Crowdsourcing Data Quality
Shift Data Maintenance to the Crowd for a Fractional Cost of Traditional Methods

June 12, 2013

© 2013 CrowdFlower
All Rights Reserved.
Do not reproduce or republish this paper.
The accuracy of data about customers and prospects is the bedrock of productive marketing and selling. The data must be just right to guarantee targeting the right people with the right product or service, to address targets by their proper name, and send offers to the correct address or social media channel. Bad contact data can derail response rates for a digital campaign. It can also prevent sales reps from having good leads or be unable to directly call a prospect.

Prior options for ensuring quality data were limited to renting lists from commercial information brokers and using typical data management techniques to consolidate, de-duplicate, match and enhance records for accuracy – either with internal (and often cost-prohibitive) resources or by outsourcing to lower-cost (but still expensive) contractors. CrowdFlower provides a new and more effective way to improve data quality. CrowdFlower combines crowdsourced human intelligence with artificial intelligence to solve large and complex data problems cost effectively and at high quality. This solutions brief explains how.

**Making Data More Effective**

For a typical company database, about a third of the data are usually correct, about a third have minor errors, and a third have major errors. With a database like this, a business faces a major impediment to successful marketing and selling. It must correct data quality to raise the odds of successful campaigns. Data quality management includes four standard processes to correct or enhance different aspects of bad or missing data. These include:

- **Data Cleansing** – Cleaning, or scrubbing, data requires detective work: finding records that are inaccurate or corrupt. A name or address may be incomplete or inaccurate. Information about which products are used by a customer might be wrong. The process of data cleansing requires replacing or modifying bad data to raise its quality. Cleaner data helps to ensure the maximum ratio of successful delivery of physical mail and email – and minimize marketing costs.

- **Data De-Duplication & Matching** – This process removes records with duplicate information. De-duplication is a form of data cleansing. It protects a company’s image by not annoying prospects with multiple duplicate contacts. It also helps save money by not sending physical mail multiple times to the same recipient. When consolidating data from multiple databases, matching identifies duplicate data from those sources to ensure the resulting database eliminates duplication.

- **Data Enrichment** – When data is enriched or enhanced, it is improved in a way that adds depth and utility. Addition of social media links such as a LinkedIn profile can help sales reps to navigate an organization and personalize calls. Enrichment may entail adding information about products currently used by a customer, or that a customer is using an older version that was replaced by a product upgrade. Additional information may include contact names, company size, products or services offered, and details on the buying process. Metadata like these enables better targeting of a marketing campaign with a Customer Relationship Management solution by renewing, cross-selling or up-selling additional products and services.

- **Data Categorization and Hierarchy** – This process adds identifier information about a business, such as a North American Industry Classification (NAIC) code to denote business type, and subsidiaries or relationship to a parent company. NAIC codes are typically used to target industries in sales and marketing initiatives. Understanding subsidiaries helps sales teams find new opportunities with related entities, or avoid wasting time selling to a subsidiary when the parent company already has an enterprise purchasing agreement, for example.
Traditionally, all of these processes for data quality management entailed high costs – especially due to the requirement for hands-on attention by people needed to make human decisions about the accuracy or state of each questionable data point. Whether those people were directly employed internally or outsourced to other providers, the outcome was the same: high cost. CrowdFlower completely changes this equation by accessing crowdsourced human intelligence via a task management platform, driving the total solution cost to a mere fraction of traditional data quality management.

**Tap into the Power of the Crowd**

CrowdFlower fulfills the process of improving data quality with microtasking. CrowdFlower takes a large, complex data quality project and divides it into small, simple tasks – microtasks – that can be performed on demand by our global virtual workforce of more than 5 million workers.

**How CrowdFlower’s Microtasking Works**

Microtasking is available as a managed service where CrowdFlower handles the details of providing workflows for a project’s specifications. Customers may also license the platform to build and run workflows as an in-house crowdsourcing capability. Either way, the platform is designed to quickly produce quality results at scale.

The mechanism to implement microtasking is the proprietary CrowdFlower platform. The cloud-based platform uses a four-step process for improving data quality:

1. **Solution Identification** – Determines the type of project and optimizes the loading of data into the platform using the API.

2. **Design** – This entails designing the workflow, decomposing and designing tasks, prepping data, and determining who should perform the microtasks.

3. **Launch** – CrowdFlower trains the workers, lets them perform the work, and manages the process. This phase enforces a high degree of quality control by having workers perform training tasks and guides them with clear, forms-based inputs. A peer review system helps to maintain accuracy.
4. **Analyze** – The last step analyzes and packages results for output to our customer’s IT systems.

