INNOTECH
INNOVATION | TECHNOLOGY | EVENTS

2019 Conference
November 19, 2019
Austin Convention Center

Data Cleansing Strategies for Your BigDataOps Team

Dr. Chuck Wiley
GM
Sr Application Architect and Dev Lead
Vehicle Data, Analytics and Decisioning

Raghu George
GM
Senior Software Engineering Manager
Big Data - Vehicle Data Factory
Twitter: @raghu_george

Dr. Yuriy Pakhotin
GM
Senior Big Data Engineer / ML Scientist
Big Data – Vehicle Data Factory

Agenda

- How Clean is Your Data? – Raghu George
- Data Cleansing Strategies – Dr. Chuck Wiley
- Demo – Dr. Yuriy Pakhotin
- Q&A

How Clean is Your Data?

Raghu George
GM
Senior Software Engineering Manager
Big Data - Vehicle Data Factory
Twitter: @raghu_george

Need Clean Data not Just Data

Mr. C: Really? You do not need data?

Mr. DS: Yes, I do not need data.

Mr. C : **Aren't you working on building an AI Model?**

Mr. DS: Yes, I do not need data. I need features.

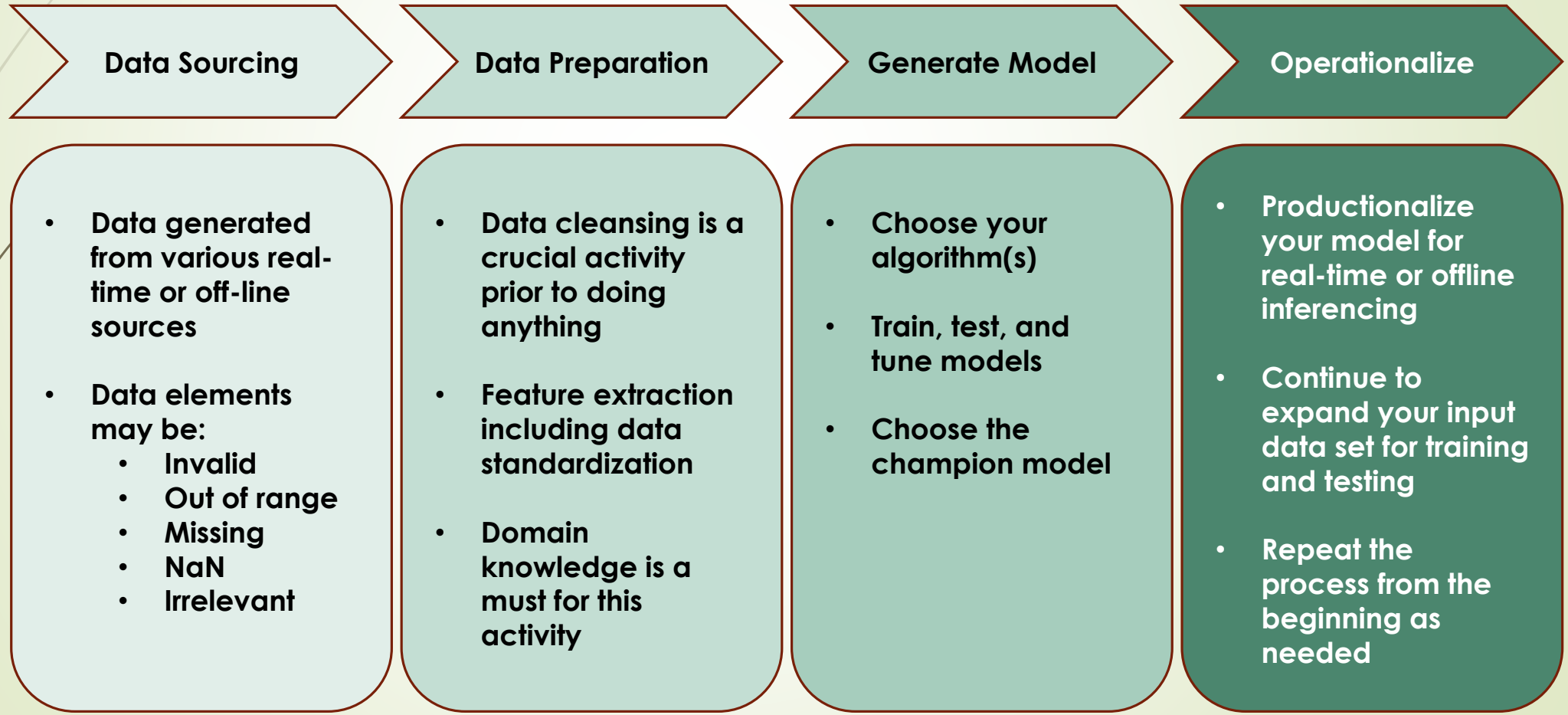
Mr. C : **Well, how you do get features without data?**

