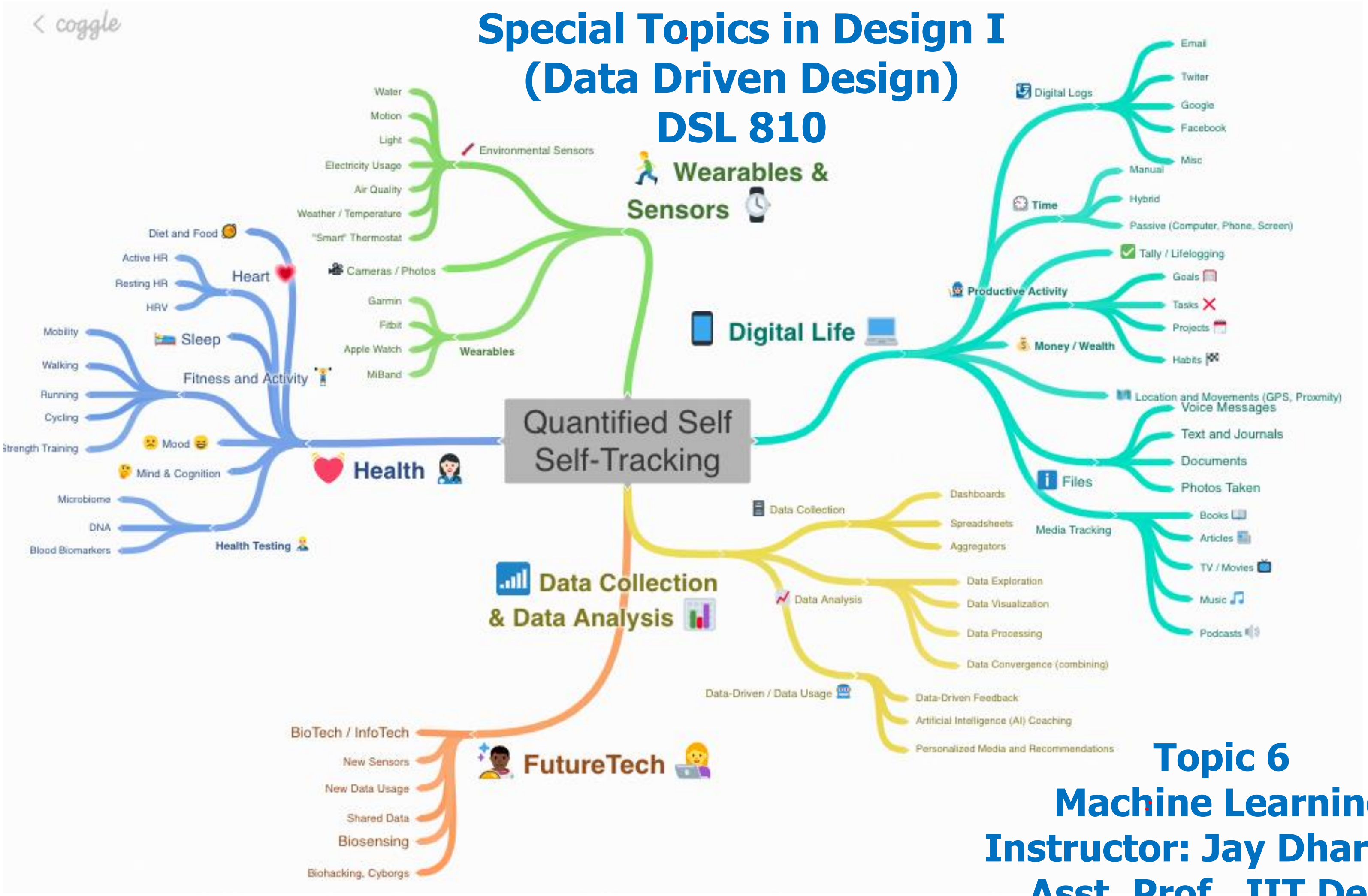
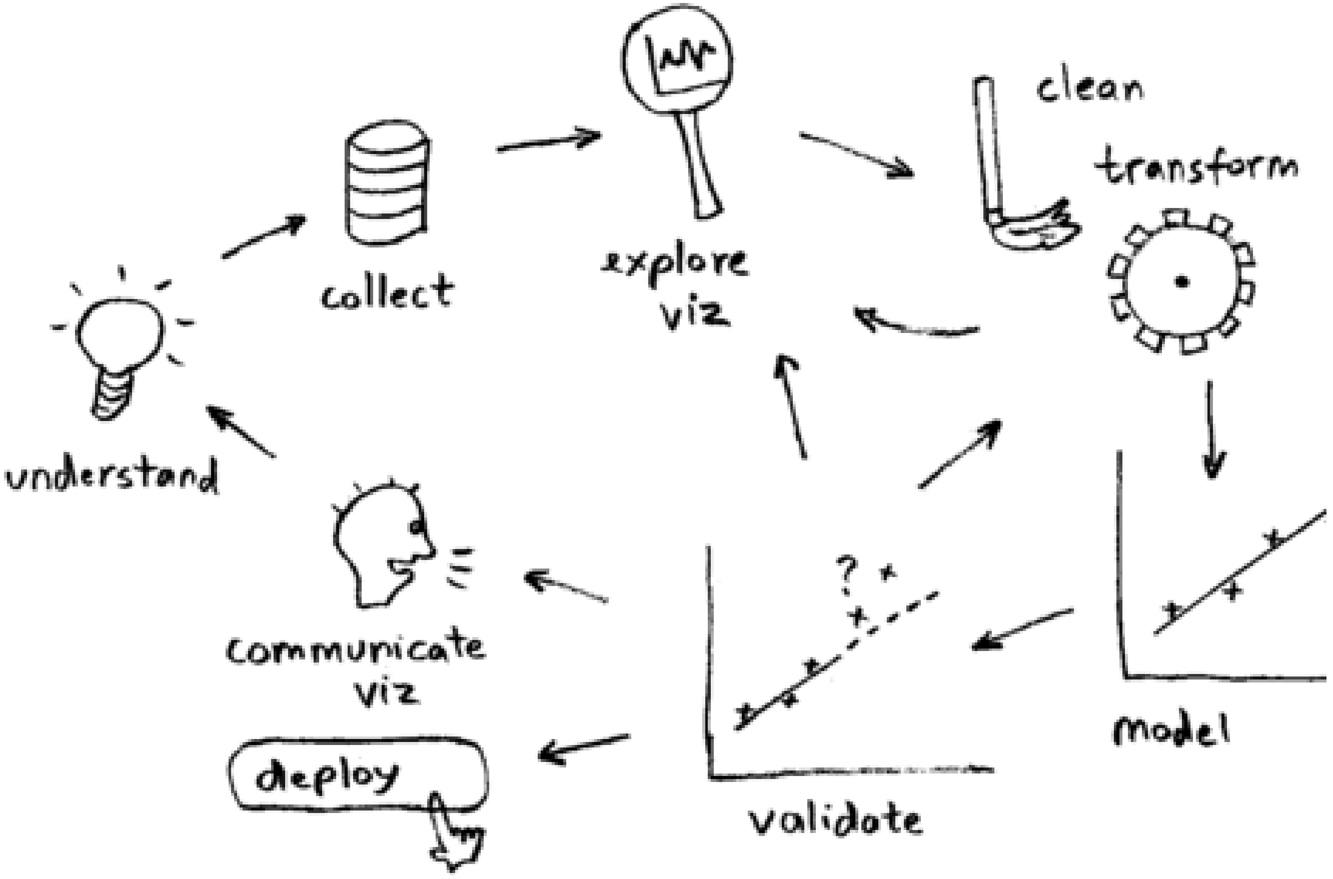
Special Topics in Design I
(Data Driven Design)
DSL 810



