AMI Conference 2020

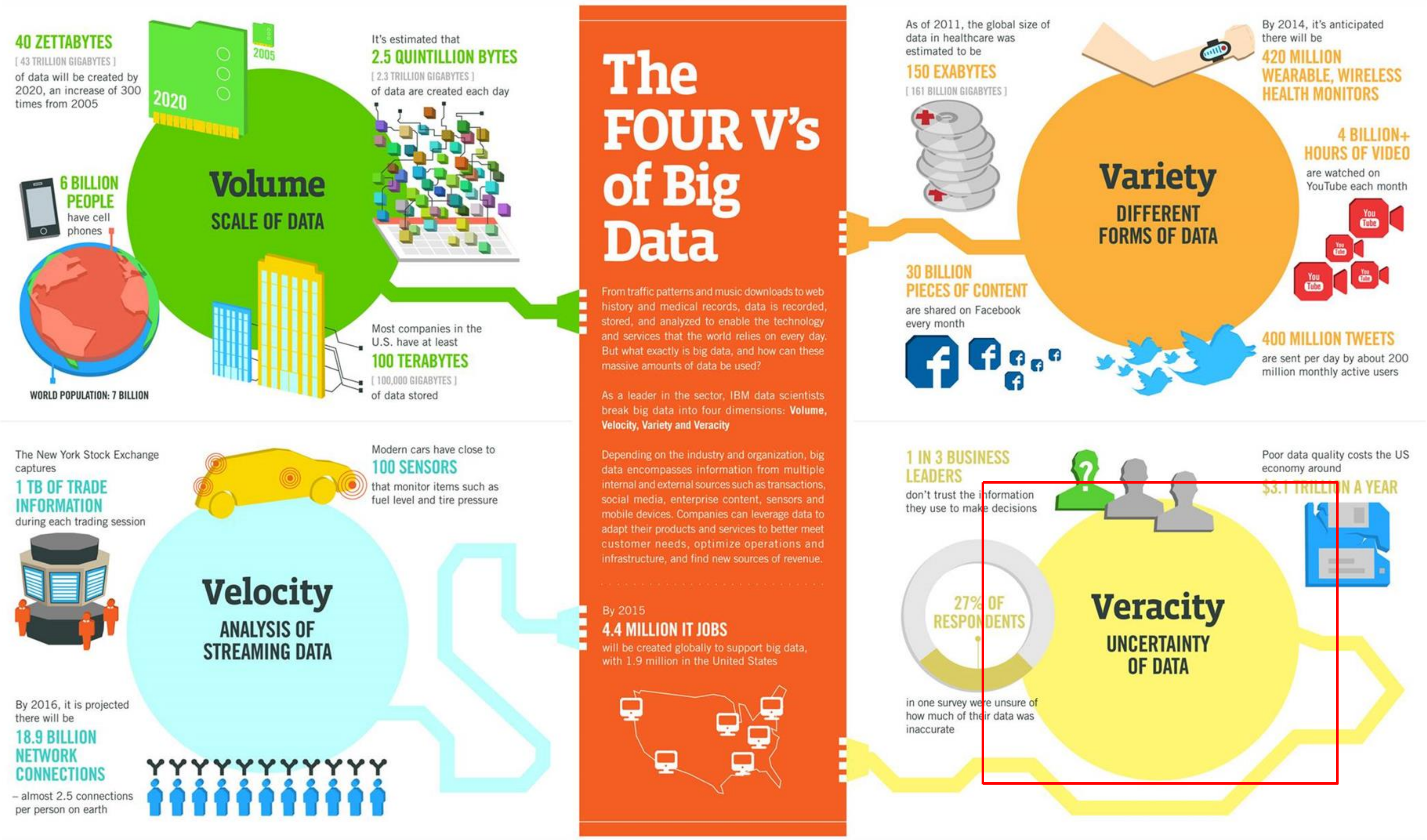
Data Cleaning: Challenges and Novel Solutions

