Automated Data Curation at Scale

Bernhard Bicher (CEO)
Dr. Noah S. Bieler (Principal Data Scientist)

Winterthur, 12th of June 2015
Data Preparation Today

Data Scientists spend up to 80% of their time preparing data. Data Preparation is no self-service activity without IT involvement. Semi-automatic integration of more than 25 data sources is unfeasible. Data origins and lineage are frequently lost during processing.
Three Options

- **Manual**
  - Hire work force
  - Unreliable
  - Not sustainable
  - Expensive

- **Rule-based**
  - ETL
  - High Maintenance
  - Completeness
  - Needs expensive IT guy

- **Probabilistic**
  - Use statistics, NLP, ML
  - Choosing and combining the right algorithms
  - Only approximate results

---

ETL  Extract Transform Load
NLP  Natural Language Processing
ML   Machine Learning
The Art of Data Integration

- Identify Sources
- Profile Data
- Clean Data
- Normalise Data
- Identify Joins
- Entity Resolution
- Deduplication
- Post-Processing

Integrated Data

Automation using Probabilistic Approaches

Automation Potential:
- Low
- Medium
- High
- Very High
Probabilistic Methods and Approaches

- Identify Sources
- Profile Data
- Clean Data
- Normalise Data
- Identify Joins
- Entity Resolution
- Deduplication
- Post-Processing

**Identify Sources**
Outlier Detection, Authoritative Data, Type Detection

**Profile Data**
Encoding Errors Fixing, Pattern Mining, Column Swap

**Clean Data**
Probability Distribution, Entropy Measurement

**Normalise Data**

**Identify Joins**
Naive vs. Advanced ML Approaches

**Entity Resolution**

**Deduplication**
Computational Complexity Reduction
Profile Data

Example: Probabilistic Schema Detection

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Premium</th>
<th>City</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hans</td>
<td>Müller</td>
<td>TRUE</td>
<td>Winterthur</td>
<td>N/A</td>
</tr>
<tr>
<td>Hans</td>
<td>Mueller</td>
<td>1</td>
<td>Winterthur</td>
<td>CH</td>
</tr>
<tr>
<td>Jan</td>
<td>Muster</td>
<td>FALSE</td>
<td>Windisch</td>
<td>CH</td>
</tr>
</tbody>
</table>

- Profiling based on Authoritative Data
- Outlier Detection based on Histograms
- Identify Missing Values
- Content Detection using Decision Trees
  - String
  - Mostly Characters
  - All Capital
  - Dates
  - Mixed
  - Phone Numbers
  - Formatted Numbers
Clean, Normalise and Impute Data

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Premium</th>
<th>City</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>Morgenthal</td>
<td>TRUE</td>
<td>Winterthur</td>
<td></td>
</tr>
<tr>
<td>Hans</td>
<td>Müller</td>
<td>TRUE</td>
<td>Winterthur</td>
<td>CH</td>
</tr>
<tr>
<td>Hans</td>
<td>Mueller</td>
<td>1</td>
<td>CH</td>
<td>Winterthur</td>
</tr>
<tr>
<td>Jan</td>
<td>Muster</td>
<td>FALSE</td>
<td>Windisch</td>
<td>CHE</td>
</tr>
</tbody>
</table>

Pattern Mining

\[
\text{city} \Rightarrow \text{Country} = \text{CH}
\]

Fix Encoding Errors
Müller → Müller

Normalisation according to a Synonym Table

<table>
<thead>
<tr>
<th>ISO2</th>
<th>ISO3</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH</td>
<td>CHE</td>
<td>Schweiz</td>
</tr>
<tr>
<td>DE</td>
<td>DEU</td>
<td>Deutschland</td>
</tr>
<tr>
<td>FR</td>
<td>FRA</td>
<td>Frankreich</td>
</tr>
</tbody>
</table>