Mr. DS: I do not need just data. I need clean data so I can extract the features for a good AI model.

Mr. C : **OK!**

Image source: https://www.clipartwiki.com/iclipmax/ihhJmx/

Data to Model to Operations



Overloaded Definition of Data Cleansing

Data cleansing or **data cleaning**

- is the process of detecting and *correcting* (or *removing*) corrupt or inaccurate records from a record set, table, or database

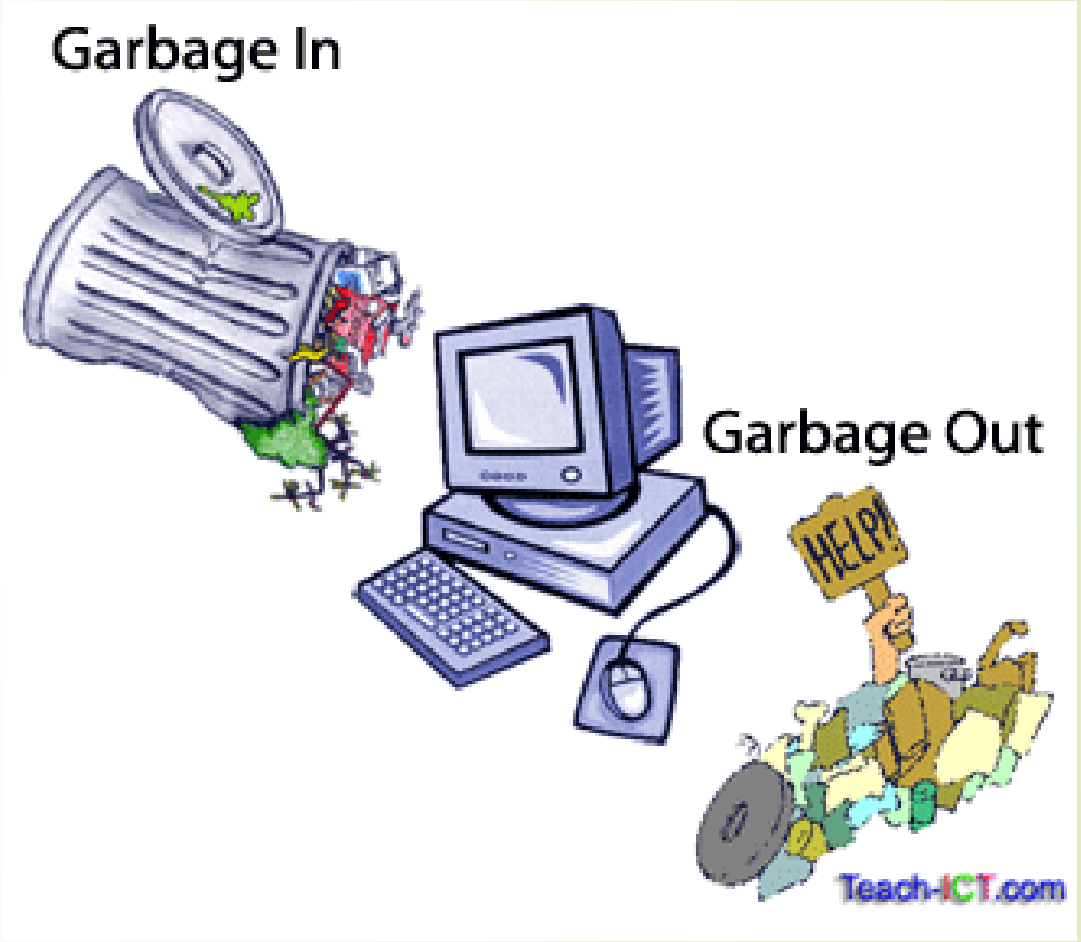
and

- refers to identifying *incomplete*, *incorrect*, *inaccurate* or *irrelevant* parts of the data and then *replacing*, *modifying*, or *deleting* the dirty or coarse data.

Quoted from: https://en.wikipedia.org/wiki/Data_cleansing

Impact of Poor Quality Data to Your AI Model

- When it comes to Machine Learning, poor quality data is enemy number one.[1]
- Poor quality leads to "Garbage In Garbage Out" scenario. [1]
