Visualization

Spiros Boosalis & Connor Gramazio
Journalism has very little in common with big data. (The data in journalism is almost entirely tiny.) But there may be some similarities, at least in spirit: we both to know things we shouldn't be able to know, depend heavily on asking the right questions and quick iteration, and prefer way more detail than we actually need, at least at the beginning.

I'll review some of the NYT graphics department's "bigger" data collaborations with academics and others, and discuss some of the broader trends in data visualization, as it is practiced by journalists trying to communicate with large audiences.

https://calendar.csail.mit.edu/events/116379
Standard Fare
Why is it important to know about visualization?
hoods, and create a noise matrix whose entries equal 
\[ \hat{p}_{s,t} = 0.5 \sum_{i=1}^{n} p_{i,t}/n + 0.5 \sum_{j=1}^{m} p_{s,j}/m. \] The action of sampling from the two source charts produces a corresponding \( P_{\text{combined}} = P_{\text{definitive}}^{k(1-\gamma)} \ast P_{\text{noise}}^{\gamma} \) where \( \ast \) is the convolution operator; mathematical derivations are provided in the supplementary materials. We compute \( \gamma \) by solving the convex optimization:

\[ \arg\min_{\gamma} KL(P_{\text{definitive}}^{k(1-\gamma)} \ast P_{\text{noise}}^{\gamma} \| P) \]

The optimal \( \gamma \) value represents the estimated amount of matches that can be attributed to noise. We then estimate the most likely distribution of topical matches \( P_{\text{denoised}} \) without the chance matches, by solving the following constrained optimization:

\[ \arg\min_{P_{\text{denoised}}} KL(P_{\text{denoised}} \ast P_{\text{noise}}^{\gamma} \| P) \]

subject to \( P_{\text{denoised}} \) being a proper probability distribution whose entries sum to 1 and are in the range [0,1]. We apply the above process to each row and column in the correspondence matrix, to obtain \( \hat{P}_{\text{denoised}} \) and \( \hat{P}_{\text{denoised}} \) from which we estimate topical misalignment scores as described previously.

4. Reference Concepts
Reference concepts in our diagnostic framework can be determined from various sources: elicited from domain

Figure 5. Topical alignment. LDA models for \( N \in [1, 80] \), \( \alpha = 5/N, \beta = 0.25 \). The y-axis shows the fraction of reference concepts that have a single matching topic (resolved), multiple matching topics (repeated) or are subsumed by one (fused) or multiple fused topics (fused & repeated). These models uncover up to 75% of the reference concepts, but coverage increases only marginally for \( N \geq 10 \).

Figure 6. Topical alignment. LDA models for \( N \in [1, 80] \), \( \alpha = 50/N, \beta = 0.001 \). This series of models uncovers up to 40% of the reference concepts. Coverage peaks at \( N=8 \). The proportion of resolved and fused topics remains stable for \( N \geq 15 \); increasing \( N \) produces only more junk topics.
The right viz for the right data

Continuous data

Discrete Data
Critiques
ECUADOR MAXIMUM PRISON SENTENCES

Drug Traffic: 16 yrs
Murder: 12 yrs

BOLIVIA MAXIMUM PRISON SENTENCES

Drug Traffic: 25 yrs
Murder: 20 yrs

MEXICO MAXIMUM PRISON SENTENCES

Drug Traffic: 25 yrs
Murder: 24 yrs
ECUADOR MAXIMUM PRISON SENTENCES
- Drug Traffic: 16yrs
- Murder: 12yrs

BOLIVIA MAXIMUM PRISON SENTENCES
- Drug Traffic: 25yrs
- Murder: 20yrs

MEXICO MAXIMUM PRISON SENTENCES
- Drug Traffic: 25yrs
- Murder: 24yrs

non grounded axis (12 ≠ \frac{1}{2} 16)
non grounded axis
inconsistent axes
non grounded axis

inconsistent axes

chartjunk
The Question:
how do the punishments for murdering someone and trafficking drugs compare?
**ECUADOR** MAXIMUM PRISON SENTENCES
- Drug Traffic: 16 yrs
- Murder: 12 yrs

