CHALLENGES OF BIOMEDICAL DATA ANALYSIS

PREMANANDA INDIC, PH.D.

DEPARTMENT OF ELECTRICAL ENGINEERING

The University of Texas at TYLER Center for Health Informatics & Analytics

ORS Research Design & Data Analysis Lab
Office of Research and Scholarship

INTRODUCTION

MathWorks®

University of Texas at Tyler

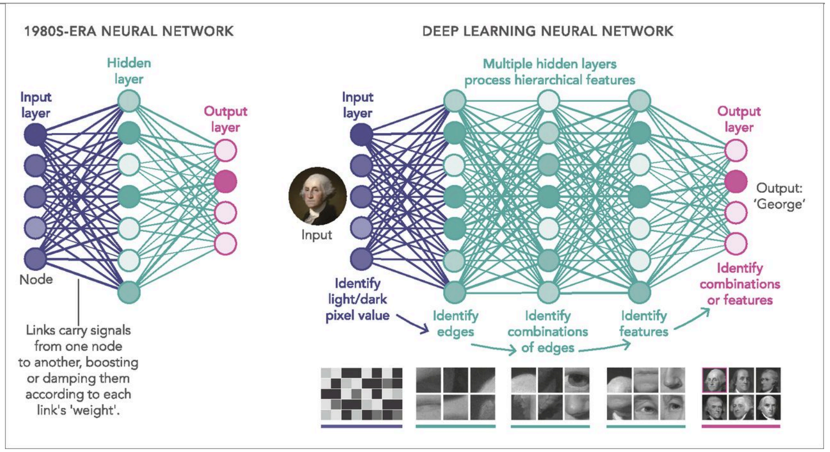
Get Software | Learn MATLAB | Teach with MATLAB | What's New

MATLAB Access for Everyone at

University of Texas at Tyler

https://www.mathworks.com/academia/tah-portal/university-of-texas-at-tyler-1108545.html

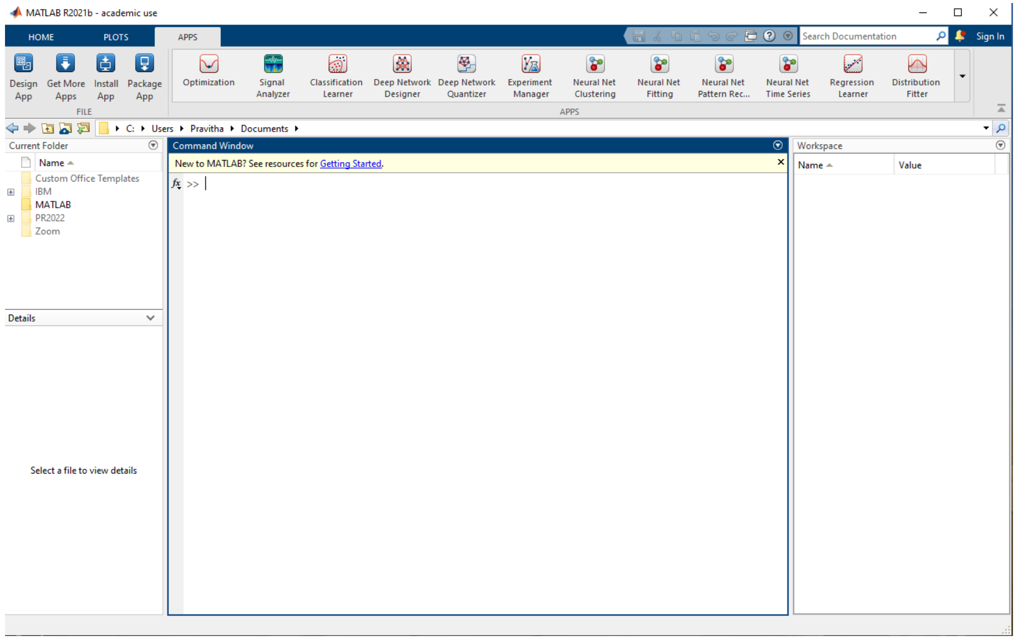
Data



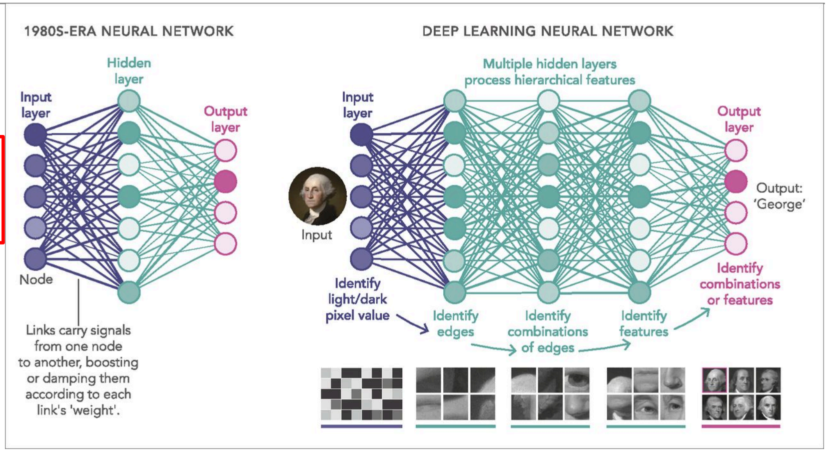
Waldrop, M.M., 2019. News Feature: What are the limits of deep learning?. *Proceedings of the National Academy of Sciences*, *116*(4), pp.1074-1077.

