CIS 602-01: Scalable Data Analysis

Data Fusion

Dr. David Koop
Technology revolutions come in measured, sometimes foot-dragging steps. The lab science and marketing enthusiasm tend to underestimate the bottlenecks to progress that must be overcome with hard work and practical engineering.

The field known as “big data” offers a contemporary case study. The catchphrase

Dirty data costs the U.S. economy $3 trillion+ per year

Estimates show dirty data is a big problem for the U.S. economy. How big? About 2x the national deficit.

Software expert Hollis Tibbets, the Global Director of Marketing at Dell, estimates that duplicate data and bad data combined cost the U.S. economy over $3 trillion every year - which is just about two times the national deficit.

In his post "$3 Trillion Problem: Three Best Practices for Today’s Dirty Data Pandemic," Hollis points to a few key facts and figures to back up his estimate.

[D. Haas et al., 2016]
In practice, **how is data cleaned** before analysis?

What are the **limitations** of existing processes?

How can **database researchers** contribute?

[D. Haas et al., 2016]
Our Survey: Tools

We're not in Kansas any more!

[D. Haas et al., 2016]

Our Misconceptions

Data cleaning is a sequential operation

[It's an] iterative process, where I assess biggest problem, devise a fix, re-evaluate. It is dirty work.
Our Misconceptions

The end-goal of data cleaning is clean data

"We typically clean our data until the desired analytics works without error."

[D. Haas et al., 2016]
Our Misconceptions

Data cleaning is a sequential operation

“[It’s an] iterative process, where I assess biggest problem, devise a fix, re-evaluate. It is dirty work.”

[D. Haas et al., 2016]
Our Misconceptions

Data cleaning is performed by one person

“...There are often long back and forths with senior data scientists, devs, and the business units that provided the data on data quality.”

[D. Haas et al., 2016]
Our Misconceptions

Data quality is easy to evaluate

“…”

I wish there were a more rigorous way to do this but we look at the models and guess if the data are correct.

[D. Haas et al., 2016]
SampleClean (and Variants)

- Dirty Data?
  - Missing Values
  - Duplicate Values
  - Incorrect Values
  - Inconsistent Values

- Estimate query results using a sample of the data

- Two ideas:
  - Direct Estimate
  - Correction
Dirty and Cleaned Data

(a) Dirty Data

<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>pub_year</th>
<th>citation_count</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>CrowdDB</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>t2</td>
<td>TinyDB</td>
<td>2005</td>
<td>1569</td>
</tr>
<tr>
<td>t3</td>
<td>YFilter</td>
<td>Feb, 2002</td>
<td>298</td>
</tr>
<tr>
<td>t4</td>
<td>Aqua</td>
<td></td>
<td>106</td>
</tr>
<tr>
<td>t5</td>
<td>DataSpace</td>
<td>2008</td>
<td>107</td>
</tr>
<tr>
<td>t6</td>
<td>CrowdER</td>
<td>2012</td>
<td>1</td>
</tr>
<tr>
<td>t7</td>
<td>Online Aggr.</td>
<td>1997</td>
<td>687</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>t10000</td>
<td>YFilter - ICDE</td>
<td>2002</td>
<td>298</td>
</tr>
</tbody>
</table>

(b) Cleaned Sample

<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>pub_year</th>
<th>citation_count</th>
<th>#dup</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>CrowdDB</td>
<td>2011</td>
<td>144</td>
<td>2</td>
</tr>
<tr>
<td>t2</td>
<td>TinyDB</td>
<td>2005</td>
<td>1569</td>
<td>1</td>
</tr>
<tr>
<td>t3</td>
<td>YFilter</td>
<td>2002</td>
<td>298</td>
<td>2</td>
</tr>
<tr>
<td>t4</td>
<td>Aqua</td>
<td>1999</td>
<td>106</td>
<td>1</td>
</tr>
<tr>
<td>t5</td>
<td>DataSpace</td>
<td>2008</td>
<td>107</td>
<td>1</td>
</tr>
<tr>
<td>t6</td>
<td>CrowdER</td>
<td>2012</td>
<td>34</td>
<td>1</td>
</tr>
<tr>
<td>t7</td>
<td>Online Aggr.</td>
<td>1997</td>
<td>687</td>
<td>3</td>
</tr>
</tbody>
</table>

[J. Wang et al., 2014]
SampleClean Framework

1. Dirty Data
   - Sample Creation
   - Data Cleaning
   - Aggregate Queries
   - Cleaned Sample

2. Result Estimation (RawSC)
   - Results with Confidence Intervals

3. Result Estimation (NormalizedSC)
   - Results with Confidence Intervals

[J. Wang et al., 2014]
Comparing the Two Approaches

Figure 1: Comparison of the convergence of the methods on two TPC-H datasets of 6M tuples with simulated errors 50% error and 5% error. On the dataset with larger errors, the direct estimate gives a narrower confidence interval, and on the other, the correction is more accurate.

