Data Management in Finance with MATLAB

Valerio Sperandeo

© 2019 The MathWorks, Inc.

80% | 20%

Agenda

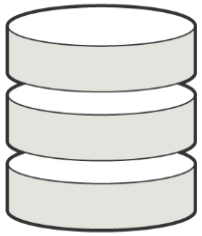
- Typical Challenges with Data Handling and Management
- A Fundamental Valuation Example
- A Text Analytics Example
- What about Cleaning Large Datasets?
- Data Cleaning by Machine Learning
- Q&A

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

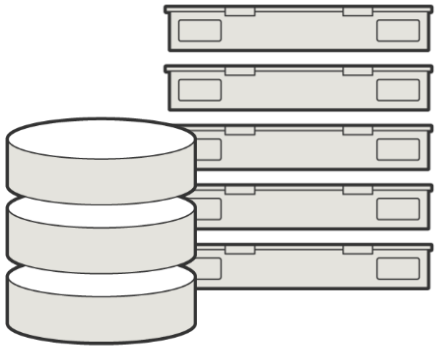
Local disk
Shared folders
Databases



Datafeeds

Webpages

Flat files/Excel



Spark+Hadoop

So many kinds of Data

person	year	income	age	sex
1	2001	1300	27	1
1	2002	1600	28	1
1	2003	2000	29	1
2	2001	2000	38	2
2	2002	2300	39	2
2	2003	2400	40	2

```
ans = 508×1 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. $GOO
    "Ad $ are going to $GOOG and $FB away from wppgy  #Advertising
    "Big bullish unusual option activity detected: $SPX, $GOOG, $C
    "REPORT: Apple to build data center in Iowa: https://t.co/jwH6
    "RT @theflynews: REPORT: Apple to build data center in Iowa: h
```

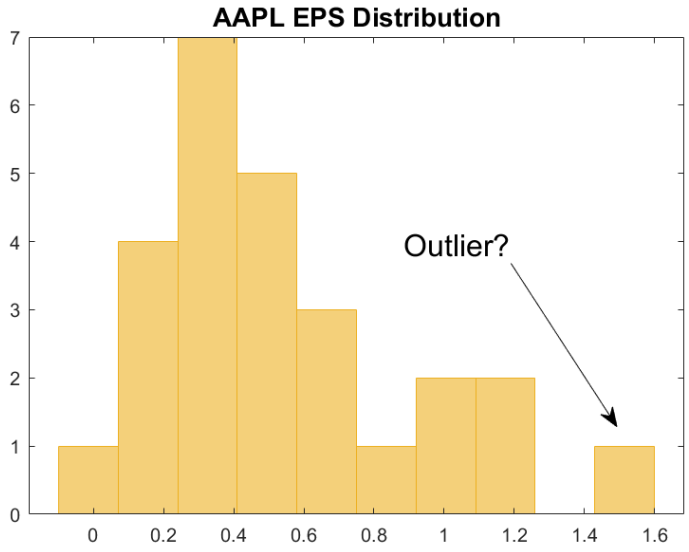
Date	Close	High	Low	Open	Volume
2017-Aug-17 09:30:00	925.87	925.87	925.78	925.78	13585
2017-Aug-17 09:31:00	923.58	925.45	923.26	925.45	5667
2017-Aug-17 09:32:00	925.2	925.48	923.985	924.42	9254
2017-Aug-17 09:33:00	925.5407	925.5407	925.27	925.27	1500
2017-Aug-17 09:34:00	924.63	925.72	924.63	925.64	1505
2017-Aug-17 09:35:00	925	925.39	924.98	925.06	1112
2017-Aug-17 09:36:00	924.7	925.22	924.11	925.195	3092
2017-Aug-17 09:37:00	924.85	924.9799	924.21	924.3	1285
2017-Aug-17 09:38:00	925.04	925.86	924.46	924.87	6171
2017-Aug-17 09:39:00	926.15	926.15	925.13	925.13	1400
2017-Aug-17 09:40:00	926.14	926.17	925.38	925.76	1000
2017-Aug-17 09:41:00	925.9364	926.295	925.93	926.16	525
2017-Aug-17 09:42:00	926.55	926.62	926.11	926.5	700
2017-Aug-17 09:43:00	926.25	926.25	925.83	925.88	700
2017-Aug-17 09:44:00	925.94	925.94	925.03	925.51	1400