Topic 6
Machine Learning
Instructor: Jay Dhariwal,
Asst. Prof., IIT Delhi

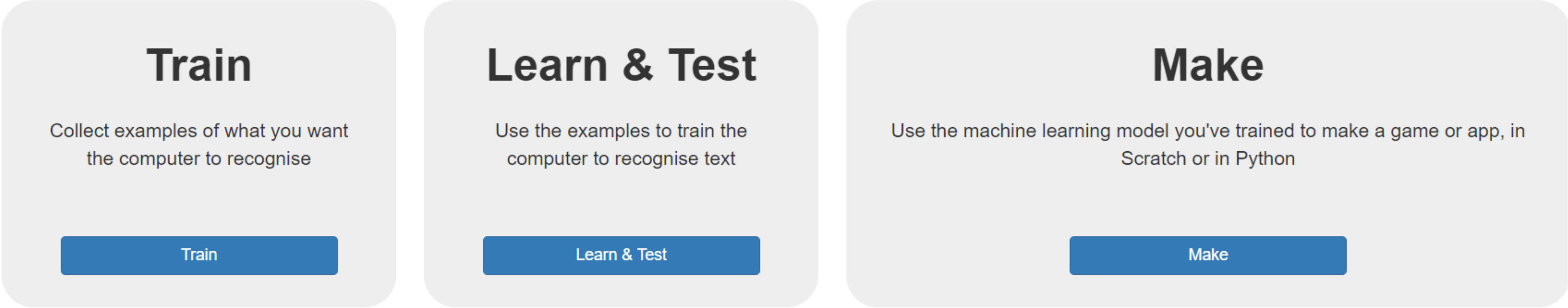
Dated: 19th November, 2020

Source: http://www.markwk.com/quantified-self-mind-map.html



Data science workflow

Source: http://datascience.la/data-science-toolbox-survey-results-surprise-r-and-python-win/



Train

Collect examples of what you want the computer to recognise

Learn & Test

Use the examples to train the computer to recognise text

Make

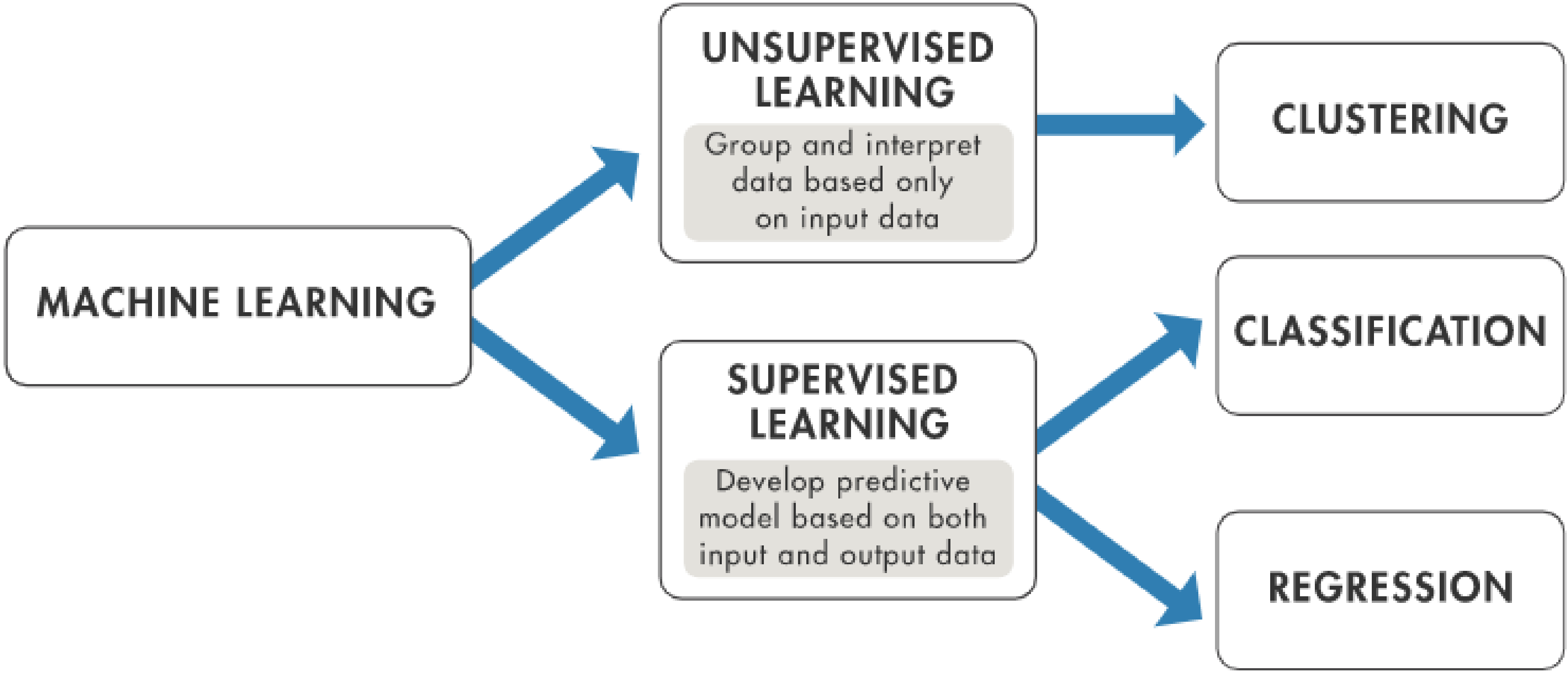
Use the machine learning model you've trained to make a game or app, in Scratch or in Python

Source: https://machinelearningforkids.co.uk/

Introduction to ML

- Google's AI AlphaGo Is Beating Humanity At Its Own Games
- Elon Musk on AI
- Eric Schimdt: AI assisted health care, Self driving cars
- Vinod Khosla: Generative Design
- Machine learning for optimization