Data → Preprocessing → Feature Engineering/ Extraction → Feature Selection → **Statistical or Machine Learning Models**

INTRODUCTION



Data



Waldrop, M.M., 2019. News Feature: What are the limits of deep learning?. *Proceedings of the National Academy of Sciences*, *116*(4), pp.1074-1077.

Statistical or Machine Learning Models

Data → Preprocessing → Feature Engineering/ Extraction → Feature Selection →

BIOMEDICAL (BIG) DATA

- NEED A SPECIFIC RESEARCH QUESTION (HYPOTHESIS)
- FROM BIG DATA TO CLINICAL IMPACT IS STILL UNCLEAR

Scientific hypothesis, an idea that proposes a tentative explanation about a phenomenon or a narrow set of phenomena observed in the natural world. The two primary features of a scientific hypothesis are falsifiability and testability

Source: https://www.britannica.com/science/scientific-hypothesis

BIOMEDICAL (BIG) DATA

- PHYSIOLOGICAL / BEHAVIORAL / DEMOGRAPHICS

  - Traditional Data Collection (Controlled Conditions)

  - Electronic Health Records (Notes, Vital Signs, Demographics, Lab Results..)

  - Sensor Data

  - Social Media Data

BIOMEDICAL (BIG) DATA

- Traditional Data Collection (Controlled Conditions)

Very expensive

Randomized Control Trials (Inclusion/ Exclusion Criteria)

Population sample must match the actual population (selection bias)

Sanson-Fisher, R.W., Bonevski, B., Green, L.W. and D'Este, C., 2007. Limitations of the randomized controlled trial in evaluating population-based health interventions. *American journal of preventive medicine*, *33*(2), pp.155-161.

BIOMEDICAL (BIG) DATA

- Traditional Data Collection (Controlled Conditions)

Very expensive

Randomized Control Trials (Inclusion/ Exclusion Criteria)

Population sample must match the actual population (selection bias)

Zadrozny, B., 2004, July. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning* (p. 114).

BIOMEDICAL (BIG) DATA

- Traditional Data Collection (Controlled Conditions)

Very expensive

Randomized Control Trials (Inclusion/ Exclusion Criteria)

Population sample must match the actual population (selection bias)

BIOMEDICAL (BIG) DATA

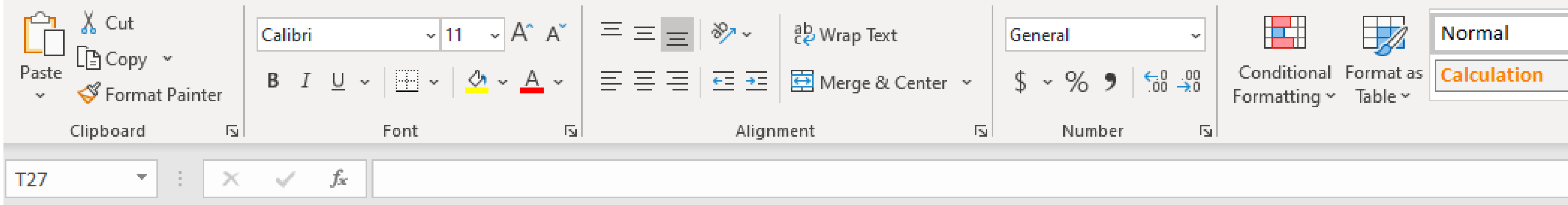
- Traditional Data Collection (Controlled Conditions)

Data: UC Irvine Machine learning Repository

https://archive-beta.ics.uci.edu/ml/datasets/heart+disease

Heart Disease from 4 databases.

