

Logit difference heatmap:  $\text{logit}('Smith') - \text{logit}('Jones')$   
clean=-4.124 corrupt=5.656