Column Swap
## Identify Join Columns

### Comparison of Probability Distribution

#### Datosilo 1

<table>
<thead>
<tr>
<th>FirstName</th>
<th>ClientID</th>
<th>Premium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Martin</td>
<td>1028934-1</td>
<td>TRUE</td>
</tr>
<tr>
<td>Sara</td>
<td>7462946-5</td>
<td>TRUE</td>
</tr>
<tr>
<td>Anna</td>
<td>9471991-3</td>
<td>FALSE</td>
</tr>
</tbody>
</table>

#### Datosilo 2

<table>
<thead>
<tr>
<th>CID</th>
<th>ProductName</th>
<th>ProductID</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-9471991</td>
<td>Monitor LCD</td>
<td>6413</td>
</tr>
<tr>
<td>C-7462946</td>
<td>Mouse Laser</td>
<td>5433</td>
</tr>
<tr>
<td>C-1028934</td>
<td>Keyboard QWERTY</td>
<td>961</td>
</tr>
</tbody>
</table>

The probability distributions $\mu_1$ and $\mu_1'$ are similar, as indicated by the visual representation.
Entity Resolution & Deduplication

Naive Approach

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Premium</th>
<th>City</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hans</td>
<td>Müller</td>
<td>TRUE</td>
<td>Winterthur</td>
<td></td>
</tr>
<tr>
<td>Hans</td>
<td>Mueller</td>
<td>1</td>
<td>Winterthur</td>
<td>CH</td>
</tr>
<tr>
<td>Jan</td>
<td>Muster</td>
<td>FALSE</td>
<td>Windisch</td>
<td>CH</td>
</tr>
</tbody>
</table>

All weights $w_i$ are the same.

$$w_i = \{0.2, 0.2, 0.2, 0.2, 0.2 \}$$

$$s = \sum_i w_i s_i$$

Advanced Approach

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Premium</th>
<th>City</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hans</td>
<td>Müller</td>
<td>TRUE</td>
<td>Winterthur</td>
<td>CH</td>
</tr>
<tr>
<td>Hans</td>
<td>Müller</td>
<td>TRUE</td>
<td>Winterthur</td>
<td>CH</td>
</tr>
<tr>
<td>Jan</td>
<td>Muster</td>
<td>FALSE</td>
<td>Windisch</td>
<td>CH</td>
</tr>
</tbody>
</table>

Adapt the weights $w_i$ using ML and optimise similarity calculations.

$$w_i = \{0.3, 0.3, 0.1, 0.2, 0.1 \}$$

De-Noising and normalisation helps to compare entities.

User feedback is incorporated into the estimate of the weights $\{w_i\}$ using ML.
Better out-of-the-box precision using ML and pre-estimated weights.

Start by initialising weights according to the column content.
For some cases, this can even eliminate the need for training at all.
Tackling Complexity in Deduplication

![Clustering Diagram]

- \( n = 10^6 \)
- \( k = 10^2 \)
- \( m = 50 \)

\[
\begin{align*}
\text{n} & \quad \text{Number of data records} \\
\text{k} & \quad \text{Number of clusters} \\
\text{m} & \quad \text{Number of iterations}
\end{align*}
\]

\[
\begin{align*}
n^2 & \rightarrow 10^{12} \\
0.5n^2 & \rightarrow 0.5 \cdot 10^{12} \\
k \cdot n \cdot m + 0.5 \cdot k(n/k)^2 & \rightarrow 10^{10}
\end{align*}
\]

Better scalability leads to faster execution.

Higher data locality, a “triangle” can run on a single node.
State-of-the-Art Infrastructure

Map-Reduce style using Apache Spark

Scalable: runs on a single Laptop as well as on a 10k-node Cluster.


Supports streaming, and provides MLlib and GraphX for machine learning and graph algorithms.
Summary

1. **Probabilistic methods save precious time**
   Decide on trade-off between fast data integration and precision.

2. **Leverage machine learning**
   Use business expert feedback to improve system precision and degree of automation.

3. **Broad data analysis**
   Mine over 100 instead of just 25 data sources.
Join us at www.meetup.com/spark-zurich!

Empowering organisations to unlock their wealth of data