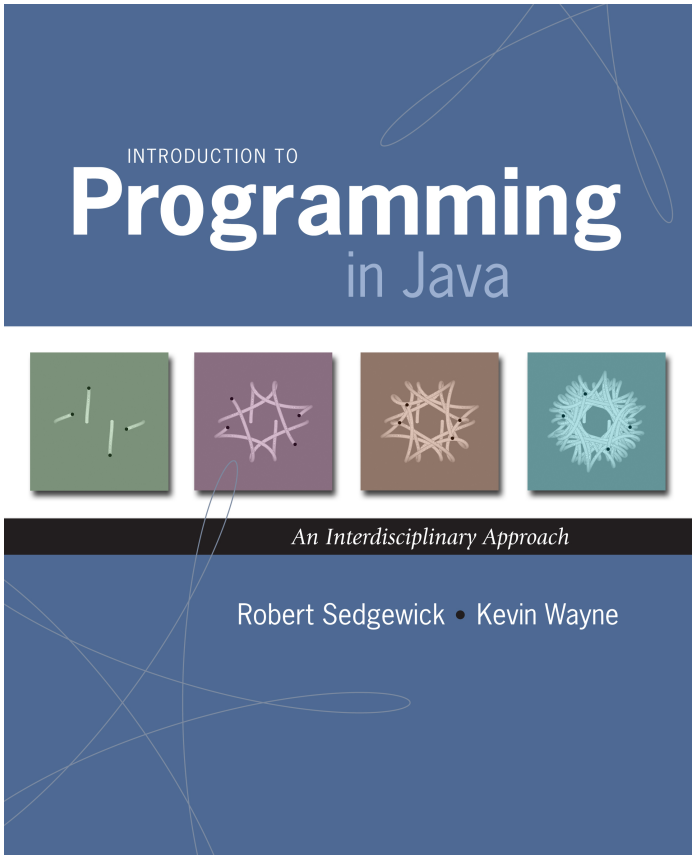
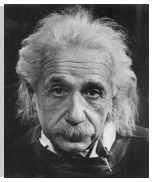
3.5 Case Study: Purple America

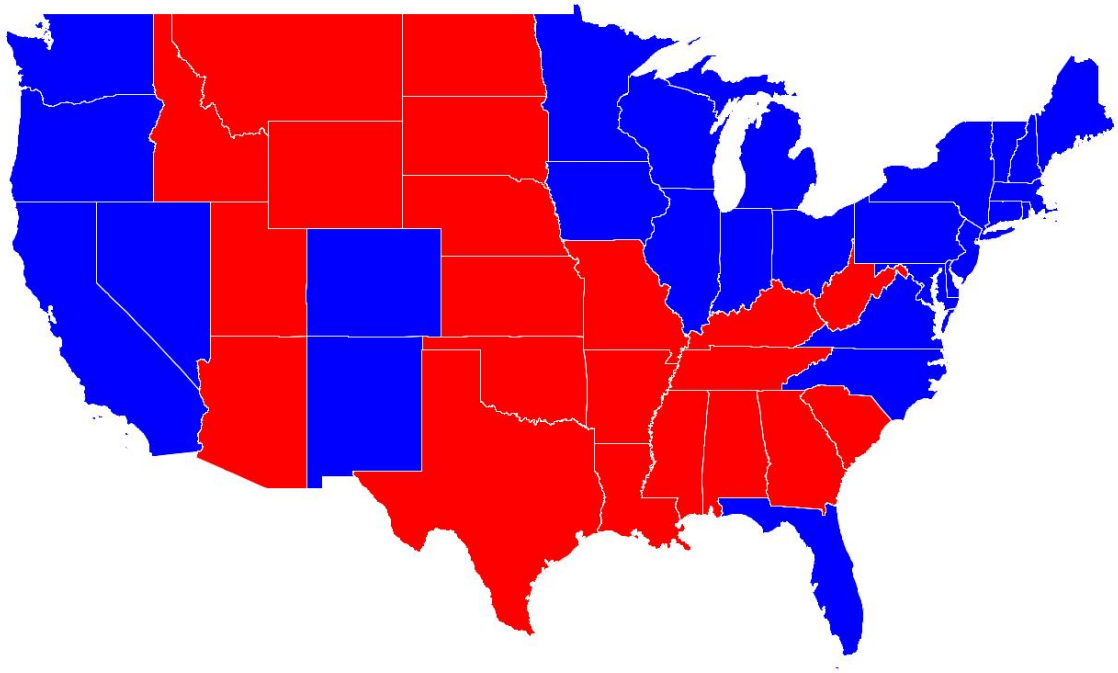


Data Visualization

Challenge. Visualize election results.

> *" If I can't picture it, I can't understand it. "*
> *— Albert Einstein*





2008 Presidential election

Modular Programming

Modular programming.

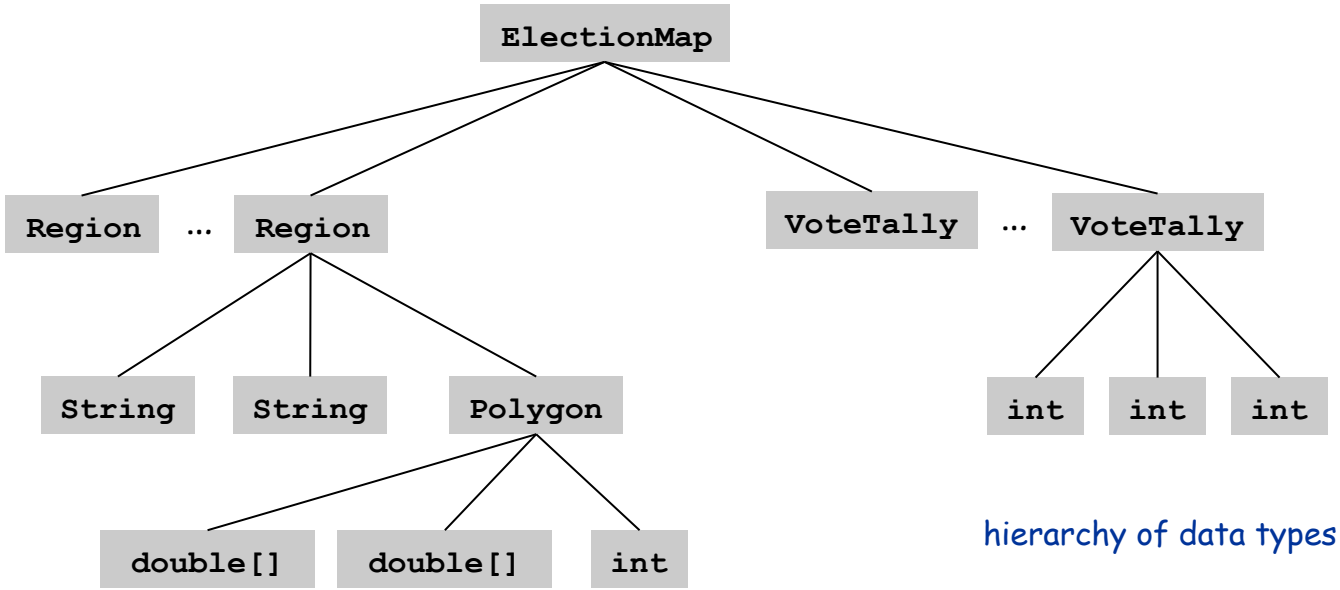
- Model problem by decomposing into components.
- Develop data type for each component.

Polygon. Geometric primitive.

Region. Name, postal abbreviation, polygonal boundary.

Vote tally. Number of votes for each candidate.

Election map. Regions and corresponding vote tallies for a given election.



hierarchy of data types

Data Sources

Boundary Data: States within the Continental US

Geometric data. [US census bureau]

- `www.census.gov/tiger/boundary`
- `NJ.txt` has boundaries of every county in New Jersey.
- `USA.txt` that has boundary of every state.

format useful for programmers

Election results. [David Leip]

- `http://uselectionatlas.org/RESULTS`
- Interactive and graphical.
- Need to screen-scrape to get data.

format useful for browsers and end-users
(need to parse to extract raw data)

Emerging standard.

- Publish data in text form on the web (like geometric data).
- Write programs to produce visuals (like we're doing!)
- Mashups.

