Literate Visualization: Making Visual Analysis Sessions Reproducible and Reusable

Alexander Lex
@alexander_lex
http://alexander-lex.net

visualization design lab

www.sci.utah.edu
Outline

What is Data Science
  Programming is essential
What’s the role of visualization in DS?
Reproducibility, Communication, Reusability
  Literate Programming is a way to achieve this
What’s literate programming for visualization?
  Provenance Tracking; extracting analysis stories.
  Inferring Analyst’s Intent
What is Data Science?

A data scientist is a statistician who lives in San Francisco.

Data science is statistics on a Mac.

A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.
What is Data Science?

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. [Wikipedia]
Data science = data analysis + programming
Why is Programming So Important?

Expressivity and Generality!

It’s hard (if not impossible) to build a higher-level system that anticipates any possible need.

But programming is difficult, not everyone can do it, and it’s slow.

Ideal World: Programming where necessary; Interactive tools where possible.
THE ROLE OF VISUALIZATION IN DATA SCIENCE
Can We Trust Statistics?

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>10 8.04</td>
<td>10 9.14</td>
<td>10 7.46</td>
<td>8 6.58</td>
</tr>
<tr>
<td></td>
<td>8 6.95</td>
<td>8 8.14</td>
<td>8 6.77</td>
<td>8 5.76</td>
</tr>
<tr>
<td></td>
<td>13 7.58</td>
<td>13 8.74</td>
<td>13 12.74</td>
<td>8 7.71</td>
</tr>
<tr>
<td></td>
<td>9 8.81</td>
<td>9 8.77</td>
<td>9 7.11</td>
<td>8 8.84</td>
</tr>
<tr>
<td></td>
<td>11 8.33</td>
<td>11 9.26</td>
<td>11 7.81</td>
<td>8 8.47</td>
</tr>
<tr>
<td></td>
<td>14 9.96</td>
<td>14 8.1</td>
<td>14 8.84</td>
<td>8 7.04</td>
</tr>
<tr>
<td></td>
<td>6 7.24</td>
<td>6 6.13</td>
<td>6 6.08</td>
<td>8 5.25</td>
</tr>
<tr>
<td></td>
<td>4 4.26</td>
<td>4 3.1</td>
<td>4 5.39</td>
<td>19 12.5</td>
</tr>
<tr>
<td></td>
<td>12 10.84</td>
<td>12 9.13</td>
<td>12 8.15</td>
<td>8 5.56</td>
</tr>
<tr>
<td></td>
<td>7 4.82</td>
<td>7 7.26</td>
<td>7 6.42</td>
<td>8 7.91</td>
</tr>
<tr>
<td></td>
<td>5 5.68</td>
<td></td>
<td></td>
<td>6.89</td>
</tr>
</tbody>
</table>

Mean x: 9 y: 7.50
Variance x: 11 y: 4.122
Correlation x – y: 0.816
Linear regression: \( y = 3.00 + 0.500x \)
Anscombe’s Quartet

Mean x: 9 y: 7.50
Variance x: 11 y: 4.122
Correlation x – y: 0.816
Linear regression: y = 3.00 + 0.500x
Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis.

Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.

Tukey and Wilk - 1966
Some implications for effective analysis are: (1) it is essential to have convenience of interaction of people and intermediate results and (2) at all stages of data analysis, the outputs need to be matched to the capabilities of the people who use it and want it.

Tukey and Wilk - 1966
What’s the Role of Visualization in Data Science?

Doing Data Science: Straight Talk from the Frontline,
by Cathy O’Neil and Rachel Schutt
Vis as commonly used by Data Scientists

Embedded, and code-generated static graphics

Great, but (commonly) static you have to write code...

Embedding ggplot into knitr document fails

We will generate a graphic in the following chunk.

```r
library(ggplot2)
mtcars_ggplot <- ggplot(mtcars) + geom_point() + geom_smooth()
```

In-line code is great for things like summary statistics. The mean miles per gallon for the mtcars data set is 20.090625.