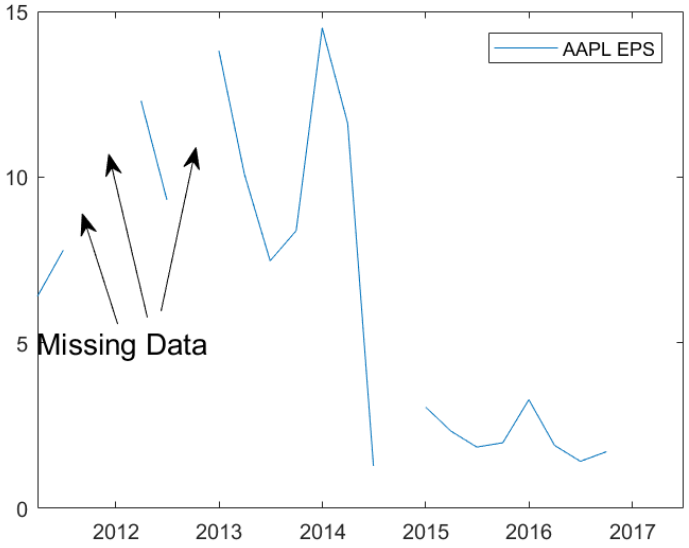
Typical Challenges in Data Cleaning, Management

- 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,,,,,,,,,,,
832488,,,,0.07,0.03,,0.05,,0.06,,0.05,0.0
1383729,,,0.07,-0.07,-0.05,-0.01,0.04,0.0
1145765,,,,,,,,,,,,,,,,,,,,,,,,,,
1023994,,,,,,,,,,,,,,,0.01,,,,-0.04,,,,,,
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?
- Data Cleaning by Machine Learning
- Q&A

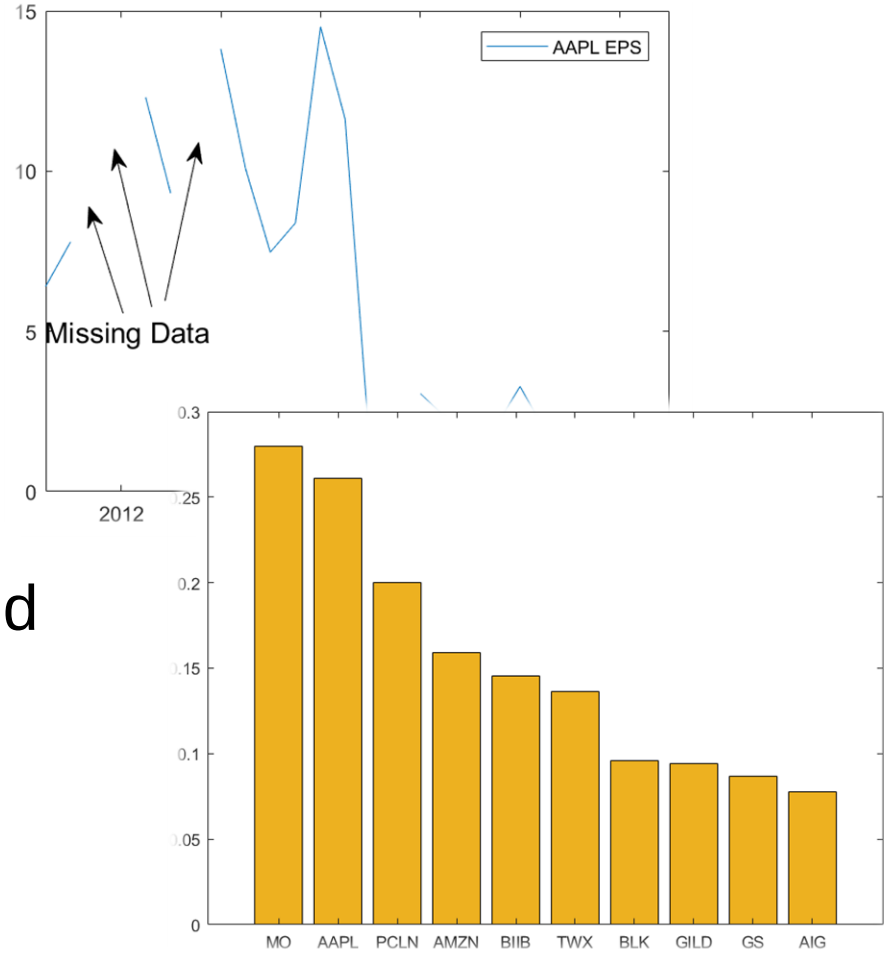
Demo: Fundamental Valuation of S&P100 securities

Goal:

- Fundamental valuation for ranking stocks based on historical EPS trends

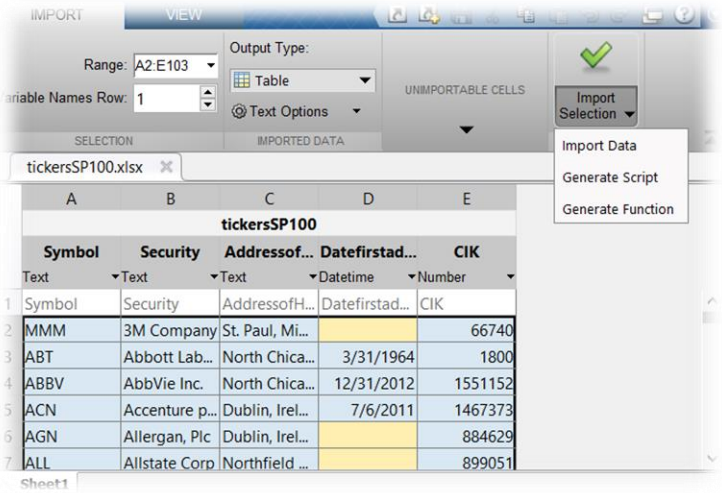
Approach

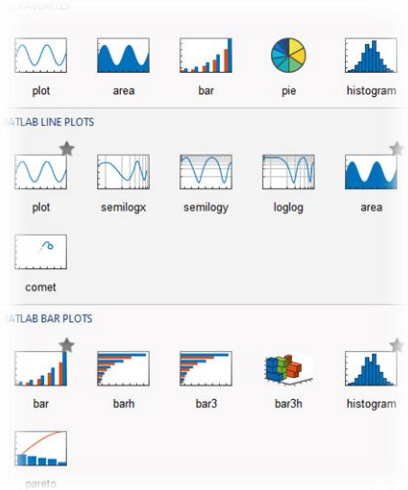
- Access data from CSV files
- Preprocess to clean-up text (missing data and outliers)
- Calculate strength of historical EPS trends



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?
- Data Cleaning by Machine Learning
- Q&A

Demo: Sentiment Analysis of SEC filings

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

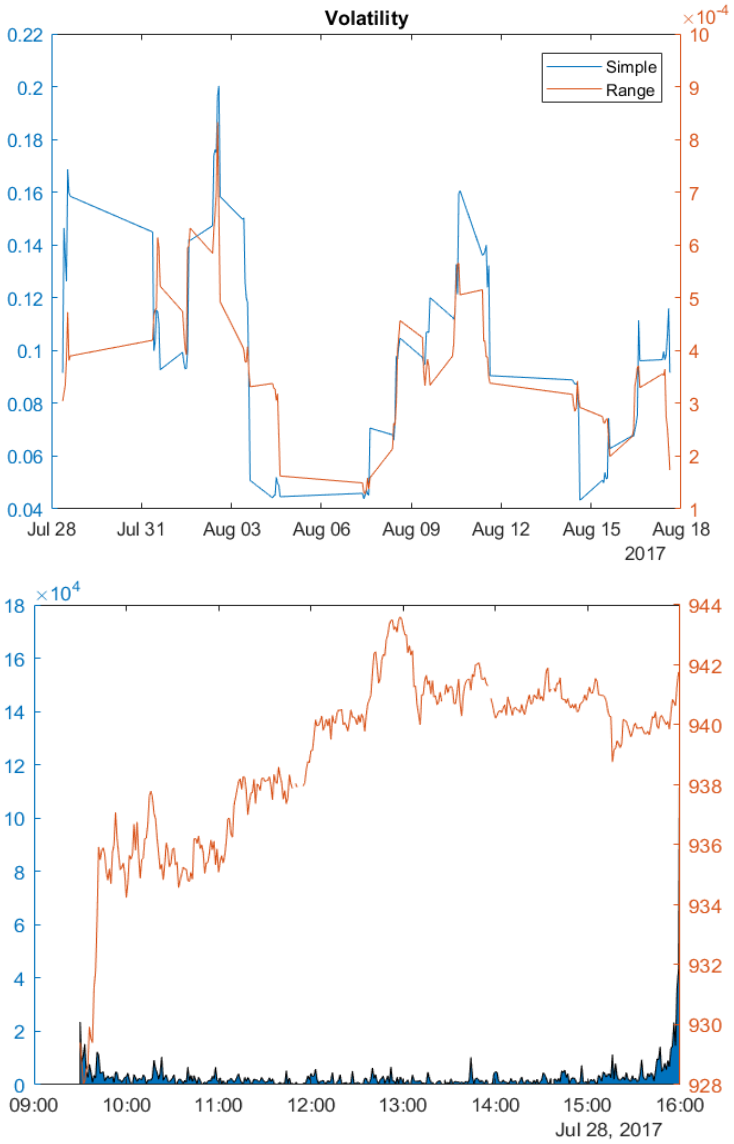
```
following discussion risk factors contains forward-looking statements
risk factors important understanding statements form 10-k
following information read conjunction part item management's discussion analy
business financial condition operating results company affected number factors
factors whole part materially adversely affect company's business financial co
following factors well factors affecting company's financial condition operati
global regional economic conditions materially adversely affect company
company's operations performance depend significantly global regional economic
uncertainty global regional economic conditions poses risk consumers businesse
worldwide regional economic conditions material adverse effect demand company'
demand differ materially company's expectations result currency fluctuations c
correspond effect strengthening
dollar
```