Vinden Wylde, Edmond Prakash, Chaminda Hewage and Jon Platts
Email: {vwylde | eprakash |chewage | jplatts} @cardiffmet.ac.uk

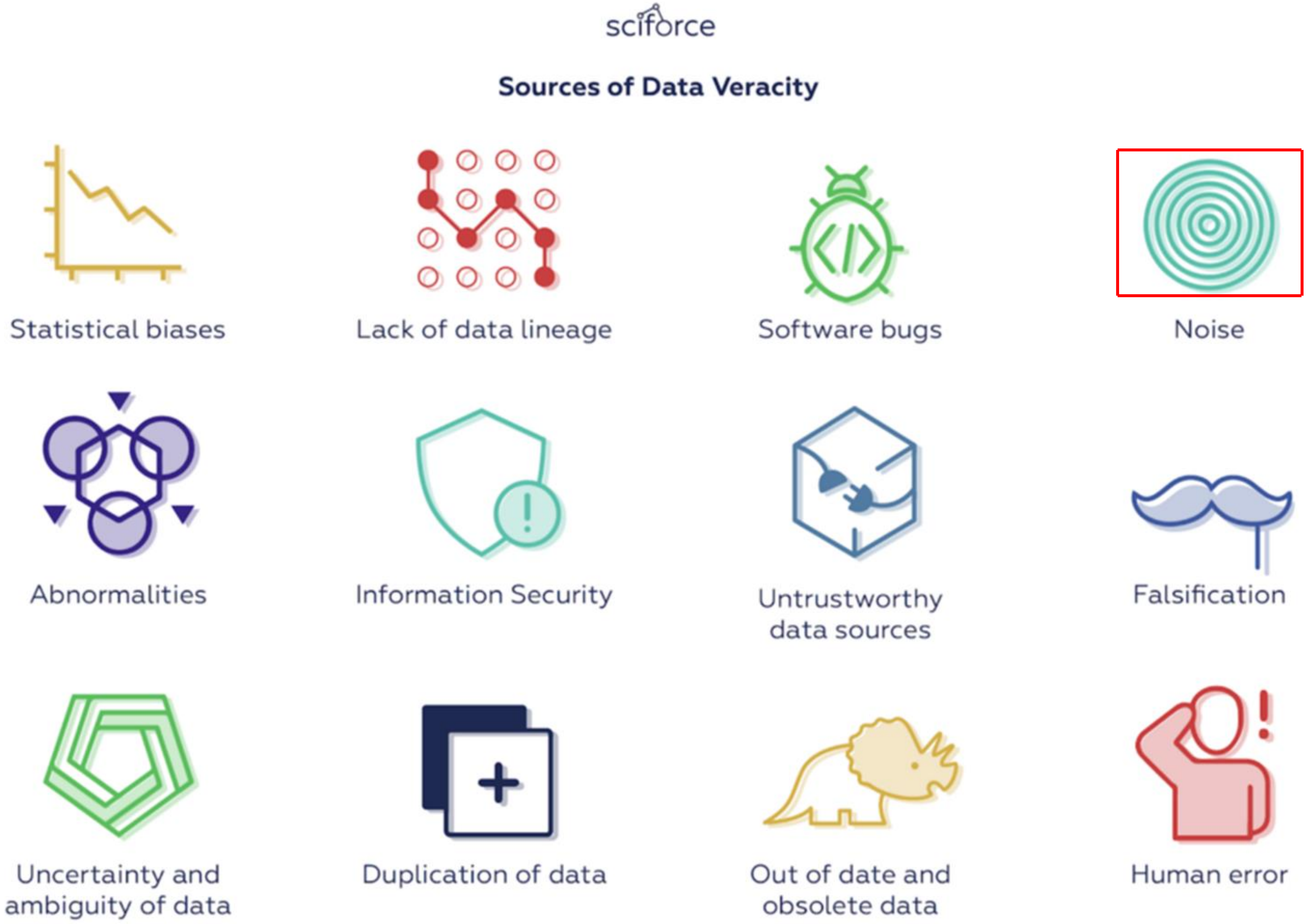
With the growth and expansion of the internet, and that the bulk of data in existence has only recently been produced, the need to define meaning and to decipher valuable truths and insights from this data plays a key role in seeking business advantage. This effort has produced a vast array of Information Technology solutions to include the use of Artificial Intelligence in creating complex mathematical frameworks and models to predict various outcomes.

