errors do not violate any visualization guidelines.

The majority of misleading charts are screenshots from reputable sources.
Sources of Deceptive Visualizations

Data exploration websites

Government reports

News media
VULNERABLE VISUALIZATIONS

According to official data, the vaccinated are the super-spreaders.
https://cdc.gov/mmwr/volumes/7...
https://it.me/EARTH20GENESIS...

Notes outside the chart
Unrestricted interaction
Added annotations

The effect of #ivermectin in #Mexico needs no explanation.

#ivermectinworks
Can we design data exploration tools that prevent cherry-picking while maintaining freedom of exploration?

PROTECTING VISUALIZATIONS

Cumulative confirmed COVID-19 deaths per million people

Due to varying protocols and challenges in the attribution of the cause of death, the number of confirmed deaths may not accurately represent the true number of deaths caused by COVID-19.
Parallel Design by 4 Authors

Cheery-Picking in Temporal Charts
(Stocks, COVID)
DESIGN SPACE

Layout

Superimposition

Juxtaposition

Context

Primary Data

Summary

A

B

C

D

E

F

A

B

C

D

E

F

A

B

C

D

E

F

A

B

C

D

E

F

A

B

C

D

E

F
CHERRYPICKING: NO GUARDRAILS

Percent change in stock price

Airline A
Airline B
Airline C
Car A
Car B
Car C
Gas A
Gas B
Gas C
Tech A
Tech B
Tech C

Car A
Airline A
Airline B

Mar 31, 2023
Jun 30, 2023
Sep 30, 2023
Dec 31, 2023
ENFORCING CONTEXT – SUPERIMPOSITION / PRIMARY DATA

Percent change in stock price

- Airline A
- Airline B
- Airline C
- Car A
- Car B
- Car C
- Gas A
- Gas B
- Gas C
- Tech A
- Tech B
- Tech C

Date:
- Mar 31, 2023
- Jun 30, 2023
- Sep 30, 2023
- Dec 31, 2023
SHOWING STATISTICAL CONTEXT – SUPERIMPOSITION / SUMMARY

Percent change in stock price
Shaded area represents the middle 50% of all values.
CONTEXT OUTSIDE – JUXTAPOSITION / PRIMARY DATA

Percent change in stock price

- Airline A
- Airline B
- Airline C
- Car A
- Car B
- Car C
- Gas A
- Gas B
- Gas C
- Tech A
- Tech B
- Tech C
**Infections per million people**

Bar on the left highlights the range of selection among all data.

- Thunderil North (Policy B)
- Eldoril West (Policy B)
- Silvoria North (Policy C)
Red-Team Study (Production):
Challenge participants to deceive w and w/o guardrails

Validation Study:
Measure how deceptive charts are
Viral Scenario:

imagine that you work for a public health agency and are in charge of promoting a health policy campaign against a viral disease that, judging by the data alone, does not appear to be the most effective policy.
Infections per million people

Add a short text caption that will go along with your visualization: *
Juxtaposed Primary is a **Cherry-Picking Machine**

Superimposed Primary makes it **harder to cherry-pick**

Others show little effect didn’t get it?

"All stocks are not doing well this year, but Airline A is the only one that actually returns some gains, whereas Gas and Car prices have dropped."
“Investment” task: make monetary decision based on information

You will need to travel to Eldoril North (Policy A) for work. You’ve come across this visualization and the accompanying caption. Please review the visualization and the caption, and then answer the questions below based solely on this information. (Please try to not rely on other visualizations you’ve seen.)

Policy A (red) is the superb policy to manage a sudden boom in infections

Infections per million people
Shaded area represents the middle 50% of all values.

Based on this information, before traveling to Eldoril North (Policy A) I would buy this much insurance:
($0 = no risk of getting sick, $100 = very high risk of getting sick)

$0  $25  $50  $75  $100
Differences by Scenario: More effective in “extreme” scenarios

Superimposed Primary Data most promising
Juxtaposed Summary no better than baseline
Superimposed Primary seems like a promising design against cherry-picking

- easy to implement
- easy to understand

Juxtaposed Primary design useful as a way to make data explorers more usable
reVISit: Empirically Evaluating Complex Interactive Visualization Techniques

[CHI 2020, CHI 2021 & VIS 2023]
Perceptual Studies

Static Visualizations

Controlled Interactions

Complex Visualization Techniques

Commonly Evaluated Using Crowdsourcing

Considered not Amenable to Crowdsourced Evaluation
Can we use empirical studies to evaluate complex interactive visualizations?
Yes We Can!