**BOLIVIA** MAXIMUM PRISON SENTENCES
- Drug Traffic: 25 yrs
- Murder: 20 yrs

**MEXICO** MAXIMUM PRISON SENTENCES
- Drug Traffic: 25 yrs
- Murder: 24 yrs

**bumps chart**
(data, not visualization, but...) why not median?
Online vs. store shopping

BY GENDER

Men
- Shop web more than stores: 25.5%
- Shop stores more than web: 53.4%
- Shop web and stores equally: 21.2%

Women
- Shop web more than stores: 17.8%
- Shop stores more than web: 51.7%
- Shop web and stores equally: 30.6%

BY INCOME

<$30,000
- Shop web more than stores: 20.0%
- Shop stores more than web: 50.3%
- Shop web and stores equally: 29.6%

$30,000-$59,999
- Shop web more than stores: 20.6%
- Shop stores more than web: 56.4%
- Shop web and stores equally: 23.0%

$60,000-$99,999
- Shop web more than stores: 19.2%
- Shop stores more than web: 52.4%
- Shop web and stores equally: 28.4%

$100,000-$149,999
- Shop web more than stores: 25.9%
- Shop stores more than web: 53.6%
- Shop web and stores equally: 20.6%

$150,000+
- Shop web more than stores: 34.5%
- Shop stores more than web: 44.0%
- Shop web and stores equally: 21.5%

Source: Compete Inc.
The Question: how does gender/wealth effect shopping online/offline?
Online vs. store shopping

BY GENDER

- Men: Shop web more than stores (53.4%), Shop stores more than web (51.7%), Shop web and stores equally (30.5%), Shop web more than stores (25.5%), Shop stores more than web (17.8%), Shop web and stores equally (20.4%)

- Women: Shop web more than stores (53.4%), Shop stores more than web (51.7%), Shop web and stores equally (30.5%), Shop web more than stores (21.2%), Shop stores more than web (17.8%), Shop web and stores equally (20.4%)

BY INCOME

- $<100K: Shop web more than stores (56.4%), Shop stores more than web (53.4%), Shop web and stores equally (51.7%), Shop web more than stores (20.4%), Shop stores more than web (18.8%), Shop web and stores equally (19.2%)

- $>100K: Shop web more than stores (53.6%), Shop stores more than web (51.6%), Shop web and stores equally (50.5%), Shop web more than stores (20.4%), Shop stores more than web (18.8%), Shop web and stores equally (19.2%)

Higher-income households and Men Are More Likely to Use Web more than Stores Compared to Lower-income Households and Women

Proportion of Sample

- Use Stores More
- Use Web More
- Both Stores and Web

Poll results from online vs. store shopping by gender and income.
What’s The Question?
What's The Question?
Life expectancy by health expenditures per capita, 2007

Health expenditures are total (public and private), in PPP-converted US dollars. Data source: OECD.
The Question:
how does the US’s life expectancy and health expenditures compare to other wealthy nations?
“US” is in bigger font

Life expectancy by health expenditures per capita, 2007

Health expenditures are total (public and private), in PPP-converted US dollars. Data source: OECD.
“US” is in bigger font
the only color is “US”
“US” is in bigger font
the only color is “US”
clean axes (just min and max)
“US” is in bigger font
the only color is “US”
clean axes (just min and max)
no other labels
Moving beyond the page:

Interaction
Interaction should

- make sense
- combat visual complexity
- improve understanding
Interaction Taxonomy

- Select
- Explore
- Reconfigure
- Encode
- Abstract/Elaborate
- Filter
- Connect

