BOSTON UNIVERSITY

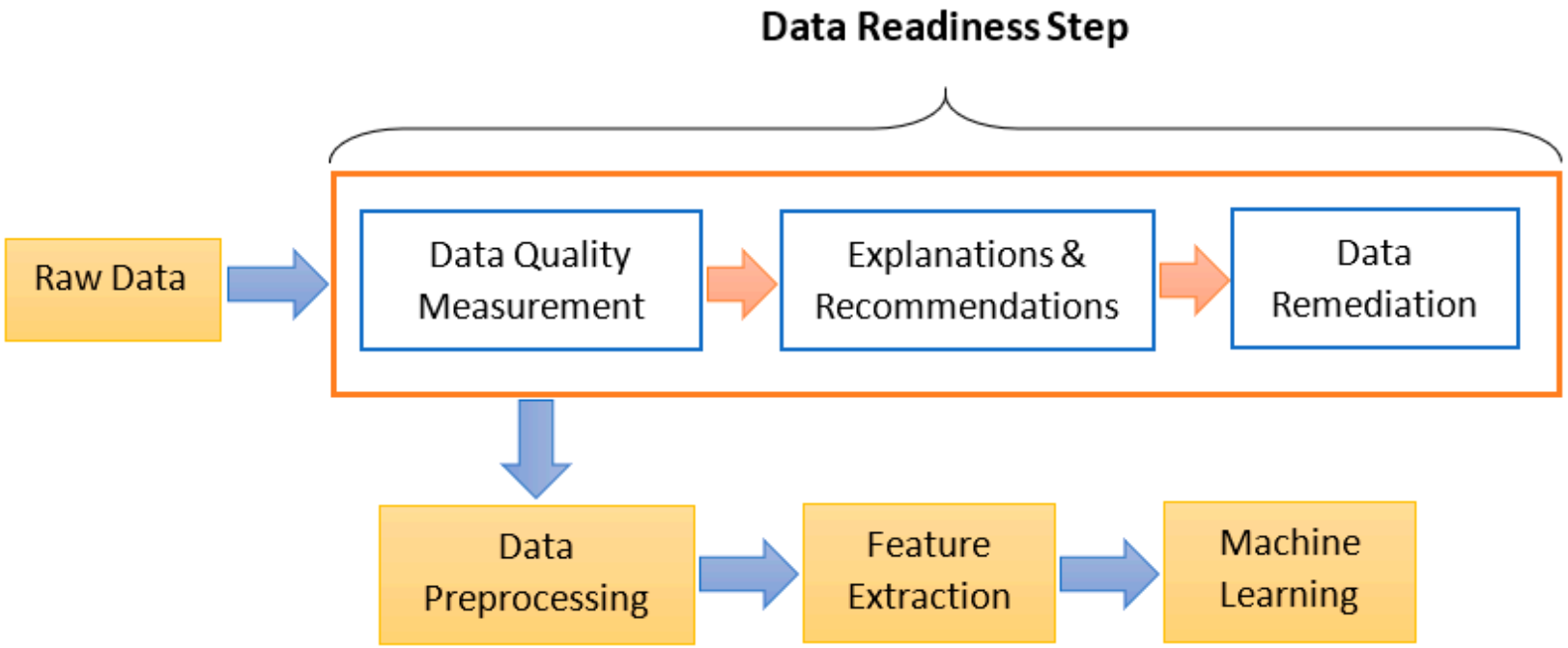
Evaluation of machine learning-readiness in Alzheimer's disease data cohorts

Vijaya B. Kolachalama, PhD

Associate Professor,
Department of Medicine, Boston University School of Medicine
Department of Computer Science, Boston University
Founding Member, Faculty of Computing & Data Sciences, Boston University
Web: http://sites.bu.edu/vkola; Twitter: @vkola_lab

Introduction

- **Machine learning:** data + model → prediction
- Quality of prediction depends on both good **models** and high-quality **data**
- Scientists spend about **3/4th** of their time in iterative pre-processing of data
- Pre-processing: **cleansing**, **validation** and **transformation**