N=120

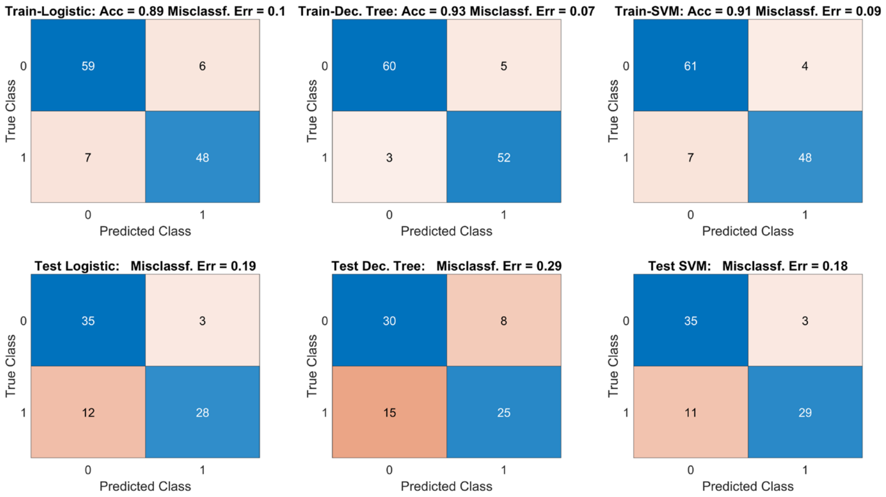


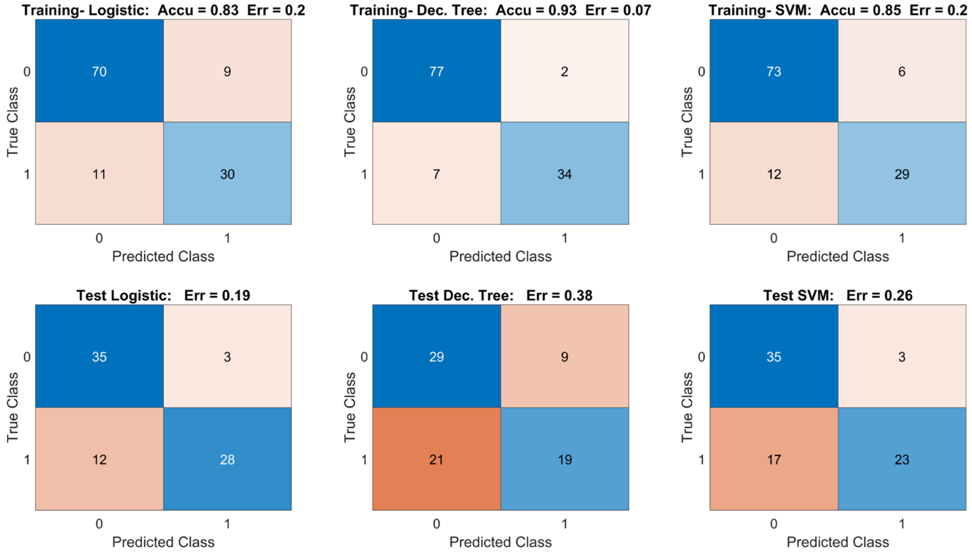
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Age	Sex	chestPainCode	restBP	Chol	FBS	restECG	MaxHR	exerAng	oldPeak	slopePeak	CA	thal	Label	
2	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0	
3	67	1	4	160	286	0	2	108	1	1.5	2	3	3	1	
4	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1	
5	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0	
6	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0	
7	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0	
8	62	0	4	140	268	0	2	160	0	3.6	3	2	3	1	
9	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0	
10	63	1	4	130	254	0	2	147	0	1.4	2	1	7	1	
11	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1	
12	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0	
13	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0	
14	56	1	3	130	256	1	2	142	1	0.6	2	1	6	1	
15	44	1	2	120	263	0	0	173	0	0	1	0	7	0	

BIOMEDICAL (BIG) DATA

- Traditional Data Collection (Controlled Conditions)

Unbiased Data





Biased Data

BIOMEDICAL (BIG) DATA

- Electronic Health Records

Demographics, current and past diagnosis, lab results, prescription drugs, notes, radiological images, ……

Subjective vs Objective

BIOMEDICAL (BIG) DATA

- Electronic Health Records

Medical Concept Extraction

Patient Trajectory Modeling

Disease Inference

Clinical Decision Support System

BIOMEDICAL (BIG) DATA

- Electronic Health Records (Notes, Vital Signs, Demographics, ….)

Missing Data

Sample Size

Miss classification error

Gianfrancesco, M.A., Tamang, S., Yazdany, J. and Schmajuk, G., 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, *178*(11), pp.1544-1547.

BIOMEDICAL (BIG) DATA

- Electronic Health Records (Notes, Vital Signs, Demographics, ….)

