Presenting computational results

E6891 Lecture 11
2014-04-09
Today’s plan

● Communicating numerical information
  ○ text (tables)
  ○ visuals (plots, images)
  ○ statistical summaries

● Much borrowing from
Why a lecture about presentation?

- Step 1 of reproducing a result:
  - what is the result?

- Reproducibility depends on clarity

- Clarity can be difficult!
Aside

- I’ll use examples mainly from my own work

- These will **not** be perfect!
  - I’m not an info-vis expert

- Let’s beat up on them together!
Communicating numerical data

- Quantitative information
- Qualitative comparisons
- Trends in data
- Statistical quantities
How should I present X?

● What should the reader take away?
  ○ Raw information? (Quantitative)
  ○ Comparisons? Trends? (Qualitative)

● Always put yourself in place of the reader

● Figures should support the text
  ○ not vice versa!
Tables

● Best for reporting small amounts of data with high precision

● Useful when data has intrinsic value
  ○ e.g., sample size, parameter range

● Not great for comparisons or large data
  ○ Trends can be obscure
  ○ Not space-efficient
<table>
<thead>
<tr>
<th>Data Set/Items</th>
<th>Songs</th>
<th>Albums</th>
<th>Artists</th>
<th>Users</th>
<th>Ratings/Evts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo! Music</td>
<td>624,961 in total</td>
<td></td>
<td></td>
<td>1,000,990</td>
<td>262,810,175</td>
</tr>
<tr>
<td>MSD</td>
<td>1,000,000</td>
<td></td>
<td></td>
<td>1,019,318</td>
<td>48,373,586</td>
</tr>
<tr>
<td>Last.fm – 360K</td>
<td></td>
<td>186,642</td>
<td></td>
<td>359,347</td>
<td></td>
</tr>
<tr>
<td>Last.fm – 1K</td>
<td></td>
<td>107,528</td>
<td></td>
<td>992</td>
<td></td>
</tr>
<tr>
<td>MusicMicro</td>
<td>71,410</td>
<td></td>
<td></td>
<td>19,529</td>
<td>594,306</td>
</tr>
<tr>
<td>MMTD</td>
<td>133,968</td>
<td></td>
<td></td>
<td>215,375</td>
<td>1,086,808</td>
</tr>
<tr>
<td>AotM-2011</td>
<td>98,359</td>
<td></td>
<td></td>
<td>16,204</td>
<td>859,449</td>
</tr>
</tbody>
</table>

**Table 1** Statistics of public data sets for music recommendation research.
## Table example (not so great)

<table>
<thead>
<tr>
<th>Data Set/Items</th>
<th>Songs</th>
<th>Albums</th>
<th>Artists</th>
<th>Users</th>
<th>Ratings/Evts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo! Music</td>
<td></td>
<td>624,961 in total</td>
<td></td>
<td>1,000,990</td>
<td>262,810,175</td>
</tr>
<tr>
<td>MSD</td>
<td>1,000,000</td>
<td></td>
<td></td>
<td>1,019,318</td>
<td>48,373,586</td>
</tr>
<tr>
<td>Last.fm – 360K</td>
<td></td>
<td>186,642</td>
<td></td>
<td>359,347</td>
<td></td>
</tr>
<tr>
<td>Last.fm – 1K</td>
<td></td>
<td>107,528</td>
<td></td>
<td>992</td>
<td></td>
</tr>
<tr>
<td>MusicMicro</td>
<td>71,410</td>
<td></td>
<td>19,529</td>
<td>136,866</td>
<td>594,306</td>
</tr>
<tr>
<td>MMTD</td>
<td>133,968</td>
<td></td>
<td>25,060</td>
<td>215,375</td>
<td>1,086,808</td>
</tr>
<tr>
<td>AotM-2011</td>
<td>98,359</td>
<td></td>
<td>16,204</td>
<td></td>
<td>859,449</td>
</tr>
</tbody>
</table>