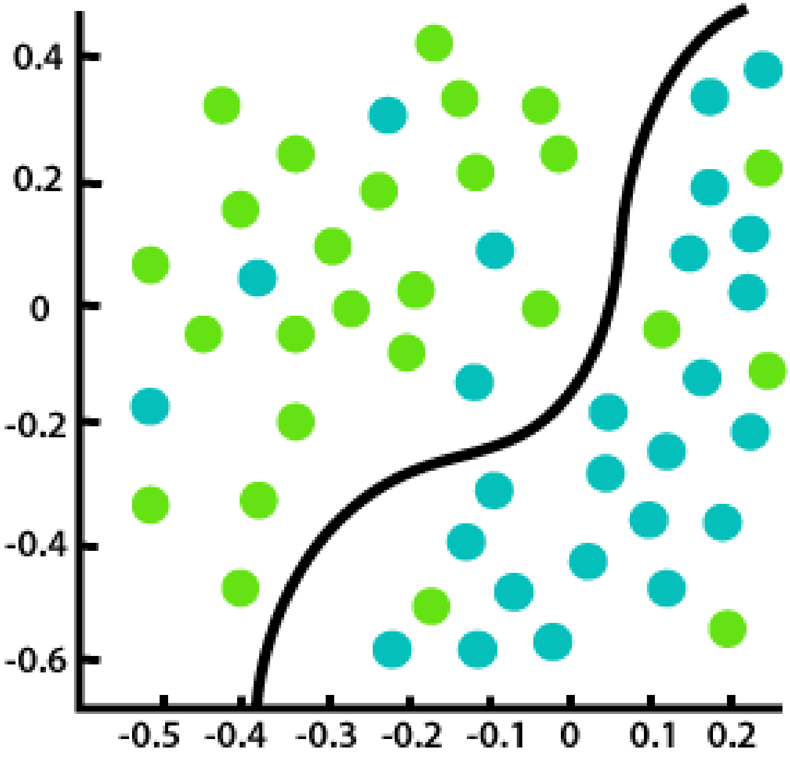
Machine Learning Techniques



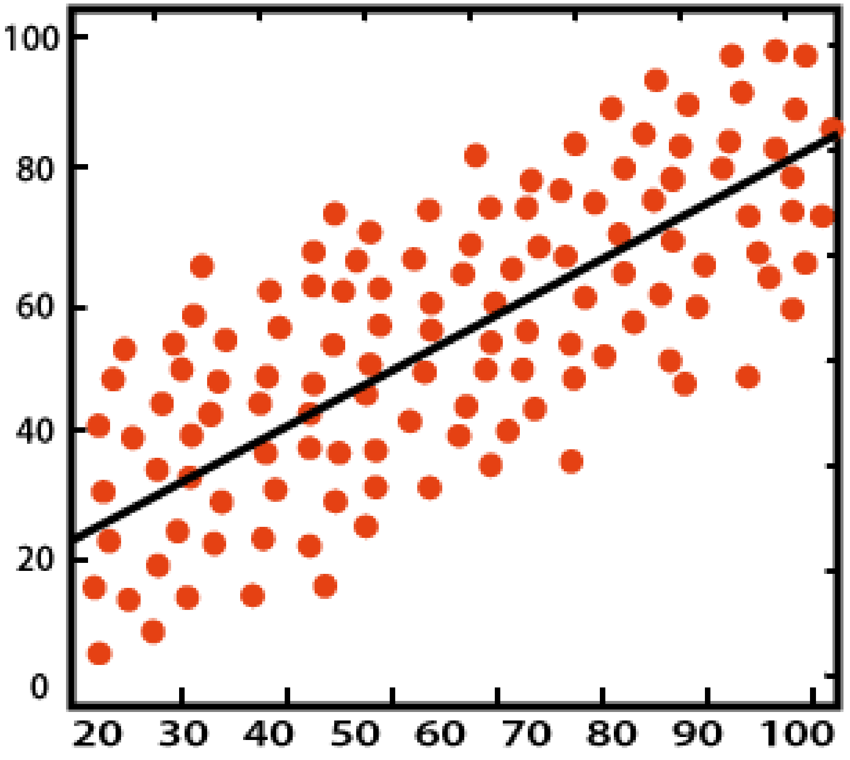
MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

Source: MATLAB

Classification vs. Regression

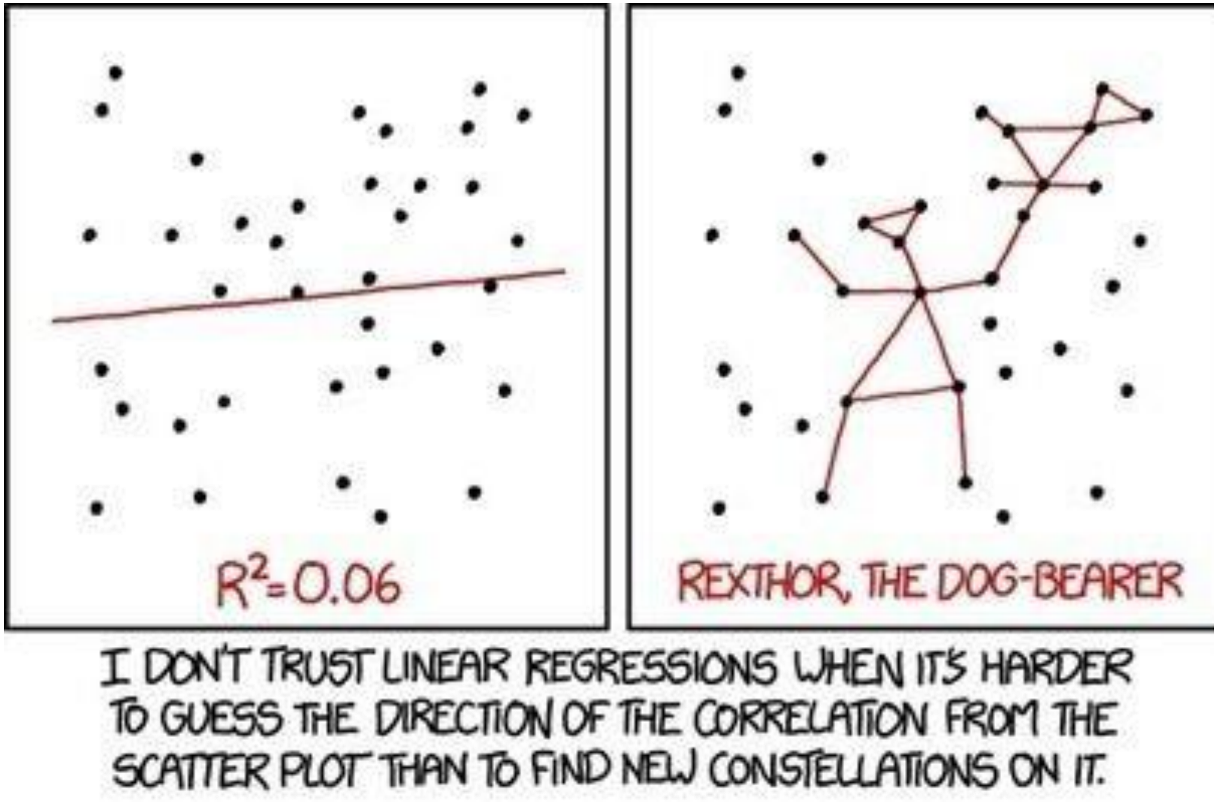


Classification



Regression

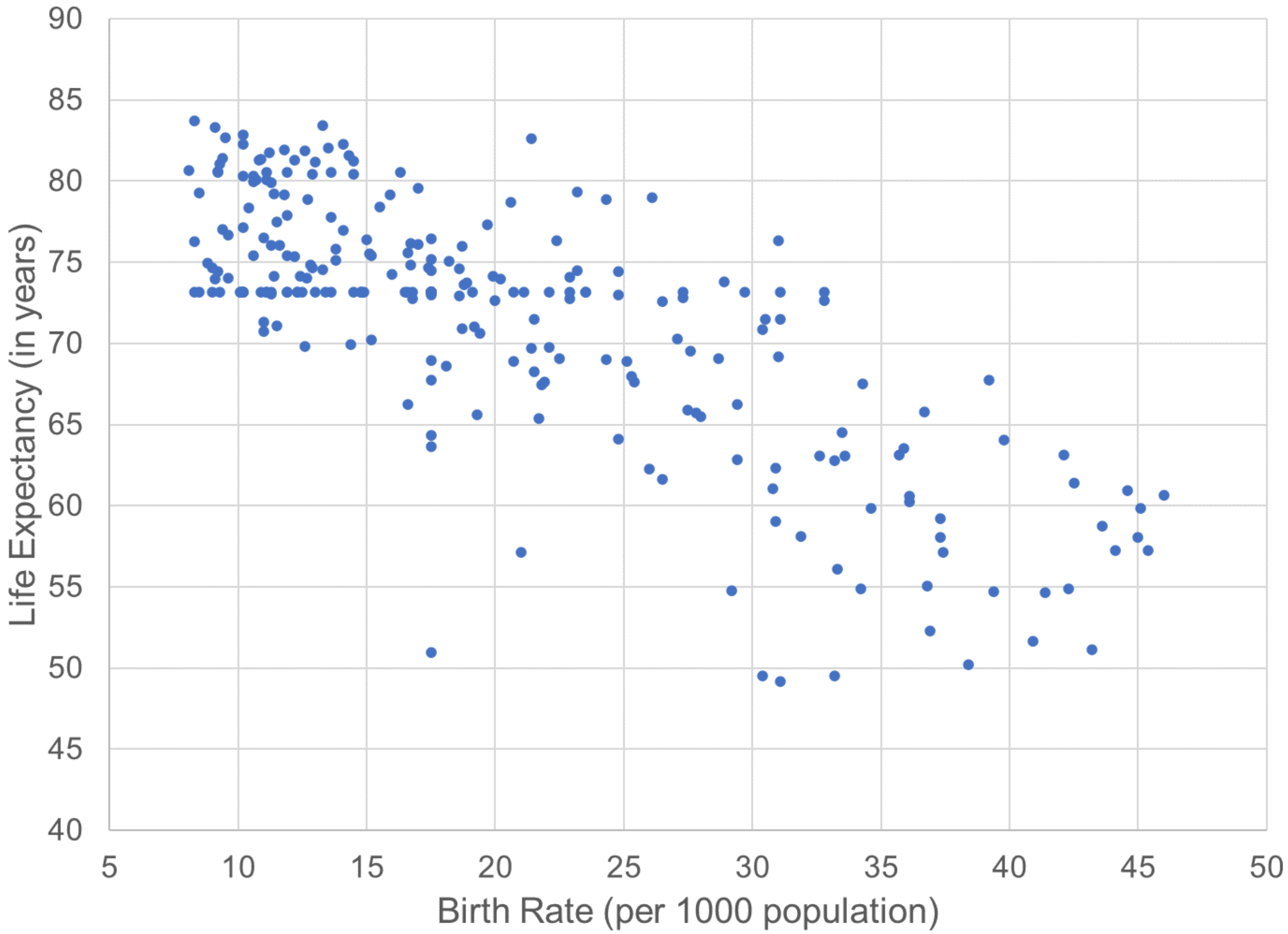
Regression


$R^2 = 0.06$
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

- Most widely used to analyze multifactor data
- Equation to express relationship between the response and predictor variables
- Elegant math and statistical theory
- Theory and practical real world applications
- Applications of regression in engineering, applied sciences, management, life sciences, social sciences, etc.

