Literate Visualization: Making Visual Analysis Sessions Reproducible and Reusable

Alexander Lex
@alexander_lex
http://alexander-lex.net
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*
SO CAN WE DO BETTER?
Universal Set
Cardinality

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
UpSet - Visualizing Intersecting Sets

Choose Dataset: Movies Genres (17 sets, 4 attributes)

Load Data | About UpSet | UpSet for R

http://vcg.github.io/upset/
The canonical way to show set data with > 3 sets
Second-most cited VIS paper of the last decade
Multiple implementations in various languages
RESEARCH AREAS
<table>
<thead>
<tr>
<th>TECHNICAL CONTRIBUTIONS</th>
<th>DOMAIN DRIVEN TECHNIQUES</th>
<th>EMPIRICAL &amp; THEORETICAL WORK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel Visualization Techniques</td>
<td>Tailored Methods and Systems for High Impact Science Problems</td>
<td>Evaluation Methodology</td>
</tr>
<tr>
<td>Visualization Process Innovations</td>
<td></td>
<td>Design Spaces / Taxonomies</td>
</tr>
<tr>
<td>Data Wrangling Methods</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Wrangling
Methods
<table>
<thead>
<tr>
<th>TECHNICAL CONTRIBUTIONS</th>
<th>DOMAIN DRIVEN TECHNIQUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Wrangling Methods</td>
<td>Tailored Methods and Systems for High Impact Science Problems</td>
</tr>
</tbody>
</table>
Tailored Methods and Systems for High Impact Science Problems

Genealogies for Clinical Data Analysis
<table>
<thead>
<tr>
<th><strong>DOMAIN DRIVEN TECHNIQUES</strong></th>
<th><strong>EMPIRICAL &amp; THEORETICAL WORK</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tailored Methods and Systems for High Impact Science Problems</td>
<td>Evaluation Methodology</td>
</tr>
<tr>
<td></td>
<td>Design Spaces / Taxonomies</td>
</tr>
</tbody>
</table>
Evaluation
Methodology

Design Spaces / Taxonomies

EMPIRICAL & THEORETICAL WORK
LITERATE VISUAL DATA ANALYSIS:
MAKING VISUAL ANALYSIS SESSIONS REPRODUCIBLE AND REUSABLE
Reproducibility Rates

Psychology: 40%
Cancer Biology: 11%

[Baker, Penny, 2016]
[Open Science Collaboration, 2016]
[Begley, Ellis, 2012]
WHY IS THERE A REPRODUCIBILITY CRISIS?

Perverse incentives (publish or perish)
Bias for “flashy” results
Problems in data analysis
Lack of data sharing
Lack of sharing the analysis process
...


Tackle problems in data analysis and enhance reproducibility: Literate Programming
LITERATE PROGRAMMING

Explain the why and how using any means necessary!

Text
Images / Visualizations
Formulas
Videos
Links
Code

[Donald E. Knuth, 1984]
LITERATE PROGRAMMING IN THE WILD

Observable

R Markdown

Jupyter Notebooks
CHARACTERISTICS OF A GOOD NOTEBOOK

Carefully curated, well narrated.

Clean code for readability

Illustrations, Formulas, Visualizations

Complex, multi-stage process
THERE IS NO STRAIGHTFORWARD WAY TO DO LITERATE DATA VISUALIZATION
Current State: no record of what was done, let alone why

Idea: make the process of an interactive, visual analysis session well reasoned and documented
CASE STUDY

Streit et al., Guided Visual Exploration of Genomic Stratifications in Cancer, Nature Methods, 2014
IDEA: USE ANALYSIS SESSION PROVENANCE AS BASIS FOR A DATA STORY
Provenance (from the French provenir, 'to come from/forth') is the chronology of the ownership, custody or location of a historical object. [Wikipedia]

In CS: a log, a record of everything that lead to a state
My story to Country Somalia
Kaplan-Meier Plot showing survival of patients in connected block

http://vistories.org/
We solved the **WHAT**, but not the **WHY**

No progress towards reusability..

So, what else can we do?
SEMANTIC SELECTIONS
WHAT ARE SELECTIONS?
FROM SELECTIONS TO ADVANCED OPERATIONS

Unassigned
Category A
Category B

Filter
Label
Categorize
Aggregate

Interesting
Follow Up
Great Example
Intent is the user’s reason for performing a brush with a visualization.

Domain Specific Intent: Capture through Annotation

Pattern-Based Intent: Capture Automatically

WHAT IS INTENT WHEN SELECTING?

Outlier

Clusters
WHY DO WE CARE?

Speed up complex selections
WHY DO WE CARE?

**ID Based Selection:**
Selected Elements: 7, 9, 13, 18, 22

**Semantic Selection:**
Elements in K-Means cluster centered at [2, 3]

**Meaningful, higher level concept:**
improves reproducibility

**Robust to changes and updates in dataset:**
enables re-usability
HOW DO WE INFER INTENT?
HOW DO WE INFER INTENT?

**Selection**

**Predictions**
- K-Means
- DBScan
- Regression
- Outlier Detection
- Skyline
- Decision Trees / Ranges
- Categories
HOW DO WE INFER INTENT?