**Table 1** Statistics of public data sets for music recommendation research.

---

**Good**

- Vertical arrangement
- Easy to interpret data
Table example (not so great)

<table>
<thead>
<tr>
<th>Data Set/Items</th>
<th>Songs</th>
<th>Albums</th>
<th>Artists</th>
<th>Users</th>
<th>Ratings/Evts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo! Music</td>
<td></td>
<td>624,961 in total</td>
<td></td>
<td>1,000,990</td>
<td>262,810,175</td>
</tr>
<tr>
<td>MSD</td>
<td>1,000,000</td>
<td></td>
<td></td>
<td>1,019,318</td>
<td>48,373,586</td>
</tr>
<tr>
<td>Last.fm – 360K</td>
<td></td>
<td>186,642</td>
<td></td>
<td>359,347</td>
<td></td>
</tr>
<tr>
<td>Last.fm – 1K</td>
<td></td>
<td>107,528</td>
<td></td>
<td>992</td>
<td></td>
</tr>
<tr>
<td>MusicMicro</td>
<td>71,410</td>
<td>19,529</td>
<td></td>
<td>136,866</td>
<td>594,306</td>
</tr>
<tr>
<td>MMTD</td>
<td>133,968</td>
<td>25,060</td>
<td></td>
<td>215,375</td>
<td>1,086,808</td>
</tr>
<tr>
<td>AotM-2011</td>
<td>98,359</td>
<td></td>
<td></td>
<td>16,204</td>
<td>859,449</td>
</tr>
</tbody>
</table>

Table 1 Statistics of public data sets for music recommendation research.

**Good**
- Vertical arrangement
- Easy to interpret data

**Bad**
- Line clutter
- Excessive detail
- Center-alignment
- Unused column
- A lot of border lines
Table example (improved)

<table>
<thead>
<tr>
<th>Collection</th>
<th>Songs</th>
<th>Artists</th>
<th>Items</th>
<th>Users</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo! Music [1]</td>
<td>—</td>
<td>—</td>
<td>625K</td>
<td>1M</td>
<td>263M</td>
</tr>
<tr>
<td>MSD [2]</td>
<td>1M</td>
<td>—</td>
<td>—</td>
<td>1M</td>
<td>48M</td>
</tr>
<tr>
<td>Last.fm – 1K [3]</td>
<td>—</td>
<td>107K</td>
<td>107K</td>
<td>1K</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 1: Statistics of public data sets for music recommendation research.

**Improvements**
- Removed clutter
- Simplified headers
- Explicit missing values
- In-place citations

**Still bad**
- “Items” may be confusing
  - but that’s the data…
  - clarify in text!
Best practices: tables

● **Do** use when numbers have intrinsic value

● **Do** arrange by column, not row

● **Do not** clutter with lines/rules/borders

● **Do not** use excessive precision

● **Do not** overload
Graphics can serve many purposes

- Space-efficient communication
- Highlight trends in data
- Help the reader form comparisons
Graphics can’t...

●  … make your point for you
  ○  *But they can help*

●  … tell the complete story
  ○  Choosing what to leave out is important!

●  … make themselves presentable
  ○  No, not even with the Matlab defaults!
How should I display my data?

● What’s the data?
  ○ Continuous
  ○ Ordered? Sequential?
  ○ Categorical? Binary?
  ○ Bounded? Non-negative? [0, 1]?

● What’s the comparison?
  ○ Absolute (e.g., classifier accuracy)
  ○ Relative (e.g., histogram data)
  ○ Something else entirely?
No one-size-fits-all solution...