Boundary Data: States within the Continental US

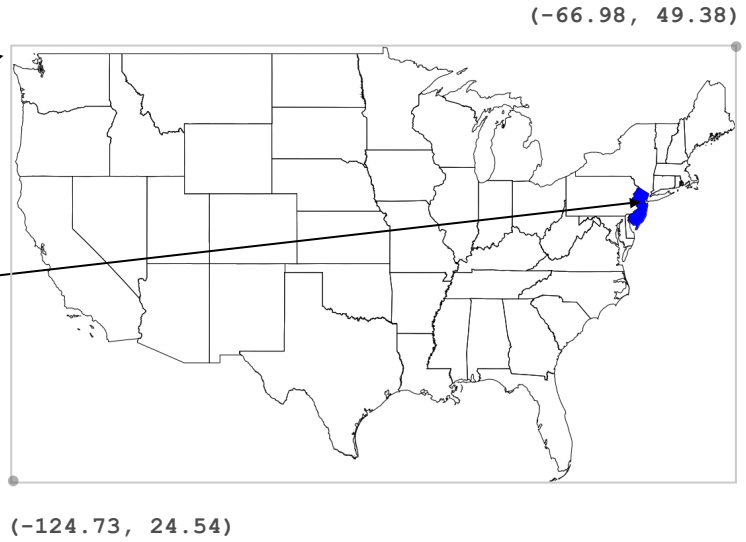
USA data file. State names and boundary points.

```
% more USA.txt
-124.731216 24.544102 -66.980385 49.384365
104
Alabama
USA
498
 -88.200027   34.995548
 -88.202919   35.007942
...
New Jersey
USA
368
 -74.695305   41.357330
 -74.461754   41.250000
 -74.366302   41.202801
...
 -74.721313   41.347294
...
```

number of regions

bounding box

368 points (longitude, latitude)



Boundary Data: Counties within a State

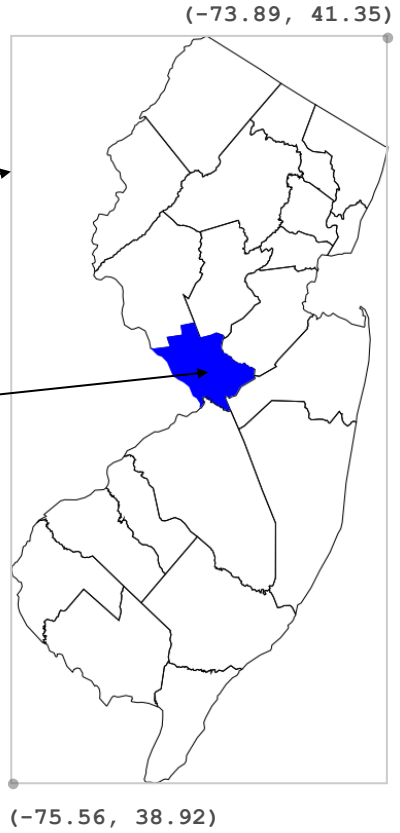
State data files. County names and boundary points.

```
% more NJ.txt
-75.560143 38.928589 -73.894402 41.35733
21
Atlantic
NJ
127
 -74.877563    39.608414
 -74.736694    39.729721
…
Mercer
NJ
88
 -74.748825    40.424248
 -74.722702    40.375301
 -74.674507    40.384399
…
 -74.808403    40.415401
…
```

number of regions

bounding box

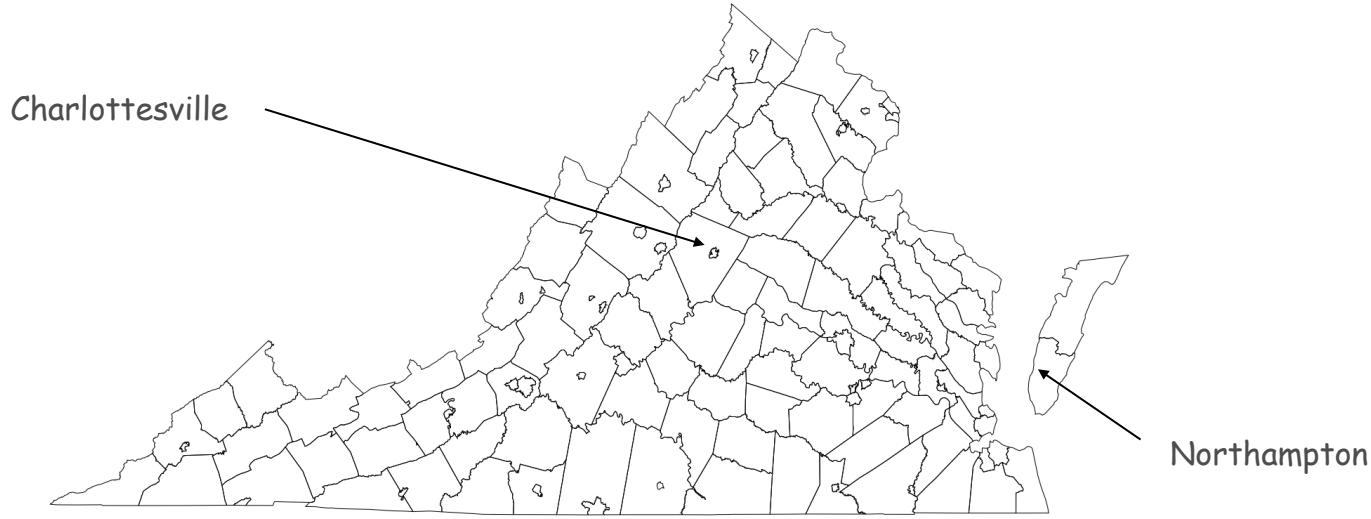
88 points (longitude, latitude)



Pitfalls: Pieces and Holes

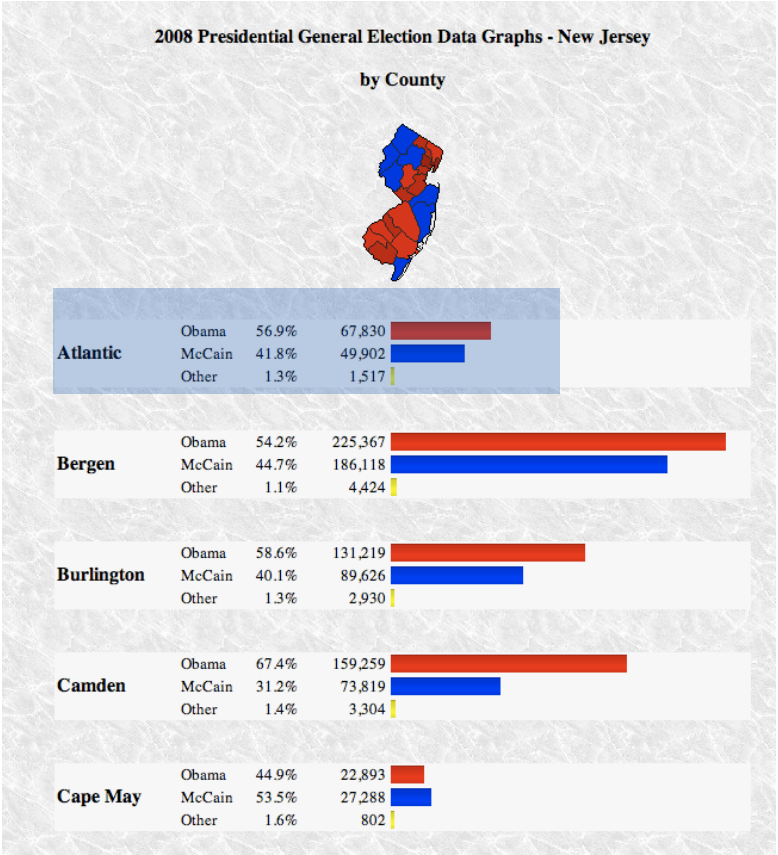
Pieces. A state can be comprised of several disjoint polygons.

Holes. A county can be entirely inside another county.



