INTRODUCTION TO DATA ANALYTICS WITH MACHINE LEARNING

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

PREREQUISITE

- NO KNOWLEDGE OF PROGRAMMING
- NO KNOWLEDGE OF ANY QUANTITATIVE METHODS
- INTEREST IN RESEARCH (and also interested in making $$$???)

http://clipart-library.com/clipart/2096816.htm

OUTLINE

- INTRODUCTION
- DATA ANALYTICS & DIFFERENT TYPES
- MACHINE LEARNING & DIFFERENT TYPES

OUTLINE

- INTRODUCTION
- DATA ANALYTICS & DIFFERENT TYPES
- MACHINE LEARNING & DIFFERENT TYPES

INTRODUCTION

- DATA ANALYTICS

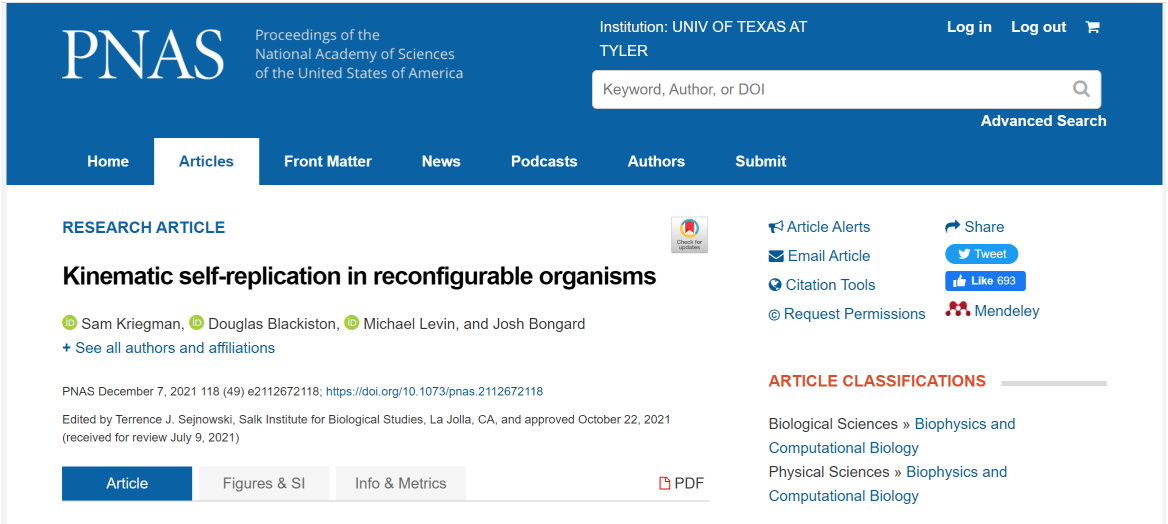
The process of analyzing raw information to get insights and draw conclusions about outcomes

https://marketoonist.com/2016/12/predictive-analytics.html

INTRODUCTION

- DATA ANALYTICS

Business, Entertainment, Sports
Healthcare & Medicine
Basic Research



https://www.uvm.edu/news/story/team-builds-first-living-robots-can-reproduce

INTRODUCTION

DATA ANALYTICS

Data Analytics Certificate Program

Data analytics is a fast-growing field in the computing sciences, and as more and more companies are recognizing the need to implement data analytics into their daily operations, employment opportunities in this industry are abundant. The Data Analytics Certificate Program is designed to broadly enhance students' opportunities in their future professional careers and/or future graduate studies.

A data analytics certificate can enhance prospects for a successful career: (1) there is a high demand for data analytics professionals, (2) job opportunities increase, (3) prospective higher wages for qualified professionals, (4) data analytics is a top priority in many organizations, and (5) there is flexibility across the professional employment sector.

Certificate Requirements

Required Courses (9 hrs.)

The certificate requires students to complete 9 semester credit hours (3 courses) from the following existing course set with a grade of C or better in each course. Prerequisites for all certificate courses selected will apply.

COSC 5347	Business Intelligence and Analysis
COSC 5371	Data Mining
CSCI 5342	Sports Data Analytics
CSCI 5350	Machine Learning

Courses completed for this certification will be listed as a milestone on an official university transcript and a certificate of completion will be awarded by the Department of Computer Science.

INTRODUCTION

- DATA ANALYTICS

Fastest growing jobs (28% according US Bureau of Labor Statistics)

Hot Jobs for Psychologists

Trend in Nursing Management

https://www.bls.gov/opub/btn/volume-7/big-data-adds-up.htm

https://www.apa.org/gradpsych/2013/01/big-data

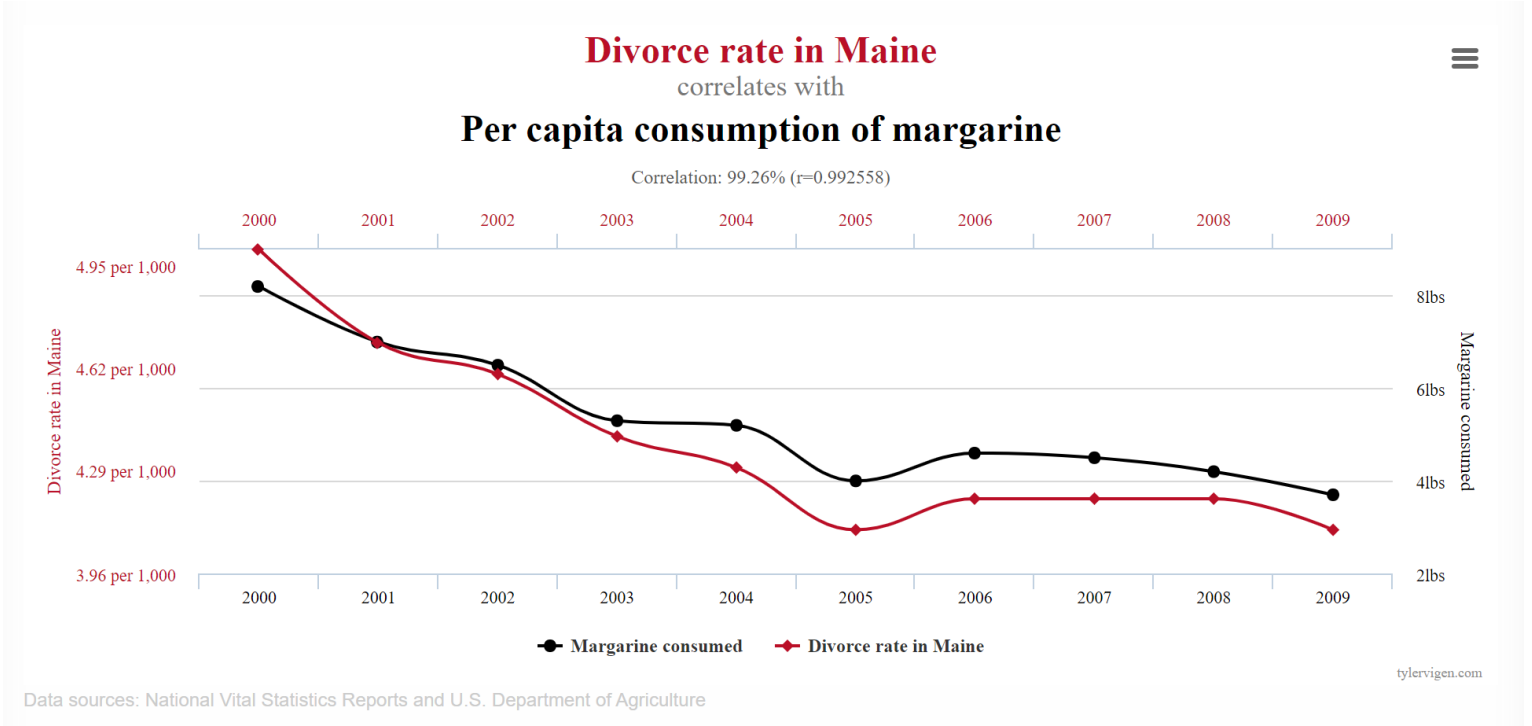
https://journals.lww.com/nursingmanagement/fulltext/2019/03000/the_synthesis_of_nursing_knowledge_and_predictive.3.aspx

INTRODUCTION

- IMPORTANT IN DATA ANALYTICS

  Knowledge about the field and understanding the features as well as outcome

- SPURIOUS RELATION



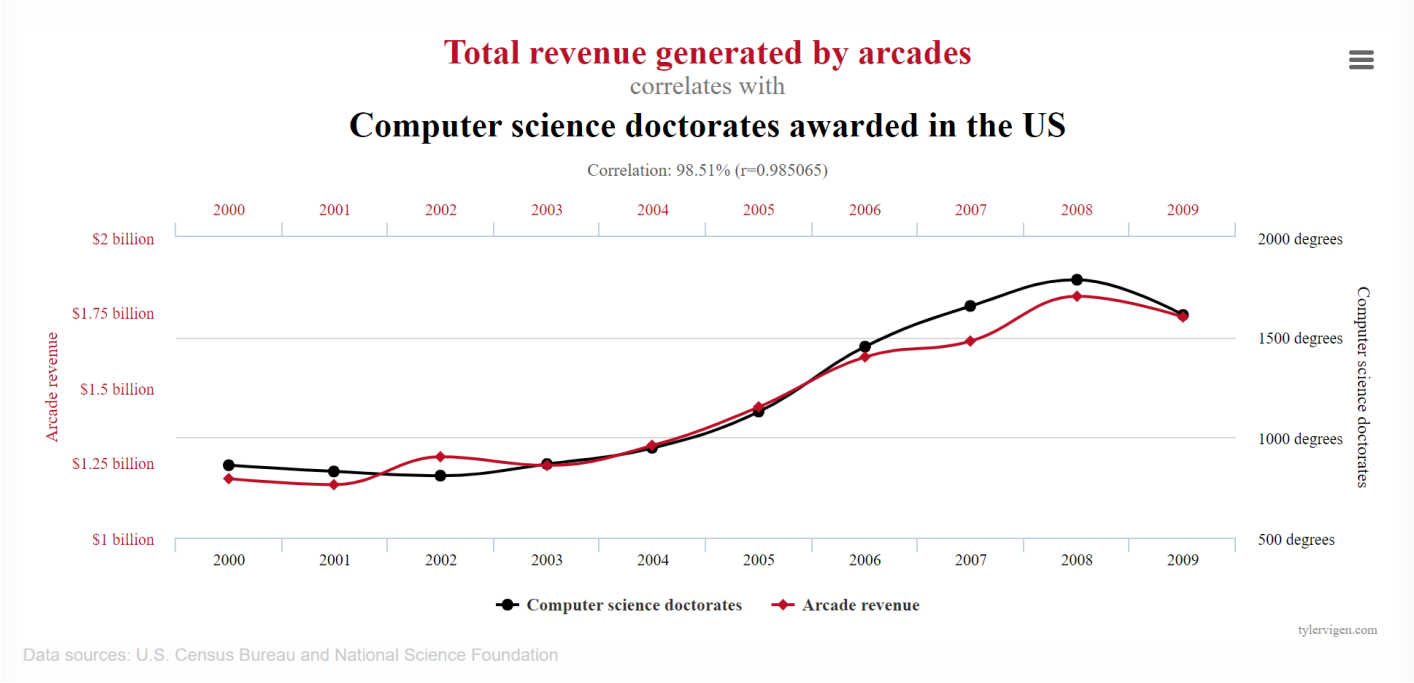
https://www.tylervigen.com/spurious-correlations