Sources of Bias Entering EHR Systems	Potentialto Differentially Affect Vulnerable Populations	Example of Biases With Respect to Clinical Decision Support Output
Missing data	Certain patients may have more fractured care and/or be seen at multiple institutions; patients with lower health literacy may not be able to access online patient portals and document patient-reported outcomes	The EHR may only contain more severe cases for certain patient populations and make erroneous inferences about the risk for such cases; conditioning on complete data may eliminate large portions of the population and result in inaccurate predictions for certain groups
Sample size	Certain subgroups of patients may not exist in sufficient numbers for a predictive analytic algorithm	Underestimation may lead to estimates of mean trends to avoid overfitting, leading to uninformative predictions for subgroups of patients; clinical decision support may be restricted to only the largest groups, spurring improvements in certain patient populations without similar support for others
Misclassification or measurement error	Patients of low socioeconomic status may be more likely to be seen in teaching clinics, where data input or clinical reasoning may be less accurate or systematically different than that from patients of higher socioeconomic status; implicit bias by health care practitioners leads to disparities in care	Algorithm inaccurately learns to treat patients of low socioeconomic status according to less than optimal care and/or according to implicit biases

BIOMEDICAL (BIG) DATA

- Electronic Health Records (Notes, Vital Signs, Demographics, ….)

From traditional machine learning to deep learning:

Features are derived directly from data

Based on Artificial Neural Networks

Shickel, B., Tighe, P.J., Bihorac, A. and Rashidi, P., 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics*, *22*(5), pp.1589-1604.

BIOMEDICAL (BIG) DATA

- Electronic Health Records (Notes, Vital Signs, Demographics, ….)

Several recent deep EHR projects.

Project	**Deep EHR Task**
DeepPatient	Multi-outcome Prediction
Deepr	Hospital Re-admission Prediction
DeepCare	EHR Concept Representation
Doctor AI	Heart Failure Prediction
Med2Vec	EHR Concept Representation
eNRBM	Suicide risk stratification

Shickel, B., Tighe, P.J., Bihorac, A. and Rashidi, P., 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics*, *22*(5), pp.1589-1604.

BIOMEDICAL (BIG) DATA

- Sensor Data

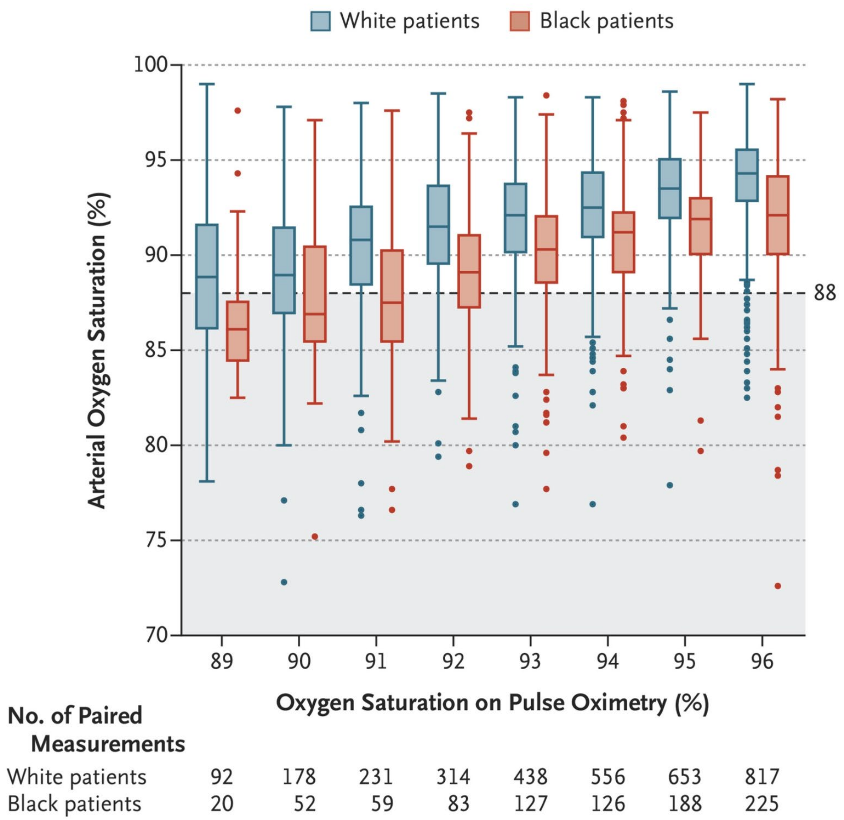
Sensor Design	: Differential characteristics
Data are nonstationary	: Feature extraction methods are stationary
Data has multiscale structure	: Analytical tools fails to capture such scales
Noise & Artifacts	: Noise/ artifacts may have useful information

