Tools for the Digital Audit
IDEA Data Analysis Software

PRESENTER:
Kevin Simon, Product Lead, Analytics
About CaseWare IDEA

• Wholly owned subsidiary of CaseWare International (Toronto, Canada) founded in 1988

• CaseWare IDEA Headquartered in Ottawa, Canada

• CaseWare IDEA brand established in 2015 to differentiate ourselves in evolving analytics marketplace

• CaseWare IDEA Offices in Netherlands, France, Belgium, Hong Kong and Costa Rica

• Over 35 Country Distribution Partners including Sama Audit Systems in India [www.samaaudit.com]
CaseWare’s R&D Teams are dedicated to validating and implementing new technologies that assist the auditor in:

- Gathering and organization of audit data
- Execution of the audit
- Presentation and use of audit results
Agenda

• Tools to Help Manage Data
• Tools to For Audit Analytics
• Other R&D Projects
A Well-Known Issue

- Getting and managing data is among the most time consuming aspects of any audit
- This area can be overlooked as ‘plumbing’
- Much like real plumbing however, it’s quite important, and problematic if not done correctly
CaseWare’s Approach

- R&D Teams are developing tools to remove some of the burden
- As CSV is a common format for ‘unknown’ data sources, try to determine the meaning of each column based on the properties available to us:
  - Anything known about the source (Journal, Payroll etc)
  - NLP on column names if they exist
  - Type/Shape of data in each column
display first 10 rows of raw data

<table>
<thead>
<tr>
<th>Jt_NO</th>
<th>Acct_NO</th>
<th>Acct_Desc</th>
<th>Value</th>
<th>Trans_Date</th>
<th>Entered_Date</th>
<th>Entered_By</th>
<th>Trans_Desc</th>
<th>Ref_Type</th>
<th>Ref_NO</th>
<th>Invoice_Amt</th>
<th>Customer_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>SJ00001</td>
<td>45100</td>
<td>-122.73</td>
<td>2016-01-01</td>
<td>2016-01-21</td>
<td>Emma Flatt</td>
<td>Sale; VUO</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>1</td>
<td>SJ00001</td>
<td>11700</td>
<td>12.27</td>
<td>2016-01-01</td>
<td>2016-01-21</td>
<td>Emma Flatt</td>
<td>Sale; VUO</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>2</td>
<td>SJ00001</td>
<td>23050</td>
<td>-12.27</td>
<td>2016-01-01</td>
<td>2016-01-21</td>
<td>Emma Flatt</td>
<td>Sale; VUO</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>3</td>
<td>SJ00002</td>
<td>11700</td>
<td>15.95</td>
<td>2016-01-03</td>
<td>2016-01-21</td>
<td>Emma Flatt</td>
<td>Sale; VUO</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>4</td>
<td>SJ00002</td>
<td>23050</td>
<td>-15.95</td>
<td>2016-01-03</td>
<td>2016-01-21</td>
<td>Emma Flatt</td>
<td>Sale; VUO</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>5</td>
<td>SJ00002</td>
<td>45100</td>
<td>-159.55</td>
<td>2016-01-03</td>
<td>2016-01-21</td>
<td>Emma Flatt</td>
<td>Sale; VUO</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>6</td>
<td>SJ00003</td>
<td>11700</td>
<td>15.95</td>
<td>2016-01-10</td>
<td>2016-01-21</td>
<td>Emma Flatt</td>
<td>Sale; VUO</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>7</td>
<td>SJ00003</td>
<td>23050</td>
<td>-15.95</td>
<td>2016-01-10</td>
<td>2016-01-21</td>
<td>Emma Flatt</td>
<td>Sale; VUO</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>8</td>
<td>SJ00003</td>
<td>45100</td>
<td>-159.55</td>
<td>2016-01-10</td>
<td>2016-01-21</td>
<td>Emma Flatt</td>
<td>Sale; The</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>9</td>
<td>SJ00005</td>
<td>11700</td>
<td>19.99</td>
<td>2016-01-28</td>
<td>2016-01-21</td>
<td>Emma Flatt</td>
<td>Sale: The</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>

apply known maps from column name to Analytics tag
CaseWare’s Approach: After Some Proprietary Math
CaseWare’s Approach: Prediction Results