However, as the volume, veracity, variety and velocity of data increases over time with the aforementioned internet growth, Data specialists such as Data Engineers and Scientists apply more and more resources to cleaning and preparing raw data prior to processing thus finding meaning from data. This presents challenges in ensuring consistency for accurate and robust results within reasonable time constraints.

By
Vinden J. Wylde
Big Data Analytics and Visualisation researcher



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS


sciforce
Sources of Data Veracity
Statistical biases
Lack of data lineage
Software bugs
Noise
Abnormalities
Information Security
Untrustworthy data sources
Falsification
Uncertainty and ambiguity of data
Duplication of data
Out of date and obsolete data
Human error

Research Purpose

After the initial proposal and PhD programme admittance in January, 2019, research gaps identified required a broad and in depth technical knowledge of algorithms and applications to include software packages and tools to gain a broad grasp of the tool-base to fully appreciate.

- For example in the recent work of Tajer et al. [3], when a country or power company needs to estimate the state of its power grid security, non-linear state recovery techniques are utilised to carry out tasks designed to assess factors such as; informing user controls, updating pricing policies, identifying structural abnormalities and predicting loads. Detecting these instances of bad data either random (sensor failures), or structured (cyber-attacks [false data injection attacks]), whilst being able to successfully recover the state of a power system, fundamental challenges remain which have been documented and formalised since the 1970's. These two operations (detecting structured and random data) utilise state estimators (recovering phase angles and bus voltages) such as algorithms that leverage data collection using multiple measurement units across the grid, to include topological and dynamic information. Identifying bad data and deciding what protocols to employ when managing distortion in the data, fundamental performance limits are presented and become unknown which limits an effective recovery from a cyber-attack or a systems infrastructure failure.
- Additionally, the network tools company CISCO for example, have included traffic management technologies to their recent network devices in response to exponential threats. This well-established technique is also facing significant challenges. Traditionally, it identifies the origin of network traffic in relation to its port number (80: HTTP), however most applications use dynamic ports therefore the pay-load-based technique is mainly adopted by business today to navigate through the traffic. Deep Packet Inspection (DPI) identifies very specific patterns contained within a payload of IP packets, however, issues such as dealing with encrypted payloads and privacy remain as a result of DPI. Other techniques such as statistical classification, which extracts sets of statistics from live traffic, utilise Machine Learning (ML) for application identification.

Research Method

Possible direction is to assess methods and tools to detect and clean data specifically for effective decision making across multiple domains.

Therefore it is proposed that creating more effective processing capabilities via providing data cleaning architectures and solutions in the form of a general and scalable framework of optimised solutions that can be intelligently deployed in parallel and proportionately to individual instances of bad data within data transmission architectures, as there is adequate opportunity for the creation of bad data (missing data, wrong information, inappropriate data [wrong column headings], duplicate data) which clearly identifies and demonstrates the need for further study with additional innovative outcomes.