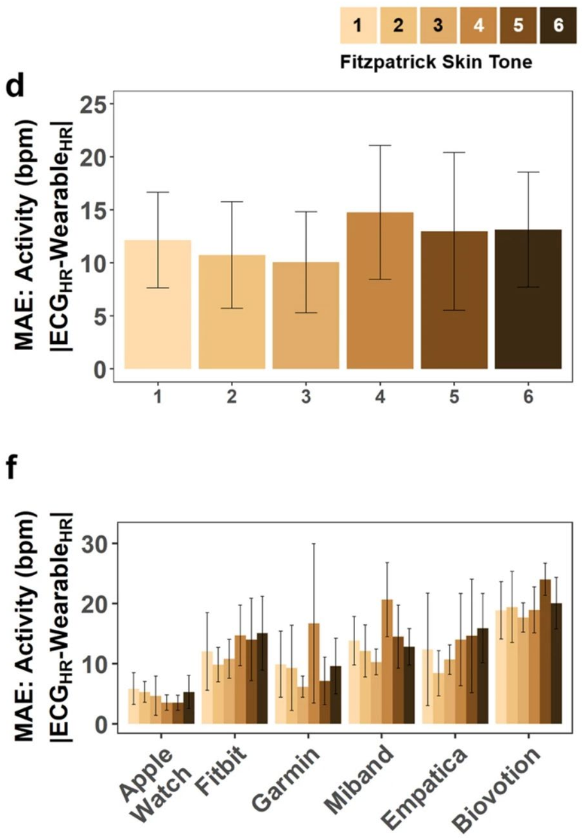
BIOMEDICAL (BIG) DATA

- Sensor Data

Sensor Design	: Differential characteristics
Data are nonstationary	: Feature extraction methods are stationary
Data has multiscale structure	: Analytical tools fails to capture such scales
Noise & Artifacts	: Noise/ artifacts may have useful information

BIOMEDICAL (BIG) DATA





Sjoding, M.W., Dickson, R.P., Iwashyna, T.J., Gay, S.E. and Valley, T.S., 2020. Racial bias in pulse oximetry measurement. *New England Journal of Medicine*, *383*(25), pp.2477-2478.

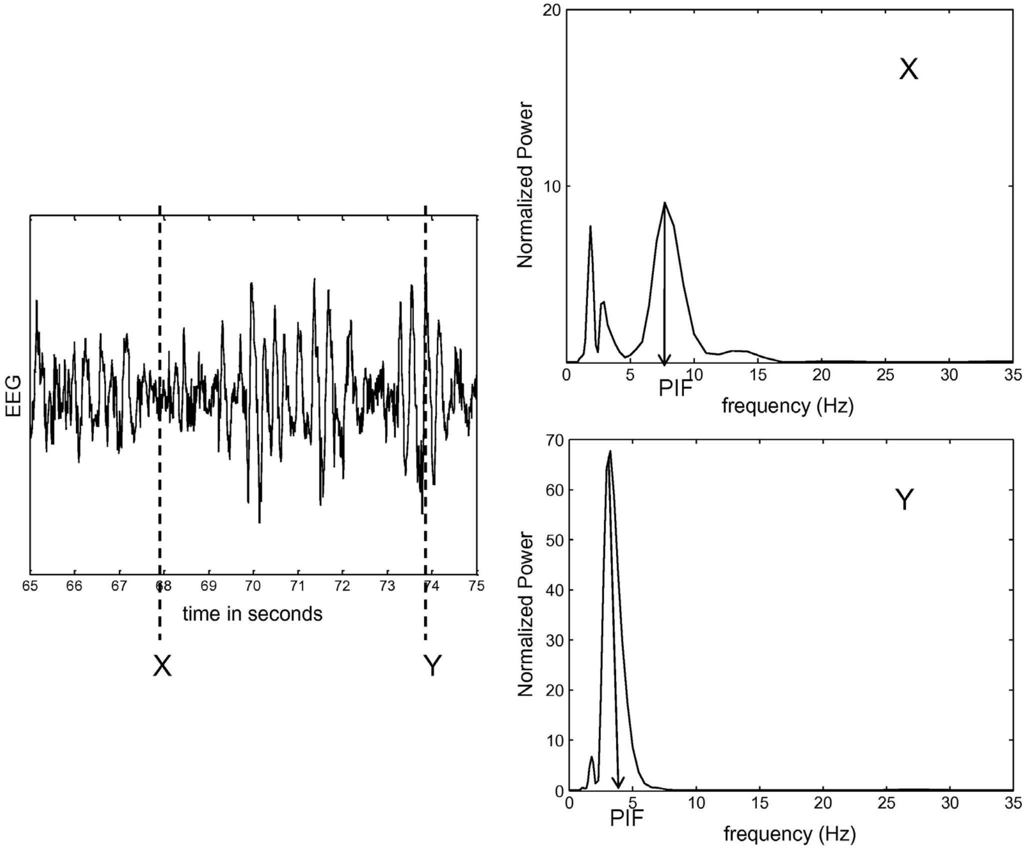
Bent, B., Goldstein, B.A., Kibbe, W.A. and Dunn, J.P., 2020. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine*, *3*(1), pp.1-9.