Source of image: https://www.pinterest.com/pin/203717583129176988/?autologin=true

Linear regression example

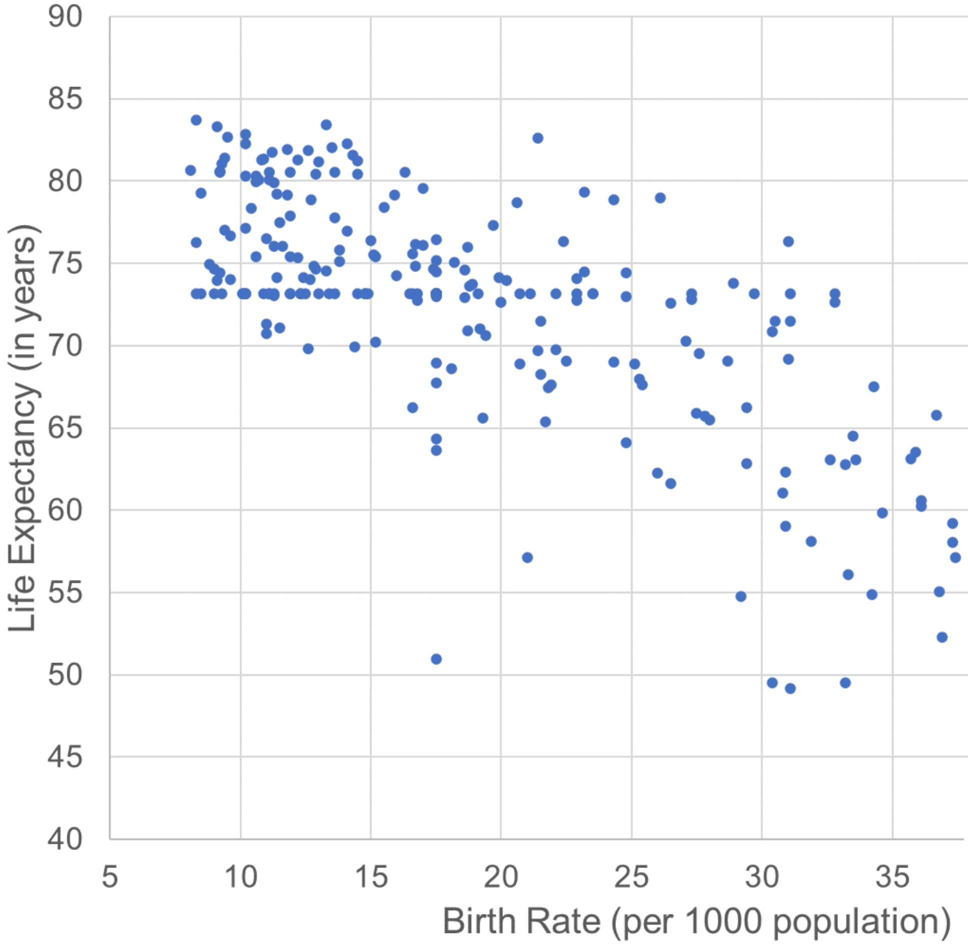


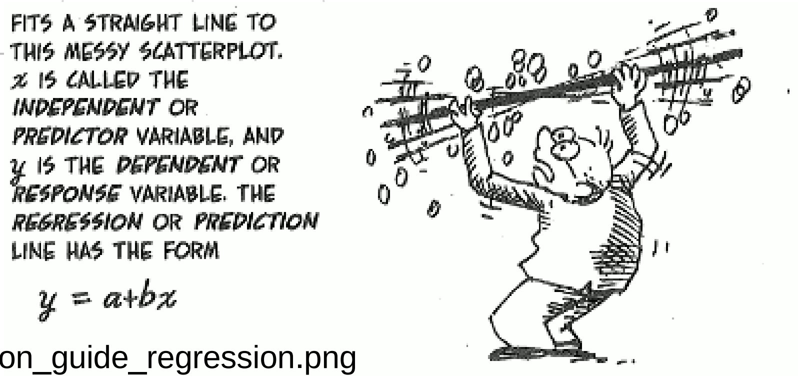
Life Expectancy = f(Birth Rate)

Regression Analysis on Life Expectancy

Dataset for Linear Regression

Linear regression example




FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE *INDEPENDENT* OR *PREDICTOR* VARIABLE, AND y IS THE *DEPENDENT* OR *RESPONSE* VARIABLE. THE *REGRESSION* OR *PREDICTION* LINE HAS THE FORM

$$y = a+bx$$

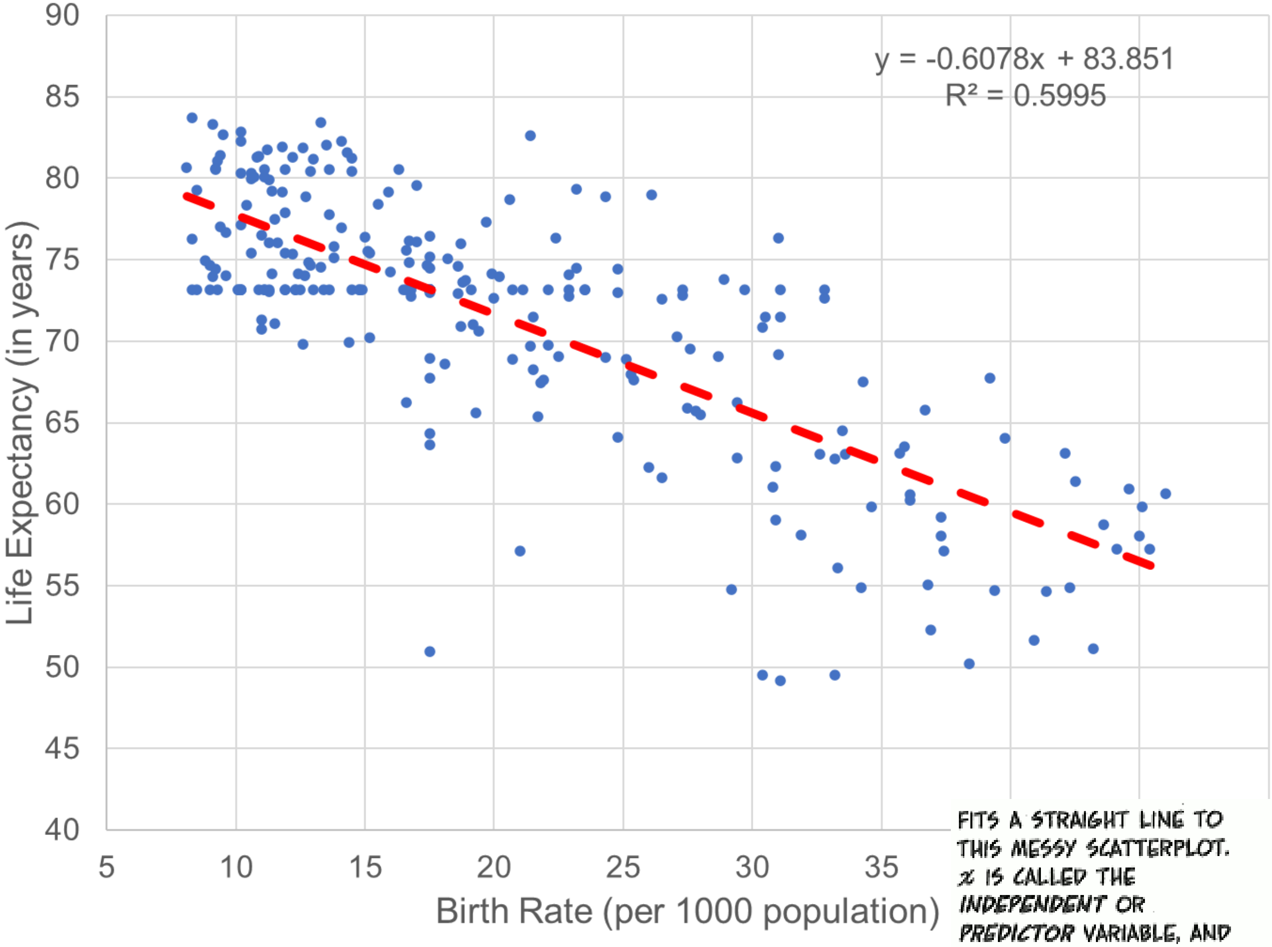
Life Expectancy = f(Birth Rate)

