Data Quality & Processing

Mingjue Wang, Ronnie Song, Frank Sun

Lessons From History

- **The Enron Accounting Scandal of 2001 - reporting problem**
  Use of accounting loopholes, special purpose entities, and poor financial reporting to hide billions of dollars in debt from failed deals and projects.

- **Tetraethyllead in Gasoline in the 1920s - processing problem**
  Industry scientists even suggested the human body naturally harbors lead, so high levels shouldn't be a health concern.

"$3.1 *trillion*, IBM's estimate of the yearly cost of poor quality data, in the US alone, in 2016." - Harvard Business Review.

Project Example

There are many ways to misuse the raw data.

The simplest way is input all the data we gathered into the machine.

- pre-processing problem

1	detectorid	starttime	volume	speed	occupancy	status	dqflags
6830	1417	9/15/2011 0:56:00	1	60	1	2	0
6831	1417	9/15/2011 0:56:20	0		0	0	0
6832	1417	9/15/2011 0:56:40	0		0	0	0
6833	1417	9/15/2011 0:57:00	0		0	0	0
6834	1417	9/15/2011 0:57:20	0		0	0	0
6835	1417	9/15/2011 0:57:40	0		0	0	0
6836	1417	9/15/2011 0:58:00	0		0	0	0
6837	1417	9/15/2011 0:58:20	1	56	0	3	0
6838	1417	9/15/2011 0:58:40	0		0	0	0
6839	1417	9/15/2011 0:59:00	0		0	0	0
6840	1417	9/15/2011 0:59:20	1	83	0	3	0
6841	1417	9/15/2011 0:59:40	0		0	0	0
6842	1418	9/15/2011 0:00:00	6	57	6	2	0
6843	1418	9/15/2011 0:00:20	0	0	1	3	72
6844	1418	9/15/2011 0:00:40	4	55	3	2	0
6845	1418	9/15/2011 0:01:00	0		0	0	0
6846	1418	9/15/2011 0:01:20	3	57	3	2	0
6847	1418	9/15/2011 0:01:40				0	0
6848	1418	9/15/2011 0:02:00	2	71	1	2	0
6849	1418	9/15/2011 0:02:20	4	53	4	2	0
6850	1418	9/15/2011 0:02:40	6	52	10	2	0
6851	1418	9/15/2011 0:03:00	3	66	6	2	0
6852	1418	9/15/2011 0:03:20	4	62	4	2	0
6853	1418	9/15/2011 0:03:40	2	56	2	2	0
6854	1418	9/15/2011 0:04:00	4	70	7	2	0
6855	1418	9/15/2011 0:04:20	1	56	0	3	0
6856	1418	9/15/2011 0:04:40	1	56	1	2	0
6857	1418	9/15/2011 0:05:00	1	84	0	3	0
6858	1418	9/15/2011 0:05:20	3	66	6	2	0
6859	1418	9/15/2011 0:05:40	1	60	5	2	0
6860	1418	9/15/2011 0:06:00	2	57	2	2	0
6861	1418	9/15/2011 0:06:20	3	56	6	2	0
6862	1418	9/15/2011 0:06:40	1	56	1	2	0
6863	1418	9/15/2011 0:07:00	4	54	2	2	0
6864	1418	9/15/2011 0:07:20	1	57	2	2	0
6865	1418	9/15/2011 0:07:40	2	70	1	2	0
6866	1418	9/15/2011 0:08:00	3	64	5	2	0

Data Quality

The general theme of data quality is around finding outliers that do not meet specific requirements and record sets that violate assumptions.

Data Quality Layers

1. **Extract Transform and Load (ETL) layer: pre-processing**
   The goal here is to check to ensure that data is not lost or degraded while moving from the source to the target system.
2. **Operation layer: processing**
   The goal here is to ensure that fundamental understandings are not violated and that the data makes sense.
3. **Reporting layer: presentation**
   This is the layer that end users interact with your data. Do you guys still remember the first section of *Calling BS*?

Data Preprocessing

More data beats clever algorithms, but better data beat more data.

-Peter Norvig

Why we care about data preprocessing

- Not ideal data
- Dirty data issues
- Poor Data -> incorrect/misleading statistics -> wrong decision
- Data Scientists spend up to 80%~90% of their time just in preprocessing the data
- Garbage in, garbage out!

Dirty Data

Raw Data is not clean!