In-line code is not good for graphics.

```r
mtcars_ggplot
```

Instead, use a chunk. This will also allow you to control how the graphic is rendered in your final document.

```r
mtcars_ggplot
```

```r
ggplot(mtcars) + geom_smooth(method = "loess")
```
Why do we need Interaction?

Explore data that is big / complex
There is too much data
There are too many ways to show it
There are too many questions to ask

Visualization = Human Data Interaction
REPRODUCIBILITY,
COMMUNICATION,
REUSABILITY
Reproducibility Crisis in Science

Psychology: 40% of experiments could be reproduced
Cancer Biology: 11%

https://science.sciencemag.org/content/349/6251/aac4716
https://www.nature.com/articles/483531a

More than 70% of researchers have tried and failed to reproduce another scientist’s experiments, and more than half have failed to reproduce their own experiments. These are some of the telling figures that emerged from Nature’s survey of 1,172 researchers who took a brief online questionnaire on reproducibility in research.

Failing to reproduce results is a rite of passage, says Marcus Munafò, a biological psychologist at the University of Bristol, U.K., who has a long-standing interest in scientific reproducibility. When he was a student, he says, “I tried to replicate what looked simple from the literature, and wasn’t able to. Then I had a crisis of confidence, and then I learned that my experience wasn’t uncommon.”
Why is there a reproducibility crisis?

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

...
Just Sharing is Not Enough

Data needs context/annotation
Analysis needs documentation and justification

Literate Programming
The past ten years have witnessed substantial improvements in programming methodology. This advance, carried out under the banner of "structured programming," has led to programs that are more reliable and easier to comprehend; yet the results are not entirely satisfactory.

My purpose in the present paper is to propose another motto that may be appropriate for the next decade, as we attempt to make further progress in the state of the art. I believe that the time is ripe for significantly better documentation of programs, and that we can best achieve this by considering programs to be works of literature.

Donald E. Knuth, 1984
Hence, my title: "Literate Programming."

Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.

Donald E. Knuth, 1984
Literate Programming

Explain the why and how using any means necessary!

Text
Images /
Visualizations
Formulas
Videos
Links
Code

All work in concert to ensure reproducibility
Literate Programming in the Wild

Observable

R Markdown

Jupyter Notebooks
Literate Visualization

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

Problem: no record of what was done (let alone why)
Aside: Literate Visualization Design

Idea: Use Literate Programming to Justify Visualization Design Choices

Design Exposition with Literate Visualization, Wood, Kachkaev, Dykes, 2018

https://www.gicentre.net/litvis
Pipeline-Based Visualization Specification

**Pro:**
Explicit, similar to source code

**Con:**
Complex
Still not “Literate”

Bowen Yu, and Claudio T. Silva, 2017
Provenance Tracking

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
Provenance Uses

Recall
Replication
Action Recovery
Collaborative Comm.
Presentation
Meta-Analysis

Heer et al., 2008
Ragan et al., 2015
How are “Literate Programs” Created?

1. Write draft code that might work for test case
2. Edit code to generalize
3. Refactor for readability, good SE practices
4. Document
5. Provide context with images/links/equations/text

-> Complex, multi-stage process
A history (e.g., version control) isn’t very relevant.

-> Visualization Provenance is necessary but insufficient
Guided Visual Exploration of Genomic Stratifications in Cancer
Nature Methods 11, 9 (2014), 884–885
Case Study presented in Nature Methods

Streit et al., Guided Visual Exploration of Genomic Stratifications in Cancer, Nature Methods, 2014
Examples

Gapminder
http://vistories.org/gapminder

StratomeX
http://vistories.org/stratomex
Kaplan-Meier Plot showing survival of patients in connected block

http://vistories.org/stratomex
Challenges

Privacy
  Curated analysis stories vs full log
Integration
  Visualization tools are only one part of the workflow
Scalability
  Provenance Graph gets fast quickly
Effort
  It’s work to create an analysis story
Re-Use
  Unlike code, can’t easily apply a visual analysis to updated data
Provenance is “dumb”

Record of what you did, but not why you did it.