Regression Analysis on Life Expectancy

Dataset for Linear Regression

Source: https://madhureshkumar.files.wordpress.com/2015/07/cartoon_guide_regression.png

Linear regression example



FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE *INDEPENDENT* OR *PREDICTOR* VARIABLE, AND y IS THE *DEPENDENT* OR *RESPONSE* VARIABLE. THE *REGRESSION* OR *PREDICTION* LINE HAS THE FORM

$$y = a + bx$$

Life Expectancy = f(Birth Rate)

[Regression Analysis on Life Expectancy Dataset for Linear Regression]

Source: https://madhureshkumar.files.wordpress.com/2015/07/car

Linear regression assumptions

Dependent Variable

Population Y intercept

Population Slope Coefficient

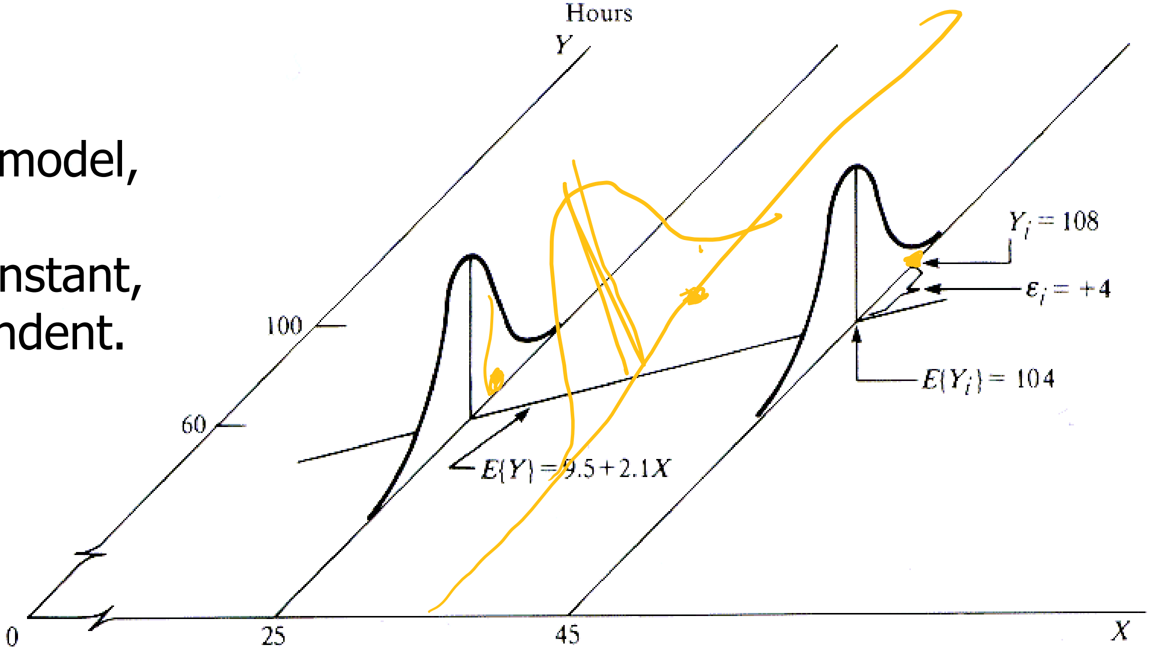
Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

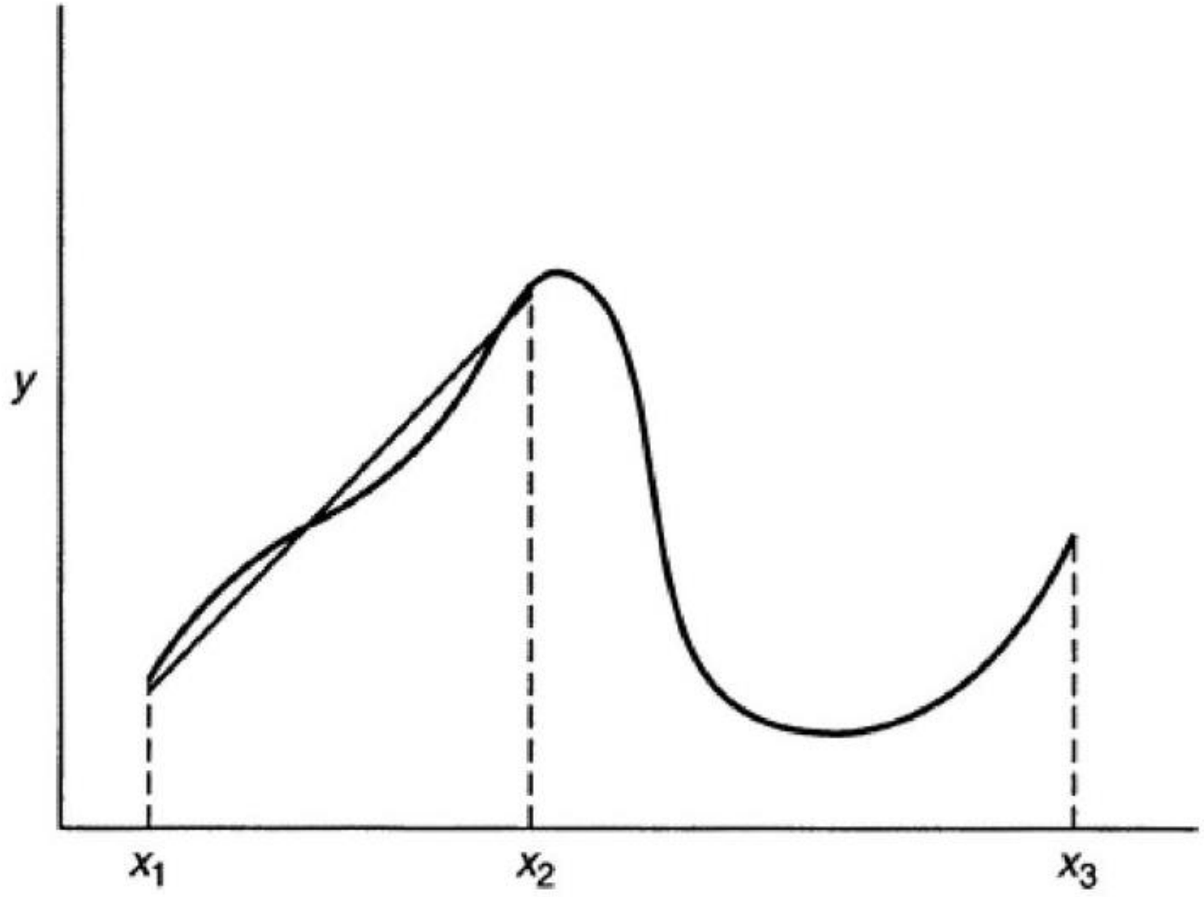
$N(0, \sigma^2)$

Assumptions:
Linearity of regression model,
Mean error is zero,
Variance of errors is constant,
Error terms are independent.



