So many types

There are so many kinds of charts, diagrams, graphs, maps
What are their features?
What tasks are they good for? – Accuracy or speed of judgment? Memorability?

Topics

• Statistical data graphs
  ▪ 1D: dotplot, boxplot, violin plot
  ▪ 1.5D: time-series plot, density plot, bar chart, pie chart
  ▪ 2D: scatterplot, ridgeline plot
  ▪ 3D: contour plot, 3D scatterplot, surface plot
• Thematic maps
  ▪ Choropleth map
  ▪ Anamorphic map
  ▪ Flow maps
• Network & tree visualization
• Animation & interactive graphics

1D: Infographic vs. Data graphic

The same data can be shown in different forms, for different purposes

One might argue that this infographic has greater impact in showing the relative size of GDP
One might argue that this statistical graph makes comparisons easier
1D: Dotplots & boxplots

What number do you give to a probability phrase?

Boxplots summarize the important characteristics of a univariate data distribution:
- center (median)
- spread (IQR)
- shape (symmetric? skewed?)
- outliers?

This example overlays the boxplot with a jittered dotplot, so we can also see the individual observations.

This visualization made the longlist for the 2015 Kantar Information is beautiful award. Data & R code: https://github.com/zonination/perceptions

1D: Density ridgeline plots

Another possible 1D display is a density estimate—a statistically smoothed histogram.

For comparing a set of them, a ridgeline plot stacks them vertically to create the impression of a mountain range.

As in the boxplot version, this uses:
- a progressive scale of colors
- transparent colors to handle overlap

Q: What features stand out here?

Software note: These figures are drawn with R, using ggplot2 and the ggridges package. See: https://cran.r-project.org/web/packages/ggridges/vignettes/introduction.html

1.5D: Time series line graphs

William Playfair (1786), The Commercial and Political Atlas, invented the time series line graph as a way to show data on England’s trade with other countries.

One curve for imports, one for exports

The balance of trade could be seen as the difference between the curves

Trade with Germany was consistently in favor of England

With North America, the balance changed back and forth over time

Economic ‘history’ could now be visualized and explained

Psychology: Distances between curves

What Playfair didn’t know is that judgments of distance between curves are biased.

We tend to see the perpendicular distance rather than the vertical distance.

Plotting balance of trade directly
Time series graphs

Things get messy when there are many series to be compared
To be fair, this was designed as timeline of history— a visual story of economics. It was Playfair’s last graph.


Parallel ranked list charts

Another solution for multiple time series is to chart the ranks of observations and connect them with lines to show changes in relative position.

Slopes of lines reflect change in rank
Red bars try to show the numbers

From: Statistical Atlas of the United States (1880)

2D: Scatterplots: Ford Nation

Who voted for Rob Ford in the 2014 Toronto mayoral election?

These simple scatterplots by data journalist Patrick Cain use simple enhancements:
• Color, for candidate (Chow, Ford, Tory)
• Overall regression line


Scatterplots: Wage gap

How to compare salaries of men & women in different occupations?

The NYT chose to plot median salaries for women against those for men, in different occupational groups

The 45° line represents wage parity
Other lines show 10, 20, 30% less for women

Alberto Cairo, The Truthful Art, Fig 9.19, from:
This graph, from fivethirtyeight.com was designed to show how some presidential candidates had shifted positions before the 2016 election.

The axes are a score on social and economic policy, but they rotate the axes by 45° to create zones related to political thought.

This info graphic is eye-catching and self-explanatory:
- colored/labeled zones
- interpretive labels on axes
- arrows showing movement to extremes

Data from the US draft lottery, 1970
- Birth dates were drawn at random to assign a “draft priority value” (1=bad)
- Can you see any pattern or trend?
  This is an example of data with a weak signal and a lot of noise

Drawing a smooth curve shows a systematic decrease toward the end of the year.
- The smooth curve is fit by loess, a form of non-parametric regression.

