In his September 12, 2016 talk at the Data Linkage: Techniques, Challenges and Applications workshop in Cambridge, Patrick Ball introduced how the Human Rights Data Analysis Group (HRDAG) carries out record linkage on databases of the casualties of violent intra-state conflict. Their current goal is to develop a database that accurately enumerates the unique individuals reported dead as part of the Syrian civil war. An all-to-all comparison across the record database is intractable, so HRDAG blocks the records into smaller sets, which are compared internally. The challenge is that they want the blocks to be as small as possible, so that they can capitalize on comptuational savings, without falsely breaking up paired records. HRDAG sorts their records into blocks by iteratively applying rules that compare features of the records. Each subsequent rule is only applied to pairs that are not covered by previous rules. Although the data itself is not feature rich, HRDAG processes the input to extract as much information as possible, in particular by converting string data into phonemes so they can quantify the distance between pairs of records.[1]

One pervasive challenge that HRDAG faces is in identifying unique reports in a noisy conflict, with few data features, and without a ground truth. The last constraint means that HRDAG can only run evaluative metrics against records that they have manually associated during a parallel processing step. After introducing HRDAG and describing their process, Ball dedicated the final third of his presentation to concerns about the accuracy of human-matched data. He contends that human matchers tend to validate more records as being unique when they are presented with records in pairs instead of in blocks of 2 or more (what Ball calls "direct clustering").[2] A bias towards marking records as "unique" when presented with pairs rather than clusters would increase the number of "unique" records, and taint the data against which HRDAG compares the outcomes of the algorithmic approaches.

Although HRDAG has worked on enumerating casualties for eight conflicts on four continents since 1999, illustrated his presentation of their workflow with reference to the Syrian civil war. While Syria is the most current conflict, Ball alluded to what may be a more subtle reasoning for focusing on the country: each conflict has a unique data-collecting environment, with unique linkage challenges. Details of these differences were outside of the scope of Ball's presentation; however, I wonder if the input datastream would be revealing about the underlying conflict processes and dynamics in each of the regions where HRDAG has worked.

[1] Do the phoneme comparisons make a metric space? There is non-negativity and an identified distance measure, but what does the triangle inequality look like in this context?

[2] It would be interesting to see the distribution of cluster sizes that they have in their data: slide 20 of Ball's presentation has most—4 out of 7— of the clusters as size 1 or 2, and only one very large cluster. How common is this in his data? Is there a Heaps' Law component, such that gathering more data increases the size of already-large clusters, rather than introducing new unique individuals or adding more data to small clusters? If the large clusters do tend to grow disproportionately when more data is added, does this provide an interesting insight into dynamics of the Syrian conflict, such as suggesting that highly-connected individuals are more likely to be killed (might increase repressive impact per display of power) or does it just reflect that highly-connected individuals are more likely to be remembered when people are interviewed by enumerators? Knowing whether particular batches of the data reflects targeted killings or indiscriminate violence might provide insight. I also wonder if studies of how crime witnesses remember events might be applicable to the question of whether people disproportionately remember and report highly-connected individuals.