Yi, Ji Soo and Kang, Youn ah and Stasko, John and Jacko, Julie: Toward a Deeper Understanding of the Role of Interaction in Information Visualization, IEEE Transactions on Visualization and Computer Graphics, 2007
Select: mark something as interesting
Explore: show me something else
Reconfigure: show me a different arrangement
Encode: show me a different representation
Abstract/Elaborate: show me more or less detail
Filter: show me something conditionally
Connect: show me related items
matplotlib

d3

gnuplot
d3.js

~ jquery (i.e. selectors and chaining)

```
svg = d3.select("#viz")
  .append("svg")
  .attr("width", 800)
  .attr("height", 800)
```
d3.js

separate data from visualization

data = [32, 57, 112, 293]

circle = svg.select(“circle”).append(“circle”)
  .data(data)

circle.attr(“r”, function(d, i){d*d})
circle.attr(“cx”, function(d, i){i*100})
callback
function(datum, index) {this

circle.data([32, 57, 112, 293])
=>
datum = 32
index = 0
this = <circle></circle>
d3.js

extra data
circle.enter()

extra elems
circle.exit()
d3.js

transition = circle
  .transition()
  .delay(100)
  .duration(1000)
d3.js

transition
  .style("color", "red")
  .attr("r", 100)

get starting values form DOM (green and 10)
get type of value (color and number)
built interpolator
schedule callbacks
Brushing
```javascript
var x = d3.time.scale().range([0, width]),
    x2 = d3.time.scale().range([0, width]),
    y = d3.scale.linear().range([height, 0]),
    y2 = d3.scale.linear().range([height2, 0]);

var brush = d3.svg.brush()
  .x(x2)
  .on("brush", actionOnBrushEventFn);

var area = d3.svg.area()
  .interpolate("monotone")
  .x(function(d) { return x(d.date); })
  .y0(height)
  .y1(function(d) { return y(d.price); });

var area2 = d3.svg.area()
  .interpolate("monotone")
  .x(function(d) { return x2(d.date); })
  .y0(height2)
  .y1(function(d) { return y2(d.price); });

var focus = svg.append("g"),
    context = svg.append("g");

var x = d3.time.scale().range([0, width]),
    x2 = d3.time.scale().range([0, width]),
    y = d3.scale.linear().range([height, 0]),
    y2 = d3.scale.linear().range([height2, 0]);

var brush = d3.svg.brush()
  .x(x2)
  .on("brush", actionOnBrushEventFn);

var area = d3.svg.area()
  .interpolate("monotone")
  .x(function(d) { return x(d.date); })
  .y0(height)
  .y1(function(d) { return y(d.price); });

var area2 = d3.svg.area()
  .interpolate("monotone")
  .x(function(d) { return x2(d.date); })
  .y0(height2)
  .y1(function(d) { return y2(d.price); });

var focus = svg.append("g"),
    context = svg.append("g");

d3.csv("sp500.csv", function(error, data) {
  x.domain(d3.extent(data.map(function(d) { return d.date; })));
  y.domain([0, d3.max(data.map(function(d) { return d.price; }))]);
  x2.domain(x.domain());
  y2.domain(y.domain());

  focus.append("path")
    .datum(data)
    .attr("clip-path", "url(#clip)"
    .attr("d", area);

  context.append("path")
    .datum(data)
    .attr("d", area2);

  context.append("g")
    .attr("class", "x brush")
    .call(brush)
    .selectAll("rect")
    .attr("y", -6)
    .attr("height", height2 + 7);
});
```
Resources

d3: d3 homepage (filled with examples).

StackOverflow: Mike Bostock, creator of d3, and other top contributors regularly answer questions.

Crossfilter.js: JavaScript library for exploring large multivariate datasets in the browser.

JunkCharts: critiques charts.

d3 helloworld

Supplementary or alternative libraries:

- dc.js: d3js + Crossfilter convenience library. Usable on top of vanilla d3.
- NVD3.js: d3 chart convenience library. Usable on top of vanilla d3.
- Vega: super simple json-based visualization generator. Built by Trifacta, a start up consisting of the Stanford viz crew (d3 and data wrangler folks) and other Bay-area viz folks.
- Chart.js: chart-only visualization library. Canvas-based (not d3), but easy and beautiful.
- CanvasJS: canvas-based visualization library.