Screen Scraping the Election Returns

Screen scrape. Download html from web and parse.



county name is text between and tags that occurs after `width:100px`

```
<div>
<br /><b>2008 Presidential General Election Data Graphs - New Jersey<br /><br />by
County</b><br /><br /><img src="img.php?
year=2008&amp;st=NJ&amp;type=map&amp;off=0&amp;fips=34&amp;elect=0" alt="Map" /><br
/><br /><div class="info"><table cellpadding="2"><tr><td style="width:100px"
rowspan="3"><b>Atlantic</b></td><td class="cnd">Obama</td><td
class="per">56.9&#37;</td><td class="dat">67,830</td><td class="bar"><div
class="bardem" style="width:28.2%"> </div></td></tr><tr><td>McCain</td><td
class="per">41.8&#37;</td><td class="dat">49,902</td><td><div class="barrep"
style="width:20.8%"> </div></td></tr><tr><td>Other</td><td
class="per">1.3&#37;</td><td class="dat">1,517</td><td><div class="baroth"
style="width:1.0%"> </div></td></tr></table><br /><br /><table
cellpadding="2"><tr><td style="width:100px" rowspan="3"><b>Bergen</b></td><td
class="cnd">Obama</td><td class="per">54.2&#37;</td><td class="dat">225,367</td><td
class="bar"><div class="bardem"
style="width:93.8%"> </div></td></tr><tr><td>McCain</td><td
class="per">44.7&#37;</td><td class="dat">186,118</td><td><div class="barrep"
style="width:77.5%"> </div></td></tr><tr><td>Other</td><td
class="per">1.1&#37;</td><td class="dat">4,424</td><td><div class="baroth"
style="width:1.8%"> </div></td></tr></table><br /><br /><table
cellpadding="2"><tr><td style="width:100px" rowspan="3"><b>Burlington</b></td><td
class="cnd">Obama</td><td class="per">58.6&#37;</td><td class="dat">131,219</td><td
class="bar"><div class="bardem"
style="width:54.6%"> </div></td></tr><tr><td>McCain</td><td
class="per">40.1&#37;</td><td class="dat">89,626</td><td><div class="barrep"
style="width:37.3%"> </div></td></tr><tr><td>Other</td><td
class="per">1.3&#37;</td><td class="dat">2,930</td><td><div class="baroth"
style="width:1.2%"> </div></td></tr></table><br /><br /><table
cellpadding="2"><tr><td style="width:100px" rowspan="3"><b>Camden</b></td><td
class="cnd">Obama</td><td class="per">67.4&#37;</td><td class="dat">159,259</td><td
class="bar"><div class="bardem"
```

...

http://uselectionatlas.org/RESULTS/datagraph.php?year=2008&fips=34

Election Scraper (sketch)

```
int year    = 2008;  // election year
String usps = "NJ";  // United States postal code for New Jersey
int fips    = 34;    // FIPS code for New Jersey

String url   = "http://uselectionatlas.org/RESULTS/datagraph.php";
In in        = new In(url + "?year=" + year + "&fips=" fips);
Out file     = new Out(usps + year + ".txt");
String input = in.readAll();

while (true) {

  // screen scrape county name
  int p = input.indexOf("width:100px", p);
  if (p == -1) break;
  int from = input.indexOf("<b>", p);
  int to   = input.indexOf("</b>", from);
  String county = input.substring(from + 3, to);

  // screen scrape vote totals for each candidate


  // save results to file
  file.println(county + "," + mccain + "," + obama + "," + other + ",");
}
```

extract text between and tags, that occurs after `width:100px`

More Pitfalls

Data sources have different conventions.

- FIPS codes: NJ vs. 34.
- County names: LaSalle vs. La Salle, Kings County vs. Brooklyn.

Plenty of other minor annoyances.

- Unreported results.
- Third-party candidates.
- Changes in county boundaries.

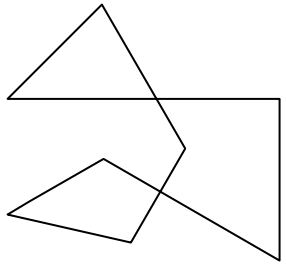
Bottom line. Need to clean up data (but write a program to do it!)

Polygons and Regions

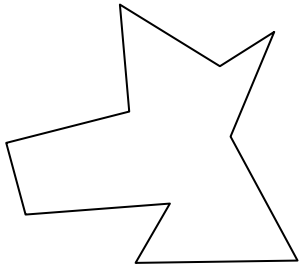
Polygon Data Type

Polygon. Closed, planar path with straight line segments.

Simple polygon. No crossing lines.



polygon
(8 points)



simple polygon
(10 points)

simple polygon
(368 points)

Polygon Data Type: Java Implementation

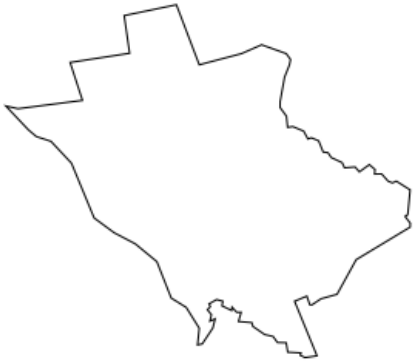
```
public class Polygon {
   private final int N;           // number of boundary points
   private final double[] x, y;   // the points (x[i], y[i])

   // read from input stream
   public Polygon(In in) {
       N = in.readInt();
       x = new double[N];
       y = new double[N];
       for (int i = 0; i < N; i++) {
          x[i] = in.readDouble();
          y[i] = in.readDouble();
       }
   }

   public void fill() { StdDraw.filledPolygon(x, y); }

   public boolean contains(double x0, double y0)    { … }
   public String toString()                         { … }
}
```

see COS 226

easy

Region Data Type

Region. Represents a state or county.



Mercer, NJ
88 point polygon

New Jersey, USA
368 point polygon

Region Data Type: Java Implementation

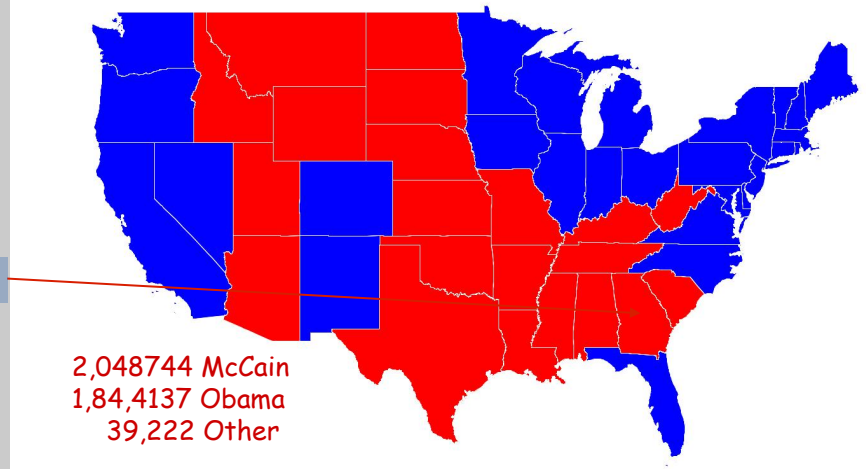
```
public class Region {
   private final String  name;  // name of region
   private final String  usps;  // postal abbreviation
   private final Polygon poly;  // polygonal boundary

   public Region(String name, String usps, Polygon poly) {
      this.name = name;
      this.usps = usps;
      this.poly = poly;
   }

   public void draw() { poly.fill(); }

   public boolean contains(double x0, double y0) {
      return poly.contains(x0, y0);
   }

   public String toString() { … }

}
```

Election Returns

Election Returns: By State

Screen-scraping results. Votes for McCain, Obama, Other by region.

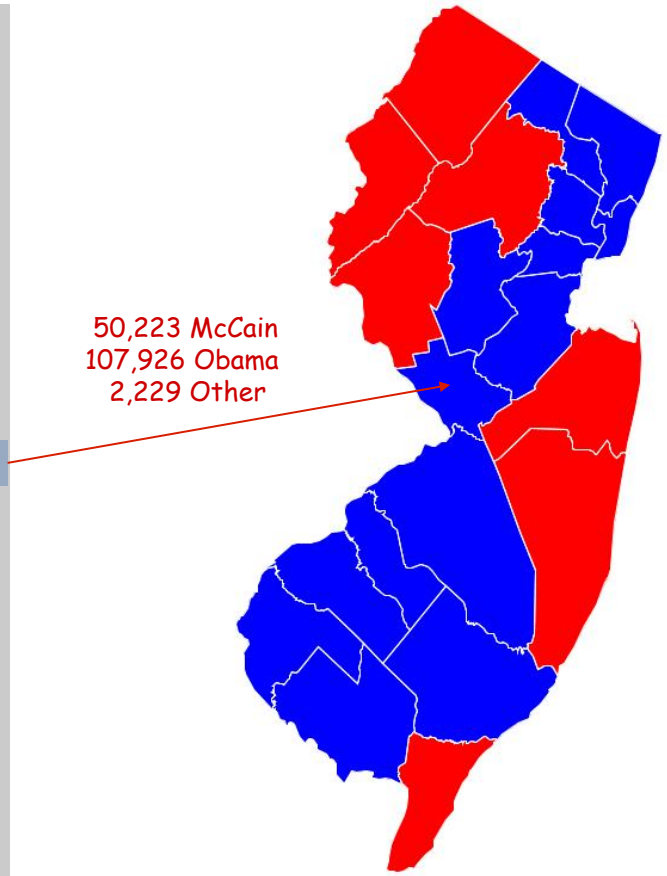
```
% more USA2008.txt
Alabama,1266546,813479,19773,
Alaska,193841,123594,8762,
Arizona,1230111,1034707,39020,
Arkansas,638017,422310,26290,
California,5011781,8274473,289260,
Colorado,1073584,1288568,39197,
Connecticut,629428,997772,19592,
Delaware,152374,255459,4579,
District of Columbia,17367,245800,2686,
Florida,4045624,4282074,82621,
Georgia,2048744,1844137,39222,
Hawaii,120566,325871,7131,
Idaho,403012,236440,17978,
Illinois,2031527,3419673,71851,
…
Virginia,1725005,1959532,38723,
Washington,1229216,1750848,68820,
West Virginia,398061,304127,12550,
Wisconsin,1262393,1677211,43813,
Wyoming,164958,82868,6832,
```