- Poor quality can be in the historical data used in training the predictive model and in the new data used for future decisions. [1]
- To fully exploit AI and its promises of truthful and correct predictions and advice, you need data with the right quality. [2]



References:
1.https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless
2.https://www.capgemini.com/2017/10/quality-data-a-must-have-for-ai/#

Data Cleansing Strategies

Dr. Chuck Wiley
GM
Sr Application Architect and Dev Lead
Vehicle Data, Analytics and Decisioning

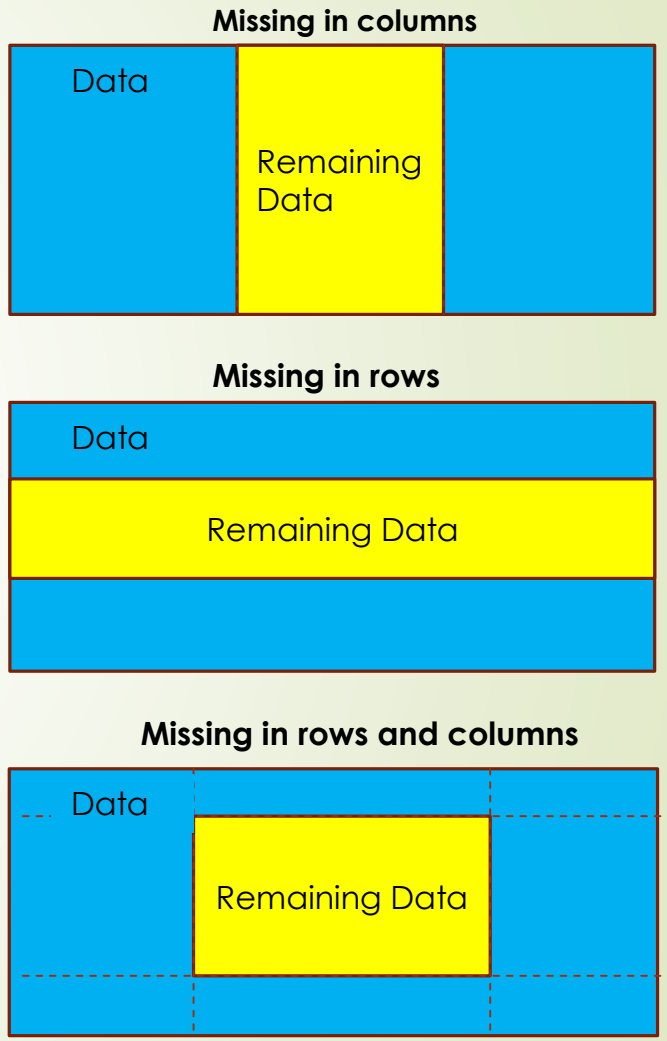
Data cleansing Strategies

- Better data will beat fancier algorithms (Garbage in Garbage out)
- Fix structural errors
  - e.g. Typos, Inconsistent capitalization
- Remove unwanted values
  - e.g. Duplicates, Irrelevant observations
- Filter outliers
- Most AI algorithms do not allow missing values
- Handle incomplete/missing data
  - Do nothing
  - Remove rows and columns with missing values
  - Impute or fill in values

