CSE 599F
Research Methods and Data Analysis in Software Systems Research
Winter 2019

Data visualization

March 05, 2019
Recap

● Monte Carlo (MC) methods
  ○ What is the key concept?
  ○ When are these applicable and useful?

● Bootstrapping (in R)
  ○ What is the key concept?
  ○ How is bootstrapping different from and related to MC methods?
  ○ What is the core assumption for bootstrapping?
Today

- From an experimental design to a research paper
- Data visualization
  - Guidelines for effective tables and graphs
  - Effective tables with \textit{booktabs}
  - Effective graphs with \textit{ggplot2}
Empirical research: the pipeline

Experimental Design → How do we get here? → Research paper
Empirical research: the pipeline

Do all results go into the paper?
Empirical research: the pipeline

- Experimental Design
- Data collection
- Data analysis
- Graphs & tables
- Research paper
- Sanity checks
- Detailed results
Effective tables
Tables vs. graphs

● When are tables useful?
  ○ Compare individual values.
  ○ Precise values are important.
  ○ Values may involve multiple units.

<table>
<thead>
<tr>
<th>Browser</th>
<th>Market share (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>June 08</td>
</tr>
<tr>
<td>Internet Explorer</td>
<td>75.4</td>
</tr>
<tr>
<td>Firefox</td>
<td>18.9</td>
</tr>
<tr>
<td>Safari</td>
<td>2.8</td>
</tr>
<tr>
<td>Chrome</td>
<td>—</td>
</tr>
<tr>
<td>Opera</td>
<td>2.1</td>
</tr>
<tr>
<td>Netscape</td>
<td>0.5</td>
</tr>
<tr>
<td>Other</td>
<td>0.2</td>
</tr>
</tbody>
</table>

● When are graphs useful?
  ○ Consider an entire set of values.
  ○ Visualize trends and patterns.
  ○ Relative differences and relationships are more important than precise values.
Effective tables: layout

A first table in LaTex:

- Recall the run-time data set
- Goal: show run times and improvements

| variant, naive, caching, forking, equivalent, first.order, run, subject |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| 11, 309.8, 157.6, 144.8, 1, 1, 1, "tax" |
| 12, 379.5, 237.4, 254.5, 0, 1, 1, "tax" |
| 13, 415.9, 225.9, 225.9, 0, 0, 1, "tax" |
| ... |

...
Effective tables: layout

A first table in LaTeX:

- Recall the run-time data set
- Goal: show run times and improvements

![Table 1: Run times and improvements.](image)

```
<table>
<thead>
<tr>
<th>Subject</th>
<th>RT-naive</th>
<th>RT-cache</th>
<th>RT-fork</th>
<th>I-cache</th>
<th>I-fork</th>
</tr>
</thead>
<tbody>
<tr>
<td>tax</td>
<td>504.11</td>
<td>247.01</td>
<td>195.42</td>
<td>51.02%</td>
<td>61.31%</td>
</tr>
<tr>
<td>tictactoe</td>
<td>17.44</td>
<td>16.32</td>
<td>15.43</td>
<td>6.31%</td>
<td>11.49%</td>
</tr>
<tr>
<td>triangle</td>
<td>3.13</td>
<td>2.79</td>
<td>1.67</td>
<td>10.91%</td>
<td>46.62%</td>
</tr>
</tbody>
</table>
```

Looks pretty good and clear, doesn’t it?
Effective tables: layout

**TABLE I**

*Run times and improvements.*

<table>
<thead>
<tr>
<th>Subject</th>
<th>RT-naive</th>
<th>RT-cache</th>
<th>RT-fork</th>
<th>I-cache</th>
<th>I-fork</th>
</tr>
</thead>
<tbody>
<tr>
<td>tax</td>
<td>504.11</td>
<td>247.01</td>
<td>195.42</td>
<td>51.02%</td>
<td>61.31%</td>
</tr>
<tr>
<td>tictactoe</td>
<td>17.44</td>
<td>16.32</td>
<td>15.43</td>
<td>6.31%</td>
<td>11.49%</td>
</tr>
<tr>
<td>triangle</td>
<td>3.13</td>
<td>2.79</td>
<td>1.67</td>
<td>10.91%</td>
<td>46.62%</td>
</tr>
</tbody>
</table>

**TABLE II**

*Run times and improvements for the naive, caching (cache), and forking (fork) strategies. Run times are given in seconds and averaged over five runs.*