Knowing why would help with scalability, effort, and re-use

So, can we make it easier to capture the why?
CAPTURING USER INTENT WHEN BRUSHING IN SCATTERPLOTS

Kiran Gadhave, Jochen Görtler, Oliver Deussen, Miriah Meyer, Jeff Phillips, Alexander Lex
What’s Brushing?

Brushable Scatterplot Matrix

This scatterplot matrix allows brushing to select data points in one cell, and highlight them across all other cells.
What is Intent when Brushing?

Intent is the user’s reason for performing a brush with a visualization.
Correlation
Multivariate Optimization
Categories
Ranges
Outlier
Clusters
Correlation
Categories
Multivariate Optimization
Ranges
How do we infer Intent?

Selection
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

1. Range
2. Cluster
3. Outlier
How do we infer Intent?

Selection

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

Ranking
1. Range
2. Cluster
3. Outlier

Confirming Intent
& Annotation

I think this cluster…
Visualization and Selection

http://vdl.sci.utah.edu/predicting-intent/
Detail: Ranges

What did I want to select?

- GDP > 61,250
- AND GDP < 111,250
- AND Child Mortality Rate > 1.5
- AND Child Mortality Rate < 13

Or

- GDP > 60k?
For the given diagram, the rules for determining whether a point is inside or outside are as follows:

- **Inside** if \( X \leq 5 \) AND \( Y \leq 6 \)
- **Outside** if \( X > 5 \) OR \( Y > 6 \)
Correcting Errors / Simplifying Selections

Did I intent to select elements > 60k?

\[ X \leq 5 \]

\[ Y \leq 6 \]

Outside

Inside

Outside

Inside

Outside
Correcting Errors / Simplifying Selections

Did I intent to select outliers?

Turn prediction into selection.
Implications

Low effort to capture intent: reduce scalability problem

Understanding intent: potential for re-use

Correcting selection: speed up analysis
RECAP & OUTLOOK
These are first steps...

**Reproducibility**
- Analysis stories and provenance make visualizations reproducible
- Annotating the Why makes them justifiable

**Communication**
- Curated stories make it easier for others (collaborators or “consumers”) to understand
- A future of “interactive figures”

**Re-Use**
- Curated, generalized visual analysis process can be applied to updated datasets
Much is left to do!

Integration
Data Provenance
Review for Re-Use
Extend Intent Inference to other Vis Types and Interaction Types
import pandas as pd

//javascript
window.asTaggle = (dataset, element) => {
  // element = JQuery output element magic by jupyter
  // increase output height
  element.parent('.output_area').css('height', '48em');
  element.css('position', 'relative');
  element.html('<main style="position: absolute; left: 0; right: 0; top: 0; bottom: 0"></main>');

  return new Promise(resolve => {
    require('./lib/TaggleJS_bundle.js'), function(taggle) {
      const parent = element.find('main')[0];
      
      const data = taggle.createLocalStorage(dataset, taggle.deriveColumnDescriptions(dataset));
      data.deriveDefault();
      instance = taggle.createTaggle(data, parent);
      instance.update();
      resolve(instance);
    }
  });
})

import json from IPython.display} import Javascript

Javascript("window.asTaggle({}, element);".format(iris.to_json(orient='records'))

<table>
<thead>
<tr>
<th>Rank</th>
<th>sepal_length</th>
<th>sepal_width</th>
<th>petal_length</th>
<th>petal_width</th>
<th>species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>virginica</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>virginica</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>virginica</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>virginica</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>virginica</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>virginica</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>virginica</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>virginica</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>virginica</td>
</tr>
</tbody>
</table>
Thanks to: Kiran Gadhave, Jochen Görtler, Oliver Deussen, Miriah Meyer, Jeff Phillips, Samuel Gratzl, Nils Gehlenborg, Nicola Cosgrove, Marc Streit

Funded by NSF, the Utah Genome Project, NIH, and DoD