Source: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html

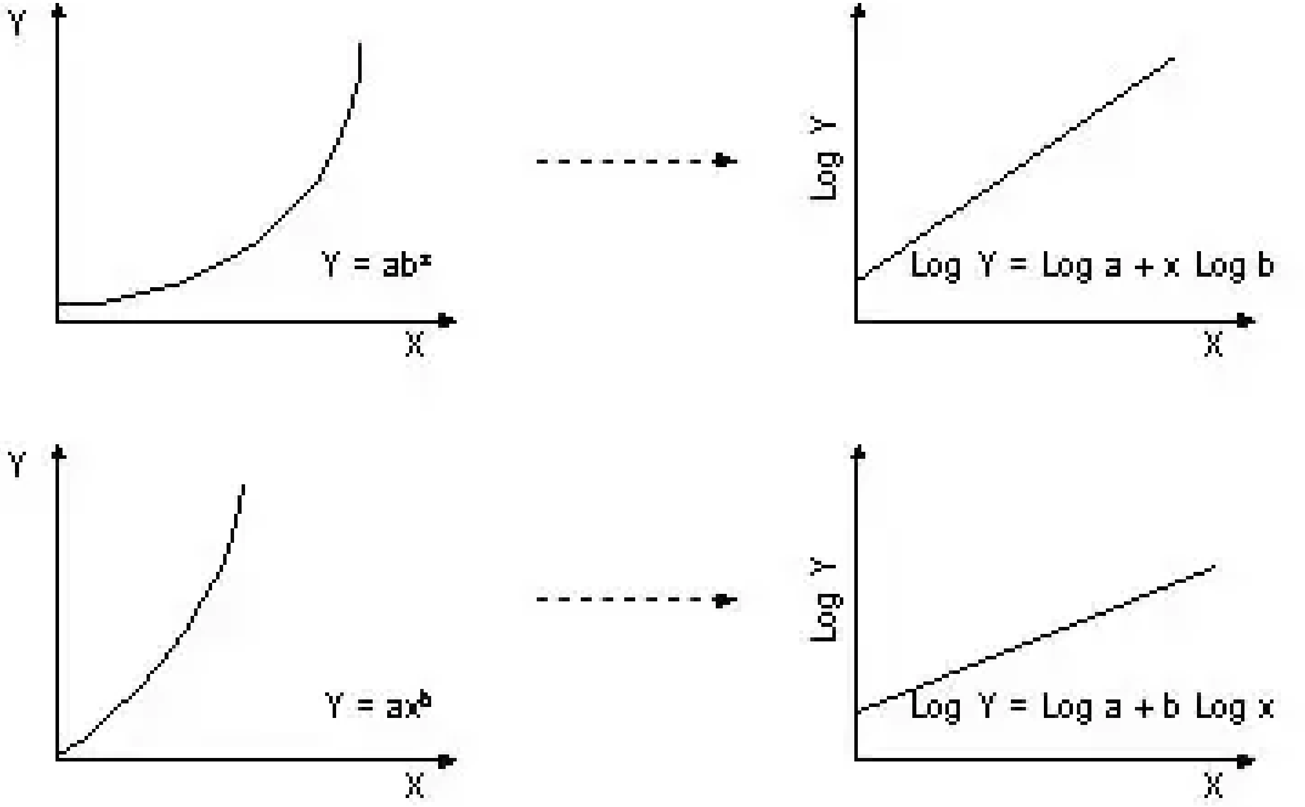
Linear regression assumptions



The danger of extrapolation in regression.

Linear Regression Analysis 5th
edition Montgomery, Peck & Vining

Linear regression transformations



Y
X
$Y = ab^x$
Log Y
X
$\text{Log } Y = \text{Log } a + x \text{ Log } b$
Y
X
$Y = ax^b$
Log Y
X
$\text{Log } Y = \text{Log } a + b \text{ Log } x$

Box cox transformation for linear regression

Source: https://www.quora.com/What-is-log-transformation-in-regression-analysis

Linearizable functions

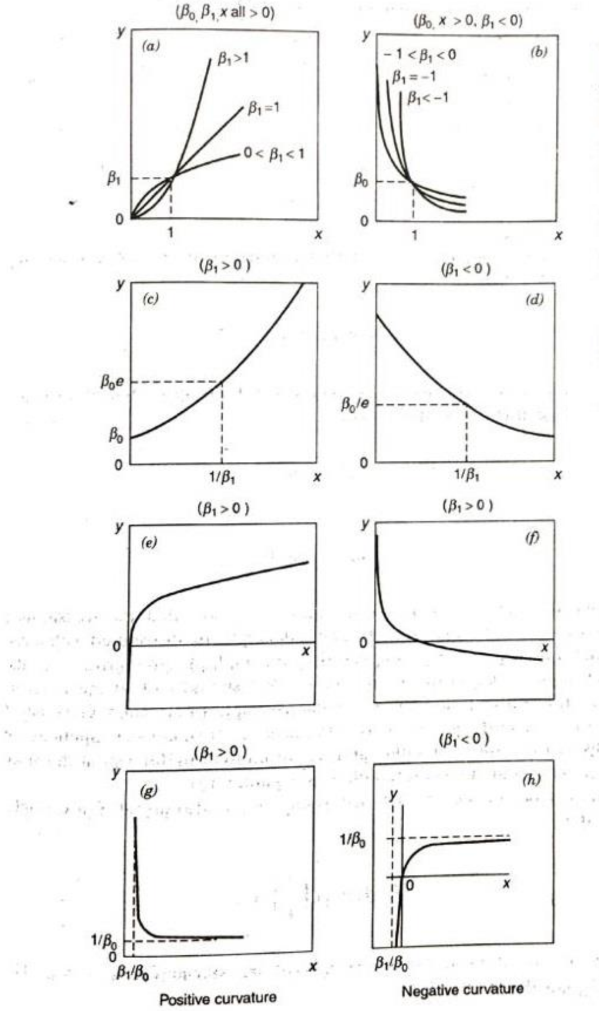


Figure 5.4 Linearizable functions. (From Daniel and Wood [1980], used with permission of the publisher.)

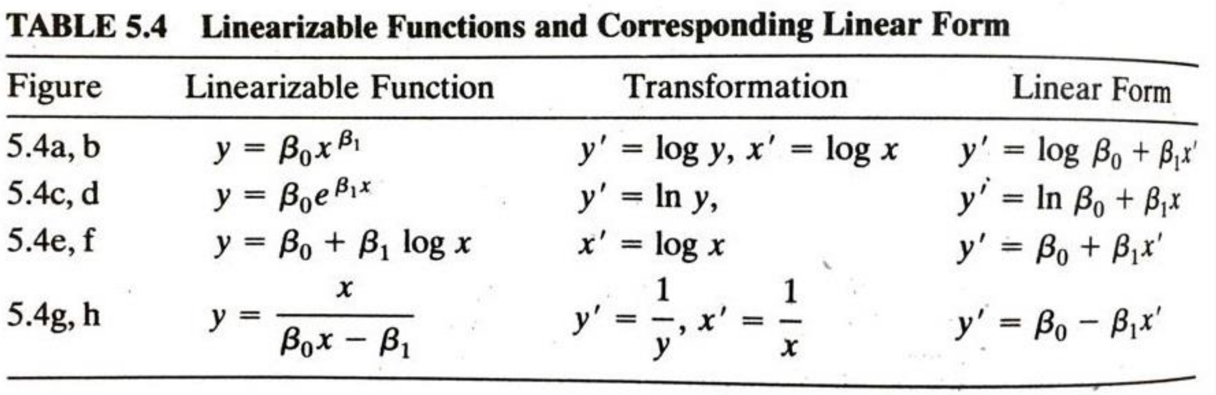


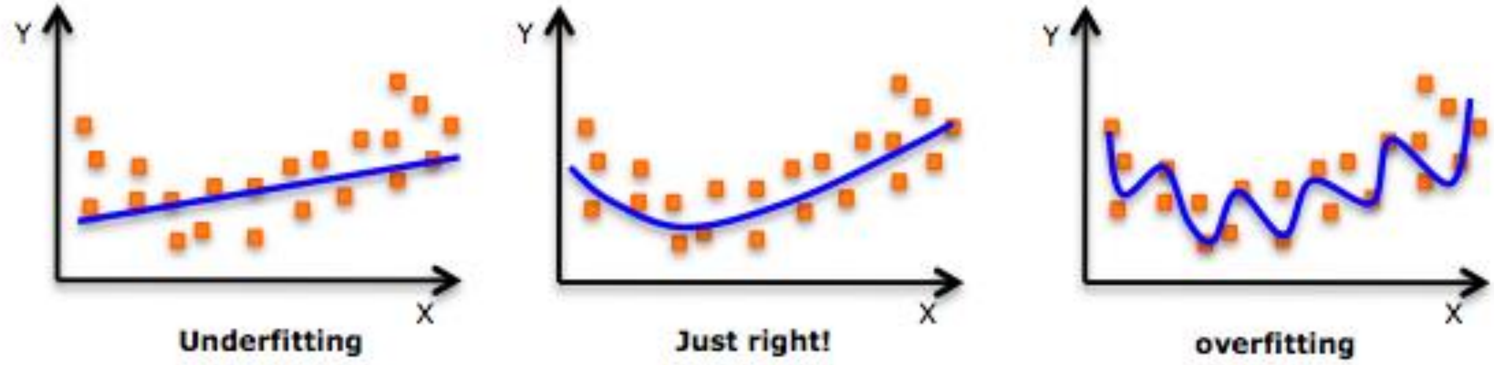
TABLE 5.4 Linearizable Functions and Corresponding Linear Form