Election Returns: By County

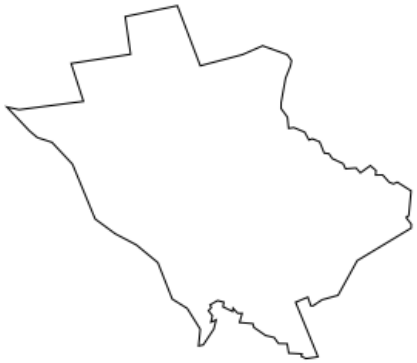
Screen-scraping results. Votes for McCain, Obama, Other by region.

```
% more NJ2008.txt
Atlantic,49902,67830,1517,
Bergen,186118,225367,4424,
Burlington,89626,131219,2930,
Camden,73819,159259,3304,
Cape May,27288,22893,802,
Cumberland,22360,34919,915,
Essex,74063,240306,2181,
Gloucester,60315,77267,1848,
Hudson,55360,154140,2116,
Hunterdon,39092,29776,1147,
Mercer,50223,107926,2229,
Middlesex,123695,193812,4283,
Monmouth,160433,148737,4244,
Morris,132331,112275,2913,
Ocean,160677,110189,4111,
Passaic,72552,113257,1904,
Salem,14816,16044,672,
Somerset,70085,79321,1672,
Sussex,44184,28840,1393,
Union,78768,141417,2241,
Warren,27500,20628,980,
```



Vote Tally Data Type

`VoteTally`. Represents the election returns for one region.



Mercer, NJ

New Jersey, USA

election returns

50,223 McCain
107,926 Obama
2,229 Other

1,613,207 McCain
2,215,422 Obama
47,826 Other

Vote Tally Data Type: Java Implementation

```
public class VoteTally {
   private final int rep, dem, ind;

   public VoteTally(String name, String usps, int year) {
      In in = new In(usps + year + ".txt");
      String input = in.readAll();
      int i0 = input.indexOf(name);
      int i1 = input.indexOf(",", i0+1);
      int i2 = input.indexOf(",", i1+1);
      int i3 = input.indexOf(",", i2+1);
      int i4 = input.indexOf(",", i3+1);
      rep = Integer.parseInt(input.substring(i1+1, i2));
      dem = Integer.parseInt(input.substring(i2+1, i3));
      ind = Integer.parseInt(input.substring(i3+1, i4));
   }

   public Color getColor() {
      if (rep > dem) return StdDraw.RED;
      if (dem > rep) return StdDraw.BLUE;
      return StdDraw.BLACK;
   }
}
```

```
% more NJ2008.txt
…
Mercer,50223,107926,2229,
…
i0    i1    i2    i3    i4
```

Election Map

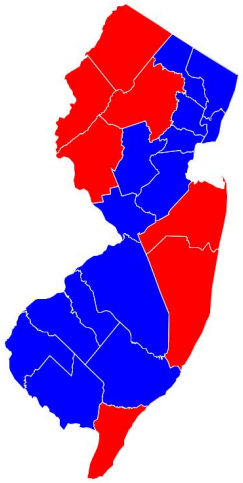
Election Map Data Type

ElectionMap. Represents the election map for a given election.

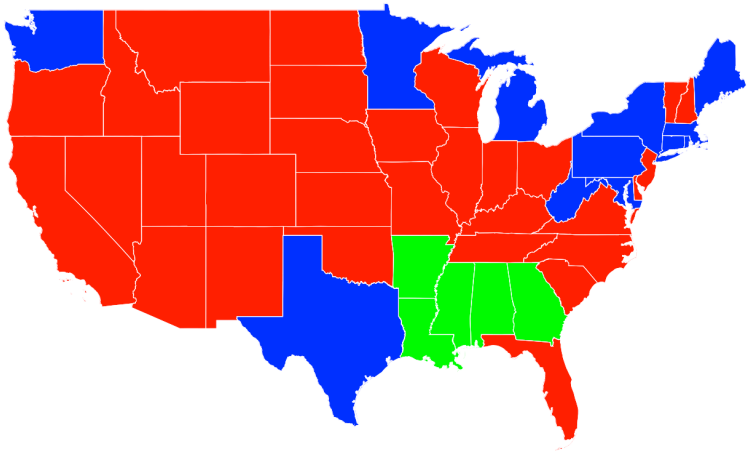
client

```
public static void main(String[] args) {
   String name = args[0];
   int year = Integer.parseInt(args[1]);
   ElectionMap election = new ElectionMap(name, year);
   election.show();
}
```

```
% java ElectionMap NJ 2008
```



```
% java ElectionMap USA 1968
```



Election Map Data Type: Java Implementation

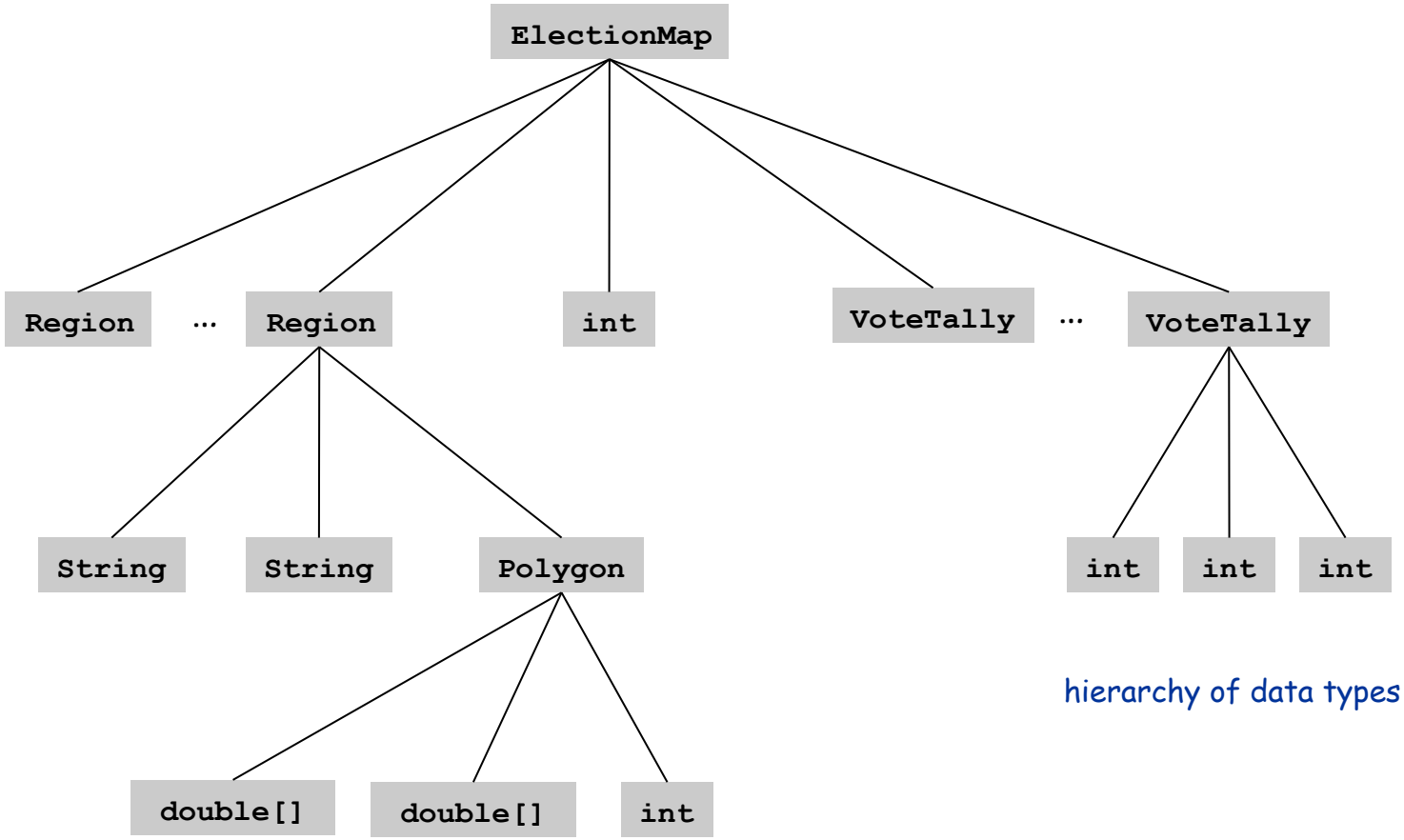
```
public class ElectionMap {
   private final int N;
   private final Region[] regions;
   private final VoteTally[] votes;

   public ElectionMap(String name, int year) {
      In in = new In(name + ".txt");

      // read in bounding box and rescale coordinates

      N       = in.readInt();
      regions = new Region[N];
      votes   = new VoteTally[N];
      for (int i = 0; i < N; i++) {
         String  name = in.readLine();
         String  usps = in.readLine();
         Polygon poly = new Polygon(in);
         regions[i]   = new Region(name, usps, poly);
         votes[i]     = new VoteTally(name, usps, year);
      }
   }

   public void show() {
      for (int i = 0; i < N; i++) {
         StdDraw.setPenColor(votes[i].getColor());
         regions[i].draw();
      }
   }
}
```