INTRODUCTION

- IMPORTANT IN DATA ANALYTICS

  Knowledge about the field and understanding the features as well as outcome

- SPURIOUS RELATION



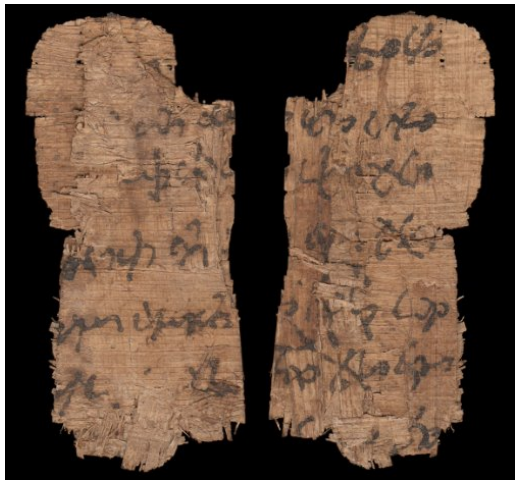
https://www.tylervigen.com/spurious-correlations

INTRODUCTION

- OLD CONCEPT IN A NEW PACKAGE

  Data gathering and analysis in ancient times

Clay Tablets



Papyrus Paper

https://www.sqlservercentral.com/articles/a-brief-history-of-data-analysis

INTRODUCTION

- OLD CONCEPT IN A NEW PACKAGE

  Number System

  Algebra

  Geometry

  Calculus

  Boolean Algebra

  ……..

  ……

  Finding Patterns, Trends,……

  Data Analytics, Machine Learning, AI,……

INTRODUCTION

- OLD CONCEPT IN A NEW PACKAGE

  Find the sum of first three odd numbers

  (1+3+5 = 9)

  Find the sum of first seven odd numbers

  (1 + 3 +5 +7 +9 + 11 +13 = 49)

  Find the sum of first N odd numbers

  (1+3+5+7+9+……+N = N^2)

INTRODUCTION

➢ What is Machine Learning ?

- Machine Learning is a field of study that gives computers the ability to “learn” without being explicitly programmed
  - Prediction
  - Classification

Samuel AL, IBM J. Research & Development, 1959, vol. 3 (3), 210-229

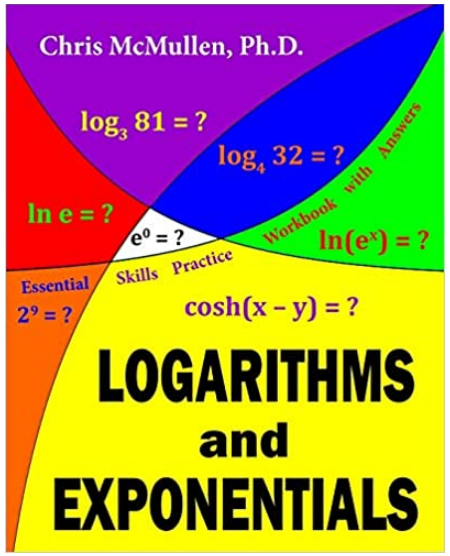
INTRODUCTION

- Too many books spoil the curiosity

- Start with Andrew Ng, Machine Learning, Stanford University available on YouTube

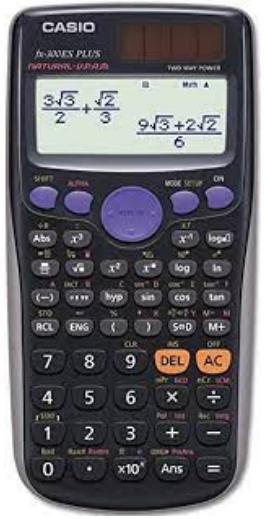
Some Statistics & Programming Knowledge Helps !

INTRODUCTION



Analytical Tools

Simple Calculator (Boolean Algebra)



Scientific Calculator (Series Expansion, Boolean Algebra)

Computer (Programming Language, Assembly Language, Series Expansion, Boolean Algebra)

Smart Devices (ML Models, Programming Language, Assembly Language, Series Expansion, Boolean Algebra)

INTRODUCTION

- Always there is a mathematical foundation

  Analytical Tools (Logarithm, Laplace Transform, Fourier Transform…….)

  Computational Tools (Boolean Algebra, Taylor Series Expansion,……..)

  Programming Languages (Basic, Fortran, C, C++, Java, …….)

  Assembly Languages (depending upon the computer processors)

  Machine Learning Models (Supervised and Unsupervised Learning

  Artificial Intelligence (Neural Network, Deep Learning)