Q\(\left( R_{\text{clean}} \right) \) = Q\(\left( R \right) \). The interesting problem is when there are systematic errors, i.e., |\(\epsilon_c\)| > 0. In other words, the corruption that is correlated with the data, e.g., where every record is corrupted with a +1.

2.2.2 Key Idea I: Direct Estimate vs. Correction

The key quantity of interest is \(\epsilon_c\), and to be able to bound a query result on dirty data, requires that \(\epsilon_c\) is 0 or bound. Direct Estimate: This technique is a direct extension of AQP to handle data cleaning. A set of \(k\) rows is sampled uniformly at random from the dirty relation \(R\) resulting in a sample \(S\). Data cleaning is applied to the sample \(S\) resulting in \(S_{\text{clean}}\). Data cleaning and sampling may change the statistical and scaling properties of the query \(Q\), so \(\hat{Q}\) may have to be re-written to a query \(\hat{Q}\). \(\hat{Q}\) is applied to the sample \(S_{\text{clean}}\) and the result is returned. There are a couple of important points to note about this technique. First, as in AQP, the direct estimate only processes a sample of data. Next, since it processes a cleaned sample of data, at no point is there a dependence on the dirty data. As we will show later in the article, the direct estimate returns a result whose accuracy is independent of the magnitude or rate of data error. One way to think about this technique is that it ensures \(\epsilon_c = 0\) within the sample.

Correction: The direct estimate suffers a subtle drawback. Suppose, there are relatively few errors in the data. The errors introduced by sampling may dominate any error reductions due to data cleaning. As an alternative, we can try to estimate \(\epsilon_c\). A set of \(k\) rows is sampled uniformly at random from the dirty relation \(R\) resulting in a sample \(S\). Data cleaning is applied to the sample \(S\) resulting in \(S_{\text{clean}}\). The difference in applying \(\hat{Q}\) to \(S\) and \(\hat{Q}\) to \(S_{\text{clean}}\) gives an estimate of \(\epsilon_c\). The interpretation of this estimate is a correction to the query result on the full dirty data. In contrast to the direct estimate, this technique requires processing the entire dirty data (but only cleaning a sample). However, as we will later show, if errors are rare this technique gives significantly improved accuracy over the direct estimates.

2.2.3 Key Idea II: Sampling to Improve Accuracy

Figure 1 plots error as a function of the cleaned sample size on a corrupted TPCH dataset for a direct estimate, correction, and AllDirty (query on the full dirty data). In both cases, there is a break-even point (in terms of number of cleaned samples) when the data cleaning has mitigated more data error than the approximation error introduced by sampling. After this point, SampleClean improves query accuracy in comparison to AllDirty. When errors are relatively rare (5% corruption rate), the correction is more accurate. When errors are more significant (50% corruption rate), the direct estimate is more accurate. Note that the direct estimate returns results of the same accuracy regardless of the corruption rate.

[S. Krishnan et al., 2015]
Assignment 3

- Analysis of the IRS Form 990 Filings Data
- Public Dataset on AWS
- Sign up for AWS Educate

Part 1: Run Spark Locally over Index
  - Process the index file (CSV), organize by return_type
  - Plot dates submitted

Part 2: Run Spark Locally over Subset
  - Find GrossReceiptsAmt max/min for IRS 990EZ filing (XML)
  - Run statistics based on specific indicators (political, school, etc.)

Part 3: Run Spark on AWS over Full Dataset (Year)
  - (Same queries as above)
  - Check your code locally first
Introduction to Data Integration

A. Doan, A. Halevy, and Z. Ives
Data Integration

• Lots of data sources, how do we answer questions where we need to access data from more than one?