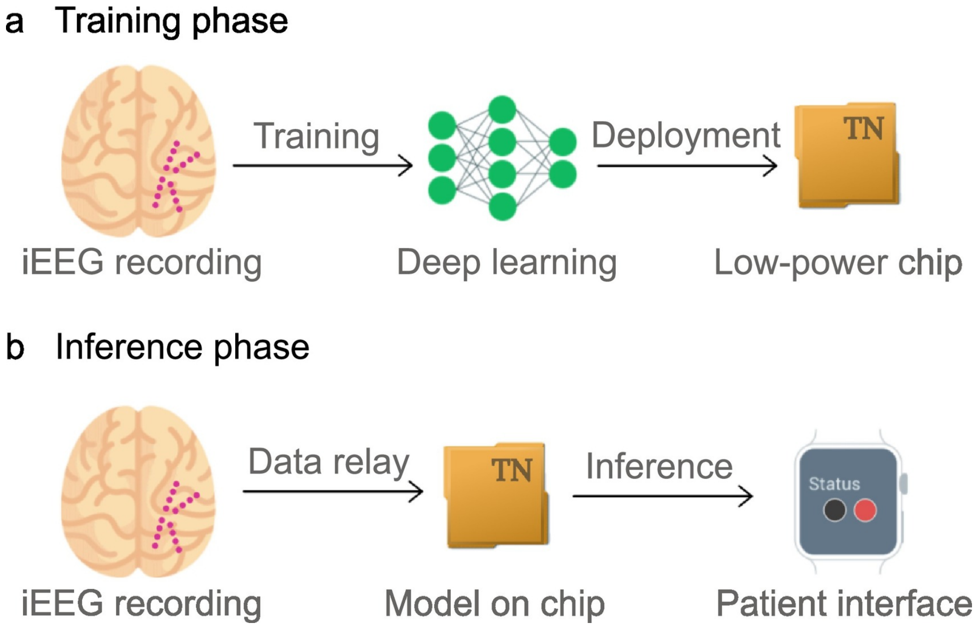
BIOMEDICAL (BIG) DATA

- Sensor Data

Sensor Design	: Differential characteristics
Data are nonstationary	: Feature extraction methods are stationary
Data has multiscale structure	: Analytical tools fails to capture such scales
Noise & Artifacts	: Noise/ artifacts may have useful information

BIOMEDICAL (BIG) DATA





Kiral-Kornek, I., Roy, S., Nurse, E., Mashford, B., Karoly, P., Carroll, T., Payne, D., Saha, S., Baldassano, S., O'Brien, T. and Grayden, D., 2018. Epileptic seizure prediction using big data and deep learning: toward a mobile system. *EBioMedicine*, *27*, pp.103-111.

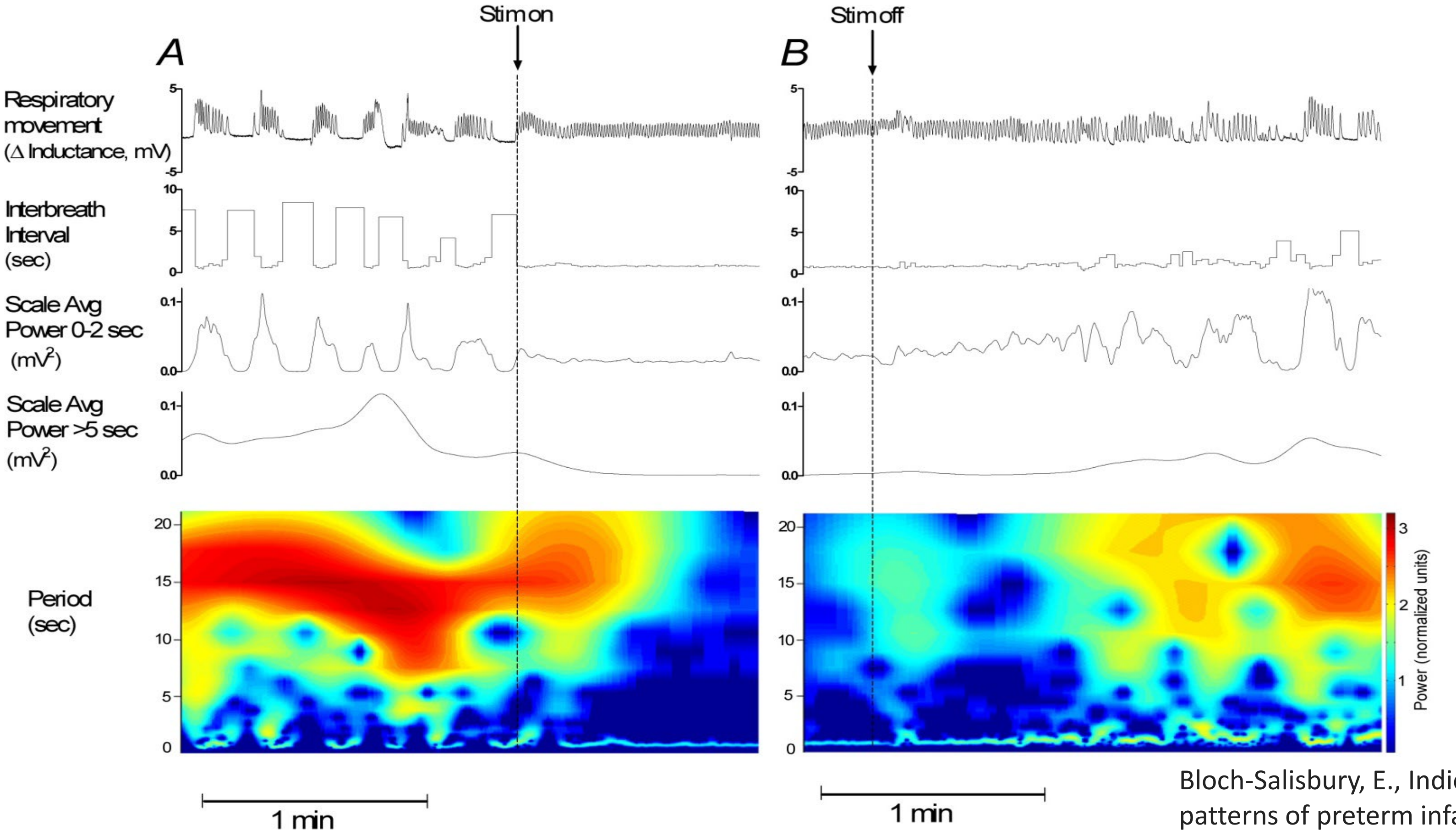
Indic, P. and Narayanan, J., 2011. Wavelet based algorithm for the estimation of frequency flow from electroencephalogram data during epileptic seizure. *Clinical neurophysiology*, *122*(4), pp.680-686.