Data quality measures

- **Class Overlap:** Overlapping regions among different classes
- **Label Purity:** Noise ratio and the number of noisy samples in the data
- **Class Parity:** Class imbalance ratio
- **Feature Relevance:** Importance of each feature with respect to the target variable (class) and other features
- **Data Homogeneity:** Transformation of data into the user's intended format
- **Data Fairness:** Identifies the bias in the dataset and returns the disparate impact score
- **Correlation Detection:** Detect correlated features
- **Data Completeness:** Detects missing values
- **Outlier Detection:** Detect and remove outliers (deviates so much from other observations) in the dataset
- **Data Duplicates:** Remove duplicate records to clean the data

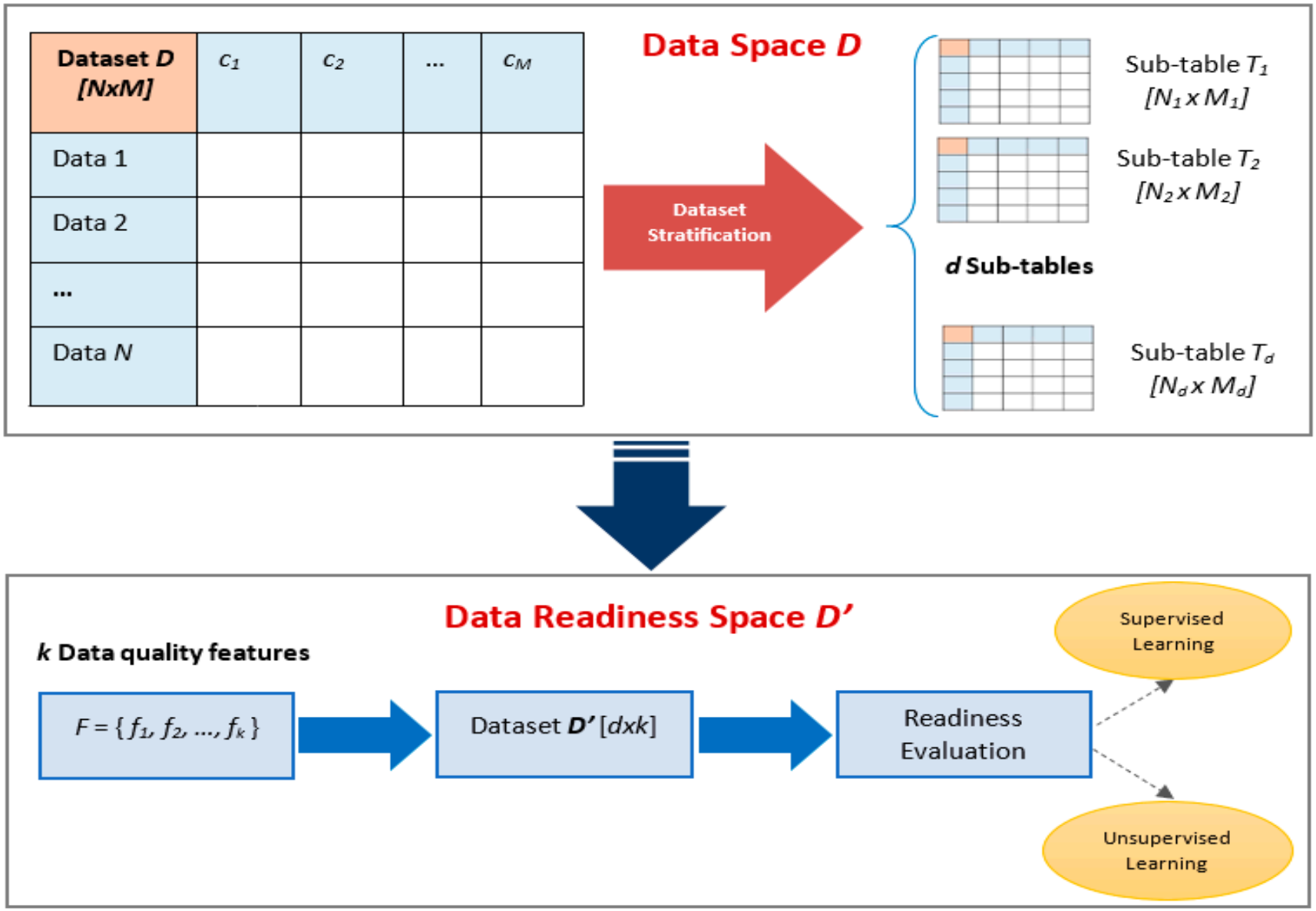
Assumptions

- **Class Overlap:**
- **Label Purity:**
- **Class Parity:**
- **Feature Relevance:**
- **Data Homogeneity:**
- **Data Fairness:**
- **Correlation Detection:**
- **Data Completeness:**
- **Outlier Detection:**
- **Data Duplicates:**