ANALYSIS PLATFORM

MathWorks®

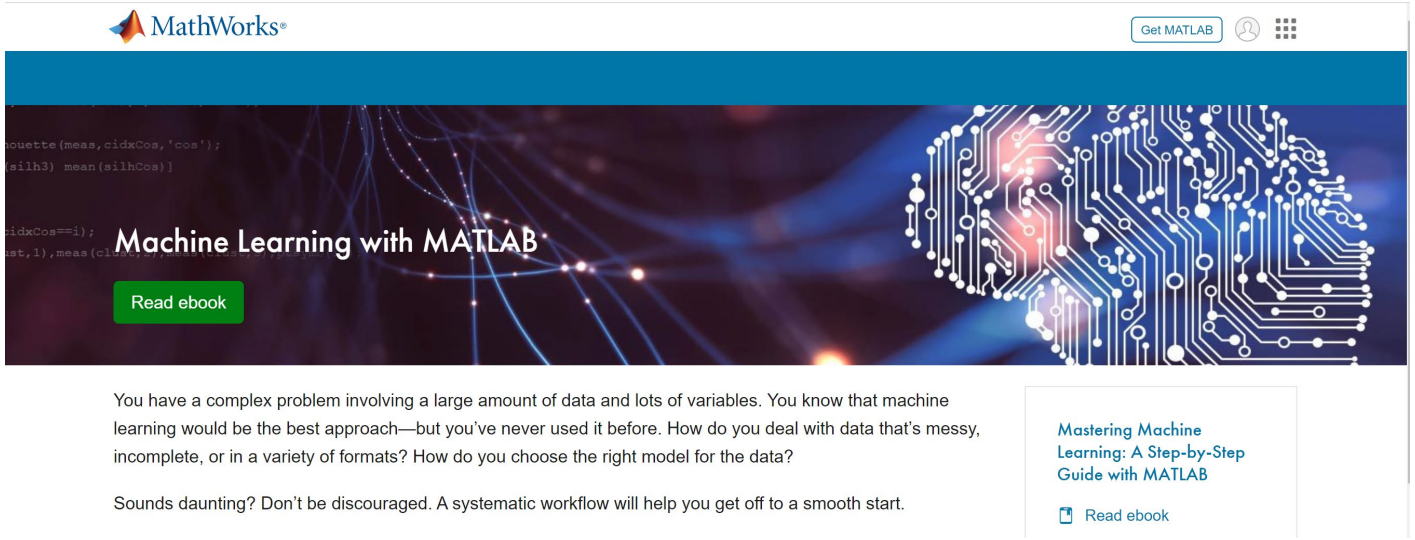
University of Texas at Tyler

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

MATLAB Access for Everyone at

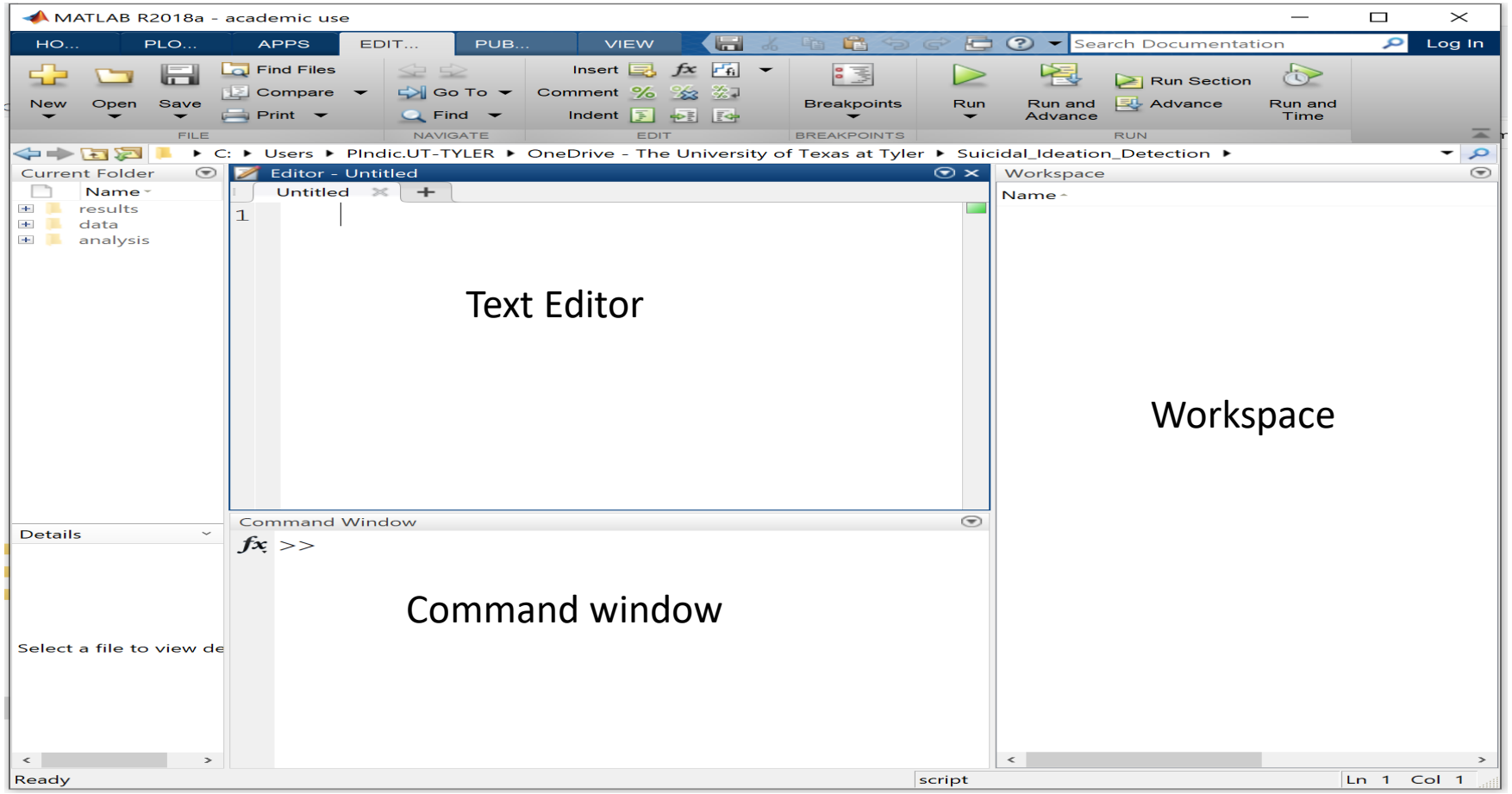
University of Texas at Tyler

Open Source Solutions:
R & Python



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

ANALYSIS PLATFORM



HYPOTHESIS

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

OUTLINE

- INTRODUCTION
- DATA ANALYTICS & DIFFERENT TYPES
- MACHINE LEARNING & DIFFERENT TYPES

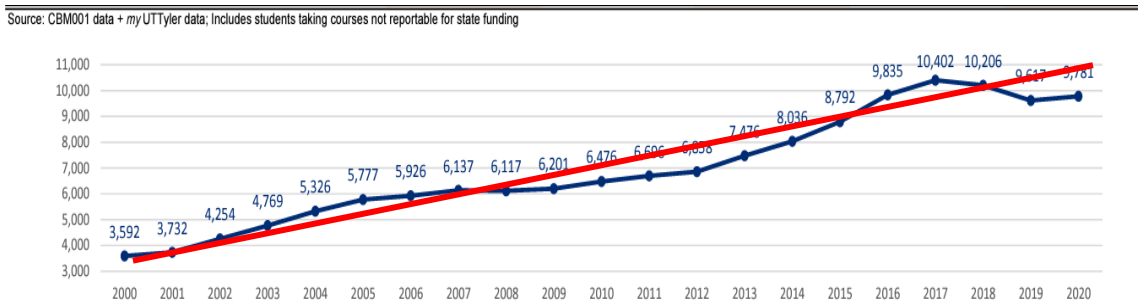
DATA ANALYTICS & TYPES

- DATA ANALYTICS

  The process of analyzing raw information to get insights and draw conclusions about outcomes

- TYPES OF DATA ANALYTICS

Describe	: Descriptive Analytics
Predict	: Predictive Analytics
Diagnose	: Diagnostic Analytics
Prescribe	: Prescriptive Analytics



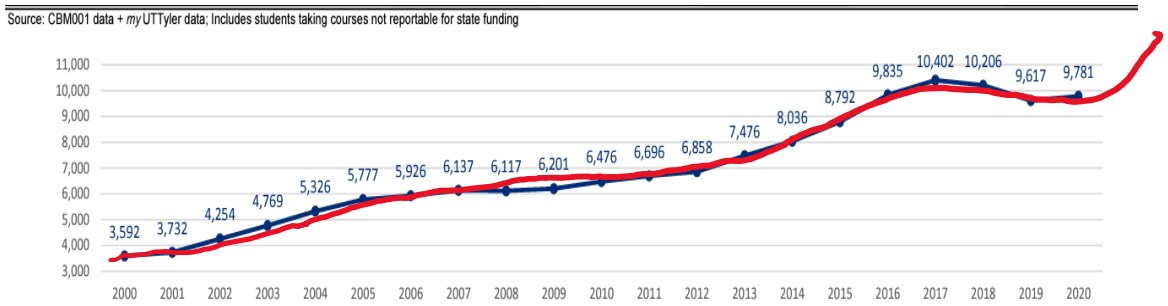
DATA ANALYTICS & TYPES

- DATA ANALYTICS

  The process of analyzing raw information to get insights and draw conclusions about outcomes

- TYPES OF DATA ANALYTICS

Describe	: Descriptive Analytics
Predict	: Predictive Analytics
Diagnose	: Diagnostic Analytics
Prescribe	: Prescriptive Analytics



DATA ANALYTICS & TYPES

- DESCRIPTIVE ANALYTICS
  - Statistical Measures (mean, variance, linear relationship,…..)
  - Graphical Representation (Bar graphs, plots, ……)

Data set	1-3	1	2	3	4	4
Variable	x	y	y	y	x	y
Obs. no. 1	10.0	8.04	9.14	7.46	8.0	6.58
2	8.0	6.95	8.14	6.77	8.0	5.76
3	13.0	7.58	8.74	12.74	8.0	7.71
4	9.0	8.81	8.77	7.11	8.0	8.84
5	11.0	8.33	9.26	7.81	8.0	8.47
6	14.0	9.96	8.10	8.84	8.0	7.04
7	6.0	7.24	6.13	6.08	8.0	5.25
8	4.0	4.26	3.10	5.39	19.0	12.50
9	12.0	10.84	9.13	8.15	8.0	5.56
10	7.0	4.82	7.26	6.42	8.0	7.91
11	5.0	5.68	4.74	5.73	8.0	6.89

TABLE. Four data sets, each comprising 11 (x, y) pairs.

Mean:

x =9

y=11

Variance:

x = 11

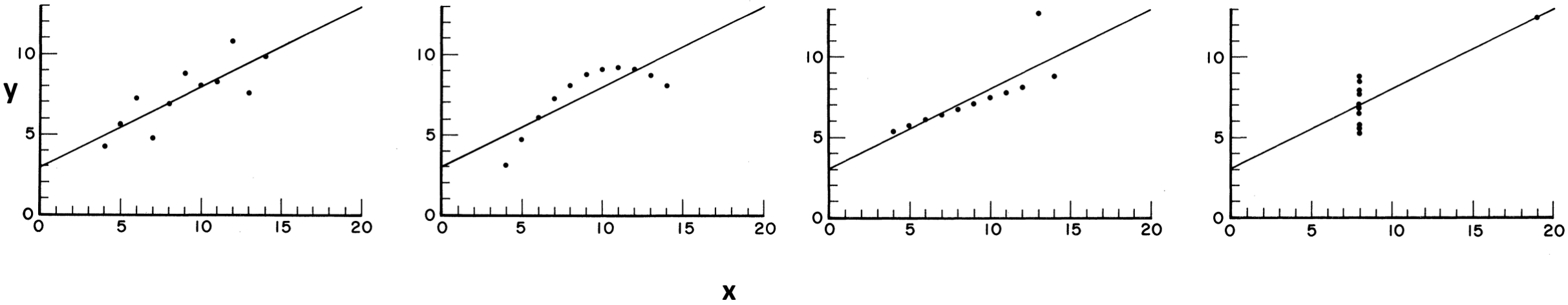
y = 4.125

y = 0.5x + 3

Anscombe, FJ (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17-21

DATA ANALYTICS & TYPES

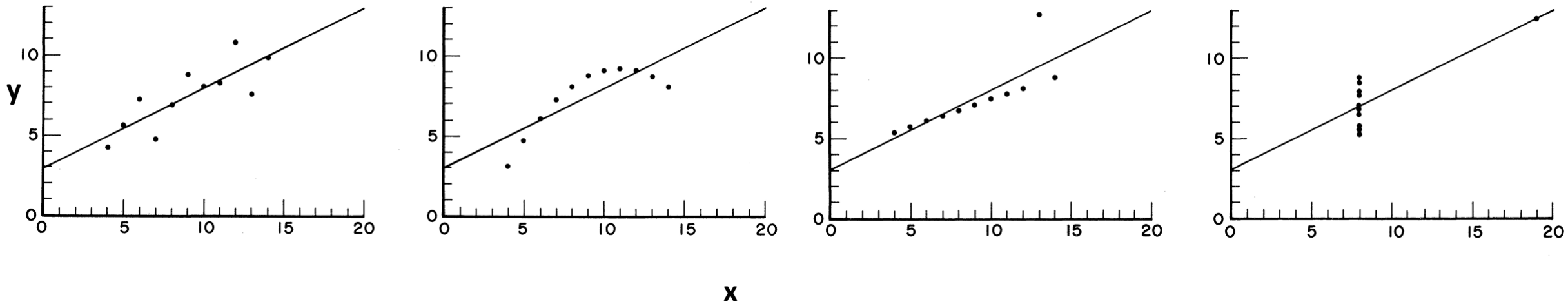
- DESCRIPTIVE ANALYTICS
  - Statistical Measures (Mean, Variance, Linear Relationship,…..)
  - Graphical Representation (Bar graphs, Plots, ……)



Anscombe, FJ (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17-21

DATA ANALYTICS & TYPES

- DESCRIPTIVE ANALYTICS
  - Statistical Measures (Assumption: Data follows normal distribution)
  - Graphical Representation (Assumption: Relationship is linear)



Anscombe, FJ (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17-21

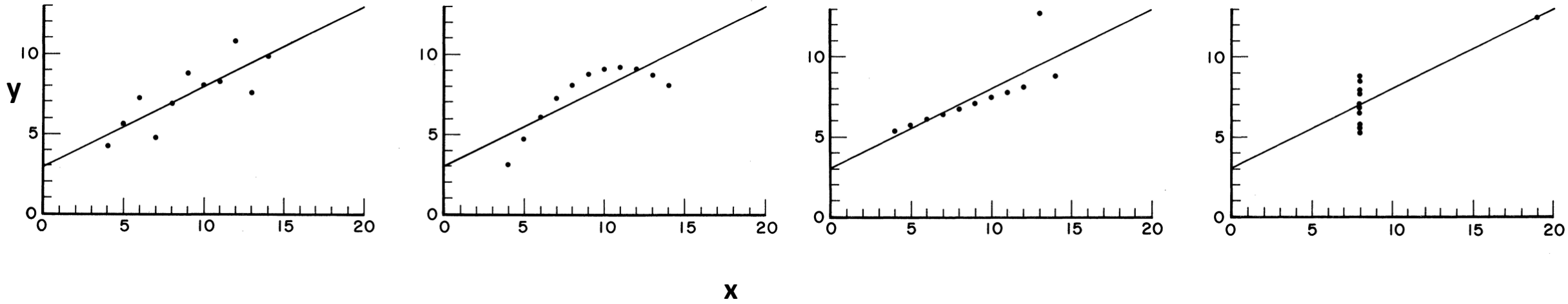
DATA ANALYTICS & TYPES

- PREDICTIVE ANALYTICS
  - Analysis of data to predict outcome

Dependent variable or outcome

$y = 0.5x + 3$

Independent variable or feature



Anscombe, FJ (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17-21

DATA ANALYTICS & TYPES

- PREDICTIVE ANALYTICS
  - Linear Regression

Weight	Acceleration	MPG
3504	12	18
3693	11.5	15
3436	11	18
3433	12	16
3449	10.5	17

```
lm = fitlm(tbl,'MPG~Weight+Acceleration')
```

```
lm =


Linear regression model:
    MPG ~ 1 + Weight + Acceleration

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                    ________    _________    _______    _________

    (Intercept)       45.155       3.4659     13.028    1.6266e-22
    Weight         -0.0082475    0.00059836   -13.783    5.3165e-24
    Acceleration      0.19694       0.14743     1.3359      0.18493

Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 4.12
R-squared: 0.743,  Adjusted R-Squared 0.738
```