Another form of smoothing is to make one variable discrete & show a graphical summary – here a boxplot
The decrease in later months becomes apparent
Perception: the boxplots form the foreground; the jittered points show the data
Scatterplot matrices

A scatterplot matrix shows the bivariate relation between all pairs of variables. Seeing these all together is more useful than a collection of separate plots.

How does occupational prestige depend on %women, education and income?

The individual plots are enhanced with linear regression lines and non-parametric smooths to show non-linearity.

This figure uses scatterplotMatrix() in the car package. There are many options.

Density plots are often more useful for showing the shapes of distributions
- women: bimodal
- income: highly skewed

A data ellipse gives a visual summary of the direction and strength of the relationship.

Again, graphical annotation provides aids for interpretation.

Scatterplot matrices hold up well with a larger number of variables

Where to live in NYC?

This SPM shows 12 variables on ~ 60 neighborhoods.

The data ellipses provide a visual summary.

This remarkable chart shows survival on the Titanic, by Class for passengers and Gender and Age.

It was drawn by G. Bron, a graphic artist, and published in The Sphere, one month after the Titanic sank.

It uses back-to-back bar charts, with area ~ frequency.

See our web page: http://datavis.ca/papers/titanic/
### Categorical data: Mosaic plots

Similar to a grouped bar chart. Shows a frequency table with tiles, area ~ frequency.

```r
> data(HairEyeColor)
> HEC <- margin.table(HairEyeColor, 1:2)
> HEC

<table>
<thead>
<tr>
<th>Hair</th>
<th>Eye</th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>68</td>
<td>20</td>
<td>15</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Brown</td>
<td>119</td>
<td>84</td>
<td>54</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>Red</td>
<td>26</td>
<td>17</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Blond</td>
<td>7</td>
<td>94</td>
<td>10</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

> chisq.test(HEC)

Pearson's Chi-squared test

data:  HEC
X-squared = 140, df = 9, p-value < 2e-16
```

How to understand the association between hair color and eye color?

### Mosaic plots

Shade each tile in relation to the contribution to the Pearson $\chi^2$ statistic

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

```r
> round(residuals(chisq.test(HEC)),2)

<table>
<thead>
<tr>
<th>Hair</th>
<th>Eye</th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>4.4</td>
<td>-3.1</td>
<td>-0.5</td>
<td>2.2</td>
<td>4.3</td>
</tr>
<tr>
<td>Brown</td>
<td>1.2</td>
<td>-1.8</td>
<td>2.1</td>
<td>-0.3</td>
<td>-0.1</td>
</tr>
<tr>
<td>Red</td>
<td>-0.1</td>
<td>-1.7</td>
<td>1.4</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Blond</td>
<td>-5.8</td>
<td>7.0</td>
<td>-2.2</td>
<td>0.8</td>
<td>-0.6</td>
</tr>
</tbody>
</table>
```

Mosaic plots extend readily to 3-way + tables. They are intimately connected with loglinear models. See: Friendly & Meyer (2016), Discrete Data Analysis with R, [http://ddar.datavis.ca/](http://ddar.datavis.ca/)

### Parallel Sets

Titanic data: Who survived?

Parallel sets use parallel coordinate axes to show the relations among categorical variables.

The frequencies of one variable (Class) are sub-divided according to the joint frequencies in the next (Sex) and shown by the width of the connecting line.