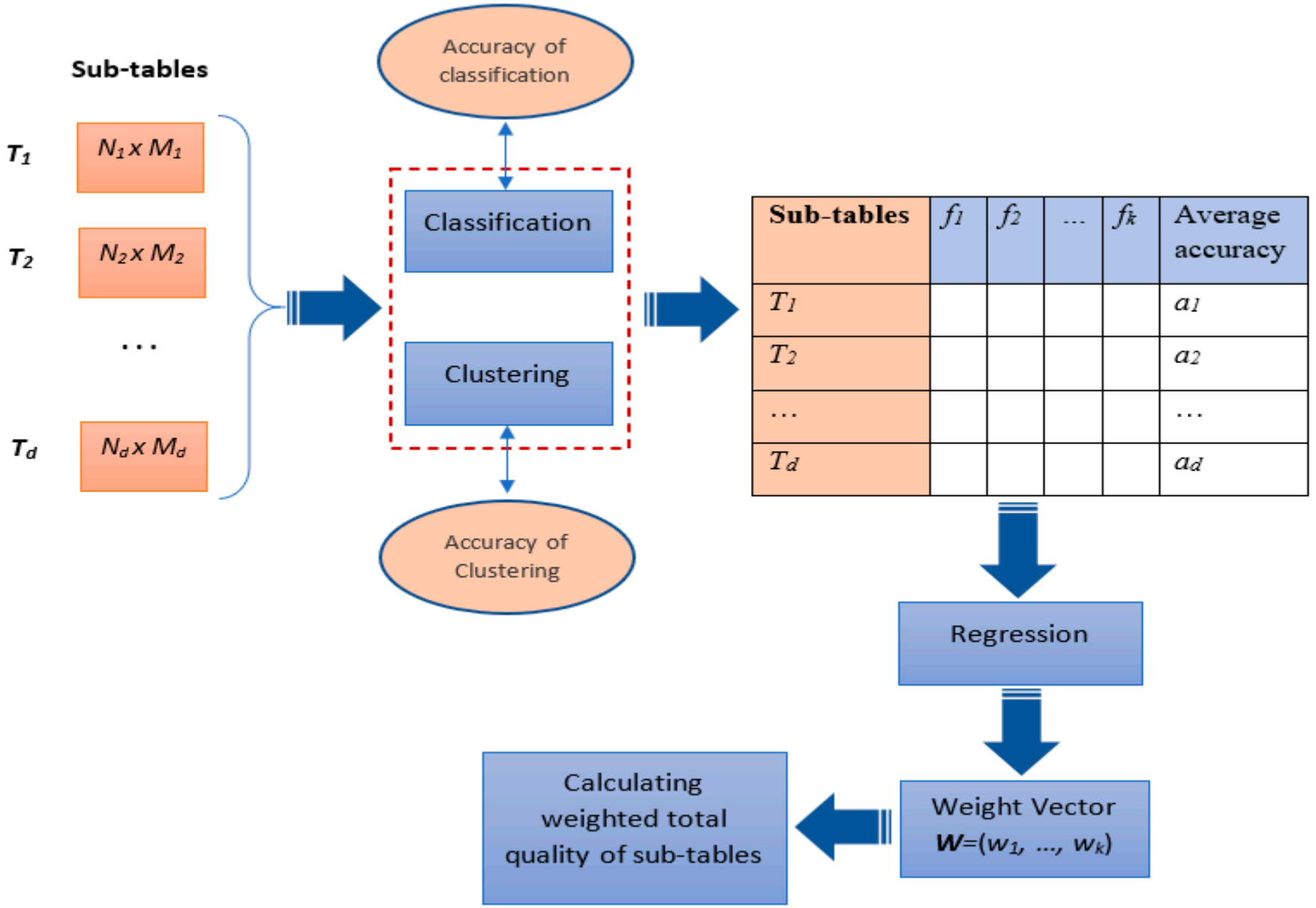
Most people evaluate these quality metrics independently and assume that poor quality affects outcomes

Framework: Evaluate the data quality by defining an objective and looking at the overall model performance.