MPG = a + b Weight + c Acceleration

DATA ANALYTICS & TYPES

- DIAGNOSTIC ANALYTICS

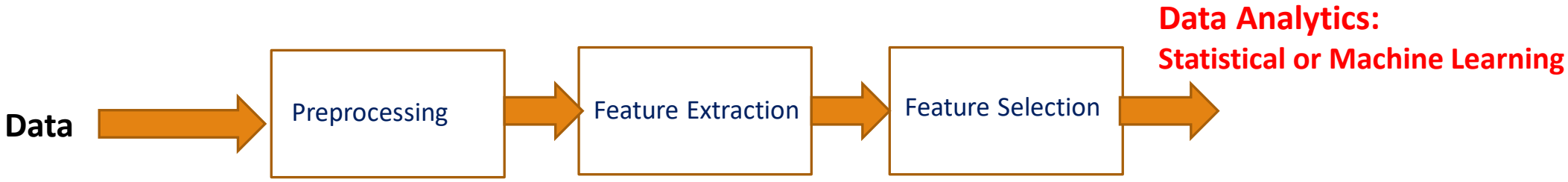
- PRESCRIPTIVE ANALYTICS

OUTLINE

- INTRODUCTION
- DATA ANALYTICS & DIFFERENT TYPES
- MACHINE LEARNING & DIFFERENT TYPES

FEATURES

- VARIABLE (example: mpg, weight)
- STATISTICAL FEATURES (example: mean, variance)
- SPECTRAL FEATURES (example: strength, timing)
- NONLINEAR FEATURES (example: irregularity, power law)



Statistical vs. Machine Learning Models

Purpose:

Statistical models are used for inference (To find association between features and an outcome). Results should be interpretable.

Machine Learning models are used for prediction (Use features that can predict an outcome). Results may not be interpretable.

- Regression, Classification, Clustering

MACHINE LEARNING & TYPES

- Supervised Learning

  Learning a relationship between features and the outcome using a training set

- Unsupervised Learning

  Learning underlying structures in features

MACHINE LEARNING & TYPES

- Supervised Learning
  - Linear Regression
  - Logistic Regression
  - Support Vector Machine
  - Artificial Neural Network
  - ……..
  - ……
  - ……

MACHINE LEARNING & TYPES

- Unsupervised Learning

  Clustering

  - Principal Component Analysis
  - Independent Component Analysis
  - Singular Value Decomposition
  - …….
  - …….

MACHINE LEARNING & TYPES

- Supervised Learning
  - Linear Regression
  - Logistic Regression
  - Support Vector Machine
  - Artificial Neural Network
  - ……..
  - ……
  - ……

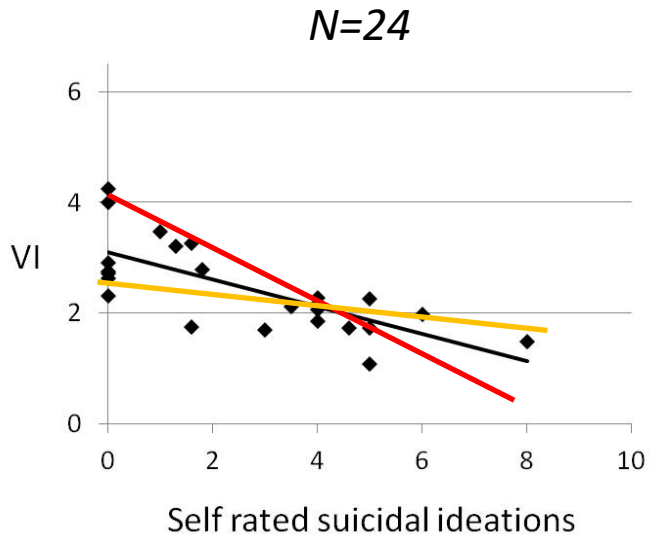
MACHINE LEARNING & TYPES

- Do machines actually “learn” ?

$$VI = m \times SI + C$$

MACHINE LEARNING & TYPES

- Do machines actually “learn” ?



$$e(N=1) = \widetilde{VI}(N=1) - VI(N=1)$$

$$e(N=2) = \widetilde{VI}(N=2) - VI(N=2)$$

……

……

$$e(N=24) = \widetilde{VI}(N=24) - VI(N=24)$$

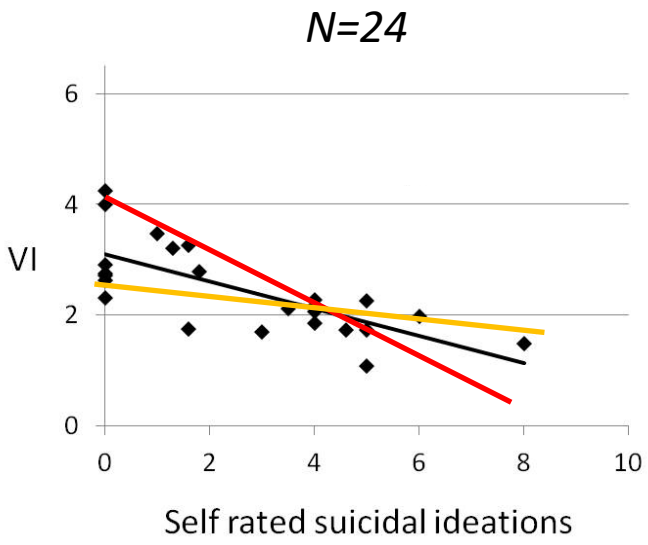
$$\widetilde{VI} = m \times SI + C$$

$$E = \sum_{n=1}^{N} e^2$$

MACHINE LEARNING & TYPES

- Do machines actually "learn" ?

How do we find minimum E ?



m →

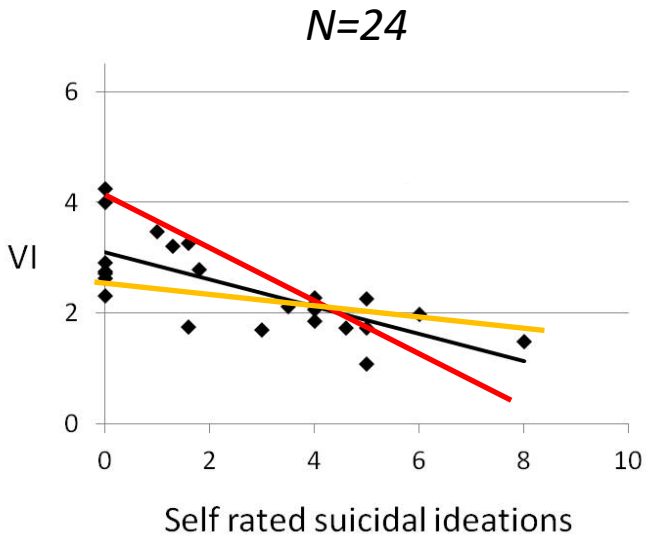
C ↓				
0.1	0.6	0.8	0.01	0.5
1	10	0.01	0.001	0.002
8	7	0.0006	0.03	0.55
100	12	0.1	12	0.89
2	1	2	0.5	0.05

$$\widetilde{VI} = m \times SI + C$$

LEARNING APPROACHES

- Do machines actually “learn” ?

How do we find minimum E ?



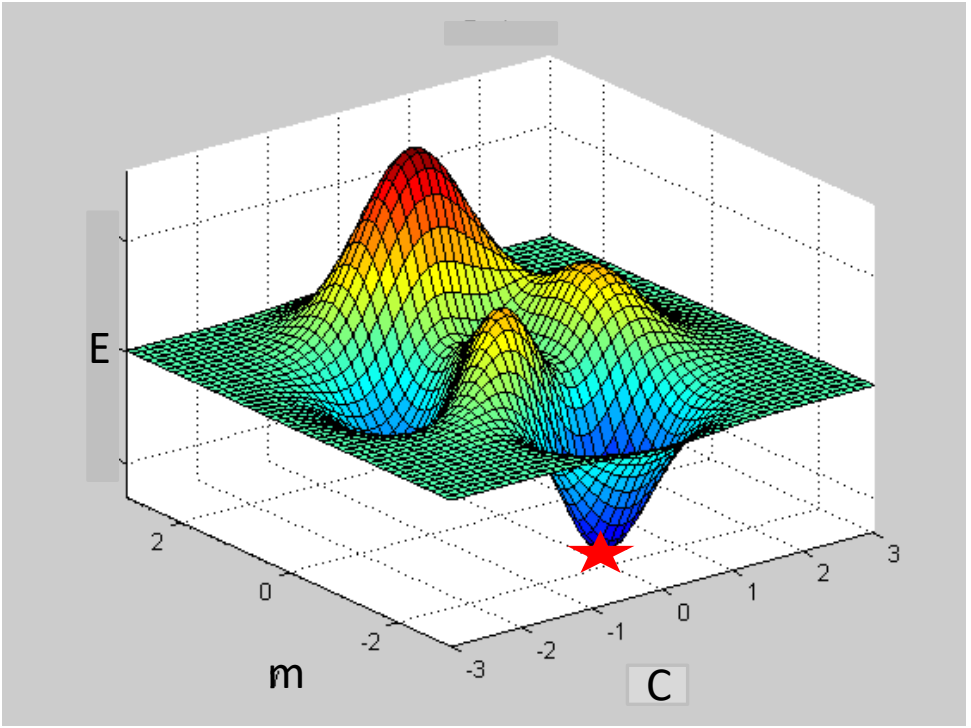
- Gradient Descent

by **Louis Augustin Cauchy** in 1847

$$\widetilde{VI} = m \times SI + C$$

Linear Regression

$$\widetilde{SI} = a \times VI + b$$



MACHINE LEARNING & TYPES

- Do machines actually "learn" ?