use polygon, region, and vote tally data types to build map

draw map

Modular Programming

Modular program. Collection of data types.

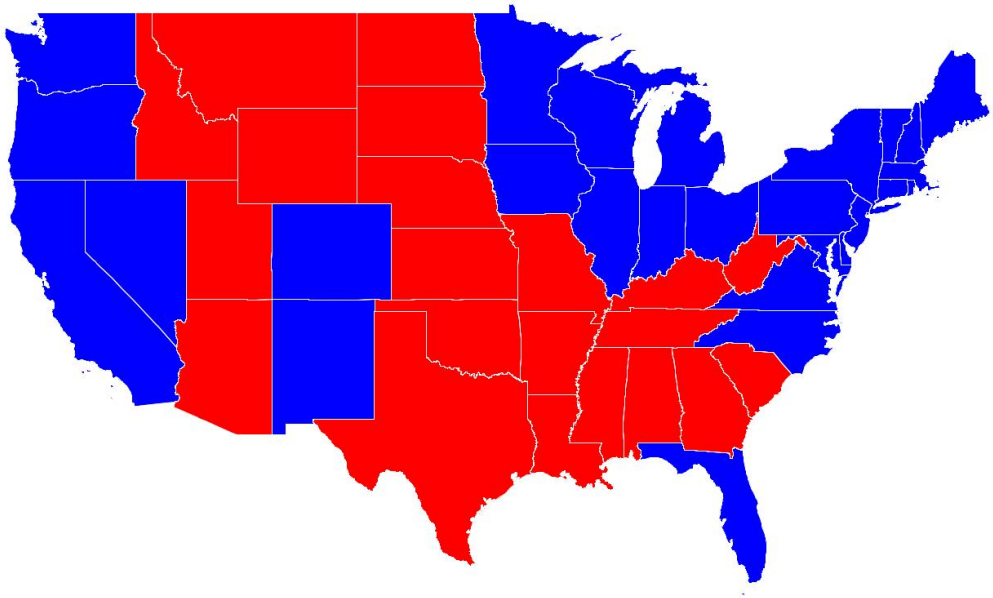


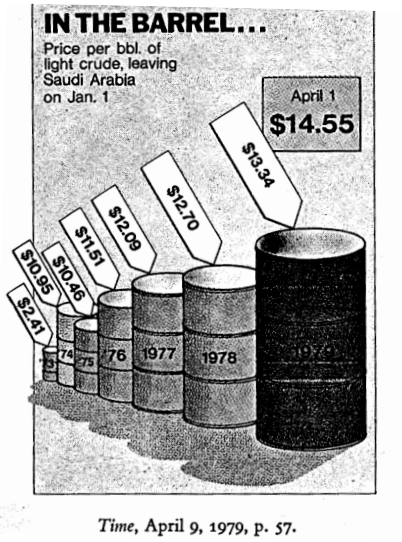
hierarchy of data types

Data Visualization

Visual Display of Quantitative Information

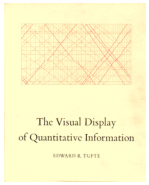
Red states, blue states. Creates a misleading and polarizing picture.

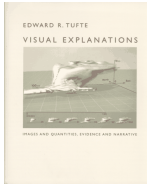


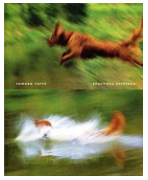


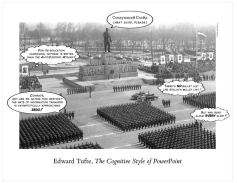
Time, April 9, 1979, p. 57.

Edward Tufte. Create charts with high data density that tell the truth.









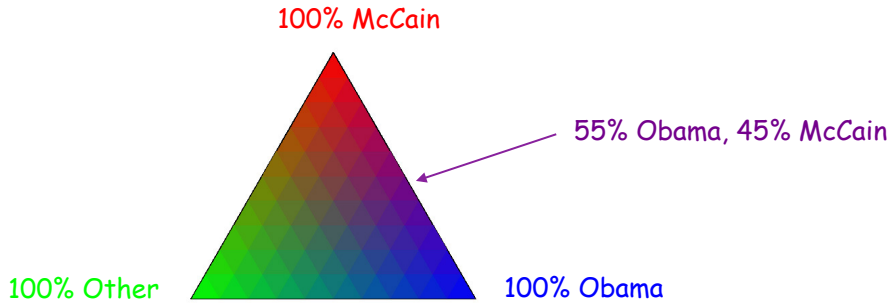
Purple America

Idea. [Robert J. Vanderbei] Assign color based on number of votes.

- a_1 = McCain votes.
- a_2 = Other votes.
- a_3 = Obama votes.

http://www.princeton.edu/~rvdb/JAVA/election2004

$$(R,\ G,\ B) = \left(\frac{a_1}{a_1 + a_2 + a_3},\ \frac{a_2}{a_1 + a_2 + a_3},\ \frac{a_3}{a_1 + a_2 + a_3} \right)$$



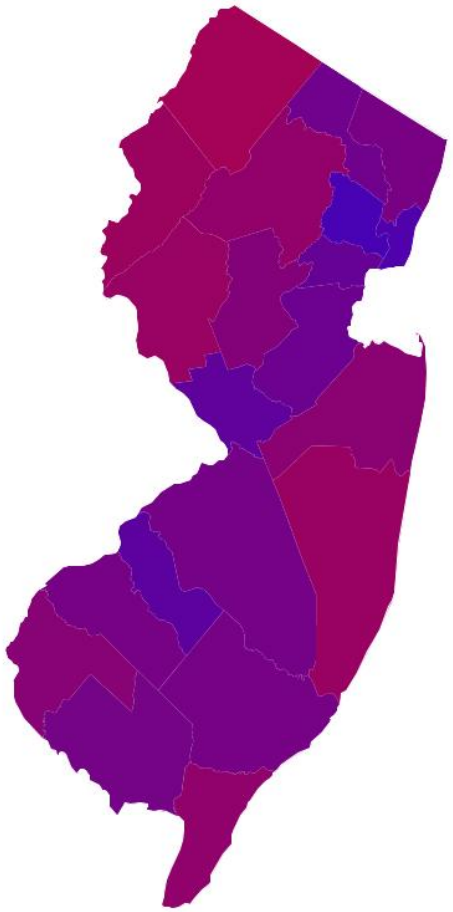
Implementation. Change one method!

