# Critical Data Exploration with Simpson's Paradox

Sarah M Brown, Chenguang Xu, Christan Grant

Brown University, University of Oklahoma

## Simpson's Paradox

*a trend reversal in subsets of the data*

We define the following components to formalize the definition

- **view**: a group of variables that define a way of analyzing the data ex: the axes of a plot or a table of a summary statistic
- **trend**: a relationship between a set of variables in a trend ex: a correlation or the ordering of rows
- **grouping variable**: a categorical variable that defines partitions of the data
- **subgroup**: a partition of the data that has a single value of a given grouping variable

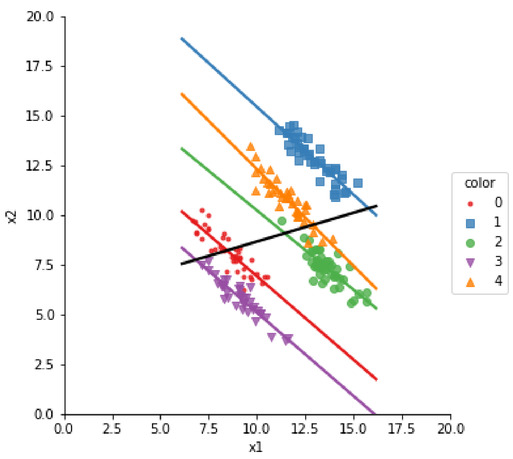
An occurrence of Simpson's Paradox is defined with respect to a given view of the data where a subgroup has the opposite trend of the whole dataset for that view

**To detect Simpson's Paradox:**

1. Iterate over views of the data
2. Compute the aggregate trend
3. iterate over the grouping variables
4. compute the subgroup trend for each value of the grouping variable
5. check for trend reversal

## Regression Type

Classic Example: drug effectiveness by dosage and gender
There's a linear correlation between two variables and grouping by another, a reversal can be in correlation coefficient




synthetic data of this form

## Rate Type

Classic example: Berkeley Grad Admissions
The trend is in the ranking of the groups divided by an 'explanatory' variable with the value of the 'outcome variable'

| | decision | | | | total |
|---|---|---|---|---|---|
| department | 0 | 1 | 2 | 3 | |
| gender | | | | | |
| F | 0.212121 | 0.136364 | 0.240506 | 0.361905 | 0.222222 |
| M | 0.024390 | 0.071429 | 0.250000 | 0.285047 | 0.242144 |

synthetic data of this form

## Generalizing Simpson's Paradox

Using the above definitions, we want to consider:

- trends where *reversal* isn't well defined
- cases where grouping variables are not all known

For more general trends we do the following:

1. consider distance between aggregate and subgroup trends
2. rank views by the distance instead of counting occurences

To include more grouping cases:

1. augment the data with clustering for a given view
2. augment the data with quantiles (distretization of a variable)

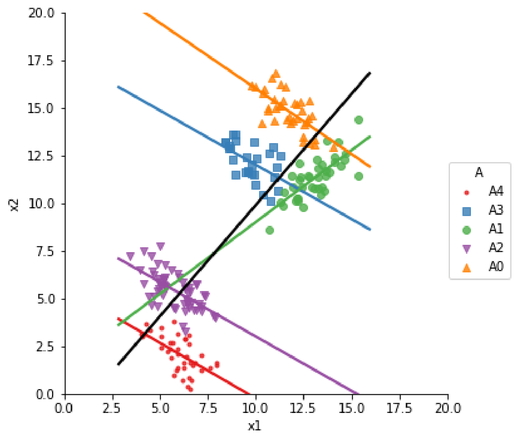
## New Detection Framework

1. Augment data with clusters
2. Augment with quantile labels
3. Compute aggregate trends for each view
4. compute subgroup trends for each categorical variable for each view
5. Compute distance between subgroups and aggregate trend for each view
6. Rank views