• Schema matching

• Problem of heterogeneity

• AI-Complete problem: difficulty is the same as making computers as intelligent as people

• Two techniques:
  - Mediation
  - Data Warehouses
Data Integration and Data Fusion

• Data Integration: focus on integrating data from different sources
• When sources are orthogonal, no problems
• What happens when two sources provide the same type of information and they conflict?
• Data Fusion: create a single object while resolving conflicting values
Data Fusion—
Resolving Data Conflicts in Integration

X. L. Dong and F. Naumann
Data Fusion Summary

- Conflict resolution strategies
- "Truth-discovery" techniques
  - Accuracy
  - Freshness
  - Dependence
- Fusion Issues
  - Accuracy
  - Efficiency
  - Usability
  - How fusion fits with the rest of data integration?
Less is More: Selecting Sources Wisely for Integration

X. L. Dong, B. Saha, and D. Srivastava
Coverage as Sources Increase

![Graph showing coverage as sources increase](image)

- **Legend**:
  - #(Returned books)
  - #Sources

---

**Figure 2: Returned correct results.**

- **Graph**:
  - X-axis: #Sources
  - Y-axis: #(Returned books)

---

**Figure 3: Different integration models.**

- **Legend**:
  - CCU
  - OTE

---

**Figure 4: Marginal cost and gain.**

- **Legend**:
  - Marginal cost
  - Marginal gain

---

**Figure 5: Error rate and accuracy.**

- **Legend**:
  - Error rate
  - Accuracy

---

**Figure 6: Efficiency of different methods.**

- **Legend**:
  - Efficiency

---

**Figure 7: Resource allocation.**

- **Legend**:
  - Resource allocation

---

**Figure 8: Scalability.**

- **Legend**:
  - Scalability

---

**Figure 9: Integration techniques.**

- **Legend**:
  - Integration techniques

---

**Figure 10: Performance comparison.**

- **Legend**:
  - Performance comparison

---

**Figure 11: Cost analysis.**

- **Legend**:
  - Cost analysis

---

**Figure 12: Revenue generation.**

- **Legend**:
  - Revenue generation

---

**Figure 13: Business models.**

- **Legend**:
  - Business models
source selection by Marginalism from data markets [1].

To balance the gain and the cost, source selection can be important when we have a larger volume of data, but also a larger number of sources and more heterogeneity, so we wish to spend the computational resources wisely. However, the Law of Diminishing Returns does not necessarily hold and sometimes “less is more”. As the research community for data integration has been focusing on improving the quality of the sources in Section 1.3. Fig. 2 plots the gain, defined as the number of correct authors after integrating the sources, as the number of sources increases.

In this example, after integrating 71 sources by applying the Marginalism principle to source selection, the result of Accu is better than the naive method, VOTE. However, the Law of Diminishing Returns does not necessarily hold in data integration, so there can be multiple marginal points. In our example (Fig. 2), after we integrate 71 sources, the gain curve flattens out and the marginal gain is much smaller than the marginal cost. The marginal gain is the difference between the gain after and before integrating the new source, while the marginal cost is the cost of integrating one source. The Law of Diminishing Returns states that the marginal cost exceeds the marginal gain for the next unit of resources.

To apply the Marginalism principle to source selection, we can use different strategies. One approach is to treat the gain and cost curves as a curve, and choose the best point on the curve. The best point is the point where the marginal gain is equal to the marginal cost. Another approach is to treat the gain and cost curves as an order, and choose the best order. The best order is the order that leads to the highest gain. Each order can lead to different gain curves. Each curve has its own marginal points, so we need to be able to compare different orders.

We propose a solution inspired by the economic theory [11]. Assuming we can measure gain and cost using economic terms, we can apply the principle of Marginalism. The principle of Marginalism states that the marginal benefit exceeds the marginal cost. The marginal benefit is the benefit of integrating one more source, while the marginal cost is the cost of integrating one more source. When the marginal benefit exceeds the marginal cost, the integration of the new source is beneficial. However, when the marginal benefit is zero, the integration of the new source is not beneficial. The principle of Marginalism can be applied to different integration techniques, which is important without a doubt. However, the principle of Marginalism does not necessarily hold and sometimes “less is more”. As the research community for data integration has been focusing on improving the quality of the sources, the principle of Marginalism is especially relevant in the big data environment: not only do we have larger volume of data, but also we have larger number of sources and more heterogeneity, so we wish to spend the computational resources wisely.

