Making Data Transformation Interactive

Jeffrey Heer, Co-Founder and CXO of Trifacta
Human judgment is at the center of successful data analysis.

Many discussions of Big Data focus on data management and machine learning methods. While these tools provide immense value, it is important to remember that they are just that: tools. A hammer does not a carpenter make — though it certainly helps.

Consider the words of John Tukey, possibly the greatest statistician of the last half-century: “Nothing — not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers — nothing can substitute here for the flexibility of the informed human mind. Accordingly, both approaches and techniques need to be structured so as to facilitate human involvement and intervention.”

To put Big Data in the hands of knowledgeable users, we must strike a novel balance of visual interfaces, algorithmic support and scalable processing.

Though Tukey and colleagues voiced these sentiments nearly 50 years ago, they ring true today. The interested analyst is at the heart of the Big Data question: how well do our tools help users ask better questions, formulate hypotheses, spot anomalies, correct errors and create improved models and visualizations? To effectively “facilitate human involvement” across all stages of analysis is a grand challenge for our age.

At Trifacta, our goal is to radically accelerate the process of Data Transformation — and good data transformation depends on well-designed interaction. In interviews with analysts, we’ve observed that their work is highly iterative, but also greatly constrained by existing tools. Anomalies uncovered using visualization, for instance, may require redoubled efforts around data acquisition and cleaning. However, many analysts find it difficult to apply interactive tools to the volumes of data they are working with. To build more effective analysis tools, deep-seated systems choices must go hand-in-hand with user interface design. To put Big Data in the hands of knowledgeable users, we must strike a novel balance of visual interfaces, algorithmic support and scalable processing.

Data Transformation: A Central Challenge for Interactive Analysis

As data processing has become more sophisticated, there has been little progress on improving the most time consuming and tedious parts of the pipeline: Data Preparation tasks including discovery, structuring, and content cleaning. In standard practice, this kind of “data wrangling” requires writing idiosyncratic scripts in programming languages such as Python or R, or extensive manual editing using interactive tools such as Microsoft Excel. The result has two significantly negative outcomes. First, people with highly specialized skills (e.g., statistics, molecular biology, micro-economics) spend far more time in tedious data wrangling tasks than they do in exercising their specialty. Second, less technical users are often unable to wrangle their own data. The result in both cases is that significant data is left unused due to the hurdle of transforming it into shape. Sadly, when it comes to standard practice in modern data analysis, “the tedium is the message.”

Big Data—rich in potential and bottom-line impact—is still more in the “intention” than “reality” phase. Analysts must regularly transform data to make it palatable to databases, statistics packages, and visualization tools. Data sets also regularly contain missing, extreme, duplicate or erroneous values that can undermine the results of analysis.

We must give equal attention to visualization and interaction techniques that help informed users better understand and transform their data.
These anomalies come from various sources, including human data entry error, inconsistencies between integrated data sets, and sensor interference. Our own interviews with data analysts have found that these types of transforms constitute the most tedious component of their analytic process. Flawed analyses due to dirty data are estimated to cost billions of dollars each year. Discovering and correcting data quality issues can also be costly: some estimate cleaning dirty data to account for 80% of the cost of data warehousing projects.

Part of the problem is that reformatting and validating data requires transforms that can be difficult to specify and evaluate. For instance, analysts often split data into meaningful records and attributes—or validate fields such as dates and addresses — using complex regular expressions that are error-prone and tedious to interpret. Converting coded values, such as mapping numerical codes to U.S. state names, requires integrating data from one or more external tables. As data sets grow in size and variety, discovering data quality issues may be as difficult as correcting them.

Despite years of computer science research on these problems, data preparation tasks have resisted automation. Different data sets may have unique attributes requiring specialized transformations. Analysts have varied questions that may be served by different preparations of the data. Moreover, human interpretation is routinely needed to make sense of anomalies and patterns in data. Statistical methods might identify outliers, but stop short of indicating if an extreme value is an error to be corrected or an important finding. Accordingly, we must give equal attention to visualization and interaction techniques that help informed users better understand and transform their data.

