Data Collection (Part-I)

BY:

DR. VIPIN KUMAR

DEPARTMENT OF COMPUTER SCIENCE &IT

MAHATMA GANDHI CENTRAL UNIVERSITY

MOTIHARI, BIHAR

Outline…

- *Motivation of Data Collection*
- *Defining Data Collection*
- *Flowchart of Data Collection Process*
- *Types of Data Collection:*
- *Data Labeling*
- *Existing Model*

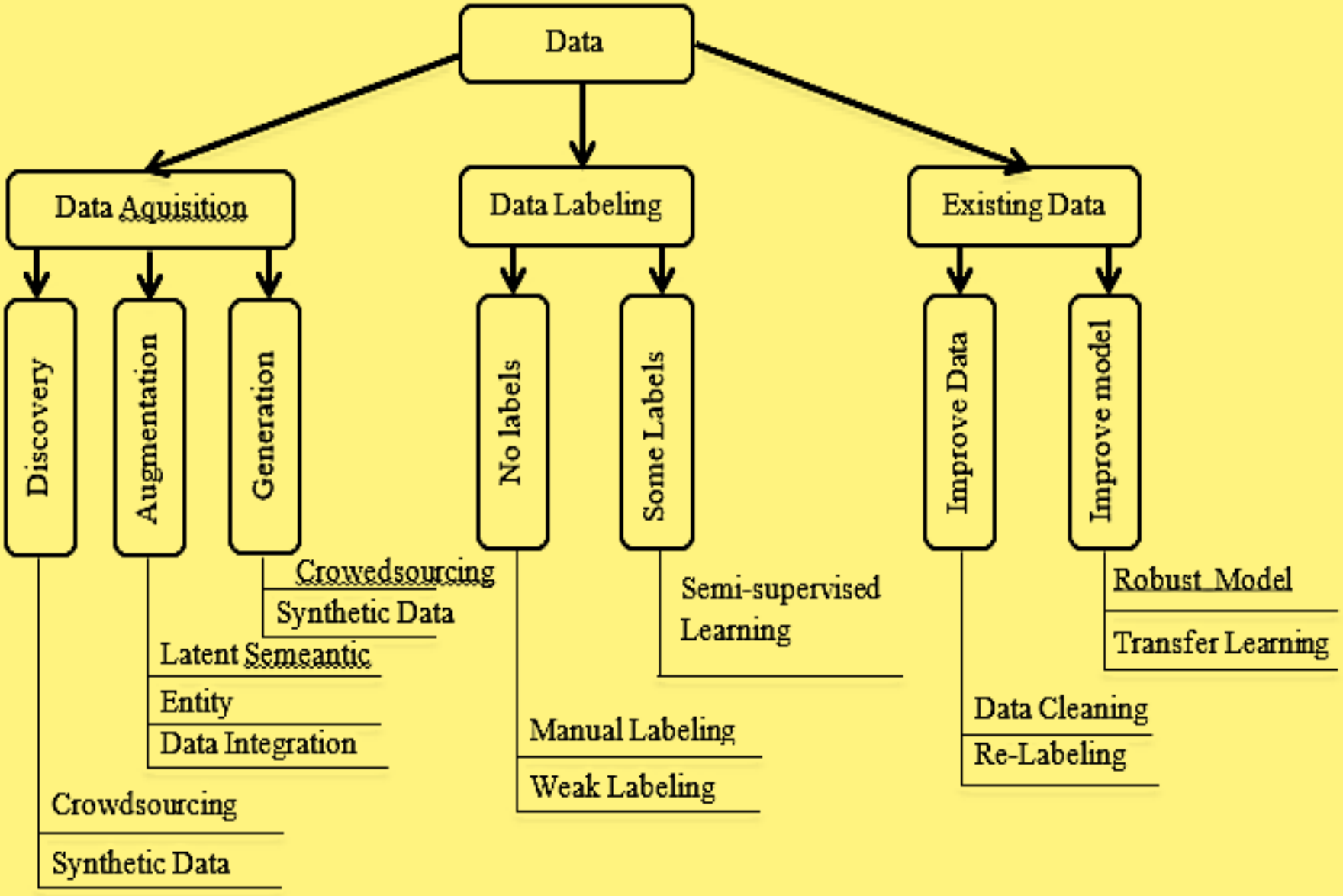
Motivation

- DATA COLLECTION IS ONE OF THE CRITICAL CHALLENGES IN DATA SCIENCE.
- ORGANIZATIONS ARE SPENDING ALMOST 80% OF THEIR WORK FOR COLLECTION OF DATA WHICH INCLUDES THE FOLLOWING TASKS:
  - Collection of raw data from multiple sources
  - Integration of data
  - Cleaning of data
  - Analyzing of data
  - Noise and inconsistency removal
  - Feature engineering of the data
  - Visualizing the data