predict most likely tag for each column

Most likely based on most common most-likely tag per cell:

```python
# Training
print("training (actual -> most likely)")
train_preds.groupby("actual").apply(lambda grp: Counter(grp["predicted"]).most_common(2)[0][0])
training (actual -> most likely)
```

```plaintext
<table>
<thead>
<tr>
<th>Column</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount</td>
<td>Amount</td>
</tr>
<tr>
<td>DocumentAmount</td>
<td>DocumentAmount</td>
</tr>
<tr>
<td>DocumentNumber</td>
<td>DocumentNumber</td>
</tr>
<tr>
<td>EffectiveDate</td>
<td>EffectiveDate</td>
</tr>
<tr>
<td>EnteredDate</td>
<td>EnteredDate</td>
</tr>
<tr>
<td>GLAccountNumber</td>
<td>GLAccountNumber</td>
</tr>
<tr>
<td>JournalEntryDescription</td>
<td>JournalEntryDescription</td>
</tr>
<tr>
<td>JournalEntryIDNumber</td>
<td>untagged</td>
</tr>
<tr>
<td>customerId</td>
<td>customerId</td>
</tr>
<tr>
<td>enteredBy</td>
<td>enteredBy</td>
</tr>
<tr>
<td>untagged</td>
<td>untagged</td>
</tr>
</tbody>
</table>
```

```python
# Testing
print("testing (actual -> most likely)")
test_preds.groupby("actual").apply(lambda grp: Counter(grp["predicted"]).most_common(2)[0][0])
testing (actual -> most likely)
```

```plaintext
<table>
<thead>
<tr>
<th>Column</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount</td>
<td>Amount</td>
</tr>
<tr>
<td>DocumentAmount</td>
<td>DocumentAmount</td>
</tr>
<tr>
<td>DocumentNumber</td>
<td>DocumentNumber</td>
</tr>
<tr>
<td>EffectiveDate</td>
<td>EffectiveDate</td>
</tr>
<tr>
<td>EnteredDate</td>
<td>EnteredDate</td>
</tr>
<tr>
<td>GLAccountNumber</td>
<td>GLAccountNumber</td>
</tr>
<tr>
<td>JournalEntryDescription</td>
<td>JournalEntryDescription</td>
</tr>
<tr>
<td>JournalEntryIDNumber</td>
<td>untagged</td>
</tr>
<tr>
<td>customerId</td>
<td>customerId</td>
</tr>
<tr>
<td>enteredBy</td>
<td>enteredBy</td>
</tr>
<tr>
<td>untagged</td>
<td>untagged</td>
</tr>
</tbody>
</table>
```
CaseWare’s Approach

- Trial balance account mapping
  - CSV, or in a known data format
- Known formats may require mapping to roll up in a consistent way
- Similar problem to CSV column tagging
- Similar approach to minimize the work the auditor needs to do
CaseWare’s Unique Value:

- 30 years of customers using our products
- Reasonably consistent data model
- Ongoing good relations
- Millions of potential audits to mine for model building
- Ability for smaller firms to opt-in and gain the benefits currently availability to the Big 4
Data Management Tools

- Management of the data is key before any analytics can be done
- CaseWare is actively working on tools to help the auditor here
- Significantly less glamorous than some other R&D areas, but huge potential to streamline the audit workflow
- Additional research focuses on other potential tools and technologies
Tools for Audit Analytics
The Classic

10.4 Release Date: November 19, 2018

- Better Python Integration

2019

- CaseWare Cloud Integration!
More IDEA

Natural Language Processing
- Opens up non-traditional data sources to the auditor
- Leverages Python integration

Outliers
- Beta now available to download
Analytics in the Cloud

Score Breakdown

<table>
<thead>
<tr>
<th>Score</th>
<th>Amount</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Transactions</td>
<td>$1,953,915</td>
<td>905</td>
</tr>
<tr>
<td>High Score</td>
<td>$419,716</td>
<td>46</td>
</tr>
<tr>
<td>Medium Score</td>
<td>$503,511</td>
<td>39</td>
</tr>
<tr>
<td>Low Score</td>
<td>$900,479</td>
<td>696</td>
</tr>
<tr>
<td>Zero Score</td>
<td>$130,207</td>
<td>124</td>
</tr>
</tbody>
</table>

Top Scored Exceptions

<table>
<thead>
<tr>
<th>Exception Type</th>
<th>Amount</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Descriptions</td>
<td>$1,351,787</td>
<td>755 transactions</td>
</tr>
<tr>
<td>Rounded Amounts</td>
<td>$983,707</td>
<td>210 transactions</td>
</tr>
<tr>
<td>High Amounts</td>
<td>$676,262</td>
<td>21 transactions</td>
</tr>
<tr>
<td>Duplicate Entries</td>
<td>$264,890</td>
<td>381 transactions</td>
</tr>
</tbody>
</table>

Top Scored Users

- Qwerty (30%)
- Adonis (20%)
- Amy (10%)
- May (5%)
- Chelsea (5%)

Top Scored Document Types

- G1
- G2
- H1
- C1
- D1
- F1

CASEWARE.
Process Mining

- Well known research topic
- CaseWare Approach:
  - Map the expected flows
  - Visually call out the exceptions
The auditor is still the key
  ○ Both the data and the automated processes need to be understood

Analytics augments the auditor for Efficiency and Accuracy

Analytics Cannot Replace the Auditor
Other R&D Projects
Prediction of Misstatement Requiring Re-filing

Machine Learning, NLP, Ensemble Predictive Engine
ai.caseware.com

Synchronoss Technologies
https://ai.caseware.com/company/0001131554
https://www.last10k.com/sec-filings/sncr
Time Series Analytics

- Various techniques to examine data sets for cycles and trends
  - Cycles can be normalized out
  - Trends can be emphasized and examined
- Heavy use of Autocorrelation
- Fourier Analysis
- Techniques that are common in other types of analysis being applied to accounting data
- Can help find related transactions vs. single transactions
Thank You

QUESTIONS?

Contact:

- Kevin Simon: kevin.simon@caseware.com
- Visit: www.casewareanalytics.com