- Picking the right techniques
- Evidence-based design
- Design validation
- Careful training
- Good compensation
- Interesting Tasks

Pushing the boundary of what can be evaluated using crowdsourcing
YET: COMPLEX STUDIES ARE HARD TO SET UP
There is no **end-to-end platform** for building **interactive experiments**

The support for **types of stimuli** in survey tools are limited

Lack of experiment **debugging** features

No built-in **data collection** and **provenance tracking** functions
REVISIT

A TOOLKIT FOR DATA VISUALIZATION EXPERIMENTS

visualization design lab

view

HIVE Human-Interactive Visualization Lab

THE UNIVERSITY OF UTAH®

WPI

UNIVERSITY OF TORONTO
The Scope of reVISIt

1. reVISIt DSL - study specification
   - Study Metadata
   - Study Component Parameters
   - Study sequence

2. Study Stimuli & Components
   - Visualization Stimuli
   - Consent Form
   - Questionnaire

3. Compile & Deploy

4. Interactive Study Web App
   - Task Instructions
   - Study Stimulus
   - Admin Panel

5. Data Management
   - Data Storage (Firebase)
   - Data Download (tidy text/csv, JSON)

6. Data Analysis
   - Study Analytics Interface
   - External Data Analysis
   - R
   - SPSS
Set up studies with all components (consent, training, trials, tasks, surveys, etc.) including sophisticated study designs (randomization)

Simple data tracking and data export

Simple to deploy and run studies

Fully reproducible and open. You can share your whole study setup for anyone to re-run, without having to have a license for software (qualtrics).
"baseComponents": {
  "bar-chart": {
    "type": "website",
    "response": [
      {
        "id": "barChart",
        "prompt": "Your selected answer: ",
        "required": true,
        "location": "belowStimulus",
        "type": "iframe"
      }
    ],
    "path": "basic-questionnaire-study/assets/bar-chart.html",
    "instructionLocation": "aboveStimulus"
  }
},

"bar-chart-1": {
  "baseComponent": "bar-chart",
  "description": "A trial for the user to click the largest bar",
  "instruction": "Click on the largest bar",
  "parameters": {
    "barData": [0.32, 0.01, 1.2, 1.3, 0.82, 0.4, 0.3]
  }
},

"bar-chart-2": {
  "baseComponent": "bar-chart",
  "description": "A trial for the user to click the smallest bar",
  "instruction": "Click on the smallest bar",
  "parameters": {
    "barData": [1.2, 1.2, 1.2, 1.3, 0.82, 0.4, 0.3]
  }
}
REVISIT SPEC

Study Design
Interactive Selections in Scatterplots

Task:
How many Adventure Movies have a production budget value greater than 100 million?

Answer: *
Documentation & Tutorials on Website
Community participation via slack etc.
Tutorial at IEEE VIS 2024
CAN WE DO QUALITATIVE EVALUATION ONLINE?

Maybe?

- Think Aloud & Provenance

Pushing the boundary of what can be evaluated using crowdsourcing
Transcript

For this task, uh, I just need to consider penguins that weigh between 4 and a half and 5.

Okay kilograms, but all that I'll flip a length between 2 and 175.

All right. I'm going to have to be quite precise so I can

I'm I'm confident about the 4 to 5K but 2 to 15.

Okay, so I'm just going to click on select.

All right, so actually 200 to 20.

So I'm just going to have to have to estimate where.

2 15 is

which is approximately there about 75% of that.

Let me just clear that selection. Let me just try this again.

Okay. Alright, so that's the 4 to 5K body mass. And then the 200215.

So that's giving me those results.

So the most common penguin There is the Gen 2

and the least common penguin with 5 is a devil if I've pronounced that correctly.

Okay clicking.
So for this task, uh, I just need to consider penguins that weigh between 4 and a half and 5.

Okay kilograms, but all that I'll flip a length between 2 and 2.15.

All right, I'm going to have to be quite precise so I can

I'm I'm confident about the 4 to 6K but 2 to 15.

Okay, so I'm just going to click on select.

All right, so actually 200 to 20.

So I'm just going to have to have to estimate where.

2.15 is

which is approximately there about 75% of that.

Let me just clear that selection. Let me just try this again.
FUTURE WORK: BETTER SCAFFOLDING FOR STUDIES

**reVISit: Scalable Empirical Evaluation of Interactive Visualizations**

<table>
<thead>
<tr>
<th>Community Input</th>
<th>User Services and Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaborators</td>
<td>Documentation / Examples</td>
</tr>
<tr>
<td>Core Community</td>
<td>Replications</td>
</tr>
<tr>
<td>Community Workshops</td>
<td>Community Engagement</td>
</tr>
<tr>
<td>Broader Community</td>
<td></td>
</tr>
</tbody>
</table>

---

**Inform Software Development**

**Synergistic Activities**

---

**Core Infrastructure: Software Components**

**Aim 1: Study Infrastructure and Multilevel Instrumentation**

- Study Scaffolding
- Integration with Crowdsourcing Platforms
- Component Registry
- Capturing Insights / Rationale
- Provenance Tracking
- Designing Trainings

**Aim 2: Data Transformation and Visualization Methods**

- Data Transformation
- Event Sequence Visualization
- Trials Visualization
- Study Results Visualization
- Study Replay Interface
- Qualitative Coding Tools
REFLECTIONS
Doing different things is interesting!
  Keeps me engaged

Role of Engineering
  Building / maintaining / documenting useful things increases impact
  reVISit, Persist, UpSet

Crossover Benefits
  Provenance -> reVISit & Persist
  reVISit -> Misinformation
  Engineering infrastructure: benefits all
A Hodgepodge of Visualization Research:
Provenance, User Studies, Misinformation

Alexander Lex
http://vdl.sci.utah.edu