Testing and static analysis tools can help root out bugs in programs, but not bugs in data. Checking data for errors is arguably as important as finding program errors, but lacks effective tool support. Currently, the only approach is manual inspection of data. Because inspection is onerous and ineffective at scale, data errors are common: for example, error rates in simple manual entry tasks are typically XXX%.

This paper introduces data debugging, an approach that combines program analysis with data analysis to locate values that have an unusual effect on the results of computations. These values simultaneously provide valuable insights into the data and can reveal errors. Data debugging is particularly promising in the context of data-intensive programming environments, where programs and data are interwined, such as databases (queries and stored procedures) and spreadsheets (formulas).

We present a data debugging tool, CheckCell, that targets spreadsheets. CheckCell builds a dependency graph of an entire spreadsheet, including formulas and charts, where the leaves are cells or ranges of cells. It then computes the influence of every cell by systematically evaluating the impact of replacing it with any other item from the same range. Because data errors only matter when they have a significant impact, CheckCell highlights important values in shades of red proportional to their influence in the spreadsheet. We perform a user study to measure the effectiveness of using CheckCell to find injected errors in spreadsheets. CheckCell users were able to find errors with XX% accuracy, while users without CheckCell were only able to achieve YY% accuracy.