Figure	Linearizable Function	Transformation	Linear Form
5.4a, b	$y = \beta_0 x^{\beta_1}$	$y' = \log y,\ x' = \log x$	$y' = \log \beta_0 + \beta_1 x'$
5.4c, d	$y = \beta_0 e^{\beta_1 x}$	$y' = \ln y,$	$y' = \ln \beta_0 + \beta_1 x$
5.4e, f	$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y' = \beta_0 + \beta_1 x'$
5.4g, h	$y = \dfrac{x}{\beta_0 x - \beta_1}$	$y' = \dfrac{1}{y},\ x' = \dfrac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

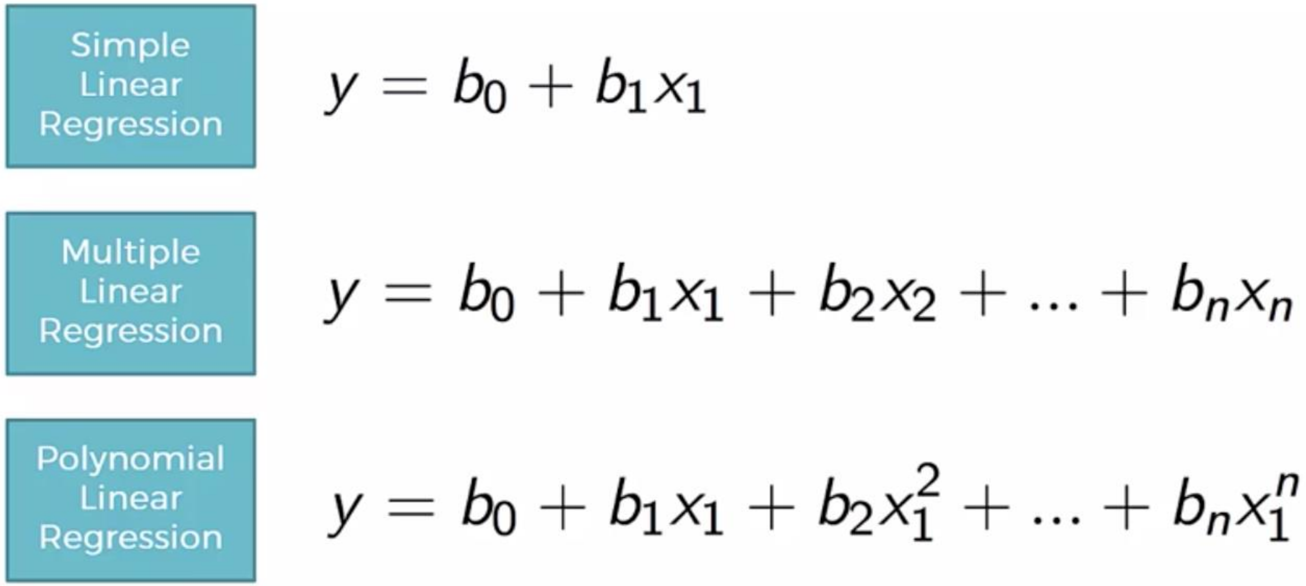
Some non-linear models can be linearized by suitable transformations. Such non-linear models are called transformably linear.

Source: Douglas C Montgomery, Elizabeth A Peck, et al. Introduction to Linear Regression Analysis, 3rd edition, Wiley, 2006

Polynomial regression



Source of image: https://mindmajix.com/polynomial-regression



Simple Linear Regression	$y = b_0 + b_1x_1$
Multiple Linear Regression	$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$
Polynomial Linear Regression	$y = b_0 + b_1x_1 + b_2x_1^2 + \dots + b_nx_1^n$

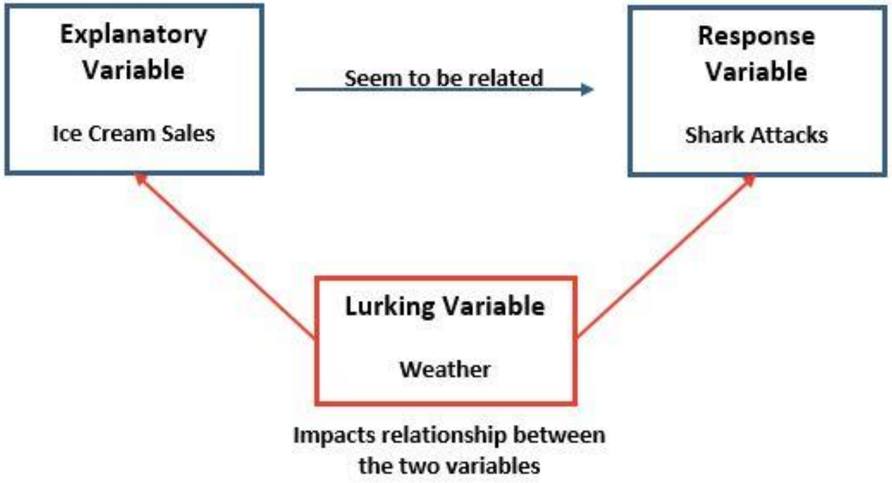
Source of image: https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18

Data collection methods

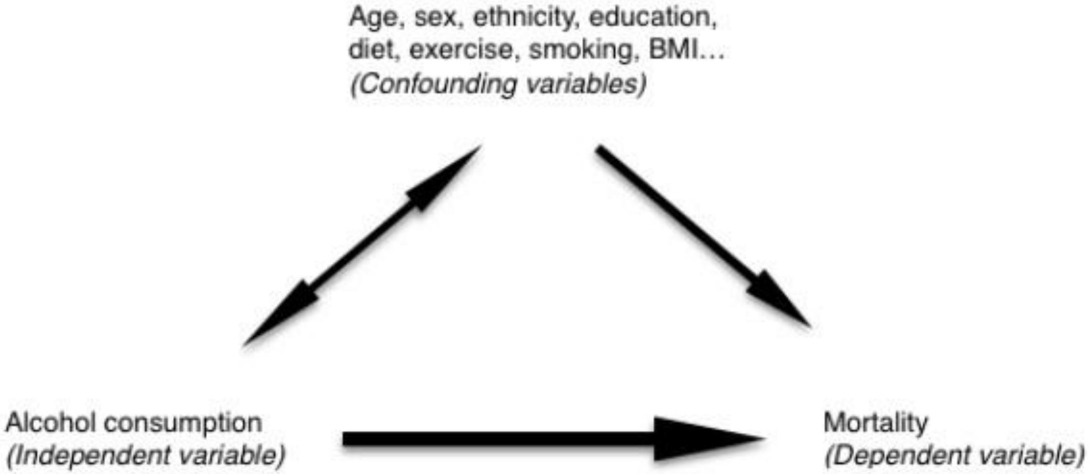
A designed experiment	An observational study
Experimental studies have higher internal validity with experiment repeated under same conditions. Participants assigned to control and treatment groups. The study can be controlled and factors not of interest can be eliminated. Experimental studies can establish causation between variables. If possible, this strategy is preferred.	Observational studies have higher external validity. Results may be applicable to typical practice. In certain cases, experimental studies not possible or not appropriate. Experimental studies are: 1) not ethical. 2) involve rare diseases 3) include variables not possible to manipulate e.g. inherent traits 4) too costly. E.g. comparing the risk for developing lung cancer between smokers vs. non-smokers.