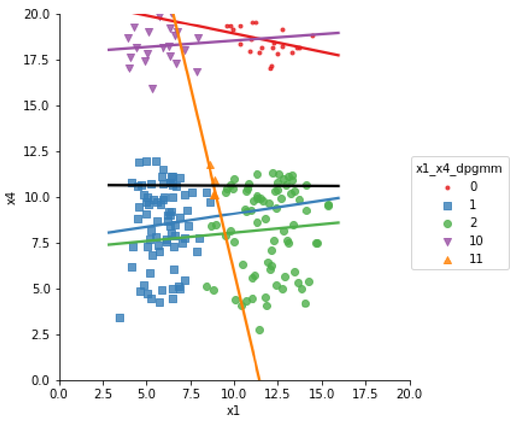
## Examples

| | feat1 | feat2 | subgroup_trend | group_feat | subgroup | trend_type | agg_trend | subgroup_slope | all_slope | angle |
|---|---|---|---|---|---|---|---|---|---|---|
| **31** | x1 | x2 | -0.241001 | x1_x4_dpgmm | 11 | pearson_corr | 0.78967 | -1.534232 | 1.165499 | 106.274373 |
| **71** | x1 | x2 | -0.426219 | x2_x7_dpgmm | 10 | pearson_corr | 0.78967 | -1.024497 | 1.165499 | 95.063612 |
| **65** | x1 | x2 | -0.506199 | x2_x6_dpgmm | 4 | pearson_corr | 0.78967 | -0.896703 | 1.165499 | 91.253016 |
| **75** | x1 | x2 | -0.511587 | x2_x8_dpgmm | 10 | pearson_corr | 0.78967 | -0.896488 | 1.165499 | 91.246209 |
| **46** | x1 | x2 | -0.511587 | x1_x8_dpgmm | 9 | pearson_corr | 0.78967 | -0.896488 | 1.165499 | 91.246209 |
| **41** | x1 | x2 | -0.556338 | x1_x7_dpgmm | 4 | pearson_corr | 0.78967 | -0.881715 | 1.165499 | 90.773477 |
| **70** | x1 | x2 | -0.541816 | x2_x7_dpgmm | 7 | pearson_corr | 0.78967 | -0.877217 | 1.165499 | 90.628143 |
| **53** | x1 | x2 | -0.412662 | x2_x4_dpgmm | 1 | pearson_corr | 0.78967 | -0.844946 | 1.165499 | 89.566377 |
| **49** | x1 | x2 | -0.396308 | x2_x3_dpgmm | 3 | pearson_corr | 0.78967 | -0.828170 | 1.165499 | 89.000889 |
| **21** | x1 | x2 | -0.405062 | x1_x2_dpgmm | 5 | pearson_corr | 0.78967 | -0.756850 | 1.165499 | 86.490601 |

Example occurence ranking




Plot of above a designed SP view




Plot of above most 'severe' view

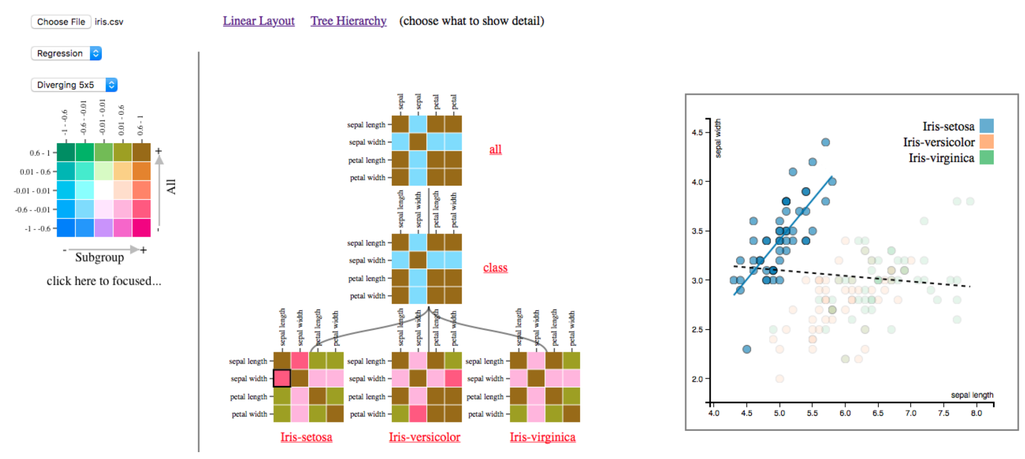
## Case Study: Racial Profiling

In a study of racial profiling and the impact of Marijuana reform on racial profiling we use our data augmentation and trend ranking tools to identify areas for qualitative analysis.

| | feat1 | feat2 | subgroup_trend | group_feat | subgroup | trend_type | agg_trend | subgroup_slope | all_slope | angle |
|---|---|---|---|---|---|---|---|---|---|---|
| **29** | search_conducted_rate | year | -0.414566 | state | CO | pearson_corr | -0.03903 | -243.731682 | -4.310706 | 12.825420 |
| **34** | search_conducted_rate | year | -0.287439 | state | NC | pearson_corr | -0.03903 | -203.568453 | -4.310706 | 12.779041 |
| **32** | search_conducted_rate | year | -0.603026 | state | MA | pearson_corr | -0.03903 | -110.008840 | -4.310706 | 12.539681 |
| **37** | search_conducted_rate | year | -0.361134 | state | TX | pearson_corr | -0.03903 | -87.062721 | -4.310706 | 12.402427 |
| **33** | search_conducted_rate | year | -0.386635 | state | MD | pearson_corr | -0.03903 | -60.410673 | -4.310706 | 12.112144 |
| **43** | search_conducted_rate | year | -0.120567 | driver_race | Asian | pearson_corr | -0.03903 | -45.782649 | -4.310706 | 11.809221 |
| **35** | search_conducted_rate | year | -0.290574 | state | RI | pearson_corr | -0.03903 | -34.565004 | -4.310706 | 11.403334 |
| **39** | search_conducted_rate | year | -0.345375 | state | WA | pearson_corr | -0.03903 | -33.844530 | -4.310706 | 11.368077 |
| **47** | search_conducted_rate | year | -0.115382 | driver_race | White | pearson_corr | -0.03903 | -33.056239 | -4.310706 | 11.327742 |
| **31** | search_conducted_rate | year | -0.199765 | state | IL | pearson_corr | -0.03903 | -22.599687 | -4.310706 | 10.526902 |

preliminary results in profiling data

## Future Work




preliminary results in profiling data

1. Integrate relaxations with visualization
2. Ranking ablation study