Proposed method



Learning weights of data quality measures



Static features of the ADNI dataset (ADNIMERGE table)

Feature Type	Column Name
Baselines (44 features)	CDRSB_bl, ADAS11_bl, ADAS13_bl, ADASQ4_bl, MMSE_bl, RAVLT_immediate_bl, RAVLT_learning_bl, RAVLT_forgetting_bl, RAVLT_perc_forgetting_bl, LDELTOTAL_BL, DIGITSCOR_bl, TRABSCOR_bl, FAQ_bl, mPACCdigit_bl, mPACCtrailsB_bl, Ventricles_bl, Hippocampus_bl, WholeBrain_bl, Entorhinal_bl, Fusiform_bl, MidTemp_bl, ICV_bl, MOCA_bl, EcogPtMem_bl, EcogPtLang_bl, EcogPtVisspat_bl, EcogPtPlan_bl, EcogPtOrgan_bl, EcogPtDivatt_bl, EcogPtTotal_bl, EcogSPMem_bl, EcogSPLang_bl, EcogSPVisspat_bl, EcogSPPlan_bl, EcogSPOrgan_bl, EcogSPDivatt_bl, EcogSPTotal_bl, ABETA_bl, TAU_bl, PTAU_bl, FDG_bl, PIB_bl, AV45_bl, FBB_bl.
Demographics (6 features)	AGE, PTGENDER, PTEDUCAT, PTETHCAT, PTRACCAT, PTMARRY
Genetics (one feature)	APOE4
Diagnosis class (one feature)	DX_bl

Simulation

- Random sub-tables were generated (R=4, d=500)
- Five diagnosis groups of patients at baseline (**CN, pMCI, sMCI,** and **AD**)
- Data quality features: **Pearson Correlation (PC), Missing Values, Spearman Correlation, Outliers,** and **Class overlap**
- **Classification method: Random Forest (RF) classifier**
- **Clustering method: Agglomerative** and **k-means clustering** (Silhouette Coefficient for calculating the accuracy of clustering)
- **Random Forest Regression** for learning the weight vector

Results

- **Mean weight vector: W* = (0.2005, 0.207, 0.2812, 0.0115, 0.2998)** for PC, Spearman Correlation, Missing Values, Outliers, and Class Overlap, respectively
- Weights of five data quality measures:

Run	PC	Spearman	Missing Values	Outliers	Class Overlap
1	0.1727	0.2331	0.2729	0.0088	0.3125
2	0.2123	0.181	0.3034	0.0042	0.299
3	0.2062	0.2177	0.2733	0.0132	0.2895
4	0.2109	0.1962	0.275	0.0197	0.2982

- Details of the best sub-table:

Run	Rows	Columns	PC	Spearman	Missing values	Outliers	Class Overlap	Classification Accuracy	Total Quality
1	1532	21	0.6708	0.6736	0.7685	1	0.495	0.7383	0.6499
2	1394	18	0.6563	0.6502	0.7277	1	0.4924	0.7188	0.6299
3	1474	16	0.7606	0.7492	0.8023	1	0.3212	0.6459	0.641
4	1994	16	0.7198	0.7163	0.6372	1	0.6433	0.7678	0.6761

Conclusion

- Data readiness is important to evaluate, and it encompasses various aspects related to data quality
- We developed a data readiness framework to evaluate AD cohorts with an objective to improve dementia assessment
- A manuscript based on these findings is in preparation
- **Limitation:** Over reliance on such measures can lead to good modeling frameworks but can limit progress towards building true AI-based systems

Acknowledgments

BU Department of Medicine
Kolachalama Laboratory

Boston University
Department of
Computer Science

Funding:

KAREN TOFFLER CHARITABLE TRUST

NIH National Institute on Aging

American Heart Association®
life is why™

NIH National Heart, Lung, and Blood Institute

Johnson & Johnson

NIH National Institute of Diabetes and Digestive and Kidney Diseases

NIH NATIONAL CANCER INSTITUTE