Selection

Predictions
- K-Means
- DBScan
- Regression
- Outlier Detection
- Skyline
- Decision Trees / Ranges
- Categories

Ranking
- Jaccard Distance
- Naive Bayes
- Classifier
- Heuristic Measures

1. Range
2. Cluster
3. Outlier

\[ J(S, C) = \frac{|S \cap C|}{|S \cup C|} \]
How do we infer intent?

Selection

Predictions
- K-Means
- DBScan
- Regression
- Outlier Detection
- Skyline
- Decision Trees / Ranges
- Categories

Ranking
- Jaccard Distance
- Naive Bayes
- Classifier
- Heuristic
- Measures

Confirming Intent & Annotation

1. Range
2. Cluster
3. Outlier

I think this cluster…
Visualization and Selection

Annotation of Intent and Predictions

http://vdl.sci.utah.edu/predicting-intent/
VALIDATION

User Driven (UD)

Computer Supported (CD)
WORKFLOWS

Based on semantic selections, we can create reusable workflows!

Kiran Gadhare, Zach Cutler, Alexander Lex
Pros and Cons

Interactive Visual Analysis

Computational Analysis

jupyter
Pros and Cons

Interactive Visual Analysis

Intuitive and Fast
Uses Human Perceptual Capabilities

Computational Analysis

jupyter
Pros and Cons

Interactive Visual Analysis

Need to redo the analysis when datasets update

Computational Analysis
Pros and Cons

Interactive Visual Analysis

Computational Analysis

Flexible, powerful, reusable.
Pros and Cons

Interactive Visual Analysis

- Can and want to program

Computational Analysis

- Time Consuming
- Require analyst can and want to program
- Difficult to see what’s in the data
IDEA: CAPTURING WORKFLOWS

“Filter Outliers” Workflow

Specify dimensions → Brush outliers → Filter out
USING WORKFLOW ON UPDATED DATA

Specify dimensions

Brush outliers

Filter out
CAPTURING SEMANTICS OF WORKFLOWS

Robust "Filter Outliers" Workflow

Original Dataset

Brush selection

Refine selection to select outliers

Filter out

Input

Output
REUSING WORKFLOWS ON **UPDATED** DATA

**Updated Dataset**

**New Data Point**

**Apply Workflow**

- [Diagram showing data points and workflow application process]
REUSING SELECTIONS ON UPDATED DATASETS

- Changed Dataset
- Tracking A Selected Cluster
- Selected Cluster on Changed Dataset
HUMAN REVIEWS
USING WORKFLOW IN A COMPUTATIONAL NOTEBOOK

# Installing the reapply-workflows adds a module called backend
# This module exposes the Reapply class which initializes the library
from backend import Reapply

# Here we load the reapply_workflows library.
r = Reapply()

# We add a workflow from our workflow database.
workflow = r.load_workflow("workflow1617808681620")

# Print the workflow name
print("Workflow: ", workflow.name, "\n")

# Description of the workflow and the operations in it
workflow.describe

Workflow: Deleting Cluster

<table>
<thead>
<tr>
<th>Root</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add Plot</td>
</tr>
<tr>
<td>Added brush to: X-Y</td>
</tr>
<tr>
<td>Cluster Selection</td>
</tr>
<tr>
<td>Filter: Out</td>
</tr>
</tbody>
</table>

# Prints the reapply results for all interactions, along with review status

# Apply the workflow to target dataset.
# apply function requires the target dataset
# and the label column as arguments.
res = workflow.apply(target, "Label")

# Results is an array of datasets for each interaction
# we grab the final one.
result_dataset = res.results[-1][\'data\']
result_dataset

This workflow has not been reviewed for all interactions.
Please go to following url: https://reapply-workflows.git/

<table>
<thead>
<tr>
<th>Label</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>P52</td>
<td>6.58351</td>
</tr>
<tr>
<td>5</td>
<td>P171</td>
<td>4.77421</td>
</tr>
<tr>
<td>8</td>
<td>P199</td>
<td>8.34966</td>
</tr>
<tr>
<td>9</td>
<td>P183</td>
<td>8.42670</td>
</tr>
<tr>
<td>10</td>
<td>P61</td>
<td>4.29760</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>141</td>
<td>P138</td>
<td>7.35179</td>
</tr>
<tr>
<td>142</td>
<td>P46</td>
<td>6.62171</td>
</tr>
</tbody>
</table>
BEFORE AND AFTER
Semantic selections & annotations allow us to have **higher-level information** about events in an analysis process.

Analysts have the **means to justify** their choices.

Makes it possible to **reuse interactive analysis processes** on updated datasets.
Thanks to: Kiran Gadhave, Zach Cutler, Marc Streit, Jochen Görtler, Oliver Deussen, Miriah Meyer, Jeff Phillips, Samuel Gratzl, Holger Stitz, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, Hanspeter Pfister, and many others!