The ParSets application is interactive:
- categories can be reordered (a, b)
- categories can be grouped (c, d)

From Kosara et al. (2006), [https://kosara.net/papers/2006/Kosara_TVCG_2006.pdf](https://kosara.net/papers/2006/Kosara_TVCG_2006.pdf)

### Sankey diagram

Pantheon, by Valerio Pellegrini

Visualizing the 100 most influential figures in History (Wikipedia visits)

Columns show occupation, country of origin and gender

Flow lines link individuals to the column variables, width ~ influence
Sankey diagram

Multiple dimensions of the most influential people in history

From: http://visualoop.com/blog/83382/pantheon-by-valerio-pellegrini

Generalized pairs plots

Generalized pairs plots from the gpairs package handle both categorical (C) and quantitative (Q) variables in sensible ways

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>Q</td>
<td>scatterplot</td>
</tr>
<tr>
<td>C</td>
<td>Q</td>
<td>boxplot</td>
</tr>
<tr>
<td>Q</td>
<td>C</td>
<td>barcode</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>mosaic</td>
</tr>
</tbody>
</table>

library(gpairs)
data(Arthritis)
gpairs(Arthritis[, c(5, 2:5)], ...)

3D: Iso-contour maps

Early attempts to show 3D data used contours of equal value on a map
The data was actually very thin; the contours the result of imaginative smoothing

Francis Galton, Isochronic chart of travel time, 1881

3D: Bivariate density estimation

John Snow’s map of cholera deaths in London, 1854
Modern statistical techniques can compute contours of constant density
Italian demographer Luigi Perozzo (1880) develops the first true 3D diagram showing the population of Sweden over years and age groups as a 3D surface.

Census counts for a given year are shown by the red lines. Survival of a given age are shown by black lines. Cohorts are shown by lines down & to the right. These 3 variables are primary in demography.

A mystery here: what caused the decline at the upper right?

How does occupational prestige depend on income & education?

This plot shows the data and a fitted multiple regression surface, connecting the points to the regression plane. It is hard to see in a static view, but easier when the plot is rotated dynamically.

This plot is produced in R, using the car and rgl packages.

data("Duncan", package="car")
scatter3d(prestige ~ income + education, data=Duncan, id.n=2)
movie3d(spin3d(c(0,1,0), rpm=6), duration=6, movie="duncan-reg3d")

Thematic maps use a wide variety of techniques to display quantitative or qualitative variables on the geographic framework of a map.

Once the domain of cartographers, these ideas are now being developed as an area of geospatial visualization and geospatial statistical methods.

Basic types of thematic maps

Most are direct mappings of numbers to visual variables.

Isopleth maps combine some analysis with display.

From: Slocum et al., Thematic cartography and geographical visualization, Fig 4.3

From: Slocum et al., Thematic cartography and geographical visualization, Fig 4.9
**Thematic maps: Theory**

Alan MacEachern (1979) classifies point, line and area symbols on thematic maps according to whether they depict **quantitative** or **qualitative** phenomena, in the physical or cultural domain.

Theories, ideas, and methods have advanced considerably since this time.

**Choropleth maps**

Balbi & Guerry (1829)

- First thematic maps of crime data
- First comparative maps ("small multiples")
- Crime against persons inversely related to crime against property
- Education: *France obscure & France éclairée*
- N. of France highest in education & also property crime

**Anamorphic maps**

- *Anamorph*: Deforming a spatial size or shape to show a quantitative variable
- Émile Cheysson used this to show the decrease in travel time from Paris to anywhere in France over 200 years

**What’s wrong with choropleth maps?**

Choropleth maps are misleading because size (area) of units dominates perception. This is particularly true for maps of the US & Canada. Not so for France (why?)

Montana looks bigger than Washington

Note use of labels for small NE states

*fivethirtyeight.com* election predictions, Oct. 13, 2017
A tilegram uses hexagonal tiles to make area proportional to a given variable. Here, the size of each state is made proportional to the number of electoral college votes. Now, it is easy to see the impact of states.

Fivethirtyeight.com election predictions, Oct. 13, 2017

Worldmapper: The world in cartograms

How to visualize social, economic, disease, ... data for geographic units?

Worldmapper.com: cartograms: area ~ variable of interest (700+ maps)

Worldmapper: Carbon emissions

These pages are well-designed according to data vis. Ideas: high impact graph + interpretive details & explanation

Worldmapper: Cholera deaths

Deaths from cholera in 2004. Territory size ~ proportion of worldwide deaths.