Image source: https://www.slideshare.net/databricks/the-key-to-machine-learning-is-prepping-the-right-data-with-jean-georges-perrin

Handling Incomplete Data

- Methods to handle missing data
  - Do nothing
  - Remove rows and columns with missing values
  - Impute or fill in values
- Removing data can result in significant reduction of your data set

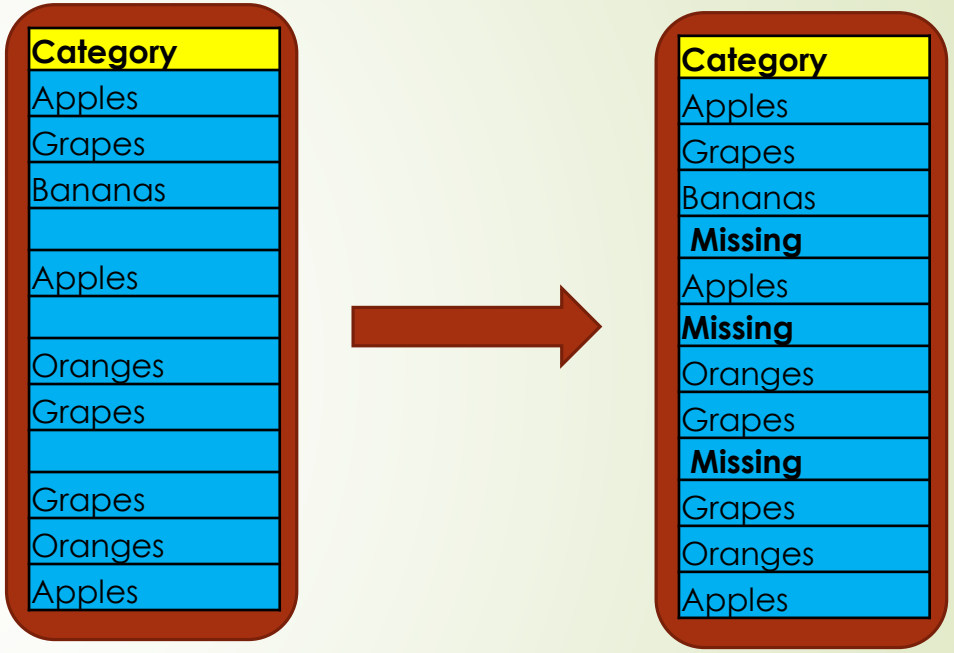


Impute Missing Values

- Many methods for filling missing values
  - Fill with an arbitrary value
  - Fill with the most frequent value (works with category and numeric features)
  - Fill with mean or average
  - K nearest neighbors – rows most similar
- "Missingness" itself is information
- No matter how complex your method of imputing missing entries, you are not adding information

Can you tell your algorithm data is missing?

- "Missingness" itself is information
- Category/enumeration feature: Just add an additional category such as "Missing"



Can you tell your algorithm data is missing?

- Numeric feature: flag and fill
  - Add an additional column with presence/absence of data
    - Doesn't have to be a Boolean indicator (Present, Missing, Suspected Outlier)
  - Fill with a constant (0) to meet requirement of having a value
- Allow your algorithm to determine the best constant for missing

number
1
3.1415
Nan
2.71
2
Inf
-Inf
6
45
1

fill

number
1
3.1415
0
2.71
0
2
0
0
6
45
1

boolean

category
TRUE
TRUE
FALSE
TRUE
FALSE
TRUE
FALSE
FALSE
TRUE
TRUE
TRUE

category
valid
valid
not a num
valid
missing
valid
infinity
-infinity
valid
valid
valid

category
valid
Valid
missing
valid
missing
valid
missing
missing
valid
valid
valid

annual income	reason missing?
10000	??
120000	valid
55000	valid
0	missing
0	missing
115000	valid
36000	valid
0	missing
110000	valid

Demo

Dr. Yuriy Pakhotin
GM
Senior Big Data Engineer / ML Scientist
Big Data – Vehicle Data Factory

https://github.com/pakhotin/Data-Cleansing-Innotech

Q & A