Data Collection

Part-II

BY:

DR. VIPIN KUMAR

DEPARTMENT OF COMPUTER SCIENCE &IT

MAHATMA GANDHI CENTRAL UNIVERSITY

MOTIHARI, BIHAR

Outlines…

2. Data Labeling

- THE NEXT TASK AFTER DATA ACQUISITION IS TO LABEL THE DATA.
- THERE ARE MANY CATEGORIES OF DATA LABELING:
  1. Use Existing labeling
  2. Crowd-based
  3. Weak-labeling

2. Data Labeling

1. Use Existing labeling:
   - It exploits the existing label to label the unlabeled data e.g. semi-supervised learning
2. Crowd-based:
   - Crowdsourcing approach get utilized to label the individual samples. e.g Active learning
3. Weak-labeling:
   - It is expensive approach for labeling. Data are labeled with less than perfect labels (weak label)

2.1 Utilization of Existing Labels

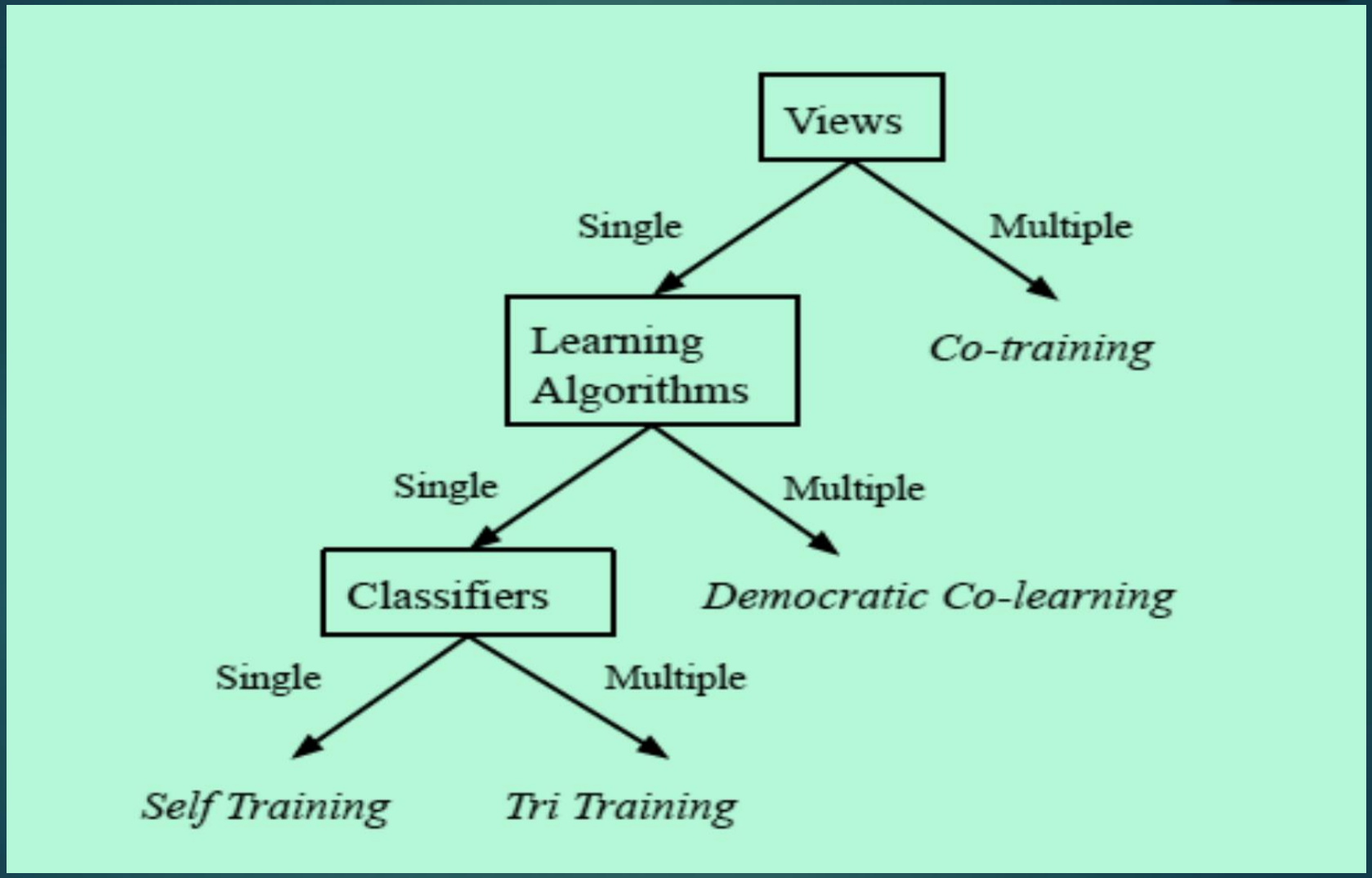
- SUPERVISED AND SEMI-SUPERVISED MACHINE LEARNING ALGORITHM CAN DIRECTLY APPLIED FOR LABELING THE DATA.
- IN SUPERVISED LEARNING:
  - Classification: to predict the categorical labels using existing labels.
  - Regression: To predict the continuous labels using existing labels.
- IN SEMI-SUPERVISED:
  - Inductive learning
  - Transductive learning

Classification of Data Labeling Techniques

Category	Approach	Machine learning task	Data types
Use Existing Labels	Self-labeled	classification	all
		regression	all
	Label propagation	classification	graph
Crowd-based	Active learning	classification	all
		regression	all
	Semi-supervised+Active learning	classification	text
			image
			graph
	Crowdsourcing	classification	all
		regression	all
Weak supervision	Data programming	classification	all
	Fact extraction	classification	text

- 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).