Data Collection

- DATA COLLECTION IS IMPORTANT STEP IN DATA SCIENCE WHICH COLLECTS THE RAW DATA AS PER PROJECT CHARTER.
- THERE ARE THREE WAY TO COLLECT THE DATA FROM THE INTERNAL SOURCES (ORGANIZATION) AND EXTERNAL SOURCES.
    1. Data acquisition
    2. Data labeling
    3. Using existing model

Flowchart of Data Collection



1. Data Acquisition:

- BASICALLY, THIS APPROACH UTILIZED TO FIND THE SUITABLE DATA FOR TRAINING THE MACHINE LEARNING ALGORITHMS.
- DATA ACQUISITION APPROACH HAS BEEN CATEGORIES IN THREE:
  - 1.1 Data discovery
  - 1.2 Data augmentation
  - 1.3 Data generation

1. Data Acquisition:

Task	Approach	Techniques
Data Discovery	Sharing	• Collaboration Analysis • Collaborative and Web
	Searching	• Web • Data Lake
Data Augmentation		• Entity latent Semantics • Deriving Latent Semantic • Data Integration
Data Generation	Crowdsourcing	• Gathering • Processing
	Synthetic Data	• Generative Adversarial Network • Policies • Image • Text

1.1 Data Discovery:

- SHARE THE GENERATED INDEXED DATA IN PUBLISH.
- META-DATA ARE CREATED WHILE GENERATING THE DATA IN POST-HOC APPROACH.
- SCALE THE SEARCHING IN DATABASE IS THE KEY CHALLENGE.
- OBTAINING THE SUITABILITY OF DATASET AFTER SEARCH IS ANOTHER CHALLENGE
- **1.1.1. DATA SHARING:**
  - **It has two approaches:**
  - **Collaborative Analysis:**
    - **Different version of dataset has to collaboratively analysised.**
    - **DataHub can be used for collaborative analysis. It is utilized to run the own version of machine learning task as team or individuals and later integrated/merge if required.**
    - **Git has one of the version of dataset apart from hosted platform.**

1.1 Data Discovery:

- **Web:**
  - Data has publish on the web.
  - Google Fusion Tables is utilized for data management and integration which includes the uploading the data that can be integrate, filter and visualize. Data can be access through web search for showing the results.
  - For finding the public data, buy and sell the data, there many market place like: Datamarket, Quandl, CKAN etc.
- **Collaborative and Web:**
  - This approach is recently immerge.
  - Kaggle provides collaborative and web-based systems. It provides shares the dataset on Web and host the competitions.

1.1 Data Discovery:

- **Web:**
  - Data has publish on the web.
  - Google Fusion Tables is utilized for data management and integration which includes the uploading the data that can be integrate, filter and visualize. Data can be access through web search for showing the results.
  - For finding the public data, buy and sell the data, there many market place like: Datamarket, Quandl, CKAN etc.
- **Collaborative and Web:**
  - This approach is recently immerge.
  - Kaggle provides collaborative and web-based systems. It provides shares the dataset on Web and host the competitions.

1.1 Data Discovery:

- 1.1.2. DATA SEARCHING:
  - It is design for the searching data by exploring the systems.
  - DATA LAKE: It is a popular searching system in corporate world.
  - It provide the environment where teams or individual are not vesting their time by re-searching or regeneration of data for machine learning task.
  - Data curation and searching method has been created by IBM earlier because data has scattered among many application.
  - It has observed by IBM that 70% time of project analyst is used for data searching, integrating the data and discovery of data.
  - IBM preform filling, creating, maintaining and governing of the data, called data wrangling.

1.1 Data Discovery:

- TECHNIQUES FOR DATA SEARCHING:
  - Google Data Search (GOODS) is a system which maintains the 10 billions meta-data of datasets.
  - GOODS provides the simple keyword queries only.
  - DATA CIVILIZER systems provides the additional power to GOODS for keyword search which makes data link graph.
  - DATARAMAN is utilized semi-structured data to obtain the structured data automatically.
  - AURUN: semantic-linked data is discovered through queries.