BIOMEDICAL (BIG) DATA

- Sensor Data

Sensor Design	: Differential characteristics
Data are nonstationary	: Feature extraction methods are stationary
Data has multiscale structure	: Analytical tools fails to capture such scales
Noise & Artifacts	: Noise/ artifacts may have useful information

BIOMEDICAL (BIG) DATA



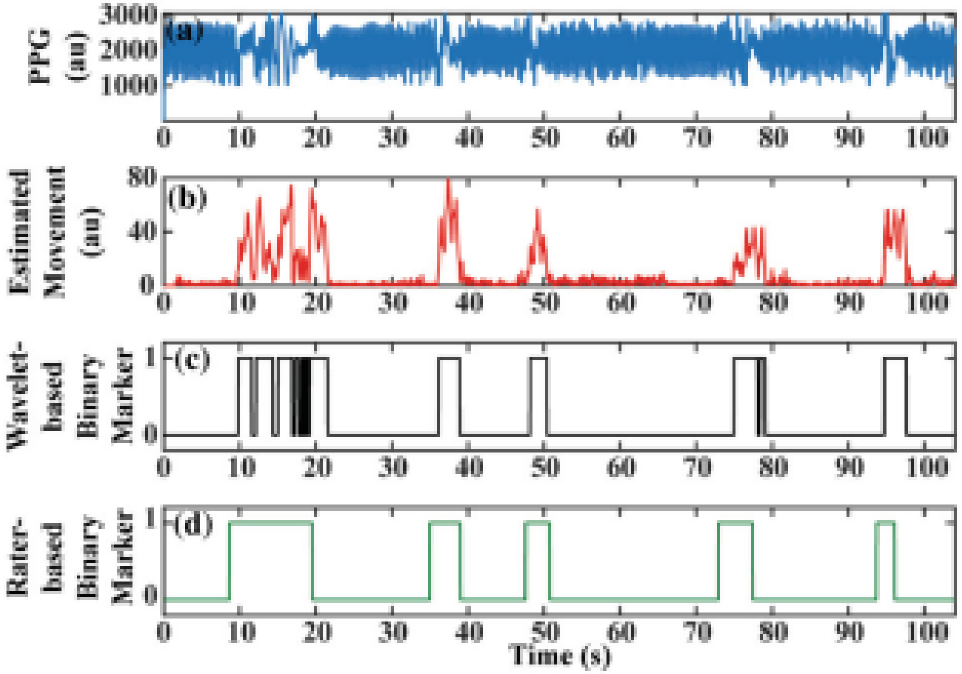
Bloch-Salisbury, E., Indic, P., Bednarek, F. and Paydarfar, D., 2009. Stabilizing immature breathing patterns of preterm infants using stochastic mechanosensory stimulation. *Journal of Applied Physiology*, *107*(4), pp.1017-1027.