Classification of Semi-Supervised Learning Techniques



- 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).

2.2 Crowd-based Techniques

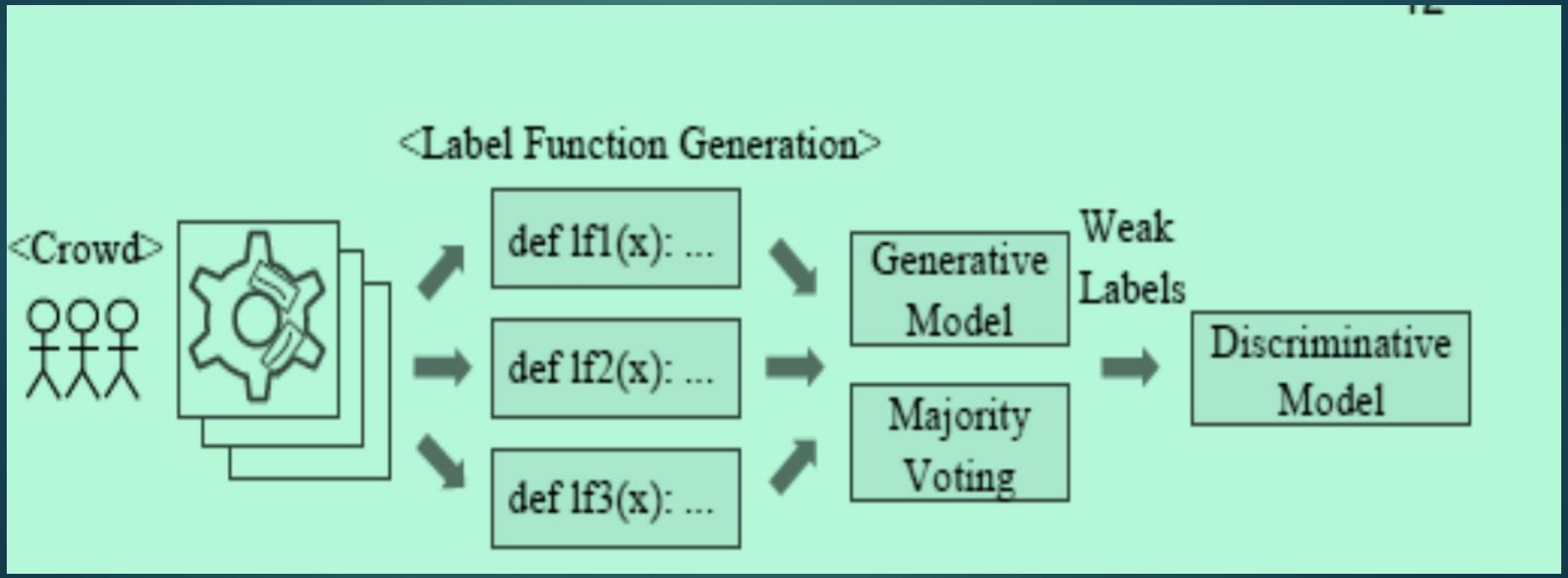
- THERE ARE TWO CATEGORIES IN CROWD-BASED TECHNIQUES:
- ACTIVE LEARNING: IT SELECTS THE MOST INTERESTING SAMPLES TO CROWD FOR LABELING. I HAS HUMAN IN LOOP FOR THE LABELING.
- CROWDSOURCING: IT UTILIZES THE HUMAN RESOURCE FOR LABELING THE DATA WHICH MAY NOT BE THE EXPERT OF THE DOMAIN.
  - Challenges: Quality Control, scalability, User interaction

2.3 Weak Supervision

- Automatically generated labels which may not be good quality of labels as crowdsourcing.
- This type of labeling is feasible for large amount of data.
- In the case of crowdsourcing, its become infeasible to label the data.
- There are two categories:
  1. Data Programing
  2. Fact extraction

2.3 Weak Supervision

- **Data Programming:**
  - It utilizes the multiple labeling function rather than one labeling function.
  - Flowchart of Data Programming:



Reference: 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).

2.3 Weak Supervision

- **Fact Extraction:**
- **Fact-extraction is utilized for labeling.**
- **It is better than manual labeling.**

Task	Techniques
Improve Data	Data Clearing
	Relabeling
Improve model	Robust Against Noise
	Transfer Learning

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