- Incomplete Data
- Inconsistent Data
- Noisy Data

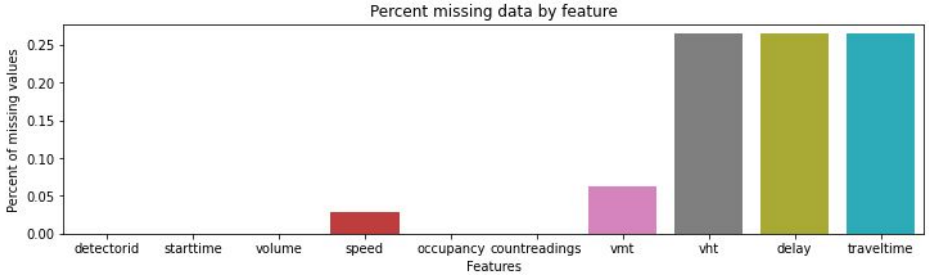
Causes:

- device malfunction
- different data sources
- measurement not possible

```
[ ]  1 dataset[dataset.isnull().any(axis=1)]
```

	detectorid	starttime	volume	speed	occupancy	countreadings	vmt	vht	delay	traveltime
92	1345	2011-10-08 01:00:00-07	0	NaN	0.0	3	0.0	NaN	NaN	NaN
94	1345	2011-10-08 04:00:00-07	0	NaN	0.0	1	0.0	NaN	NaN	NaN
135	1345	2011-10-19 20:00:00-07	0	NaN	0.0	1	0.0	NaN	NaN	NaN
316	1346	2011-10-08 04:00:00-07	0	NaN	0.0	1	0.0	NaN	NaN	NaN
2049	1348	2011-10-08 03:00:00-07	0	NaN	0.0	1	0.0	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
65854	1956	2011-11-15 19:00:00-08	449	0.0	0.0	180	NaN	NaN	NaN	NaN
65855	1956	2011-11-15 20:00:00-08	407	0.0	0.0	180	NaN	NaN	NaN	NaN
65856	1956	2011-11-15 21:00:00-08	364	0.0	0.0	180	NaN	NaN	NaN	NaN
65857	1956	2011-11-15 22:00:00-08	220	0.0	0.0	180	NaN	NaN	NaN	NaN
65858	1956	2011-11-15 23:00:00-08	138	0.0	0.0	180	NaN	NaN	NaN	NaN

25074 rows × 10 columns

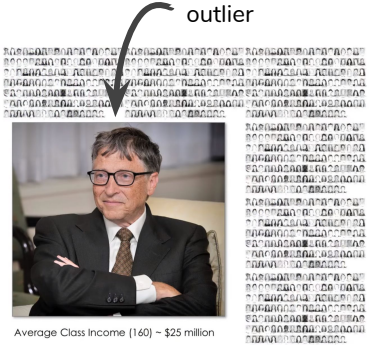


Data Cleaning: Incomplete Data

Replace With Zero, Mean, Median, or mode.

- **Mean**: Suitable for continuous data without outliers
- **Median**: Suitable for continuous data with outliers
- **Mode**: Suitable for a categorical data

detectorid	starttime	speed
1345	NaN	60.85
1346	2011-09-16	60.85
1348	2011-09-16	NaN
...	...	...
1956	2011-09-16	60.85



Example from Calling BS 4.2

Data Cleaning: Inconsistent Data

Recalculate the values

Ensure consistent units

- **Data:** July 4, 2020, 07/04/2020 -> 2020/07/04

Decimal scaling

- **Speed:** 93.342km/h, 93.342kmh, 58mph -> 93.342
- **Temperature:** 80F, 80f, 27C -> 28

Nominalization/standardization

- fall between [0, 1], or [-1, 1].

detectorid	starttime	speed
1345	2011-09-16	60.85
1346	09/16/2011	58.1
1348	2011-09-16	62.3mph
...	...	...
1956	Sep 16	60.31

Data Cleaning: "Nosy" Data

Drop an observation (row) with missing values

Replace with attribute means or median if it is continuous data

Substitute with a value from a similar instance

detectorid	starttime	speed
1345	2011-09-16	0.12
1346	2011-09-16	60.85
1348	2011-09-16	61.52
1349	2011-09-16	59.12
1350	2011-09-16	30000
1351	2011-09-16	62.54
...	...	...
1956	2011-09-16	60.85

A Simple Data Cleaning Pipeline

```
[ ]  1 dataset[dataset.isnull().any(axis=1)]
```

	detectorid	starttime	volume	speed	occupancy	countreadings	vmt	vht	delay	traveltime
92	1345	2011-10-08 01:00:00-07	0	NaN	0.0	3	0.0	NaN	NaN	NaN
94	1345	2011-10-08 04:00:00-07	0	NaN	0.0	1	0.0	NaN	NaN	NaN
135	1345	2011-10-19 20:00:00-07	0	NaN	0.0	1	0.0	NaN	NaN	NaN
316	1346	2011-10-08 04:00:00-07	0	NaN	0.0	1	0.0	NaN	NaN	NaN
2049	1348	2011-10-08 03:00:00-07	0	NaN	0.0	1	0.0	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
65854	1956	2011-11-15 19:00:00-08	449	0.0	0.0	180	NaN	NaN	NaN	NaN
65855	1956	2011-11-15 20:00:00-08	407	0.0	0.0	180	NaN	NaN	NaN	NaN
65856	1956	2011-11-15 21:00:00-08	364	0.0	0.0	180	NaN	NaN	NaN	NaN
65857	1956	2011-11-15 22:00:00-08	220	0.0	0.0	180	NaN	NaN	NaN	NaN
65858	1956	2011-11-15 23:00:00-08	138	0.0	0.0	180	NaN	NaN	NaN	NaN