CrowdFlower has built-in layered security to ensure the confidentiality of an enterprise’s data. CrowdFlower’s platform and databases are housed in Amazon EC2 data centers, which are certified compliant with standards designed to secure data. To address data theft by crowd workers, CrowdFlower disaggregates data in customer records by only revealing a tiny subset to each contributor. No contributor ever sees a complete record, just the part they are working on (such as the address or phone number). CrowdFlower also limits the number of records worked on by each contributor, who processes a single field in a small subset of those records. No contributor ever knows the company name of our client’s data they are processing. CrowdFlower does not perform jobs containing personally identifiable information (PII) so that information is never at risk. Finally, a subset of our crowd workers has signed a non-disclosure agreement and can be assigned especially sensitive jobs.

**Quality Results Add Value to Business Data**

To illustrate how crowdsourcing helps improve data quality, consider results experienced by CrowdFlower customers.

**Case Study: Data Enrichment for Groupon**

Groupon is an Internet company that offers daily deals on the best stuff to do, see, eat, and buy in more than 500 markets in 48 countries. It negotiates deep discounts at local businesses in return for a predetermined number of customers. But first, Groupon must find potentially profitable partners for its daily deals. The essence of targeting these prospects is crowdsourcing data for its listing enrichment system. These data serve a dual purpose, for Groupon’s customers also depend on accuracy of business details published on its website. With CrowdFlower, Groupon’s system has achieved 98% accuracy.

Each month, CrowdFlower enriches Groupon’s list of 50,000 local business leads with star ratings and updated real-time information from Yelp, Google Places, and other crowdsourced online review sites. The diagram shows results of data quality management work performed during a typical month, starting with cleaning bad data, matching and de-duplicating multiple entries (in red), and enriching each contact with business hours and customer reviews from social media (in blue).
Case Study: Data Enrichment for Autodesk

Autodesk is one of the world’s 25 largest software companies by revenue. The company provides design, engineering and entertainment software to customers in architecture, manufacturing, building, and media and entertainment. In a move from selling individual products to end-to-end solutions, Autodesk needed a better way to identify its most promising sales leads and incentivize its sales team to pursue solution sales.

For every lead, the company’s CRM system needed complete data: company industry, size, parent and child companies, website URL, executive team bios and contact information. This data came from multiple sources with varying quality. One source was Dun & Bradstreet, but it could provide enhancement for just 70% of the database. Autodesk’s business problem was almost a third of potential sales were not being incentivized.

For every lead, the company’s CRM system needed complete data: company industry, size, parent and child companies, website URL, executive team bios and contact information. This data came from multiple sources with varying quality. One source was Dun & Bradstreet, but it could provide enhancement for just 70% of the database. Autodesk’s business problem was almost a third of potential sales were not being incentivized.
To boost data quality, Autodesk turned to CrowdFlower. Autodesk automatically funnels leads missing key data into the CrowdFlower platform via a direct API connection. Answers from multiple CrowdFlower workers start by cleansing bad data. Workers then enrich leads with company hierarchy and categorization information to provide critical support for solution selling. They also cross-check and match entries with the different data sources, de-duplicate redundant information, and categorize by NAIC code, allowing accurate reporting of sales by industry. Workers then transmit results back into Autodesk’s CRM system. CrowdFlower is currently processing about 100,000 records a month for Autodesk. As a side benefit, instead of licensing data from providers as it did in the past, Autodesk is now able to purchase data from CrowdFlower – allowing Autodesk to own its data instead of renting it each year.

Key results for Autodesk by using CrowdFlower included:

• Increased incentivized leads provided to sales by 36%.

• Percentage of leads provided to sales with complete information rose from 70% to 85%.

• Autodesk was able to file financial reports based on customer industry for the first time.

• Using the crowd to perform company research on demand has cost 75% less than paying an outsourcing company.

Learn How the Crown Can Add Quality to your Data

We invite you to learn more about how your enterprise can use CrowdFlower and the power of the crowd to inject quality into your business data. To learn more, please visit our website at crowdflower.com/solutions or call us at (415) 471-1920.

About CrowdFlower

CrowdFlower combines human intelligence with the scalability and efficiency of computer algorithms to offer quality-ensured processing of business information. CrowdFlower’s platform provides solutions to a wide variety of data needs for enterprises such as product catalog enhancement, content generation, image moderation, and business listing enrichment. With 5 million Crowd Contributors completing millions of judgments each month for over 500 customers, CrowdFlower is the leader in enterprise crowdsourcing. The company has successfully worked with well-respected enterprises including Apple, AT&T, Autodesk, eBay, Ford, LinkedIn, Microsoft, Sears, Toshiba and Twitter.

For more information, visit www.crowdflower.com.