Cleaning Up and Managing Dirty Data in MATLAB

Siddharth Sundar, Application Engineer
Agenda

- Typical Challenges with Data Handling and Management
  - A Fundamental Valuation Example
  - A Text Analytics Example
  - What about Cleaning Large Datasets?
  - Summary and Resources
Typical Challenges in Data Cleaning, Management

- We are Drowning in Data
  - Data Volume and Variety
  - Different sources, types, sizes
  - Garbage-in garbage-out
So many Data Sources

Local disk
Shared folders
Databases

Flat files/Excel

Datafeeds

Spark+Hadoop

Webpages
So many kinds of Data

<table>
<thead>
<tr>
<th>person</th>
<th>year</th>
<th>income</th>
<th>age</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2001</td>
<td>1300</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2002</td>
<td>1600</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2003</td>
<td>2000</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2001</td>
<td>2000</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2002</td>
<td>2300</td>
<td>39</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2003</td>
<td>2400</td>
<td>40</td>
<td>2</td>
</tr>
</tbody>
</table>

```
ans = 508x1 string array

"Walmart: "you wanna destroy Amazon?" Google: "bet" $WMT $GOOG
$WMT wants next level customer service w/highly personalized
"Ironic prelude to $DIS buying $TWTR soon IMO $AAPL $GOOG $SPY
"$AMZN the $WMT threat grows each and every day https://t.co/
"MU Investments Co. Ltd. Sells 30 Shares of Alphabet Inc. $GOOG
"Ad $ are going to $GOOG and $FB away from wppg #Advertising
"Big bullish unusual option activity detected: $SPX, $GOOG, $FB
"REPORT: Apple to build data center in Iowa: https://t.co/jwH
"RT @theflynews: REPORT: Apple to build data center in Iowa: I
```
So many kinds of Data

LET’S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
Typical Challenges in Data Cleaning, Management

- Drowning in Data
  - Data Volume and Variety
  - Different sources, types, sizes
  - Garbage-in garbage-out

- Poor Data Quality
Poor Data Quality

SEC ID, 2011Q1, 2011Q2, 2011Q3, 2011Q4, 2012Q1
1572422, 1510333, 1417664, 721683, 0.25, 0.28, 0.3, 0.32, 0.3, 0.35, 0.32, 0.1175029, 0.01, 0.03, 0.03, 0.08, 0.832488, 0.07, 0.03, 0.05, 0.06, 0.05, 0.01383729, 0.07, -0.07, -0.05, -0.01, 0.04, 0.0, 0.1145765, 0.01, -0.04, 0.1023994, 0.01, -0.04, 0.1130464, 0.68, 0.19, -0.27, 0.64, 0.68, -0.31, 0
Typical Challenges in Data Cleaning, Management

- **Drowning in Data**
  - Data Volume and Variety
  - Different sources, types, sizes
  - Garbage-in garbage-out

- **Poor Data Quality**
  - Poorly formatted files
  - Irregularly sampled data
  - Redundant, Missing data, Outliers

- **Need for more customized analytics**
  - No one size fits all
Agenda

- Typical Challenges with Data Handling and Management
- A Fundamental Valuation Example
- A Text Analytics Example
- What about Cleaning Large Datasets?
- Summary and Resources
Demo: Fundamental Valuation of S&P100 securities

**Goal:**
- Fundamental valuation for ranking stocks based on historical EPS trends

**Approach**
- Access data from CSV files
- Preprocess to clean-up text (missing data and outliers)
- Calculate strength of historical EPS trends
How do we handle Missing Data?

Does missing data have meaning?

Yes

Type of data

Numerical

Convert missing values to meaningful number

Categorical

Missing values become their own category

No

Remove instances with missing data

Does the data follow a simple distribution?

Yes

Replace value with value of preceding instance

No

Impute with simple ML model

Large, temporarily ordered data

Impute missing values with column mean

Does the data have meaning?

Yes

Remove instances with missing data

No

Impute missing values with column median

Dataset is big and little data is missing at random

Impute with simple ML model

Otherwise

Impute missing values with column median

Numerical

Categorical

Missing values become their own category
Summary: Fundamental Valuation Example

- Interactive tools to import, visualize data
- Code generation from interactive tools
- Built-in clean up functions
- Align and calculate group stats
- Save time
Agenda

- Typical Challenges with Data Handling and Management
- A Fundamental Valuation Example
- A Text Analytics Example
- What about Cleaning Large Datasets?
- Summary and Resources
Goal:

- Analyze the sentiment of SEC filings for S&P 100 companies to use as a stock picking/ranking indicator

Approach

- Access data directly from HTML/PDF
- Preprocess to clean-up text and deal with domain-specific terms
- Predict sentiment
Summary: Sentiment Analysis Example

- String Class in MATLAB
- Easy to use/read functions to do text processing
- Text visualization functions
- Less regexp, more built-in commands
- More processing, less time
Agenda

- Typical Challenges with Data Handling and Management
- A Fundamental Valuation Example
- A Text Analytics Example
- What about Cleaning Large Datasets?
- Summary and Resources
Demo: Technicals calculation to time the market

- **Objective**
  - Calculate technical indicators on Big intraday data

- **Data**
  - Intraday tick data scraped from the web
  - Missing data, outliers etc.

- **Approach**
  - Preprocess data
  - Explore data
  - Calculate technicals
How do you work with tall arrays in MATLAB?

- **datastore**
  - Points to the data
  ```matlab
  >> fileLoc = '.\datasets\*.csv';
  >> ds = datastore(fileLoc);
  ```

- **Tall array**
  - Variable representation of the data in your workspace
  ```matlab
  >> tt = tall(ds);
  ```

- **Functions**
  - Operate on tall arrays
  ```matlab
  >> tt = fillmissing(t,'nearest');
  ```
Summary: Technicals Demo

- Big Data handled just like data that fits in memory (Tall)
- No need for use of Mapreduce or other Big Data technologies/frameworks
- Easy Big Data visualization
- Scalability of MATLAB models
Agenda

- Typical Challenges with Data Handling and Management
- A Fundamental Valuation Example
- A Text Analytics Example
- What about Cleaning Large Datasets?
- Summary and Resources
Revisiting the Challenges with Handling Dirty Data

- **Drowning in Data**
  - Data Volume and Variety
  - Different sources, types, sizes
  - Garbage-in garbage-out

- **Poor Data Quality**
  - Poorly formatted files
  - Irregularly sampled data
  - Redundant, Missing data, Outliers

- **No one size fits all solution for Data cleaning**
Get Training

Accelerate your learning curve:
- Customized curriculum
- Learn best practices
- Practice on real-world examples

Options to fit your needs:
- Self-paced (online)
- Instructor led (online and in-person)
- Customized curriculum (on-site)
Training Roadmap

MATLAB for Financial Applications

Data Analysis and Modeling
- Statistical Methods
- Machine Learning
- Time-Series Modeling (Econometrics)
- Risk Management
- Optimization Techniques

Application Development
- Programming Techniques
- Interactive User Interfaces
- Parallel Computing

Content for On-site Customization
- Asset Allocation
- Interfacing with Databases
- Interfacing with Excel
Want to Learn More?

Engineers

Siddharth Sundar
Finance Application Engineer
Siddharth.Sundar@mathworks.com

Marshall Alphonso
Senior Finance Engineer, New York
Marshall.Alphonso@mathworks.com

Account Managers

Chuck Castricone
Account Manager
Chuck.Castricone@mathworks.com

Mike DeLucia
Senior Account Manager
Mike.DeLucia@mathworks.com