BIOMEDICAL (BIG) DATA

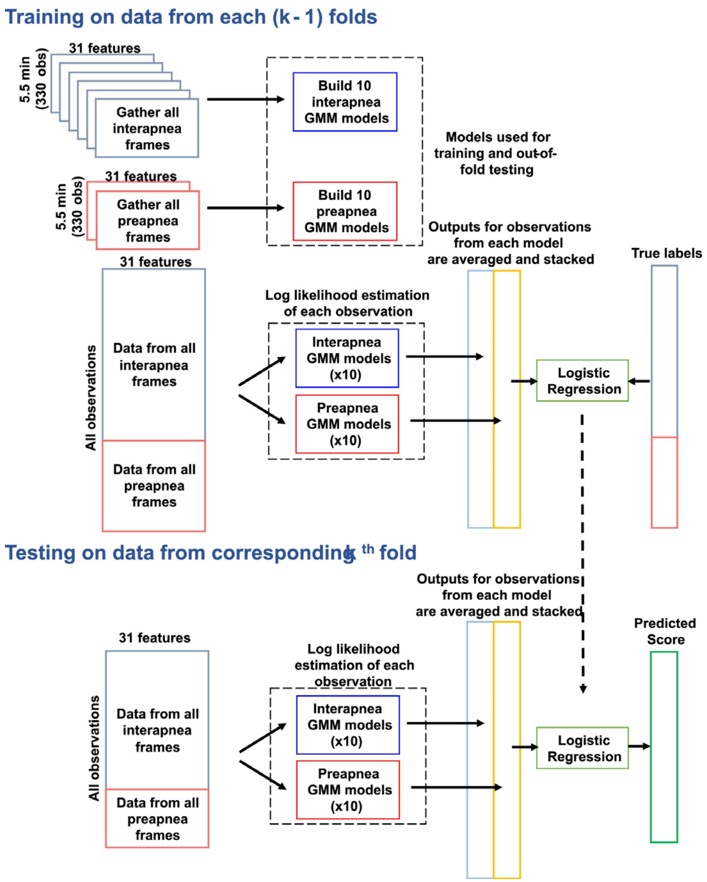
- Sensor Data

Sensor Design	: Differential characteristics
Data are nonstationary	: Feature extraction methods are stationary
Data has multiscale structure	: Analytical tools fails to capture such scales
Noise & Artifacts	: Noise/ artifacts may have useful information

BIOMEDICAL (BIG) DATA



Zuzarte, I., Sternad, D. and Paydarfar, D., 2021. Predicting apneic events in preterm infants using cardio-respiratory and movement features. *Computer Methods and Programs in Biomedicine*, *209*, p.106321.



SUMMARY

Bias:

Historical bias	: Structural issue with data collection
Representation bias	: Effect of sampling
Measurement bias	: Measurement of a specific feature
Evaluation bias	: Model iteration and evaluation
Aggregation bias	: Flawed assumptions

Suresh, H. and Guttag, J.V., 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, *2*.

SUMMARY

Challenges:

Conceptual	: Standards, Physician's intuition, Reasoning
Technical	: Limitations in sensors & algorithms
Humanistic	: Values & duties, ethics
Interpretability	: From black box to clinical inference

Quinn, T.P., Senadeera, M., Jacobs, S., Coghlan, S. and Le, V., 2021. Trust and medical AI: the challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association*, *28*(4), pp.890-894.

THANK YOU

ORS Research Design & Data Analysis Lab

Office of Research and Scholarship

SBIR: RAE (Realize, Analyze, Engage) - A digital biomarker based detection and intervention system for stress and carvings during recovery from substance abuse disorders.
PIs: M. Reinhardt, S. Carreiro, P. Indic

STARs Award

The University of Texas System
P. Indic (PI, UT Tyler)

Department of Veterans Affairs

Design of a wearable sensor system and associated algorithm to track suicidal ideation from movement variability and develop a novel objective marker of suicidal ideation and behavior risk in veterans.
Clinical Science Research and Development Grant (approved for funding),
P. Indic (site PI, UT-Tyler)
E.G. Smith (Project PI, VA)
P. Salvatore (Investigator, Harvard University)

Design of a wearable biosensor sensor system with wireless network for the remote detection of life threatening events in neonates.

National Science Foundation Smart & Connected Health Grant
P. Indic (Lead PI, UT-Tyler)
D. Paydarfar (Co PI, UT-Austin)
H. Wang (Co PI, UMass Dartmouth)
Y. Kim (Co PI, UMass Dartmouth)

Pre-Vent

National Institute Of Health Grant
P. Indic (Analytical Core PI, UT-Tyler)
N. Ambal (PI, Univ. of Alabama, Birmingham)

ViSiOn
P. Indic (site PI, UT-Tyler)
P. Ramanand (Co-I, UT Tyler
N. Ambal, (PI, Univ. of Alabama, Birmingham)

QUESTIONS