In this example, after integrating 71 sources, the gain curve flattens out and the marginal gain is much smaller than the marginal cost. The marginal gain is the difference between the gain after and before integrating the new source, while the marginal cost is the cost of integrating one source. The marginal gain is much smaller than the marginal cost. When the marginal gain is small, the integration of the new source is not beneficial. The marginal cost is the cost of integrating one source. When the marginal cost is large, the integration of the new source is not beneficial. The marginal cost is larger than the marginal benefit.
Evaluating Fusion Models

![Graph showing accuracy vs. number of sources](image-url)

**Figure 4:** Model monotonicity.

The transformation between different cases is shown in Fig. 7. For example, if the most popular false value, denoted by $\bar{S}$, has much higher probability to provide the true value than any particular false value. Assume there is a single false value.

We can estimate the accuracy of $V$ according to Eq.(9-12). Empirically the difference between the estimated accuracy and the true one is typically small, as we show in the next example.

**Example 4.3.** Assume $V$ contains 41 sources with accuracy $0.7$ and $\bar{S}$ contains 5 sources with accuracy $0.9$. Assume there is a single false value.

We can estimate the accuracy for $V$ by dynamic programming. Then, $A_1 = 0.7$, $A_i = \max(0, A_{i-1} + 0.2)$ for $2 \leq i \leq 40$. Therefore, $A_{41} \approx 0.92$.

The hardness of accuracy estimation remains an open problem. The evaluation of fusion models is #P-hardness is a complexity class for hard counting problems, believed to be harder than #NP.

**Theorem 4.1.** Assume the values provided by each source are uniformly distributed and each source has only a slightly decreases by 1; otherwise, $p_1 = \frac{1}{2}$.

**Theorem 4.2.** Assume $m, d$.

Table 1 shows computa-

\[
\begin{array}{|c|c|c|}
\hline
\text{Value} & \text{Probability} & \text{Accuracy} \\
\hline
1 & 0.7 & 0.9 \\
2 & 0.8 & 0.95 \\
3 & 0.9 & 1.0 \\
\hline
\end{array}
\]

**Theorem 4.3.** In case the most popular false value, denoted by $\bar{S}$, is provided. According to our analysis, $V$ contains 41 sources with accuracy 0.7 and $\bar{S}$ contains 5 sources with accuracy 0.9. Assume there is a single false value.

We can estimate the accuracy by dynamic programming. Then, $A_1 = 0.7$, $A_i = \max(0, A_{i-1} + 0.2)$ for $2 \leq i \leq 40$. Therefore, $A_{41} \approx 0.92$.
Source selection by Marginalism from data markets [1].

Here, the marginal gain is the difference in gain resulting from integrating an additional source. The marginal cost is the cost of integrating that source. We envision an optimization problem: finding the subset of sources that maximizes the result. There are two standard ways to formalize the problem: finding the subset of sources that maximizes the result.

1.2 Source selection by Marginalism

Source selection falls outside the scope of traditional integration tasks, such as mapping schemas, linking records that refer to the same real-world entity, and resolving conflicts. On the one hand, it is not as good as that of VOTE and similar for marginal cost. In our example, if the gain of finding one author list is 1 while the cost of integrating one source is .1, then we may select the first 520 sources, obtaining 65 correct lists.

However, if we instead select 526 sources, we introduce 1% more cost but can obtain 81 correct lists (improving by 25%); arguably, this is not as good as that of VOTE and similar for marginal cost.

In [1], the gain curve flattens out and the marginal gain is much larger. Here, the marginal gain is the difference in gain resulting from integrating an additional source. The marginal cost is the cost of integrating that source. We envision an optimization problem.
Different Fusion Models

![Graph showing recall vs. number of book-sources]

- **SrcRecall**
- **Vote**
- **Accu**
- **PopAccu**

[Source: X. L. Dong et al., 2012]
Marginalism selects 18 sources (47%) for OP. Fig.9 compares different source selection methods in solving the M\text{\textasciitilde}C by 7.5% and over A\text{\textasciitilde}A. C by 55% and beats M\text{\textasciitilde}C by 10%. We implemented in Java and experimented on a Linux server.

Table 2: Estimated recall vs. real fusion recall averaged on each data set.