VoteTally.java

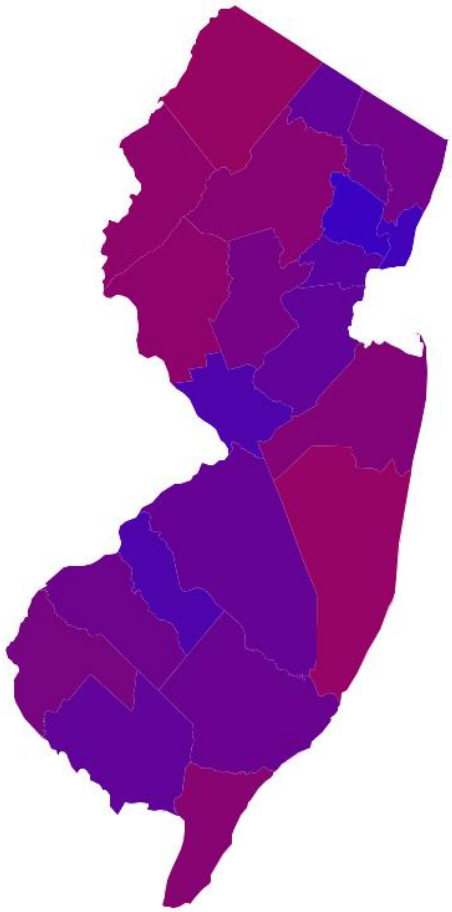
```
public Color getColor() {
   int tot = dem + rep + ind;
   return new Color((float) rep/tot, (float) ind/tot, (float) dem/tot);
}
```

Purple New Jersey

```
% java ElectionMap NJ 2004
```

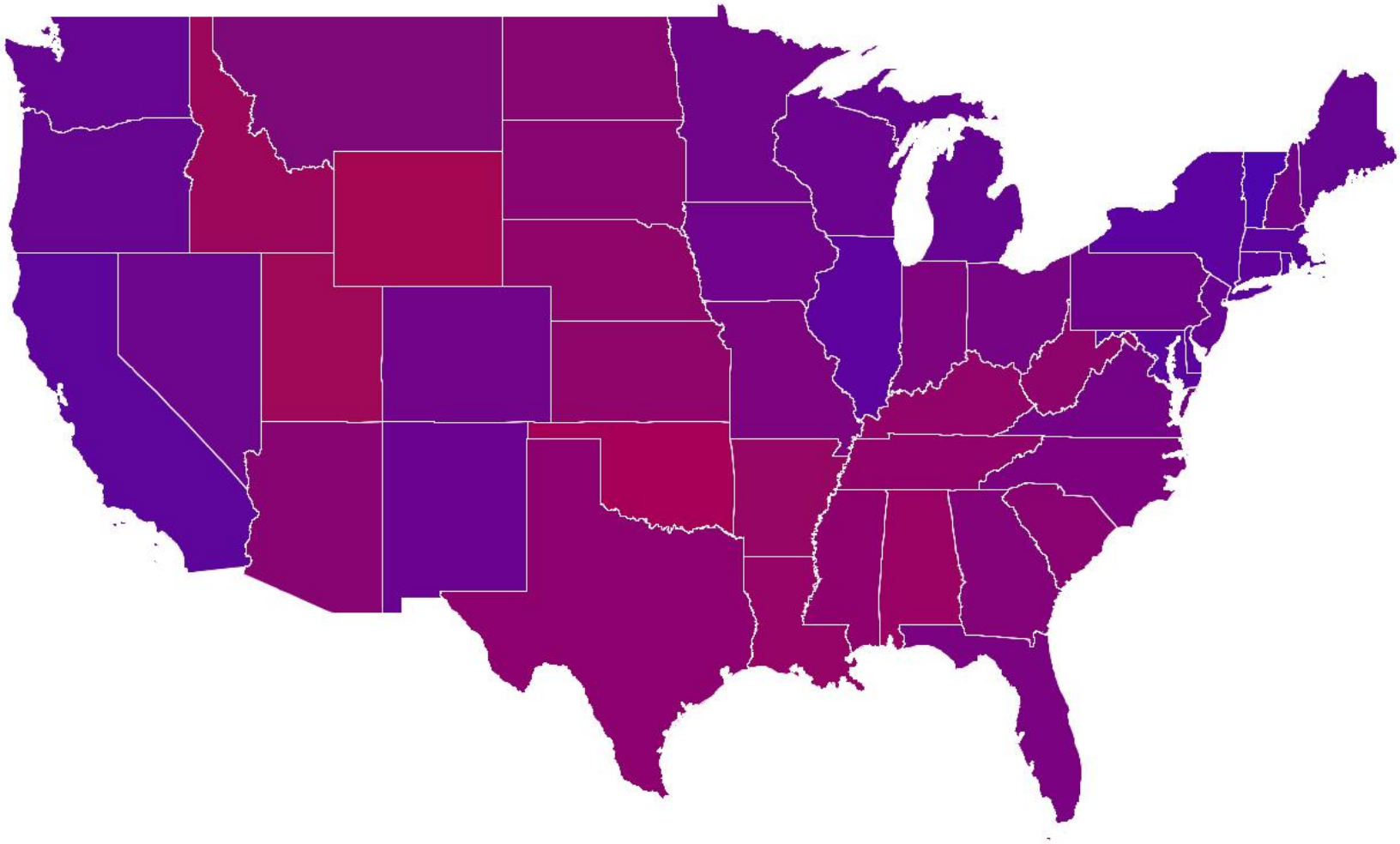


```
% java ElectionMap NJ 2008
```



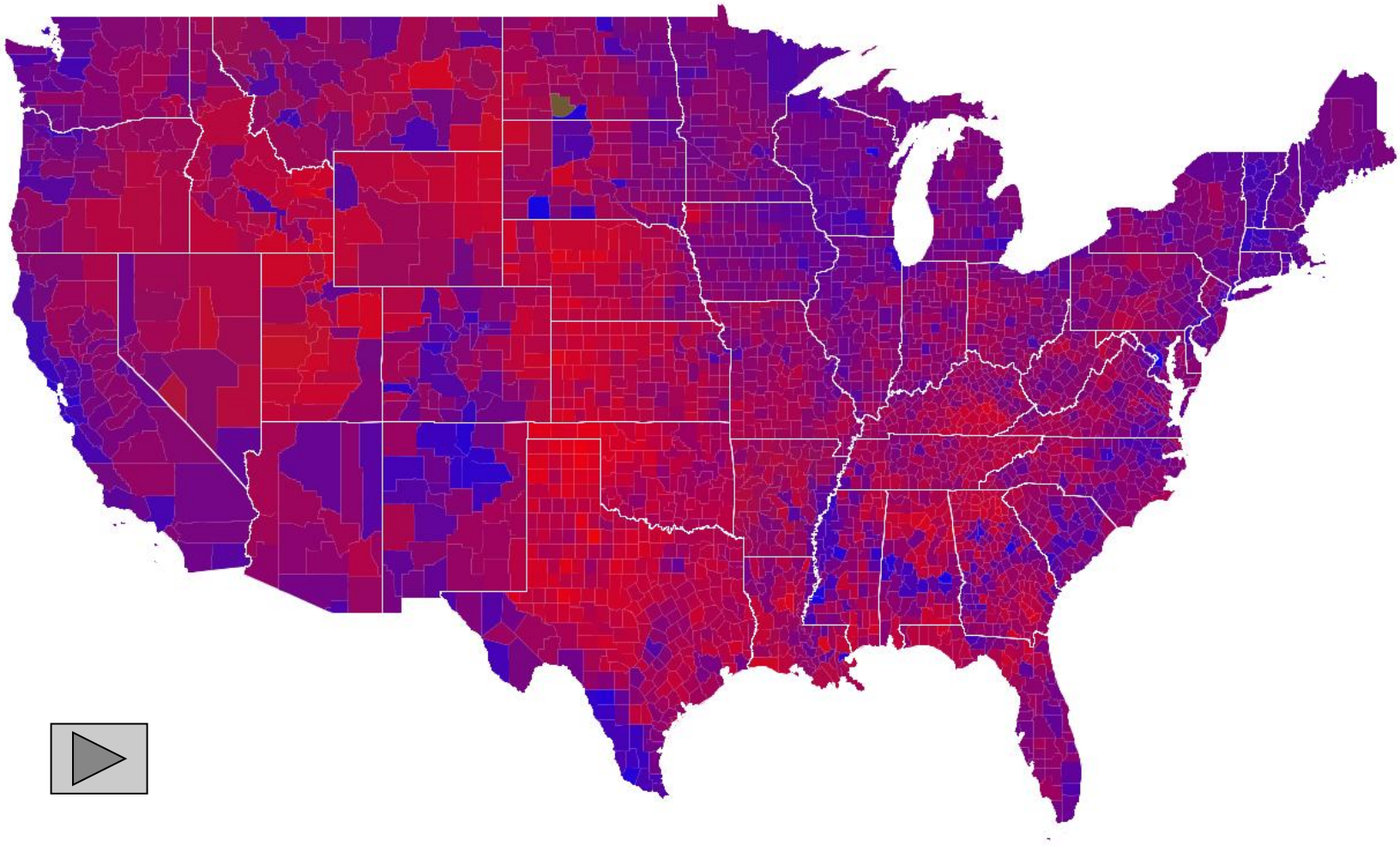
Purple America

```
% java ElectionMap USA 2008
```