25074 rows × 10 columns

Dirty Data

data cleaning

detectorid	starttime	speed
1345	2011-09-16	60.14
1346	2011-09-16	60.85
1348	2011-09-16	61.52
1349	2011-09-16	59.12
1350	2011-09-16	60.14
1351	2011-09-16	62.54
...	...	...
1956	2011-09-16	60.85

Clean Data

store

Data warehouse
or Database

Data Processing

The conversion of data into usable and desired form

Data Processing

(1)	{ 17 fields }
sharded	true
capped	false
wiredTiger	{ 14 fields }
ns	dms.history
count	1301472270.0
size	168735402751.0
storageSize	70561280000.0
totalIndexSize	99157499904.0
indexSizes	{ 3 fields }
avgObjSize	129.0
maxSize	0
nindexes	3
nchunks	7548

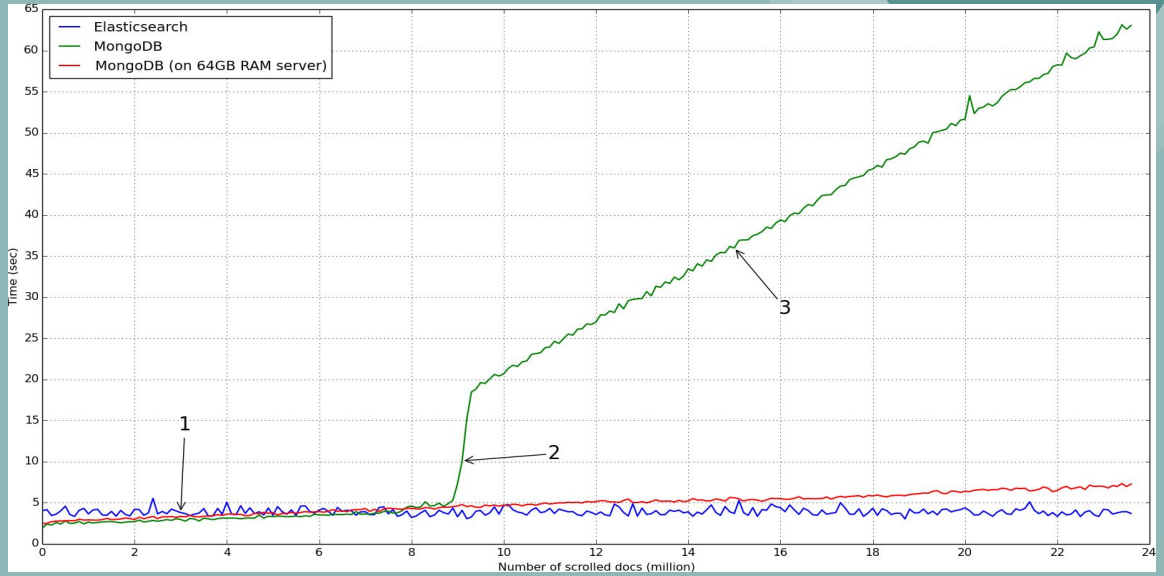
Response(20017ms)

```
HTTP/1.1 200 OK
X-Powered-By: Express
Access-Control-Allow-Origin: *
Content-Type: application/json; charset=utf-8
Content-Length: 192
ETag: W/"c0-411kKg+urQ38/sfkVJDHp/0sFq4"
Date: Thu, 16 Jul 2020 03:48:01 GMT
Connection: close

[
  {
    "_id": {
      "detectorid": 1347
    },
    "totalvolume": 13208
  },
```

Elasticsearch

- **Distributed text search engine**
- **Try to resolve the performace issue**



Thank you

Q&A

Reference:

Thomas, Redman, (September 22, 2016). Bad Data Costs the U.S. $3 Trillion Per Year. Source from: https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year#comment-section.

Matthew, Zajechowski, (May 25, 2017). The Lessons We can Learn from Bad Data Mistakes Made Throughout History. Source from: https://www.smartdatacollective.com/lessons-can-learn-bad-data-mistakes-made-throughout-history/.

Gary, Cheung, (January 3, 2019). A Deep Dive Into Data Quality. Source from: https://towardsdatascience.com/a-deep-dive-into-data-quality-c1d1ee576046.