Classification of High Risk (n=43) vs. Low Risk (n=95)

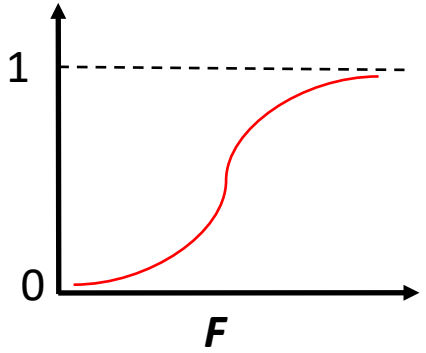
0 = Low Risk, 1 = High Risk (SI>4)

VI
Mean
Variance
Skewness
Kurtosis
Power
Period

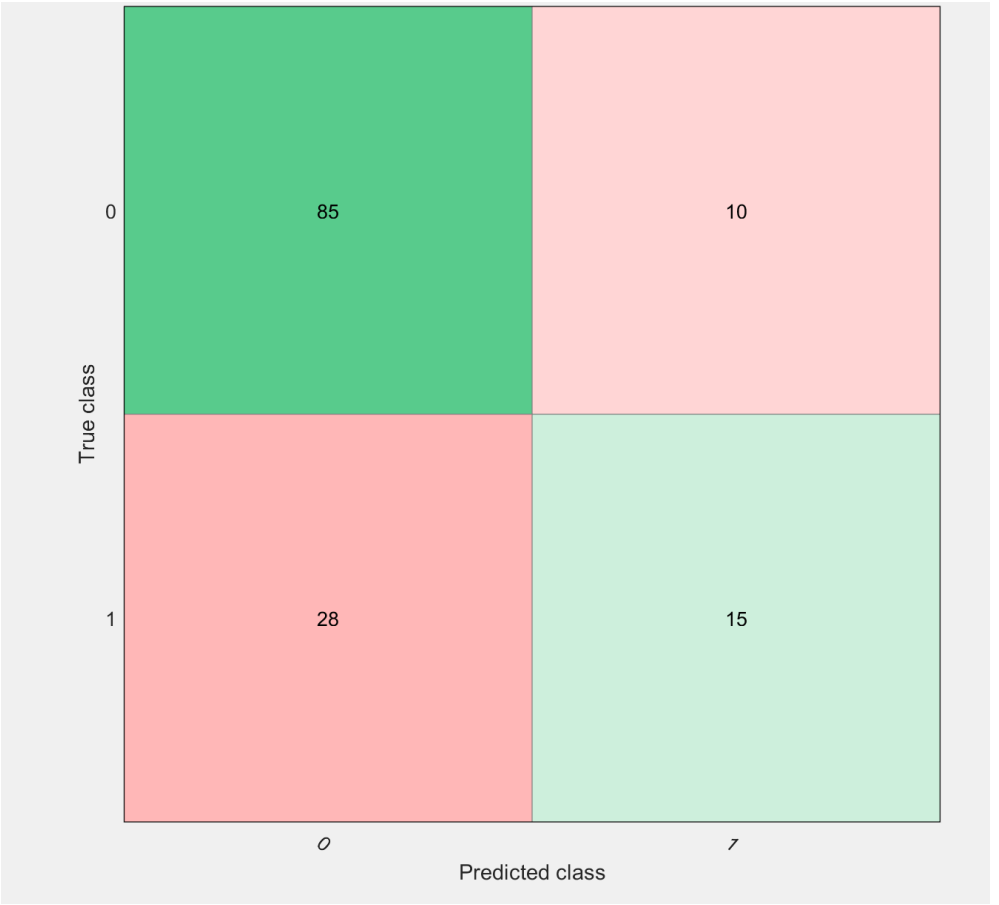
$$p = \frac{1}{1 + e^{-(a \times VI + b)}}$$

Linear Regression

$$P = \frac{1}{1 + e^{-(A^T F + B)}}$$



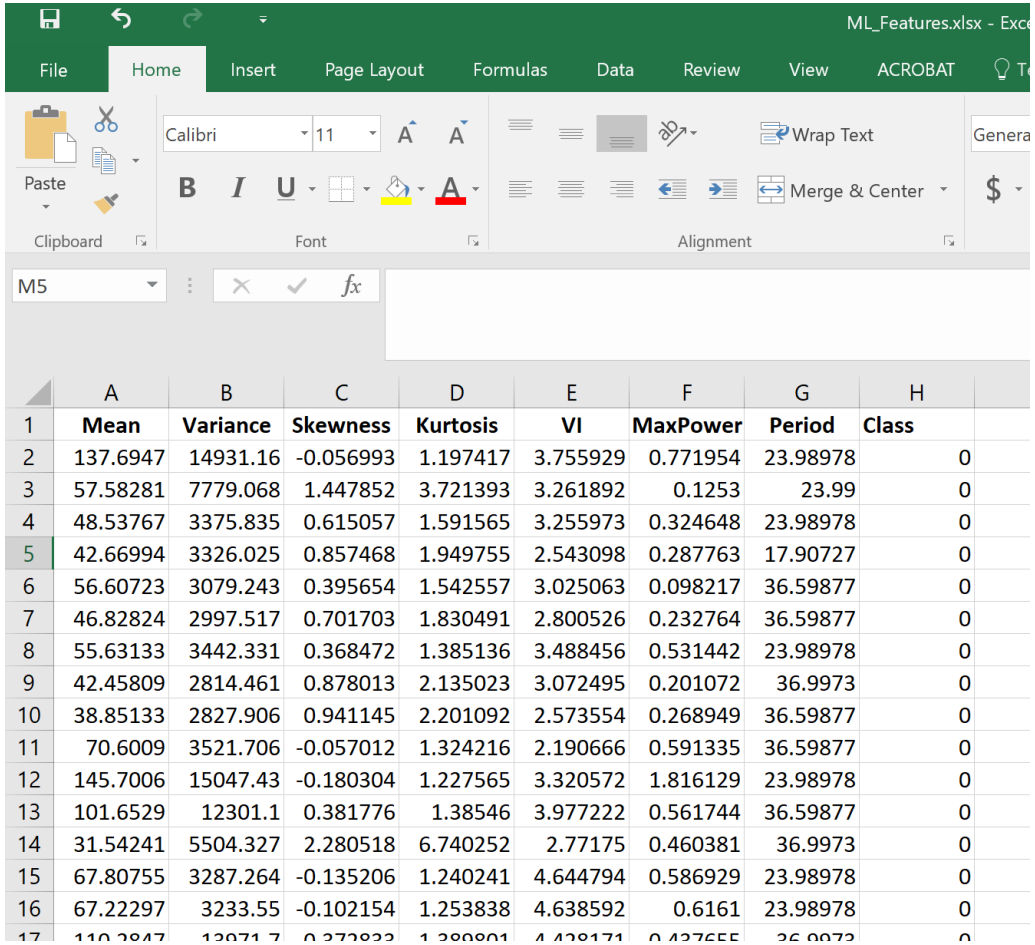
Logistic Regression
Accuracy ~73%



MACHINE LEARNING & TYPES

- How to implement in MATLAB ?

Step 1: Create an excel sheet with features with class assignments

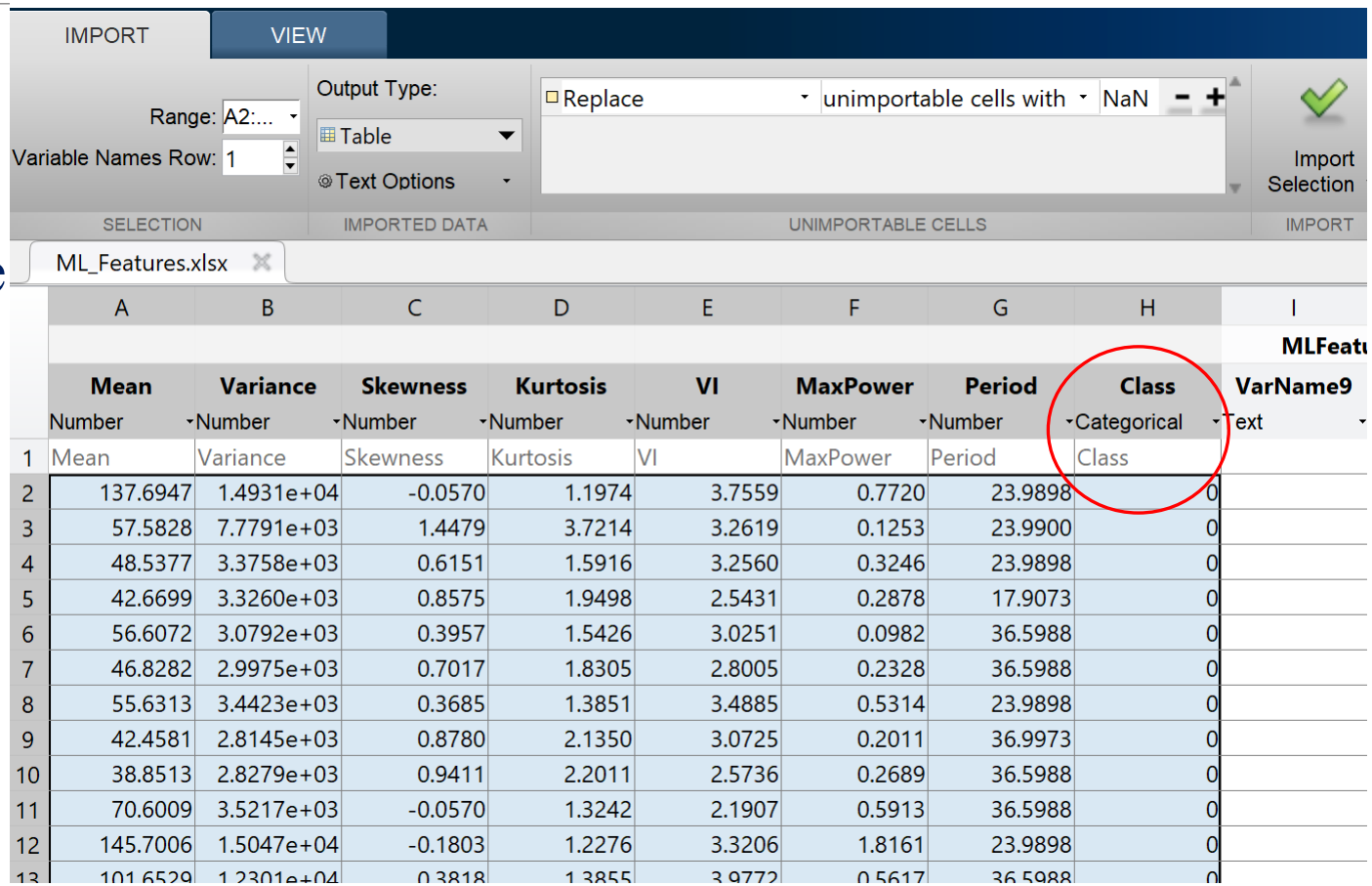


	A	B	C	D	E	F	G	H
1	Mean	Variance	Skewness	Kurtosis	VI	MaxPower	Period	Class
2	137.6947	14931.16	-0.056993	1.197417	3.755929	0.771954	23.98978	0
3	57.58281	7779.068	1.447852	3.721393	3.261892	0.1253	23.99	0
4	48.53767	3375.835	0.615057	1.591565	3.255973	0.324648	23.98978	0
5	42.66994	3326.025	0.857468	1.949755	2.543098	0.287763	17.90727	0
6	56.60723	3079.243	0.395654	1.542557	3.025063	0.098217	36.59877	0
7	46.82824	2997.517	0.701703	1.830491	2.800526	0.232764	36.59877	0
8	55.63133	3442.331	0.368472	1.385136	3.488456	0.531442	23.98978	0
9	42.45809	2814.461	0.878013	2.135023	3.072495	0.201072	36.9973	0
10	38.85133	2827.906	0.941145	2.201092	2.573554	0.268949	36.59877	0
11	70.6009	3521.706	-0.057012	1.324216	2.190666	0.591335	36.59877	0
12	145.7006	15047.43	-0.180304	1.227565	3.320572	1.816129	23.98978	0
13	101.6529	12301.1	0.381776	1.38546	3.977222	0.561744	36.59877	0
14	31.54241	5504.327	2.280518	6.740252	2.77175	0.460381	36.9973	0
15	67.80755	3287.264	-0.135206	1.240241	4.644794	0.586929	23.98978	0
16	67.22297	3233.55	-0.102154	1.253838	4.638592	0.6161	23.98978	0

MACHINE LEARNING & TYPES

- How to implement in MATLAB ?