Predictive Interaction™: A Model for Interactive Data Transformation

Understanding that people are critical to the Data Transformation process, Trifacta has developed a new interaction technology called Predictive Interaction. Rather than expect users to author programs, Predictive Interaction elevates data manipulation into a visual experience, using predictive machine learning algorithms to translate interactions with visualized data back down to specific, scalable data transformations. The result is a lightweight interaction for the user, grounded by concrete, verifiable transformation steps generated by software. Moreover, this experience improves over time, as more people and data in an organization work in this environment. Predictive Interaction transforms low-level programming tasks into high-level visual interaction in the following interactive loop:

**VISUALIZE**: Intelligent visualization software presents the user with overviews and details of their data, including examples of raw content, and charts that are automatically derived from the data.

**INTERACT**: The user interacts directly with data and charts in the visual interface, highlighting values, structures or trends of interest.

**PREDICT**: The user is presented with a ranked list of the transforms predicted by the algorithms, and can quickly browse visual previews of the outcome of each suggested transform to choose or adapt the best suggestion.
For example, a marketing analyst may wish to extract all email addresses in an unstructured text field. First, a sample of data records is presented to the user in a visual, interactive table display (Visualize). The user then selects a few email addresses to provide examples of the desired pattern (Interact). Algorithms then generalize from these selections to produce candidate selection rules for automatic pattern extraction (Predict). These ranked suggestions—along with visualizations of their effects on the data—are then presented back to the user for review and potential modification.

At Trifacta, we use Predictive Interaction to accelerate a broad swath of data transformation tasks, from text extraction to data type inference to filtering and aggregation. By lifting programming tasks into visual interactions, we can make data transformation faster and more accessible to a broad base of users. Using algorithms to help map user interactions to possible transformation provides both automated assistance and helps new users learn available actions. The result is massive productivity improvements for authoring transformations that can run on scalable infrastructures such as Hadoop. In existing deployments, we’ve found that Trifacta’s Predictive Interaction model helps analysts implement effective transformation workflows ten times faster than before.

People Transforming Data

Many Big Data technologies focus on either the precursors (data storage and processing) or later stages (business intelligence, dashboards) of the analysis process. Trifacta focuses on the central yet time-consuming process of transforming raw data into actionable information. By lifting low-level transformations into a visual, interactive experience, Trifacta puts the analyst at the center of the transformation process. Users view and interact with data, transformations and intermediate results. Meanwhile, supporting algorithms suggest transformations and provide scalable processing to both inform and magnify user actions.

As the diversity, size and availability of relevant data sources continues to grow, Predictive Interaction provides a framework for user centered tools that better align Big Data systems with the capabilities of analysts seeking to make sense of it all.

Trifacta, the pioneer in data transformation, significantly enhances the value of an enterprise’s Big Data by enabling users to easily transform raw, complex data into clean and structured inputs for analysis. Leveraging decades of innovative work in human-computer interaction, scalable data management and machine learning, Trifacta’s unique technology creates a bi-directional partnership between user and machine, with each component learning from the other and becoming smarter through use. Trifacta is backed by venture capital firms Greylock and Accel and is headquartered in San Francisco. Its founders and technical advisors include global leaders in data science, interaction design and big data. To learn more about Trifacta, please visit http://www.trifacta.com/resources for more information, or contact team@trifacta.com with any questions or comments.

Jeffrey Heer is Trifacta’s Chief Experience Officer and a Professor of Computer Science at the University of Washington, where he directs the Interactive Data Lab. Jeff’s passion is the design of novel user interfaces for exploring, managing and communicating data. The data visualization tools developed by his lab (D3.js, Protovis, Prefuse) are used by thousands of data enthusiasts around the world. In 2009, Jeff was named to MIT Technology Review’s list of “Top Innovators under 35”.