Purple America

```
% java ElectionMap USA-county 2008
```

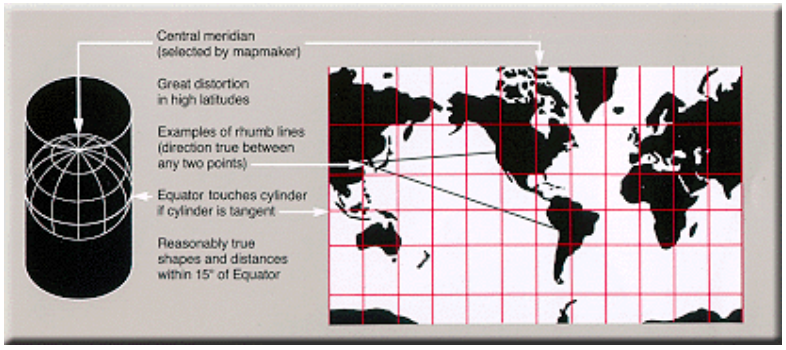


Data Visualization: Design Issues

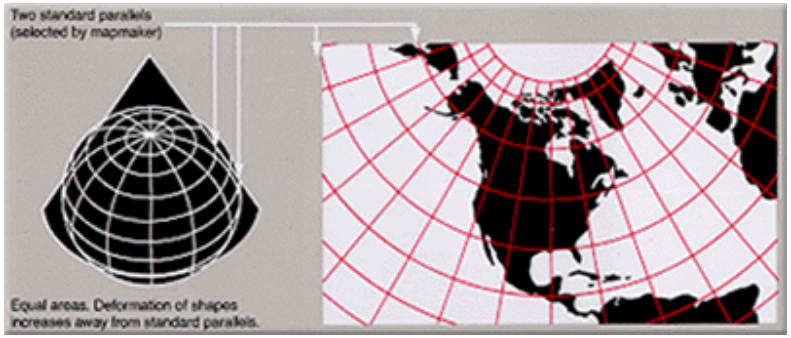
Remark. Humans perceive red more strongly than blue.

Remark. Amount of color should be proportional to number of votes, not geographic boundary.

Remark. Project latitude + longitude coordinates to 2d plane.