Step 2: Open MATLAB and drag the excel file to workspace

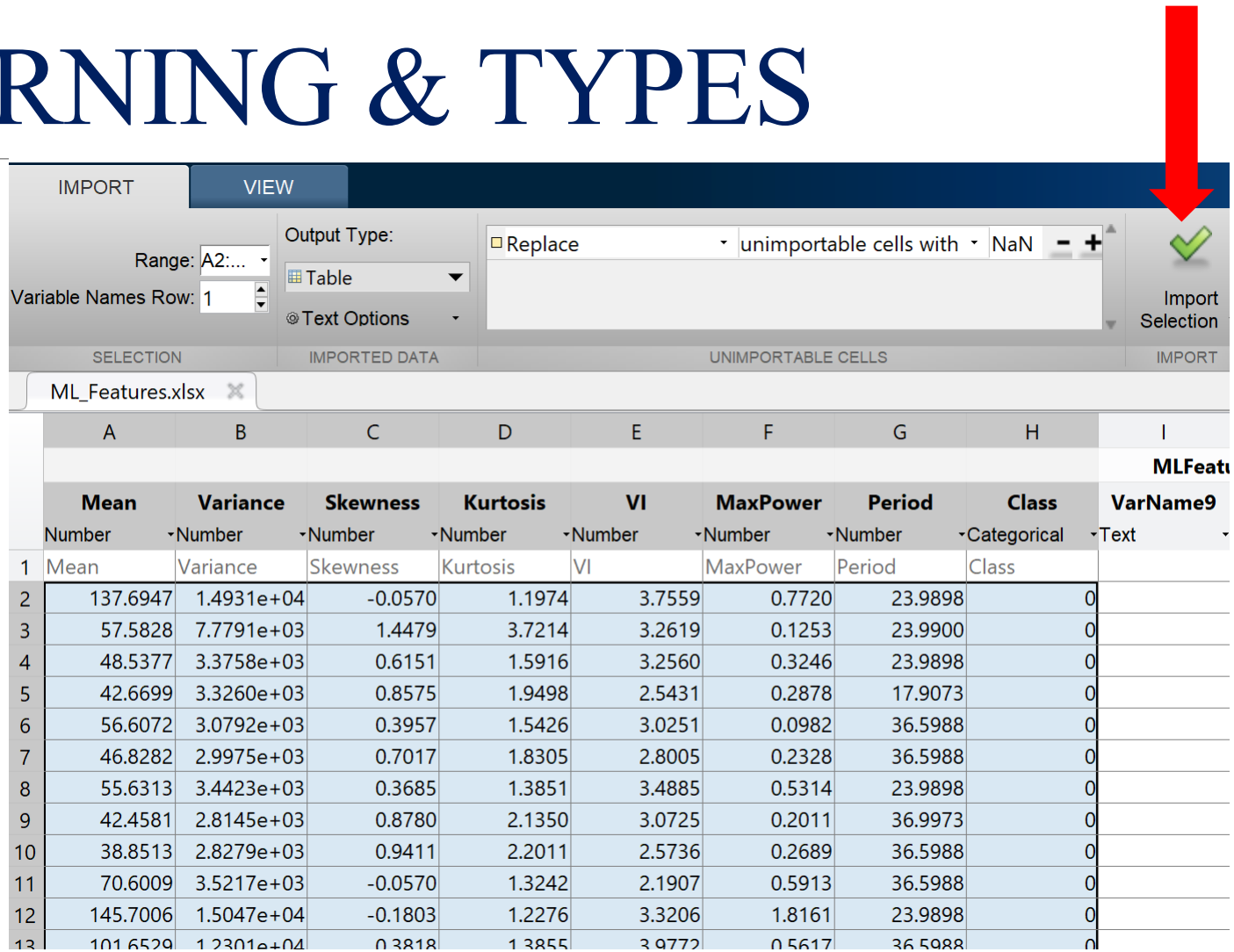


	A	B	C	D	E	F	G	H	I
									MLFeatu
	Mean	Variance	Skewness	Kurtosis	VI	MaxPower	Period	Class	VarName9
	Number	Number	Number	Number	Number	Number	Number	Categorical	Text
1	Mean	Variance	Skewness	Kurtosis	VI	MaxPower	Period	Class	
2	137.6947	1.4931e+04	-0.0570	1.1974	3.7559	0.7720	23.9898	0	
3	57.5828	7.7791e+03	1.4479	3.7214	3.2619	0.1253	23.9900	0	
4	48.5377	3.3758e+03	0.6151	1.5916	3.2560	0.3246	23.9898	0	
5	42.6699	3.3260e+03	0.8575	1.9498	2.5431	0.2878	17.9073	0	
6	56.6072	3.0792e+03	0.3957	1.5426	3.0251	0.0982	36.5988	0	
7	46.8282	2.9975e+03	0.7017	1.8305	2.8005	0.2328	36.5988	0	
8	55.6313	3.4423e+03	0.3685	1.3851	3.4885	0.5314	23.9898	0	
9	42.4581	2.8145e+03	0.8780	2.1350	3.0725	0.2011	36.9973	0	
10	38.8513	2.8279e+03	0.9411	2.2011	2.5736	0.2689	36.5988	0	
11	70.6009	3.5217e+03	-0.0570	1.3242	2.1907	0.5913	36.5988	0	
12	145.7006	1.5047e+04	-0.1803	1.2276	3.3206	1.8161	23.9898	0	
13	101.6529	1.2301e+04	0.3818	1.3855	3.9772	0.5617	36.5988	0	

MACHINE LEARNING & TYPES

- How to implement in MATLAB ?

Step 3: Click Import Selection and import data

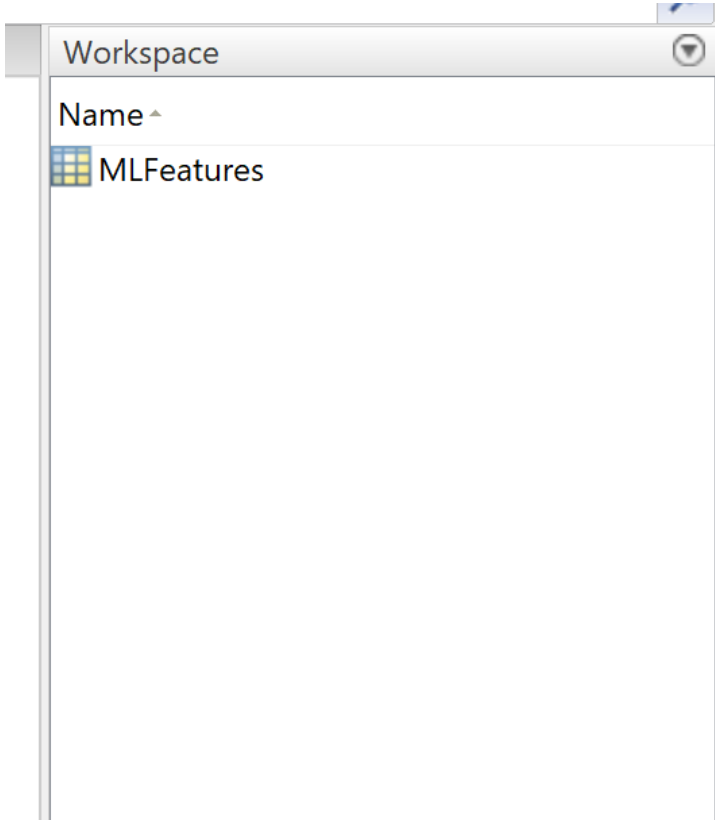


	A	B	C	D	E	F	G	H	I
									MLFeatu
	Mean	Variance	Skewness	Kurtosis	VI	MaxPower	Period	Class	VarName9
	Number	Number	Number	Number	Number	Number	Number	Categorical	Text
1	Mean	Variance	Skewness	Kurtosis	VI	MaxPower	Period	Class	
2	137.6947	1.4931e+04	-0.0570	1.1974	3.7559	0.7720	23.9898	0	
3	57.5828	7.7791e+03	1.4479	3.7214	3.2619	0.1253	23.9900	0	
4	48.5377	3.3758e+03	0.6151	1.5916	3.2560	0.3246	23.9898	0	
5	42.6699	3.3260e+03	0.8575	1.9498	2.5431	0.2878	17.9073	0	
6	56.6072	3.0792e+03	0.3957	1.5426	3.0251	0.0982	36.5988	0	
7	46.8282	2.9975e+03	0.7017	1.8305	2.8005	0.2328	36.5988	0	
8	55.6313	3.4423e+03	0.3685	1.3851	3.4885	0.5314	23.9898	0	
9	42.4581	2.8145e+03	0.8780	2.1350	3.0725	0.2011	36.9973	0	
10	38.8513	2.8279e+03	0.9411	2.2011	2.5736	0.2689	36.5988	0	
11	70.6009	3.5217e+03	-0.0570	1.3242	2.1907	0.5913	36.5988	0	
12	145.7006	1.5047e+04	-0.1803	1.2276	3.3206	1.8161	23.9898	0	
13	101.6529	1.2301e+04	0.3818	1.3855	3.9772	0.5617	36.5988	0	

MACHINE LEARNING & TYPES

- How to implement in MATLAB ?

Step 4: Features are in workspace and ready



MACHINE LEARNING & TYPES

➢How to implement in MATLAB ?

