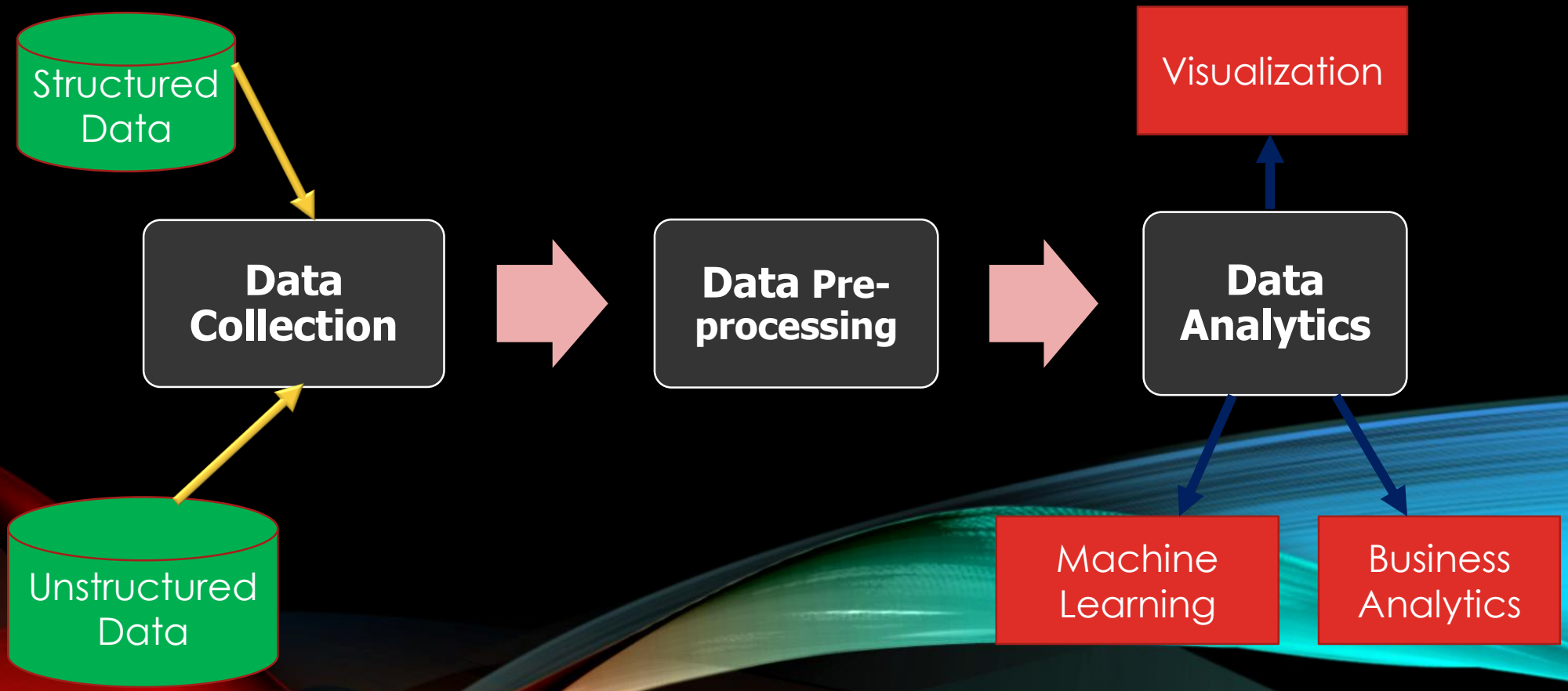
DATA PRE-PROCESSING

Mukund G Kallapur

AGENDA FOR TODAY'S LECTURE

- Data Flow
- Types of Data
- Why data pre-processing ?
- What is data pre-processing ?
- Steps in data pre-processing
- Final Words

DATA FLOW


Structured Data
Unstructured Data
Data Collection
Data Pre-processing
Data Analytics
Visualization
Machine Learning
Business Analytics

TYPES OF DATA

Structured data

- Government sources.
- Banks and financial institutions.
- Hospitals, diagnostic centres.
- Private and public industries.
- Schools, colleges and universities.
- Many more !!!

Unstructured data

- Audio and Video streaming applications and websites
- Images generated from various sources such as cameras, CCTVs, mobile phones.
- Text data generated from mobile phones.

WHY DATA PRE-PROCESSING?

- Data Quality
  - Data is received from multiple sources and in multiple formats.
  - Ensure that the data is consistent.
  - Uniformity of data.
- Outliers
  - Data might contain certain records which do not fit. Treat them.

DATA PRE-PROCESSING

- First and foremost step for any data analytics or machine learning activity.
- It is the process in which the raw data is cleaned to transform it into a more meaningful and insightful mine of information which can be seamlessly used by Machine Learning models.
- Approximately, 70-80% of the time during entire Machine Learning life cycle is (should be) spent on Data pre-processing.

STEPS IN DATA PRE-PROCESSING

- Data Cleaning
- Feature engineering/Dimensionality reduction
- Standardization

DATA CLEANING

- Missing Values
- Outlier detection and treatment
- Uniformity of categorical values
- Date conversions

FEATURE ENGINEERING

- Merge two columns into a single column.
- Dropping of columns.
- Dimensionality reduction.

STANDARDIZATION

- Standardize or normalize the data to bring them to same units.

- Different ways to standardize.
  - Min-Max normalization
  - Z-Score standardization

FINAL WORDS

- Data pre-processing – An important phase in the data analytics pedagogy.
- Different phases in data pre-processing.
- Should be done on a case by case basis and there is no single solution.

Questions ?

Thank You