Objectives

Key focus is bad data and methods to clean data

- Carry out preliminary research survey
- Analyse knowledge gaps that have been identified
- Experiments to validate and compare algorithms within these different domains.
- Focus on process methodologies and resources to monitor and analyse data quality methodologies for preventing and/or detecting and repairing dirty data.
- Produce Data Cleaning Framework
- Organise a sequence of data cleaning activities
- Minimise exceptions

Example ‘Dirty’ Data Set

color	director_name	duration	gross	movie_title	anguage	country	budget	title_year	imdb_score
Color	Martin Scorsese	240	116866727	The Wolf of Wall StreetÂ	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3Â	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful EightÂ	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	MargaretÂ	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of SmaugÂ	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of JusticeÂ	English	USA	250000000	202	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected JourneyÂ	English	USA	180000000	2012	7.9
Color	Edward Hall	180		RestlessÂ	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud AtlasA	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon TattooÂ	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of GodÂ	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Battle of the Five ArmiesÂ	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6

(Medium, 2019a)

Processing

```python
# -*- coding: utf-8 -*-
"""
Created on Tue Dec  3 22:21:50 2019

@author: Vinden Wylde
"""


# Import python libraries
import numpy as np
import pandas as pd


# Import data
dataset = pd.read_csv('movie_sample_dataset.csv', encoding='utf-8')

# Drop useless attributes
dataset.drop(['color','language'], axis=1, inplace=True)

# Handle text attributes
dataset['director_name'].fillna('', inplace=True)

# Handle numeric attributes
dataset['gross'].fillna(0, inplace=True)
# dataset['gross']=pd.to_numeric(dataset['gross']).astype('float64')
dataset['budget'].fillna(0, inplace=True)

# Unify countries names
dataset['country']=dataset['country'].str.upper()
dataset['country'] = np.where(dataset['country']=='UNITED STATES','USA', dataset['country'])

# Bad data entry
dataset['director_name'] = np.where(dataset['director_name']=='N/A','', dataset['director_name'])
dataset['director_name'] = np.where(dataset['director_name']=='Nan','', dataset['director_name'])
dataset['director_name'] = np.where(dataset['director_name']=='Null','', dataset['director_name'])
dataset['movie_title'] = dataset['movie_title'].str.replace('Â', '')

# Handling outliers
dataset["gross"]=dataset["gross"].astype(float)
dataset["duration"]=dataset["duration"].astype(float)
dataset["budget"]=dataset["budget"].astype(float)

dataset['duration'] = np.where(dataset['duration']<=10,0, dataset['duration'])
dataset['duration'] = np.where(dataset['duration']>300,0, dataset['duration'])
dataset['imdb_score'] = np.where(dataset['imdb_score']<=0,0, dataset['imdb_score'])

# Normalize data

# spliting actors
actor_list = dataset["actors"].str.split(",", n = 2, expand = True)
dataset["actor1"]= actor_list[0]
dataset["actor2"]= actor_list[1]
dataset["actor3"]= actor_list[2]
dataset.drop(columns=['actors'], inplace=True)

# Adding new feature

# Add a new metric GOB(Gross over Budget)
dataset['GOB'] = dataset.apply(lambda row: row['gross']/row['budget'] if row['budget']!=0 else 0, axis=1)
top_GOB=dataset.sort_values('GOB',ascending=False).head(15)

# dataset['title_year'] = dataset['title_year'].apply(np.int64)
# dataset['duration'] = dataset['duration'].apply(np.int64)

dataset.to_csv('output_IMDB.csv')
```

‘Clean’ Data Set

director_name	duration	gross	genres	actor1
Martin Scorsese	240	116866727	Biography|Comedy|Crime|Drama	Leonardo DiCaprio
Shane Black	195	408992272	Action|Adventure|Sci-Fi	Robert Downey Jr.
Quentin Tarantino	187	54116191	Crime|Drama|Mystery|Thriller|Western	Craig Stark
Kenneth Lonergan	186	46495	Drama	Matt Damon
Peter Jackson	186	258355354	Adventure|Fantasy	Aidan Turner
	183	330249062	Action|Adventure|Sci-Fi	Henry Cavill
Peter Jackson	0	303001229	Adventure|Fantasy	Aidan Turner
Edward Hall	180		Drama|Romance	Rufus Sewell
Tom Tykwer	172	27098580	Drama|Sci-Fi	Tom Hanks
	158	102515793	Crime|Drama|Mystery|Thriller	Robin Wright
Christopher Spencer	170	59696176		Roma Downey
Christopher Nolan	169	187991439	Adventure|Drama|Sci-Fi	Matthew McConaughey
F. Gary Gray	167	161029270	Biography|Crime|Drama|History|Music	Aldis Hodge
Richard Linklater	165	25359200	Drama	Ellar Coltrane
Quentin Tarantino	0	162804648	Drama|Western	Leonardo DiCaprio
Michael Bay	165	245428137	Action|Adventure|Sci-Fi	Bingbing Li
Christopher Nolan	164	448130642	Action|Thriller	Tom Hardy