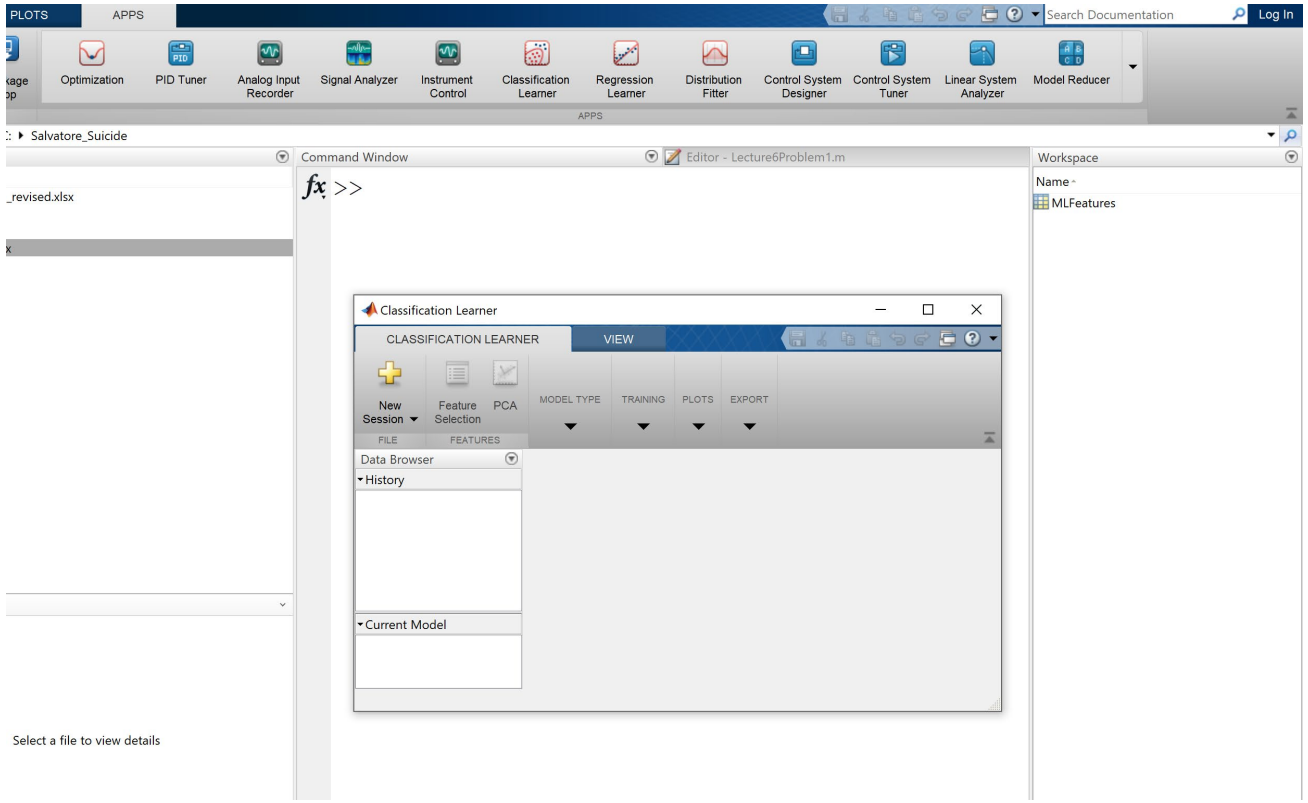
Step 5: Go to Apps,

-click classification learner,

- -select Logistic Regression from Model Type

-click New Session,

-select from Workspace

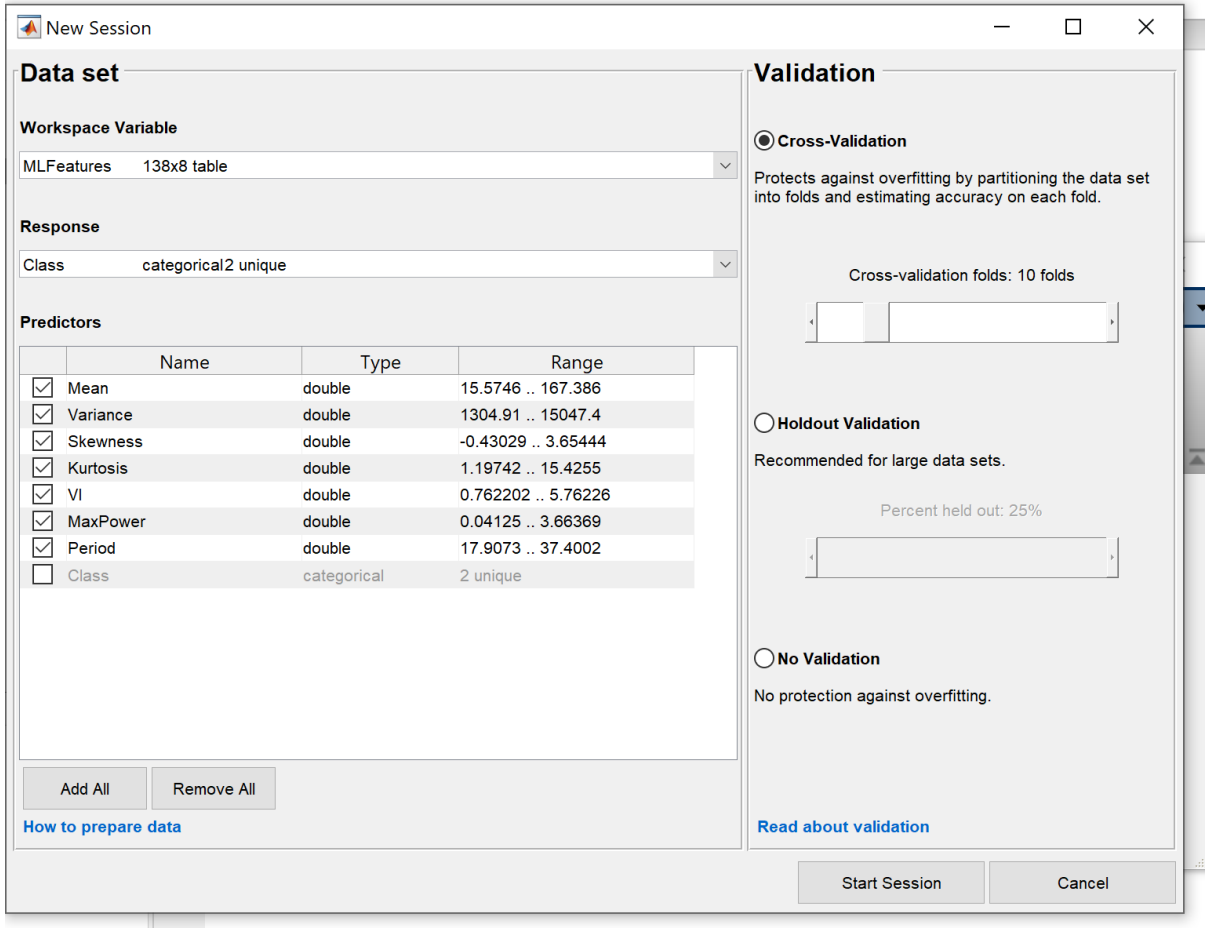


MACHINE LEARNING & TYPES

- How to implement in MATLAB ?