<table>
<thead>
<tr>
<th>Subject</th>
<th>Run times</th>
<th>Improvements</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>naive</td>
<td>cache</td>
<td>fork</td>
<td>cache (vs. naive)</td>
<td>fork (vs. naive)</td>
</tr>
<tr>
<td>Tax</td>
<td>504</td>
<td>247</td>
<td>195</td>
<td>51.0%</td>
<td>61.3%</td>
</tr>
<tr>
<td>TicTacToe</td>
<td>17.4</td>
<td>16.3</td>
<td>15.4</td>
<td>6.31%</td>
<td>11.5%</td>
</tr>
<tr>
<td>Triangle</td>
<td>3.13</td>
<td>2.79</td>
<td>1.67</td>
<td>10.9%</td>
<td>46.6%</td>
</tr>
</tbody>
</table>

Which table looks better?
Effective tables: content

Still room for improvements -- keep it simple:

- Avoid mixing higher-is-better and lower-is-better values.
- Allow for easy comparisons, primarily by row.
- Summarize the table (what is the bottom line?).

**TABLE II**

RUN TIMES AND IMPROVEMENTS FOR THE NAIVE, CACHING (CACHE), AND FORKING (FORK) STRATEGIES. RUN TIMES ARE GIVEN IN SECONDS AND AVERAGED OVER FIVE RUNS.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Run times</th>
<th>Improvements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>naive</td>
<td>cache</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(vs. naive)</td>
</tr>
<tr>
<td>Tax</td>
<td>504</td>
<td>247</td>
</tr>
<tr>
<td></td>
<td>51.0%</td>
<td>61.3%</td>
</tr>
<tr>
<td>TicTacToe</td>
<td>17.4</td>
<td>16.3</td>
</tr>
<tr>
<td></td>
<td>6.31%</td>
<td>11.5%</td>
</tr>
<tr>
<td>Triangle</td>
<td>3.13</td>
<td>2.79</td>
</tr>
<tr>
<td></td>
<td>10.9%</td>
<td>46.6%</td>
</tr>
</tbody>
</table>
Effective tables: summaries

<table>
<thead>
<tr>
<th>Subject</th>
<th>LOC</th>
<th>Speed up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tax</td>
<td>8900</td>
<td>10.2%</td>
</tr>
<tr>
<td>TicTacToe</td>
<td>120</td>
<td>54.2%</td>
</tr>
<tr>
<td>Triangle</td>
<td>80</td>
<td>60.9%</td>
</tr>
<tr>
<td>Average</td>
<td>3393</td>
<td>41.8%</td>
</tr>
</tbody>
</table>

Total vs. Average

How to properly summarize a table?
Effective tables

Effective tables in LaTeX:
● Use the **booktabs** package!
● Use descriptive (hierarchical) headers.
● Make each table self-contained (content and caption).
● Don’t use horizontal lines between related rows.
● Don’t use vertical lines between related columns.
● Right align numbers.
● Summarize with meaningful totals or weighted averages.
● Think about precision vs. significant digits (be consistent)!
Effective graphs
Effective graphs: taming complexity

Way too many details!
How can this plot be simplified and improved?
Effective graphs: axes

Truncated axes are misleading and not a proper way to “demonstrate” effect size.
Effective graphs: point plots vs. line plots

Point plots (scatter plots):
- Good visual summary of point clouds, trends, and relationships.
- May obscure relevant trends (overlapping points).
- Hard to reason about density (without adding transparency).
Effective graphs: histogram vs. kernel density

Histograms:
- Good visual summary of count data.
- Binning may lead to misleading results.
- Kernel density overlay can provide information about adequate binning.

Adequate binning

Changed binning

Kernel density
Effective graphs: box plots vs. violin plots

**Box plots:**
- Good visual summary for continuous data.
- Nicely complements hypothesis tests.
- May be misleading for multimodal data.
Effective graphs: facet plots
# Point plot of the raw data
```r
ggplot(data=result, aes(x=Step, y=Ratio, color=Strategy)) + geom_point() + xlab("Work") + ylab("Test completeness")
```

# Aggregated line plot of the same data
```r
ggplot(data=result, aes(x=Step, y=Ratio, color=Strategy)) + stat_summary(fun.y="mean", geom="line") + xlab("Work") + ylab("Test completeness")
```
Effective graphs

Effective graphs in R:
- Use the **ggplot2** package!
- Make each plot self-contained (content and caption).
- Relate tables and graphs to tell a coherent story.
- Avoid multiple, unrelated axes.
- Consistently put the DV on the vertical axis (by default).
- Choose an appropriate graph for the data (don’t connect unrelated data points).
- Reduce complexity with facet plots.