color	director_name	duration	gross	movie_title	anguag	country	budget	title_year	imdb_score
Color	Martin Scorsese	240	116866727	The Wolf of Wall StreetÂ	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3Â	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful EightÂ	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	MargaretÂ	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of SmaugÂ	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of JusticeÂ	English	USA	250000000	202	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected JourneyÂ	English	USA	180000000	2012	7.9
Color	Edward Hall	180		RestlessÂ	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud AtlasÂ	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon TattooÂ	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of GodÂ	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Battle of the Five ArmiesÂ	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6
Color	Kathryn Bigelow	157	95720716	Zero Dark ThirtyÂ	English	USA	40000000	2012	7.4
Color	Ridley Scott	156	105219735	Robin HoodÂ	English	USA	200000000	2010	6.7
Color		156	183635922	The RevenantÂ	English	USA	135000000	2015	8.1
Color	Michael Bay	154	352358779	Transformers: Dark of the MoonÂ	English	USA	195000000	2011	6.3
Color	Denis Villeneuve	153	60962878	PrisonersÂ	English	USA	46000000	2013	8.1

(Medium, 2019)

Future Challenges

Alongside developing a fundamental and more precise understanding of algorithm concepts, components and deployment, inherent challenges exist that further justify the need for research into data cleaning as detecting and repairing dirty data.

- In recent times, the continual surge of interest from industry and academia on data cleaning problems and solutions has provided new abstractions, approaches for scalability, interfaces and statistical techniques. To thoroughly understand these new advances, a taxonomy of the data cleaning literature will be produced and examined to highlight issues such as constraints, rules and patterns to detect quantitative errors.
- State-of-the-art techniques also highlight their limitations, whilst traditionally such approaches are distinct from quantitative approaches such as outlier detection, recent work that casts such approaches into a statistical estimation framework including: using Machine Learning to improve the efficiency and accuracy of data cleaning and considering the effects of data cleaning on statistical analysis.
- The methods and applications involved in data cleaning are vast, it is with hope that the proposal and ongoing project can indeed generate original work and with innovative ideas and solutions.

References

1. Christy, A., Gandhi, G. and Vaithyasubramanian, S. (2015). Cluster Based Outlier Detection Algorithm for Healthcare Data. Procedia Computer Science, 50, pp.209-215.

2. Mishra, B., Rath, A., Nanda, S. and Baidyanath, R. (2019). Efficient Intelligent Framework for Selection of Initial Cluster Centers. International Journal of Intelligent Systems and Applications, 11(8), pp.44-55.

3. Tajer, A., Sihag, S. and Alnajjar, K. (2019). Non-linear state recovery in power system under bad data and cyber-attacks. Journal of Modern Power Systems and Clean Energy, 7(5), pp.1071-1080.

4. Wang, X. and Wang, C. (2020). Time Series Data Cleaning: A Survey. IEEE Access, 8, pp.1866-1881.

5. Wang, B., Zhang, J., Zhang, Z., Pan, L., Xiang, Y. and Xia, D. (2019). Noise-Resistant Statistical Traffic Classification. IEEE Transactions on Big Data, 5(4), pp.454-466.

6. ScienceDIrect (2019) *The cost, coverage and rollout implications of 5G infrastructure in Britain* [Online] Available at: https://www.sciencedirect.com/science/article/pii/S0308596117302781 [Accessed 01/12/2019]