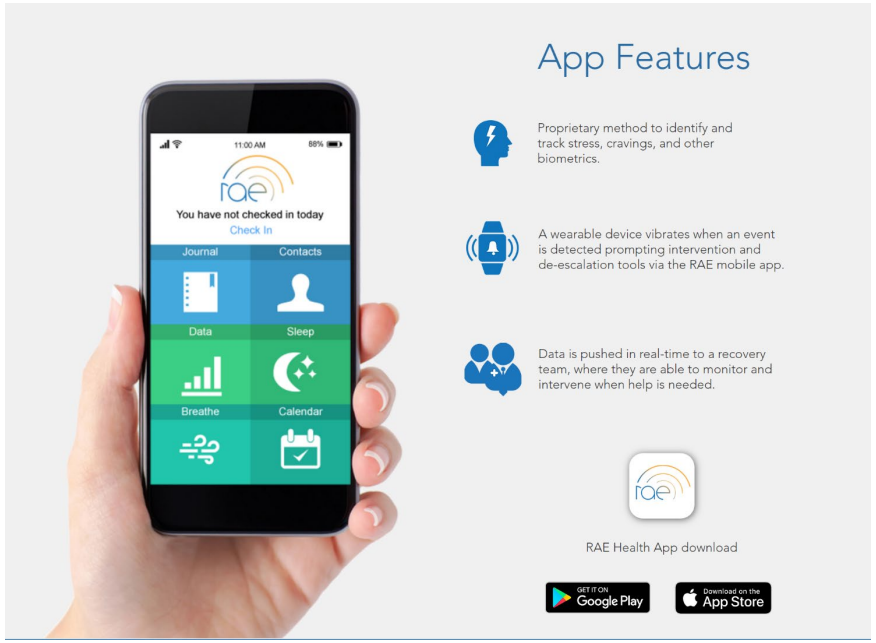
Step 6: Set 10-fold Cross validation

- Start the session

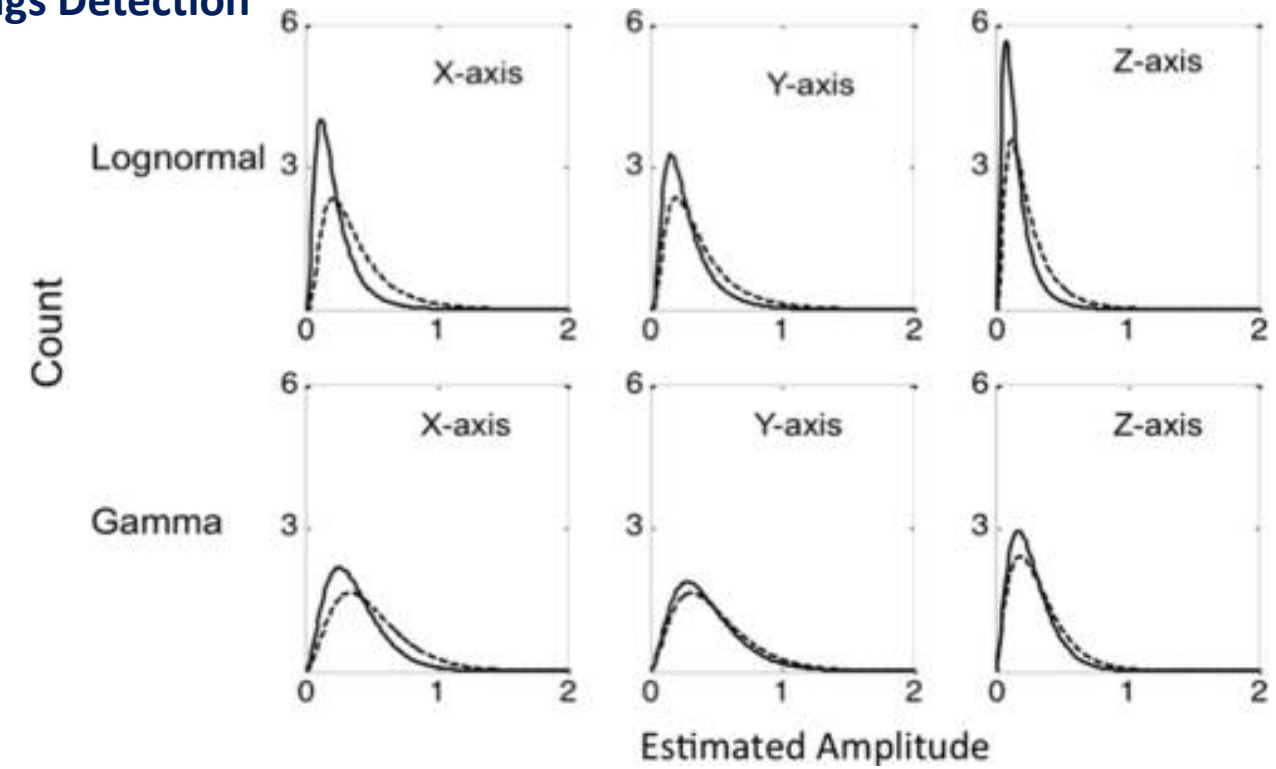


MACHINE LEARNING & TYPES

- NONLINEAR FEATURES



Cravings Detection

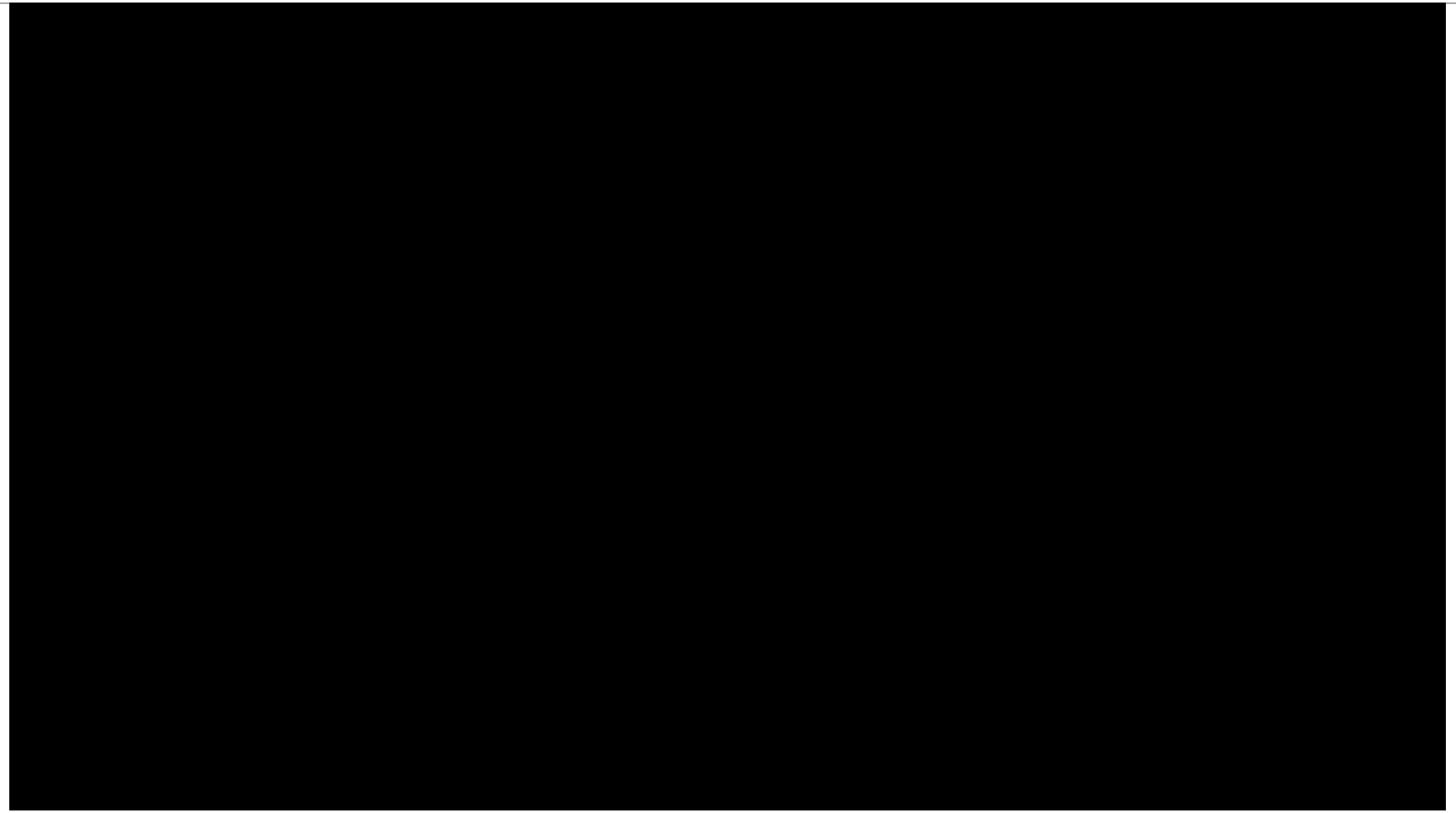


Carreiro, S, Chintha KK, **Shrestha S**, Chapman B, Smelson D, Indic P. Wearable sensor based detection of stress and craving in patients during treatment for substance use disorder: A mixed methods pilot study. Drug and Alcohol Dependence. 2020, 107929

MACHINE LEARNING & TYPES

2 : No Stress

3: Stress



Stress Detection Algorithm

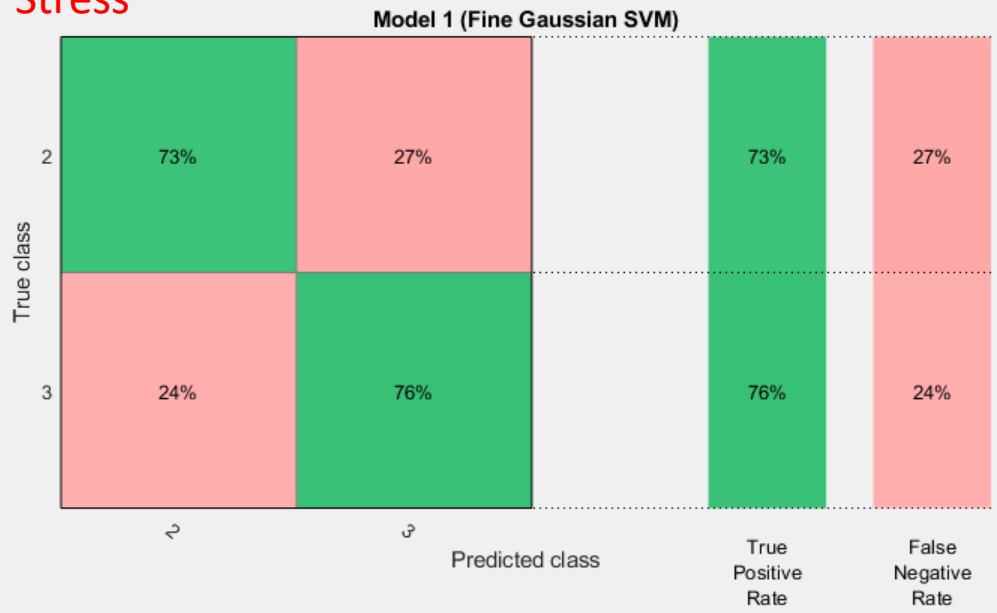
Sloke Shrestha

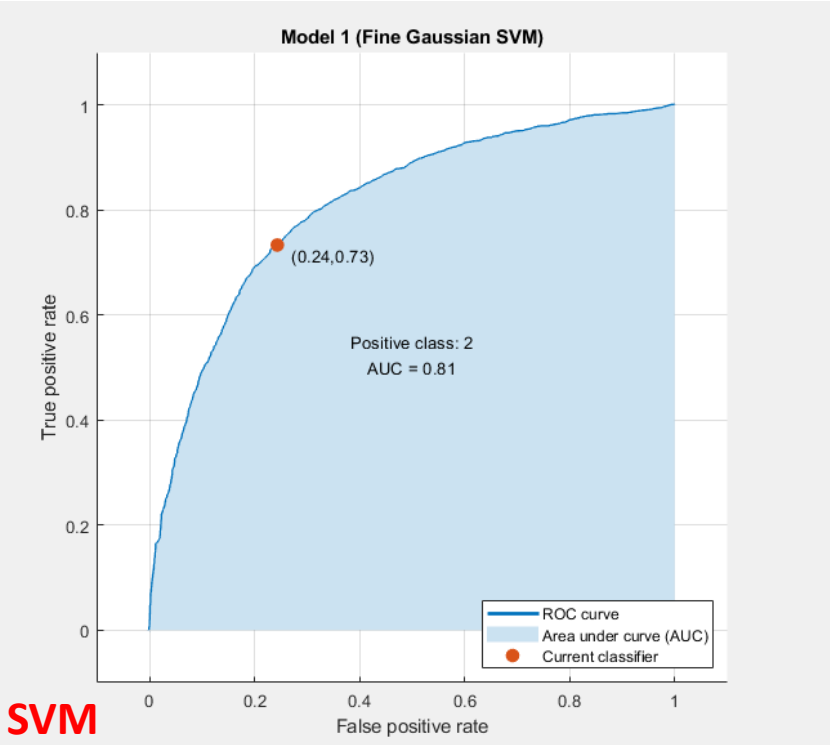
MACHINE LEARNING & TYPES

Stress Detection Algorithm

2 : No Stress

3: Stress





Fine Gaussian SVM

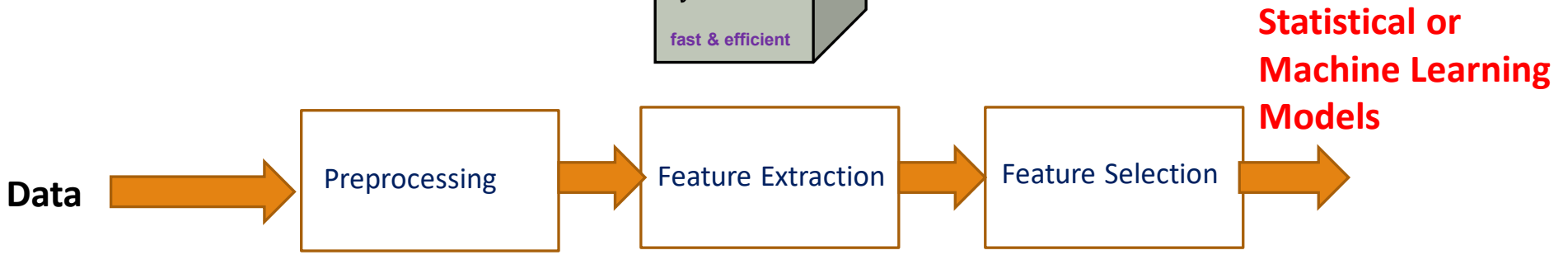
Accuracy ~75%

SUMMARY

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

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

- Ready to go features and machine learning models



https://calendly.com/pindic/30min?month=2021-12

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