Agenda

- Typical Challenges with Data Handling and Management
- A Fundamental Valuation Example
- A Text Analytics Example
- What about Cleaning Large Datasets?
- Data Cleaning by Machine Learning
- Q&A

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
- Tall array
  - Variable representation of the data in your workspace
- Functions
  - Operate on tall arrays

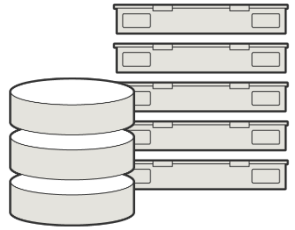
```
>> fileLoc = '.\datasets\*.csv';
>> ds = datastore(fileLoc);
```

```
>> tt = tall(ds);
```

`tall`

```
>> tt = fillmissing(t,'nearest');
```

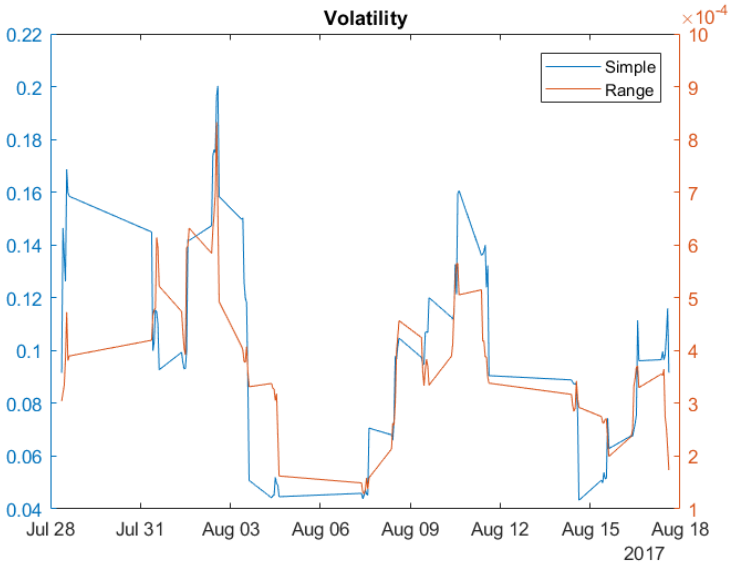
Local disk
Shared folders
Databases

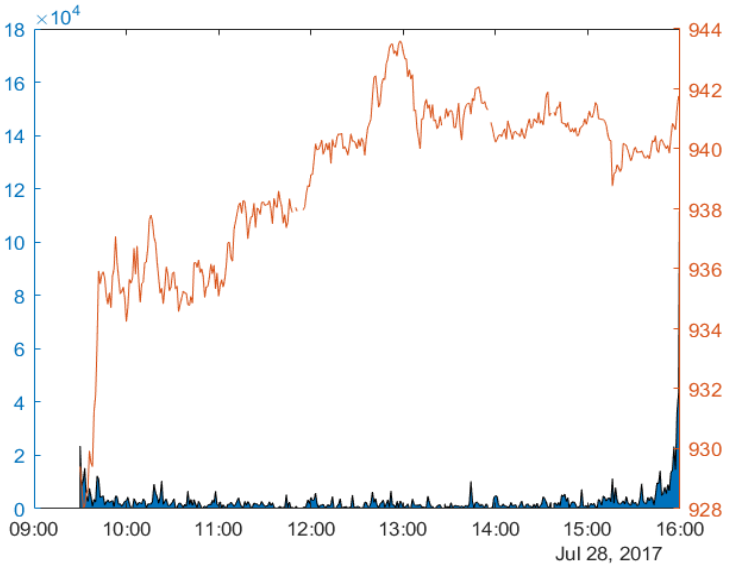


Spark + Hadoop

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?
- Data Cleaning by Machine Learning
- Q&A

Agenda

- Typical Challenges with Data Handling and Management
- A Fundamental Valuation Example
- A Text Analytics Example
- What about Cleaning Large Datasets?
- Data Cleaning by Machine Learning
- Q&A