In the *Cambridge Online Survey of World Englishes*, Bert Vaux and Marius L. Jøhndal surveyed 11,500 people to study the ways people use English words. NC State Univ. student Joshua Katz turned the US data into shaded kernel density maps.

Take the survey: [http://www.tekstlab.uio.no/cambridge_survey](http://www.tekstlab.uio.no/cambridge_survey)
Programming in R: [http://blog.revolutionanalytics.com/2013/06/r-and-language.html](http://blog.revolutionanalytics.com/2013/06/r-and-language.html)

**Flow maps**

Flow maps show movement or change in a geographic framework. The master work is this image by Charles Joseph Minard (1869):

- Marey (1878): “defies the pen of the historian in its brutal eloquence”
- Tufte (1983): “the best statistical graphic ever produced”

**Effect of US civil war on cotton trade**

Note the deformation of the map to accommodate the data.
The Great Migration

In a graphic tribute to C. J. Minard and W. E. B. Du Bois, Raymond Andrews & Howard Wainer tell the story of the migration of blacks from the southern US after freedom from slavery.

Network visualization

Once the domain of mathematicians & computer scientists, graph theory and network visualization turn out to have surprising & interesting applications.

Network visualization: Transport maps

How do I get from Chigwell to Charing Cross? How much will it cost?

This route map shows the connections and fare zones

The first one was designed by Henry Beck in 1931.

The modern version is zoomable and available on your phone.

Network visualization: Shakespeare tragedies

A new form of literary criticism?

Martin Grandjean looked at the structure of Shakespeare tragedies through character interactions.

Each circle (node) represents a character, and an edge represents two characters who appeared in the same scene.

The structural characteristics of the graphs have meaningful interpretations.
Various tasks can be used to assess the relations among words/concepts in our semantic memory. The data can be used to calculate measures of similarity, and be shown in network or other diagrams.

**Verbal fluency task**: Say/write all the names of [animals, countries, …] you can in 1 minute.

**Similarity ratings**: For each pair, indicate how similar they are.

From: Wulff et al. (2018), Structural differences in the semantic networks of younger and older adults.

---

**Love, Actually: Interactive app**

Interactions among characters in Love, Actually

Data:

Interactive Shiny app: https://dgrtwo.shinyapps.io/love-actually-network/

From: http://jonathanstray.com/a-full-text-visualization-of-the-iraq-war-logs

---

**WikiLeaks Iraq war logs**

Johnathan Stray & Julian Burgess analyzed > 11,000 documents for SIGACT (“significant action”) reports from the 2006 Iraqi civil war made available by WikiLeaks.

Each report is a dot. Each dot is labelled by the three most “characteristic” words in that report.

Documents that are “similar” have edges drawn between them, width ~ similarity.

The graph-drawing algorithm placed similar nodes together.

From: http://jonathanstray.com/a-full-text-visualization-of-the-iraq-war-logs
**WikiLeaks Iraq war logs**

Certain themes became clear, and could be studied in rich detail. The underlying methods use “term frequency–inverse document frequency” measures of text-mining.

Murder cluster. All contain the word “corpse”

Torture-abduction cluster

```
library(rtweet)
followers <- get_followers("datavisFriendly")
```


**Twitter network of R users**

Perry Stephenson explores the connections among the top 50 R users on Twitter.

```
library(rtweet)
followers <- get_followers("datavisFriendly")
```


**Tree diagrams**

Trees are natural, organic visual metaphors for branching processes and space-filling designs.

Ramon Llull’s tree of science, showing roots and branches of knowledge

Charles Darwin’s first visual sketch of the evolution of species

**History as a Tree:**


- The entire history of Europe in one diagram
- space-filling design: resolution ~ time²
- natural metaphors for roots, branches

History as a Tree

- Branches for countries & domains of thought
- Leaves for all the details
- Linear horizontal scale $\rightarrow$ area $\sim$ time^2

Where has my hard disk space gone?