- But you can get really far with:
  - **line** (grouped data)
  - **scatter** (ungrouped data)

- Primary goal: *simplicity*

- Prefer many simple plots to one complex plot
Lines

- Line grouping helps illustrate trends
- Quantity to be compared is on the vertical axis
Information overload

- Too many comparisons for one figure:
  - (4 methods) $\times$ (4 VQ values) $\times$ (4 t values)
Multiple plots

- Some redundancy is okay
- Restrict intended comparisons to lie within one subplot
- Minimize inter-plot comparisons
Scatter

- Why not lines?
  - no meaningful ordering
  - clutter
Why not **lines**?
- no meaningful ordering
- clutter

Why not **bars**?
- obscures error bars
- invisible baseline
- fractional comparisons aren’t relevant
Scatter

- Why not **lines**?
  - no meaningful ordering
  - clutter

- Why not **bars**?
  - obscures error bars
  - invisible baseline
  - fractional comparisons aren’t relevant

**Bad**
- [0.65, 0.85]?
- Maybe overloaded
- Bright green can be hard to see
Best practices: plots / subplots

● Label all axes

● Quantity of comparison on the y-axis

● Use meaningful limits when possible
  ○ Be consistent when multi-plotting

● Be consistent with markers/styles

● Don’t rely too much on color
If using a legend, match the ordering to the visualization

Better yet, label points/curves directly
  - As long as it’s still readable...

Use captions to resolve ambiguities

Empty space can be ok, if it’s meaningful
About color...

- Color is the easiest thing to get wrong

- Things to watch out for:
  - printer-friendly
  - projector-friendly
  - colorblind-friendly
  - unintended (dis)similarity
Example: spectrogram

- *Jet* colormap provides **false contrast**
- Does not translate to grayscale
Example: spectrogram

- But the data is bounded: $(-\infty, 0]$  
- Use a sequential gradient  
- Observe conventions as far as possible
Example: signed data
Example: signed data

- Divergent colormaps visualize both magnitude and direction (sign)
What makes color difficult?

● Numerical data -> RGB HSV

● Input data can be multi-dimensional
  ○ Sequential data is 1d (distance from boundary)
  ○ Divergent data is 2d (magnitude, direction)

● Color parameters are non-linear
  ○ … so is human perception

● Physical and perceptual constraints
Choosing a colormap 1

Color Brewer
Choosing a colormap 2

Color-blind simulator
Best practices: colormaps

- **Sequential**
  - OrRd
  - Greys
  - (or any single-hue gradient)

- **Divergent**
  - PuOr

- **Never use jet**
  - Rainbow maps can be ok for categorical data...
  - … but continuous rainbow maps are dangerous
Statistical quantities

- Results are typically statistical, e.g.:
  - classifier accuracy on a test sample
  - $P[\text{sample data} \mid \text{model}]$

- We use finite-sample approximations to estimate unobservable quantities
  - e.g., true accuracy of the classifier

- Approximations imply **uncertainty**
  - this should be reported too!
Error bars

- Repeating an experiment with random sampling helps us to quantify uncertainty
  - leave-one-out, k-fold cross-validation, etc.

- Depending on the statistic being reported, different notions of uncertainty make sense
  - standard deviation
  - quantiles/inter-quartile range
Hypothesis testing

● Somewhat dicey territory these days…

● Quantify confidence in a statistical claim
  ○ e.g., difference in accuracy between two classifiers
  ○ are they actually different?

● Does the data support my hypothesis?
  ○ Assume the contrary: the null hypothesis
  ○ Use data to refute the null hypothesis
p-values

The p-value is the probability (under [the null hypothesis]) of observing a value of the test statistic the same as or more extreme than what was actually observed.


- **NOT** $P[\text{null hypothesis} \mid \text{data}]

- A p-value can be high if
  - the null hypothesis is true (and it almost never is!)
  - the test statistic has low power
Pitfalls of p-values

- Rejection threshold is arbitrary
  - 0.05 vs 0.051?
  - It’s better to report values directly than claim significance against a fixed threshold

- p-value does not measure effect size
  - with enough samples, any difference is “significant”
  - but is it meaningful?

- We usually already know the null hypothesis is false
Discussion