Creative data visualisation: A structured approach

Dec 2, 2013

When faced with a mass of data, how do we turn this into a data visualisation?

This is what I asked myself when I first got hold of Tennis-Data's file of Wimbledon 2013 match results. I had no idea what I'd create when I began with this data-set, but I followed a well-defined path that helped me get creative with the data:

It's worth saying that I'm particularly interested in charts off the beaten path. Bar charts, pie charts and suchlike have their place but I want to give opportunity to other, original layouts, especially when tools such as D3 allow such creativity.

I've identified four distinct stages, or milestones, in the process:

1. Raw Data
2. Questions
3. Structured Data and
4. The Visualisation.

4 Stages of Creative Data Visualisation

1. Raw Data

It seems obvious, but first you need your data. Don't underestimate this task. Perhaps you've already been given some data to visualise. Or you have an idea but no data, in which case you'll need to find suitable data sources. (Google can help out here.)

I was lucky with the Wimbledon visualisations. I knew I wanted to make some visualisations of Wimbledon and I quickly found a very well formed dataset at Tennis-Data:

```
ATP,Location,Tournament,Date,Series,Court,Surface
37,London,Wimbledon,24/06/13,Grand Slam,Outdoor,
37,London,Wimbledon,24/06/13,Grand Slam,Outdoor,
37,London,Wimbledon,24/06/13,Grand Slam,Outdoor,
37,London,Wimbledon,24/06/13,Grand Slam,Outdoor,
37,London,Wimbledon,24/06/13,Grand Slam,Outdoor,
37,London,Wimbledon,24/06/13,Grand Slam,Outdoor,
etc.
```

2. Questions

Once you have your data, you need to ask it questions. Give it the 3rd degree. With the Wimbledon data I asked:

- who was the overall winner?
- who was runner up etc.?
- who did we expect to win?
- did they win?

- who lost out unexpectedly?
- any major surprises?
- did anyone punch above their weight?
- what was the overall picture of the tournament?
- who played whom?
- what were the most exciting matches?
- who won the most games, the most sets etc.?

Don’t worry about answering them right now, but never stop asking questions of your data. You’ll also find that more questions will come once you’ve produced some visualisations so this can be an iterative process.

3. Structured data

Given our raw data and a number of questions, the next step is to structure the data in such a way that the questions can be answered. A knowledge of data structures is necessary here, but in my experience it boils down to just a few:

- arrays (a list of data)
- networks (a group of things, some of which are connected to each other)
- trees (a group of things, where each thing has other things as its ‘children’)

Arrays

Generally speaking, arrays can be used in most cases. An array is simply a list of data, where each data item (a datum) can consist of parameters such as name, age & height.

Here's the first 3 elements of an array of players I created for the Wimbledon visualisations. Each element represents a player:

```
[
  {
    "name": "Melzer J."
    "matchesWon": 3,
    "setsWon": 11,
    "gamesWon": 95,
    "ranking": 37,
    "points": 1085,
    "roundReached": 4,
    "heroScore": 260,
  },
  {
    "name": "Fognini F."
    "matchesWon": 0,
    "setsWon": 1,
    "gamesWon": 17,
    "ranking": 30,
    "points": 1345,
    "roundReached": 1,
    "heroScore": -260,
  },
  {
    "name": "Reister J."
    "matchesWon": 1,
    "setsWon": 4,
    "gamesWon": 49,
    "ranking": 121,
    "points": 461,
    "roundReached": 2,
    "heroScore": 630,
  },
  etc.
]
```

Network

If the data is connected, then a network structure (a list of nodes and a list of links) is required.

For example, we could create a network structure of players and add a link if two players have played one another:

```
{ "nodes" :
  [
    "Stakhovsky S.",
    "Federer R.",
    "Murray A.",
    "Djokovic N.",
    "Janowicz J."
    etc.
  ],
  "links" :
  [
    { "source": 0, "target": 1 },
    { "source": 2, "target": 3 },
    { "source": 2, "target": 4 }
    etc.
  ]
}
```

Tree

If the data is connected in such a way that it forms a hierarchy, then a tree is probably required.

In the case of the Wimbledon data we could create a tree structure with the overall winner at the root and their children being the players they have defeated. It’s recursive i.e. each of the defeated players might have players that they’ve defeated:

```
{
  "name": "Murray A.",
  "children": [
    {
      "name": "Djokovic N.",
      "children": [
        etc.
      ]
    },
    }
      "name": "Janowicz J.",
      "children": [
         etc.
      ]
    },
    etc.
}
```

Choosing your data structures

Deciding which data structures to use is a bit of an art and gets easier with experience. It requires a degree of imagination and creativity and it’s often one of the most enjoyable parts of the process.

The process is: for each question, decide what data structure would help answer that question. There’s often more than one.

There’s often a degree of experimentation and iteration at this stage, so don’t expect to get the correct data structures straightaway.

Checking your data structures

Once you’ve determined your data structures it’s useful to check that your questions can be answered. For example, the tree data allows us to determine the overall winner.

How about the ‘who’s punched above their weight’ question? We can determine this from the player’s array by calculating the ATP points / games won ratio.

Data wrangling

The process of transforming the raw data into useful, structured data is often referred to as data wrangling. In practical terms there’s a number of techniques for producing the structured data. I like to use Node.js with a little help from underscore.js whilst many others like to use Python.

4. The visualisation

We've got our structured data, whether it be a single array or a mix of arrays, graphs and trees. At this point we can select an off-the-shelf chart type such as a bar chart, pie chart or line chart and plug our data into it:

This is the approach taken by the likes of Excel, Google Chart Tools and Highcharts. It's quick and easy, but not so creative. Nor does it guarantee the most effective visualisation.

Instead we'll use an approach taken by the web-based library D3 where each bit of data (a datum) is assigned to simple graphical elements such as lines, circles or rectangles. We can then set parameters such as the length, width, radius and colour according to the data on each graphical element. This approach is outlined in a paper by D3's authors.

For example, if we have an array of Wimbledon players, we can assign a rectangle to each player and set its length proportional to the number of matches the player has won. In other words, a bar chart:

Just as easily we could assign each datum to a circle and set the radius:

We could even assign circles to each player and set the x and y positions according to ATP points and games won, respectively. We could then take an array of matches and assign an arrow for each match pointing from the winning player to the loser. This would give us a hybrid scatter/network visualisation.

Instead of asking ‘what chart shall we use’, we’re asking ‘what shapes (circles, lines, rectangles, curves) shall we use to represent each element of data and which variables shall we assign to the shape’s parameters (position, height, width, radius, colour, texture, opacity etc.)?’. This approach gives us an incredible number of options for displaying our data. Imagination is the limit!

This is a process that involves a lot of creativity, experimentation and it’s often the most enjoyable part as this is where your hard work of structuring the data manifests into a beautiful visual.

Summing up

I’ve presented the approach I used for creating 10 different visaulisations of a single data-set from this year’s Wimbledon tournament. I’m sure there are many other approaches and I wouldn’t take this approach as gospel. However when feeling the onset of blank page syndrome I hope that this approach will help.

Work with me

If you’re interested in working with me to create a data visualisation please get in touch and I’ll be happy to discuss.