Treemaps display hierarchical data as a set of nested rectangles.

- Each node (leaf) has an area $\sim$ size (file space)
- The construction makes efficient use of space
- Nesting shows relative size at multiple levels

Treemaps: Google Newsmap

They turn out to be useful in a wide range of applications

Google NewsMap shows top news stories with
- Size $\sim$ popularity
- Color: domain–world news, sports, national, ...
- Shades: recency

Interactivity: Hover, click to show details

Radial trees: Visual Thesaurus

The Visual Thesaurus, from Thinkmap was the first application to make word meanings visual and interactive.

- They used a radial layout to show the various related senses of given focus word.
- This application was incisive in promoting ideas of interaction with tree-based data: query, zoom, tool-tips, ...

This fig from Manuel Lima, The Book of Trees, p. 127
Animation & Interactive Graphics

- Origins: Visualizing motion
- Animated graphics
- Dynamically updated graphics
- Linking views
- Interactive application development frameworks

E.-J. Marey: A science of visualizing motion

- Physiology: How to make internal physiological processes subject to visual analysis?
  - Invented many graphic recording devices (heart rate, blood pressure, muscle contraction, etc.)
  - “Every kind of observation can be expressed by graphs”

Marey’s sphygmograph, recording a visual trace of arterial blood pressure

Animation: Chronophotography

Marey pioneered the study of human and animal motion photographically

Animated graphics

Animated graphics, like movies are just a series of frames strung together in a sequence

The data for this animation come from human figures in motion-capture suits dancing the Charleston.

The Carnegie-Mellon Graphics Lab maintains a Motion Capture Database, http://mocap.cs.cmu.edu/

From: http://blog.revolutionanalytics.com/2017/08/3-d-animations-with-r.html
Statistical animations

Statistical concepts can often be illustrated in a dynamic plot of some process.

This example illustrates the idea of least squares fitting of a regression line.

As the slope of the line is varied, the right panel shows the residual sum of squares.

This plot was done using the animate package in R.

Animated graphics

Hans Rosling captivated audiences with dynamic graphics showing changes over time in world health data

Video: Hans Rosling, “The best stats you’ve ever seen,”
https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

Animation & Interactivity

The Gapminder “moving bubble chart” was the vehicle.

• Choose \((x, y)\) variables
• Choose bubble size variable
• Animate this over time

Liberating the \(X\) axis from time opened new vistas for data exploration

Software made this available as a general tool

Dynamically updated data visualizations

You don’t need a weatherman to know which way the wind is blowing. The wind map app, http://hint.fm/wind/ is one of a growing number of R-based applications that harvests data from web sources, and presents a visualization
Linking animated views

This example links a **dendrogram** to a **grand tour** and map of the USArrests data to visualize a classification in 5 dimensions.

The grand tour animates a series of 2D projections of the 5D data.

The image is recorded as a GIF.


---

Interactive application frameworks

shiny for R makes it easy to create interactive applications

https://walkerke.shinyapps.io/neighborhood_diversity/

---

shiny gallery

There is now a large collection of shiny applications, https://shiny.rstudio.com/gallery/

These integrate other interactive web software: d3, Leaflet, Google Charts, ...

---

Summary

- The topics here were largely about data graphs, for analysis & presentation. Mainly not Info-graphics
  - Quantitative data: different forms for 1D, 1.5D, 2D, 3+D data
  - Categorical data: often best shown as areas ~ frequency (bar plots, mosaic plots)
- Thematic maps: visualizing spatially varying data
  - Raw data with different visual encodings
  - Spatial statistical models provide some smoothings
- Networks/trees: visualizing connections
- Animation: show changes over time or space
- Interaction: allow the viewer to explore the data