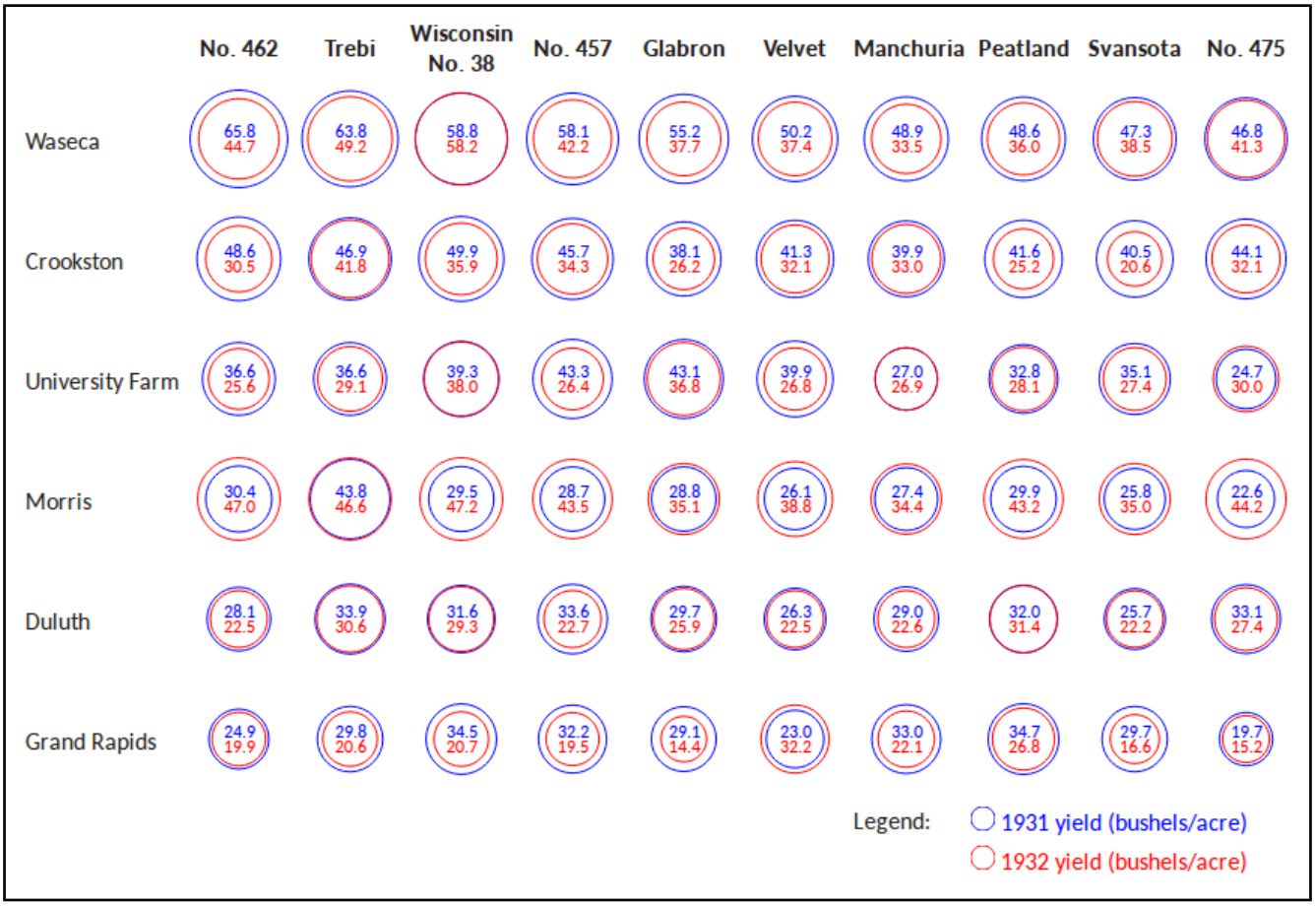
CS 448B (Fall 17) Assignment 1
Jianqing Yang (yangjq)



This visualization is meant to give an overview of the entire dataset as a starting point for drill-down inquiry, thus all 4 variables and 120 data points are represented. The nominal variables for site and variety are mapped to the Y and X positions respectively. Typically when making a scientific analysis we would want to fix some variables and observe the changes for the rest, thus the intention is to facilitate the comparison of yields for different varieties for the same site by scanning rows or vice versa by scanning columns.

Next, the yields of each site-variety pair are represented by the area of the circles drawn. The yields for both years for each site-variety are drawn as concentric circles and color coded according to the legend, which allows us to quickly observe which is greater or whether they are roughly equal. For example, one story which emerges here is that most yields drop from 1931 to 1932 except at Morris, where the red circles are larger than the blue. Circles allow us to visually compare adjacent values both vertically and horizontally, unlike bars for example, which must be aligned towards one direction. One weakness here is that comparing circle areas across sites and varieties is less precise than comparing between years, especially for minor differences in yield, but this is a trade-off for reserving the 2D axes for site and variety to give these variables clear separation

instead of using one 2D position for yield which may result in high visual overlaps for either site or variety (i.e. the variable not using position). As a mitigating measure and for avoiding ambiguity, the yield values for both years are printed within the circles using the same color coding, rounded to 1 decimal place to avoid clutter and font-sized to fit within the smallest circle in this dataset.

A second attempted mitigating measure is in the sorting. With the biggest yield value in the dataset belonging to (Waseca, No. 462, 1931), the grid is sorted in descending yields in 1931 on the Y axis for variety No. 462 in 1931 and on the X axis for site Waseca in 1931. While this sorting does not hold true for every other site and variety, it reduces the amount of jitter when scanning rows and columns and we can observe a general trend that the higher yields are on the top-left, and the lower yields on the bottom-right – it is easily apparent that Waseca produced higher yields than Grand Rapids across all varieties, for example.

Tools: the data was formatted using LibreOffice Calc and plotted in HTML using SVG elements.