

|          |   |
|----------|---|
| $\Omega$ | Universe of all valid entities (unknown size)   |
| $r$      | A valid unique entity or data item  |
| $D$      | Ground truth or the underlying population   |
| $S$      | Observed sample of size $n =  S $ , with duplicates   |
| $K$      | Integrated database with only unique entities from $S$  |
| $U$      | <i>Unknown unknowns</i> that exist in $D$ , but not in $S$ or $K$                                   |
| $M_0$    | <i>Unknown unknowns</i> distribution mass in $D$  |
| $c$      | The number of unique data items in $S$ ; $c =  K $  |
| $s_j$    | Source $j$ with $n_j =  s_j $ data items  |
| $N$      | The size of the ground truth; $N =  D $   |
| $\phi$   | The aggregated query result: e.g., $\phi_D$ (over $D$ )   |
| $\Delta$ | <i>The impact of unknown unknowns</i> : $\Delta = \phi_D - \phi_K$                                  |
| $f_j$    | A frequency statistic, i.e., the number of data items with exactly $j$ occurrences in $S$ .         |
| $F$      | The set of frequency statistics, $\{f_1, f_2, \dots, f_n\}$   |
| $\rho$   | The correlation between publicity and value distributions, i.e., <i>publicity-value correlation</i> |
| $\gamma$ | Coefficient of variance (data skew measure)   |
| $C$      | Sample coverage, also $C = 1 - M_0$   |