1.2 Data Augmentation:

- Data augmentation is a process to enrich the information within the dataset.
- There are many methods for data augmentation:
  - Deriving Latent Semantics
  - Entity Augmentation
  - Data Integration

1.2 Data Augmentation:

- **DERIVING LATENT SEMANTICS:**
  - It fills the missing information like entire features or values.
  - InfoGather is a method to search information using Web tables.
- **ENTITY AUGMENTATION:**
  - Latent semantics is derived from the give data.
  - Word2vec method is to convert text corpus to real number vector that captures the linguistic context.
- **DATA INTEGRATION:**
  - This process extend the data set by integrating acquired dataset to the existing one.
  - Hamlet and Hamlet++ are the methods utilized Key-foreign key (KFK) joins. It reduces the total time of user significantly.

1.3 Data Generation:

- If there are no data available for the training of model to perform the specific task.
- then we have two option to generate the data: manually and automatically.
- There are two approach to generate data:
  - Crowdsourcing
  - Synthetic Data Generation

1.3 Data Generation:

- CROWDSOURCING:
  - This approach utilizes the human as workers to complete the task of gata generation.
  - Amazon Mechanical Turk is a popular platform where humans are compensated for their work.
  - Crowdsourcing can be divided into two categories:
    - Data gathering: it has utilized one of the task from procedural or declarative.
    - Data Preprocessing: This approach is used for data preprocessing after data gathering approach. Some relevant preprocessing task has been identified primarily like joining the data, entity resolution and data curation.

1.3 Data Generation:

- SYNTHETIC DATA GENERATION:
  - Synthetic data generation provides the flexible and low cast data to perform machine learning tasks.
  - It utilizes probability distribution to generate samples with labels (if required)
  - There are many methods to generate data synthetically:
    - Generative Adversarial Networks (GAN)
    - MEDGAN
    - TABLE-GAN
    - Synthetically Controlled Paraphrase Networks (SCPNs)
    - Semantically Equivalent Adversarial Rules (SEARs)

Bibliography:

- **ROH, YUJI, GEON HEO, AND STEVEN EUIJONG WHANG. "A SURVEY ON DATA COLLECTION FOR MACHINE LEARNING: A BIG DATA-AI INTEGRATION PERSPECTIVE." *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* (2019).**
- DEEP LEARNING FOR DETECTION OF DIABETIC EYE DISEASE," HTTPS://RESEARCH.GOOGLEBLOG.COM/2016/11/DEEP-LEARNINGFOR-DETECTION-OF-DIABETIC.HTML .
- I. GOODFELLOW, Y. BENGIO, AND A. COURVILLE, DEEP LEARNING. THE MIT PRESS, 2016.
- A. Y. HALEVY, "DATA PUBLISHING AND SHARING USING FUSION TABLES," IN CIDR, 2013.
- H. GONZALEZ, A. Y. HALEVY, C. S. JENSEN, A. LANGEN, J. MADHAVAN, R. SHAPLEY, W. SHEN, AND J. GOLDBERG-KIDON, "GOOGLE FUSION TABLES: WEB-CENTERED DATA MANAGEMENT AND COLLABORATION," IN SIGMOD, 2010, PP. 1061–1066.
- M. J. CAFARELLA, A. Y. HALEVY, H. LEE, J. MADHAVAN, C. YU, D. Z. WANG, AND E. WU, "TEN YEARS OF WEBTABLES," PVLDB, VOL. 11, NO. 12, PP. 2140–2149, 2018.
- R. BAUMGARTNER, W. GATTERBAUER, AND G. GOTTLOB, "WEB DATA EXTRACTION SYSTEM," IN ENCYCLOPEDIA OF DATABASE SYSTEMS, SECOND EDITION, 2018.
- L. XU AND K. VEERAMACHANENI, "SYNTHESIZING TABULAR DATA USING GENERATIVE ADVERSARIAL NETWORKS," CORR, VOL. ABS/1811.11264, 2018.
- I. J. GOODFELLOW, "NIPS 2016 TUTORIAL: GENERATIVE ADVERSARIAL NETWORKS," CORR, VOL. ABS/1701.00160, 2017.
- E. D. CUBUK, B. ZOPH, D. MAN´E, V. VASUDEVAN, AND Q. V. LE, "AUTOAUGMENT: LEARNING AUGMENTATION POLICIES FROM DATA," CORR, VOL. ABS/1805.09501, 2018.
- J. MALLINSON, R. SENNRICH, AND M. LAPATA, "PARAPHRASING REVISITED WITH NEURAL MACHINE TRANSLATION," IN EACL. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2017, PP. 881–893.
- M. IYYER, J. WIETING, K. GIMPEL, AND L. ZETTLEMOYER, "ADVERSARIAL EXAMPLE GENERATION WITH SYNTACTICALLY CONTROLLED PARAPHRASE NETWORKS," CORR, VOL. ABS/1804.06059, 2018.

Thank You