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg est.</th>
<th>Avg real</th>
<th>Abs diff</th>
<th>Rel diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>A\text{\textasciitilde}A</td>
<td>.924</td>
<td>.933</td>
<td>.043</td>
<td>4.7%</td>
</tr>
<tr>
<td>CCU</td>
<td>.975</td>
<td>.971</td>
<td>.064</td>
<td>2.9%</td>
</tr>
<tr>
<td>OP</td>
<td>.964</td>
<td>.960</td>
<td>.040</td>
<td>4.2%</td>
</tr>
</tbody>
</table>

Also, M\text{\textasciitilde}C by 7.2% and over A\text{\textasciitilde}A is monotonic (under the independence assumption), M\text{\textasciitilde}C by 72%. Among the 894 sources in the gold standard, each method we reported the percentage of times that the best selection is returned and for returned sub-optimal selections we reported the recall. For fusion results, we compared the returned results from Section 1 that with top-46 with top-46. We observed that (1) G\text{\textasciitilde}C by 13.8% and over A\text{\textasciitilde}A, which in turn took 1 order of magnitude longer time.

Fig.10 compares different source selections and sub-optimal selections when we applied V\text{\textasciitilde}C by 46% and over A\text{\textasciitilde}A. Recall of P\text{\textasciitilde}C by 2.8% and over A\text{\textasciitilde}A. Also, M\text{\textasciitilde}C by 2.9% and over A\text{\textasciitilde}A. Table 2 compares the estimated recall with the real one. The difference is quite small and is the smallest for P\text{\textasciitilde}C by 19.7% and over A\text{\textasciitilde}A. We ordered the sources such that the selected sources are ordered in decreasing order of their recall. Fig.8 shows the recall of P\text{\textasciitilde}C by 7.5% and over A\text{\textasciitilde}A.

For fusion results, we compared the returned results with the gold standard and reported the recall. For quality estimation, we reported the absolute and relative difference between the estimated recall and the fusion recall. For source selection we select 165 (72.4%) sources, taking considerably longer time (3 orders of magnitude). There are very few decreases for P\text{\textasciitilde}C by 2.26 GHz Intel Xeon Processor X7560 and 24M Cache.
Efficiency for Accuracy Estimation

![Graph showing efficiency for accuracy estimation]

- Vote
- Accu
- PopAccu

Estimation time (ms) vs. #Flight-Sources

1 4 7 10 13 16 19 22 25 28 31 34 37

[10^0] 10 100 1000 10000 100000

[Source: X. L. Dong et al., 2012]
Scalability of Source Selection

- PopAccu_ random
- Accu_ random
- Vote_ random
- Vote_dec
- Vote_inc

![Graph showing scalability of source selection](image)

- Execution time (sec)
- #Book-Sources

Table 4: Profit difference for various quality measures.

<table>
<thead>
<tr>
<th>Sources</th>
<th>POP</th>
<th>ACCU</th>
<th>VOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book</td>
<td>0.001</td>
<td>0.1</td>
<td>1%</td>
</tr>
<tr>
<td>Sources</td>
<td>0.1</td>
<td>10</td>
<td>1000</td>
</tr>
<tr>
<td>OP</td>
<td>0.3%</td>
<td>50</td>
<td>110</td>
</tr>
<tr>
<td>UAD</td>
<td>0.5%</td>
<td>70</td>
<td>130</td>
</tr>
<tr>
<td>TEP</td>
<td>0.9%</td>
<td>90</td>
<td>150</td>
</tr>
<tr>
<td>OTE</td>
<td>3.9%</td>
<td>110</td>
<td>170</td>
</tr>
</tbody>
</table>

- Scalability: source selection is much more sensitive because the gain is not continuous with fusion.
- Robustness: source selection is quite robust with respect to cost and gain.
- Scalability: source selection may even drop: in random order when we increased the number of sources from 50 to 228 (3.56 times more), the execution time increased when presence of high-quality and low-cost sources, source selection is much more sensitive for GRASP; source selection is quite robust with respect to cost and gain.
- Scalability: source selection is quite robust with respect to cost and gain.
- Recommendations: not surprisingly, repeating the random search and so there are 1.6%, .1%, 1% and .1% respectively. This is not surprising because no matter which fusion model we apply, our algorithm tends to find a better solution in the random search and so there are 1.6%, .1%, 1% and .1% respectively. This is not surprising because no matter which fusion model we apply, our algorithm tends to find a better solution in the random search and so there are 1.6%, .1%, 1% and .1% respectively.
Scalability of Source Selection

- **PopAccu_random**
- **Accu_random**
- **Vote_random**
- **Vote_dec**
- **Vote_inc**

![Graph showing scalability of source selection](image)

- Execution time (sec) vs. #Flight-Sources
- [X. L. Dong et al., 2012]
Discussion