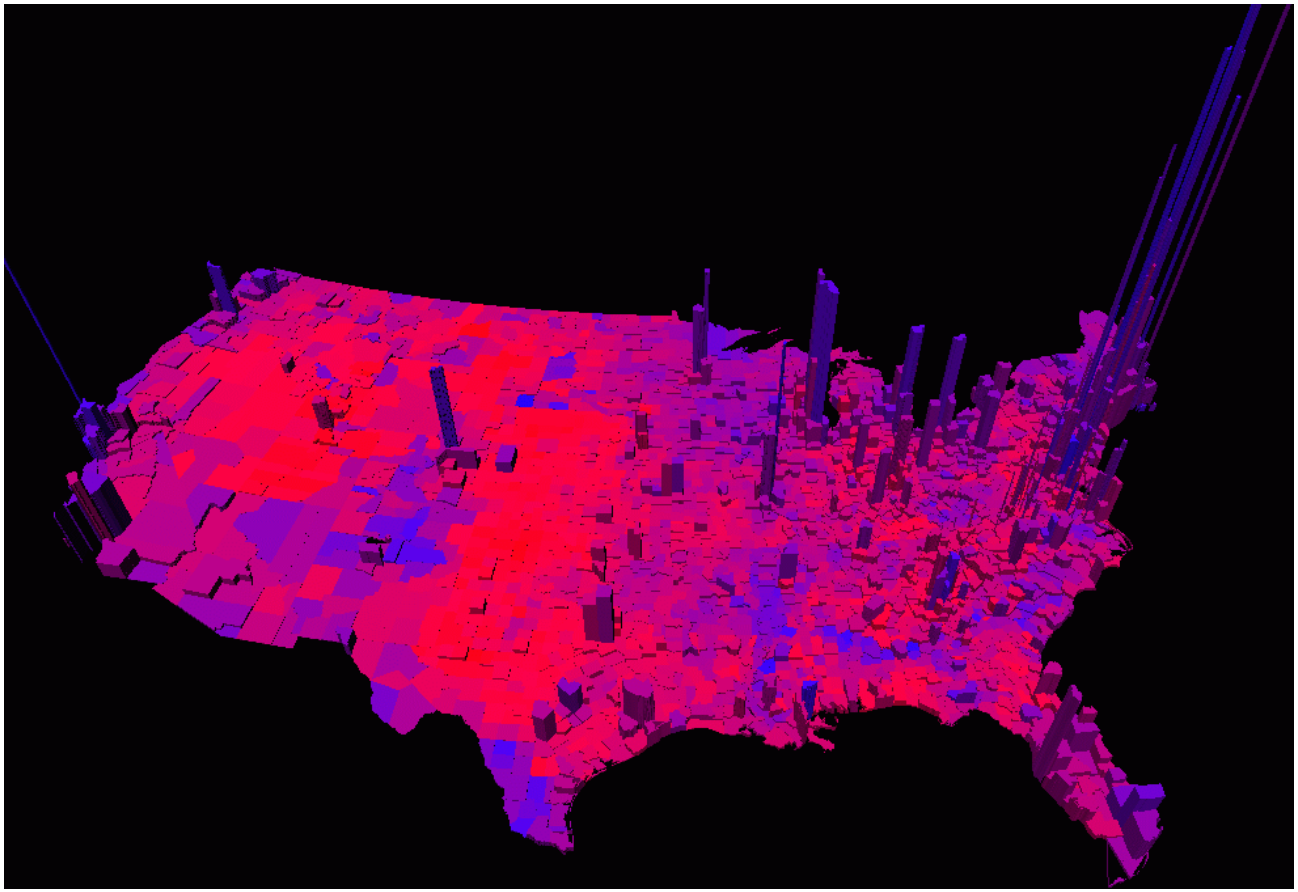
Mercator projection



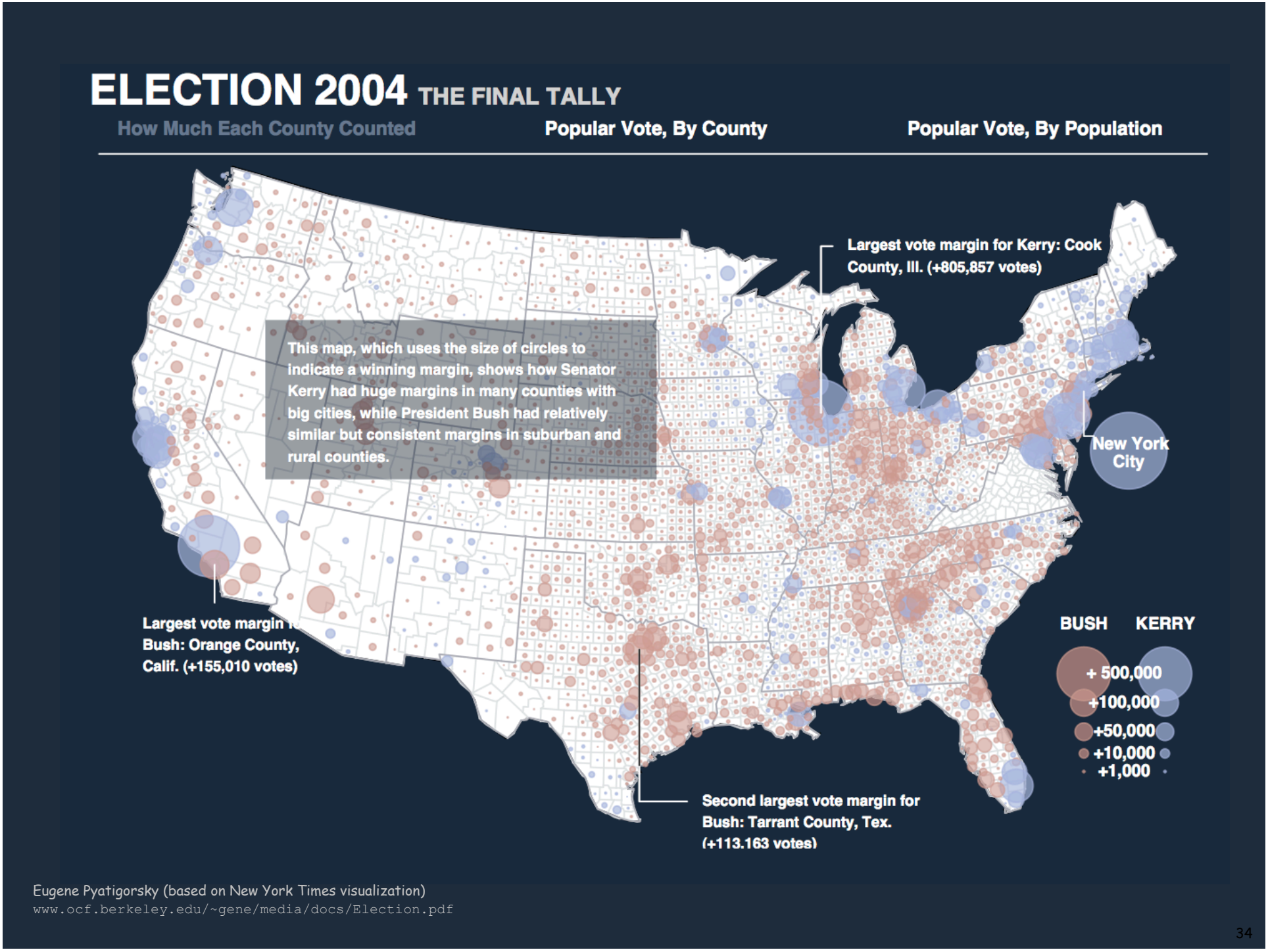
Albers projection

3D Visualization

3D visualization. Volume proportional to votes; azimuthal projection.



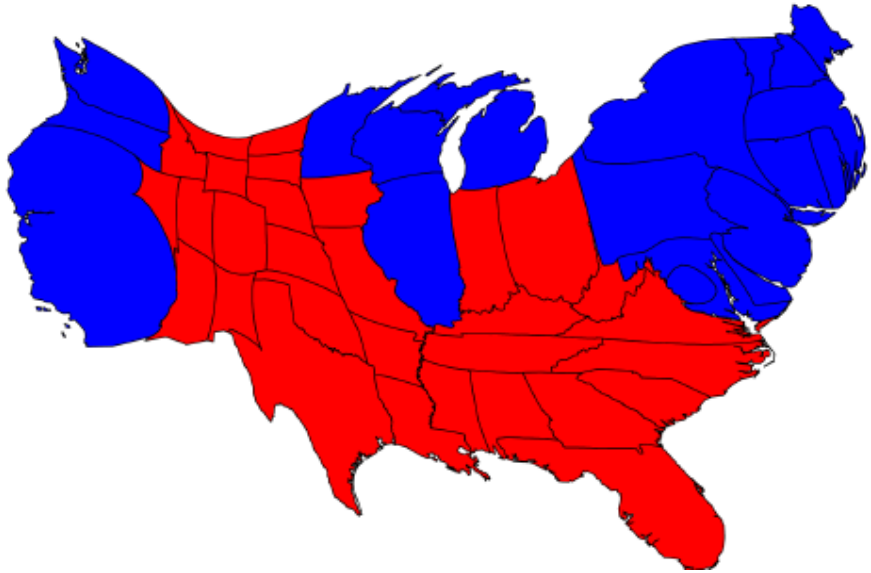
Robert J. Vanderbei
www.princeton.edu/~rvdb/JAVA/election2004



Eugene Pyatigorsky (based on New York Times visualization)
www.ocf.berkeley.edu/~gene/media/docs/Election.pdf

Cartograms

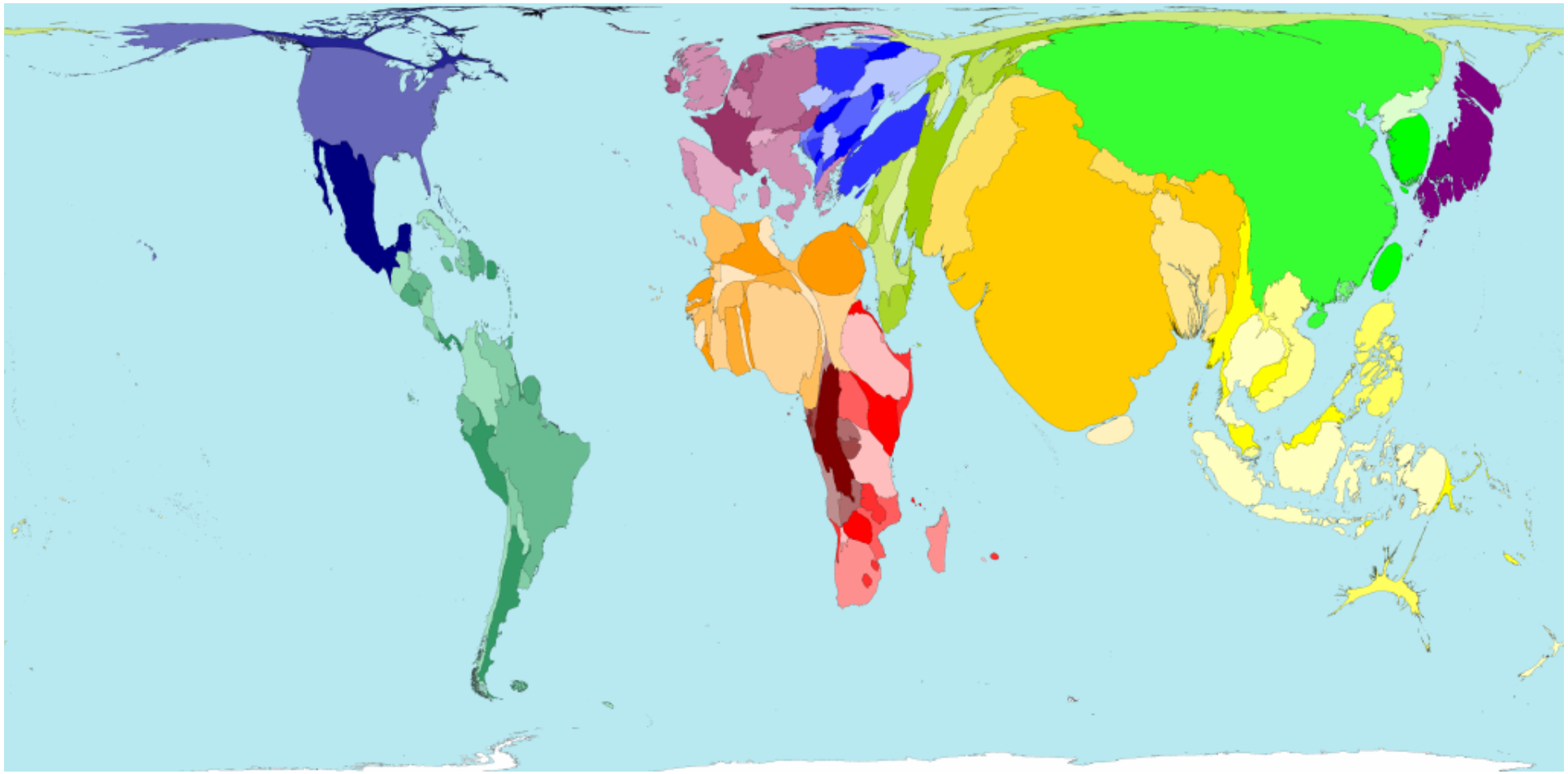
Cartogram. Area of state proportional to number of electoral votes.



Michael Gastner, Cosma Shalizi, and Mark Newman
www-personal.umich.edu/~mejn/election

Cartograms

Cartogram. Area of country proportional to population.



Summary

Modular programming.

- Break a large program into smaller independent components.
- Develop a data type for each component.
- Ex: `Polygon`, `Region`, `VoteTally`, `ElectionMap`, `In`, `Out`.

Ex 1. Build large software project.

- Software architect specifies API.
- Each programmer implements one module.
- Debug and test each piece independently. [unit testing]

Ex 2. Build reusable libraries.

- Language designer extends language with new data types.
- Programmers share extensive libraries.
- Ex: `In`, `Out`, `Draw`, `Polygon`, ...

Data visualization. You can do it! (worthwhile to learn from Tufte)