Kang, Hyun. Appropriate design of research and statistical analyses: observational vs. experimental studies. Korean J Anesthesiol 2013 August 65(2): 105-107. http://dx.doi.org/10.4097/kjae.2013.65.2.105

Lurking and confounding variables

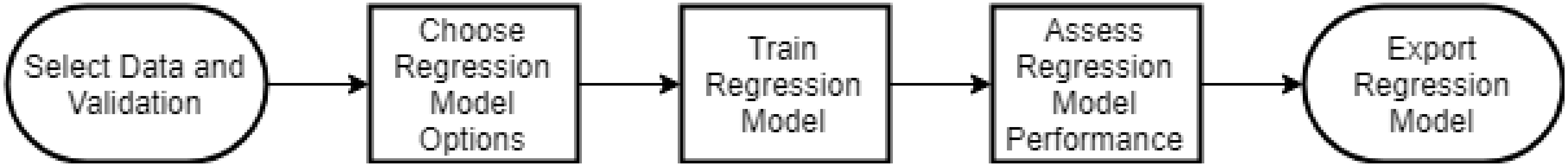


https://www.statology.org/lurking-variables/



https://s4be.cochrane.org/blog/2018/10/01/a-beginners-guide-to-confounding/

Regression Learner Toolbox in MATLAB



Select Data and Validation → Choose Regression Model Options → Train Regression Model → Assess Regression Model Performance → Export Regression Model

Workflow

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

Dependent variable (DV) Independent variables (IVs)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

Residuals, Error, ϵ: $N(0, \sigma^2)$

Regression Learner Toolbox in MATLAB

Regression Analysis on Life Expectancy

Dataset for Linear Regression

Life expectancy = f(Birth Rate, Cancer Rate, Dengue Cases, Environmental Performance Index (EPI), Gross Domestic Product (GDP), Health Expenditure, Heart Disease Rate, Population, Area, Population Density, Stroke Rate)

Model:

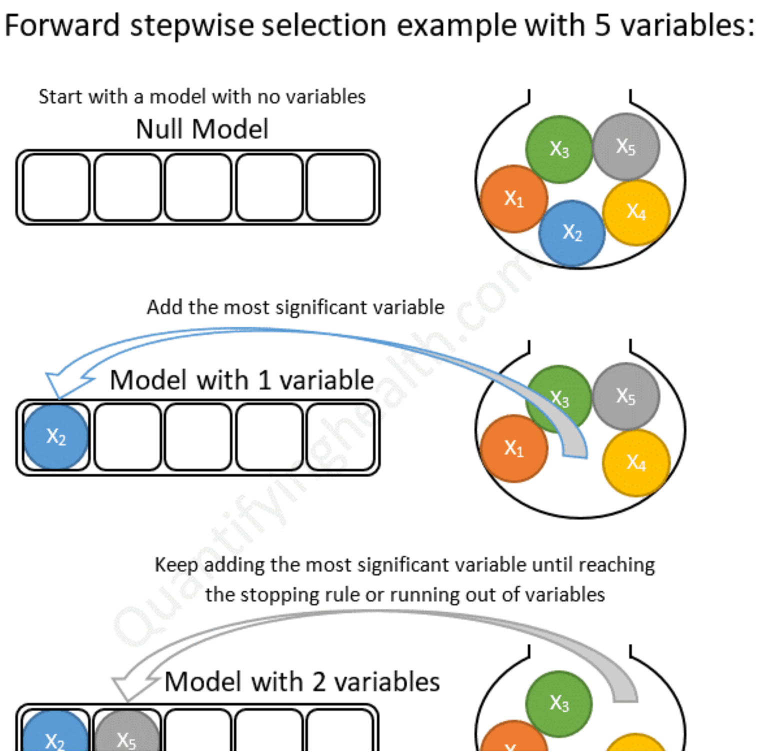
Life expectancy = 74.3305 – 0.3982*Birth Rate + 0.1953*EPI – 0.0627*Stroke Rate

Improvement Focus based on the Model Predictions

Country	Birth Rate	EPI	Stroke Rate	Life expectancy
India	21.8	30.57	71.48	67.14
India	16	36	60	71.23

Stepwise Regression

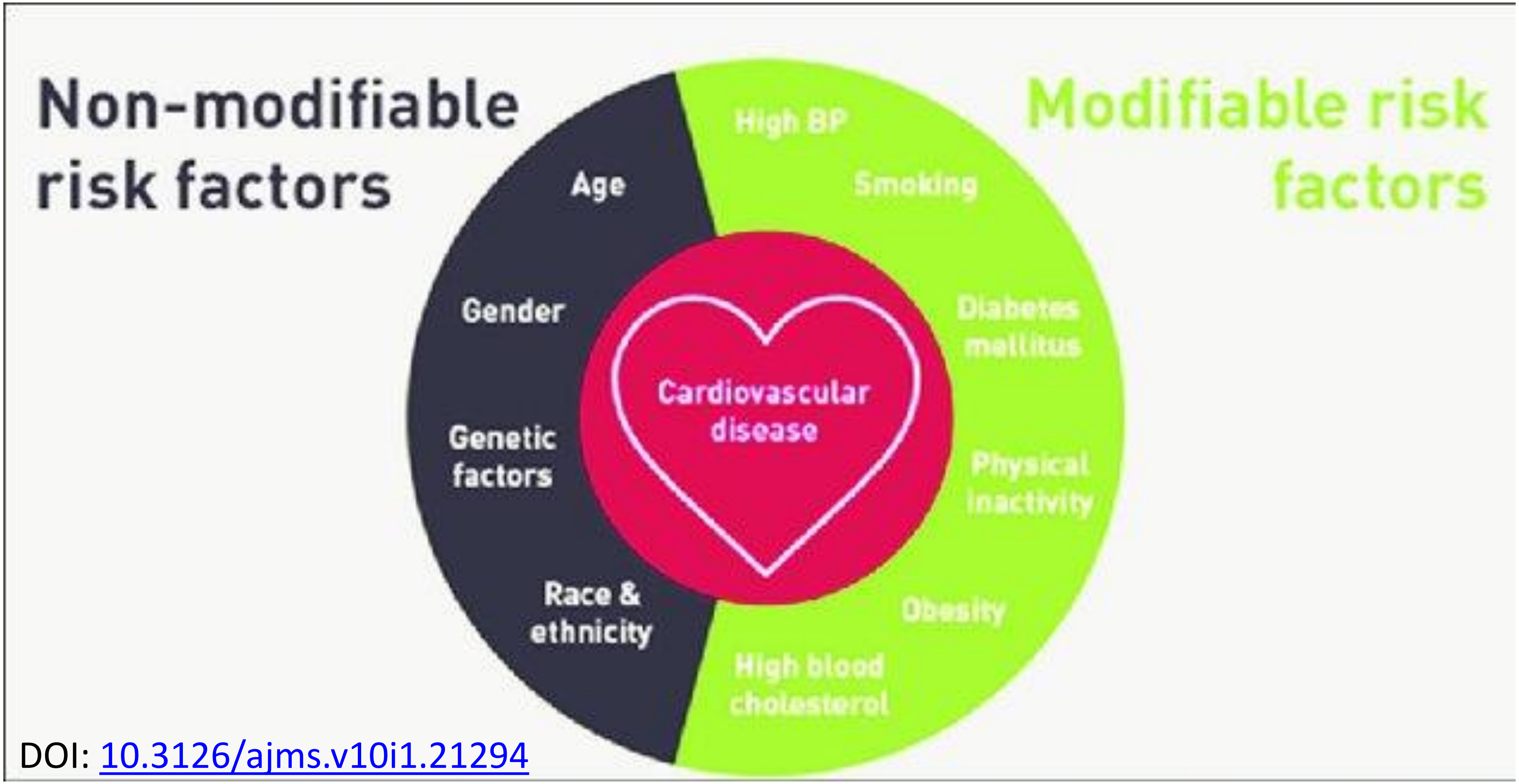
- Y may depend on many independent variables X
- How to find the subset of X's which best predict Y
- There are several criteria such as adjusted R-squared for model selection and many algorithms such stepwise regression
- Stepwise regression is most commonly used
- Higher sample size for stability of the model.


Forward stepwise selection example with 5 variables:
Start with a model with no variables
Null Model
X_1
X_2
X_3
X_4
X_5
Add the most significant variable
Model with 1 variable
X_2
X_1
X_3
X_4
X_5
Keep adding the most significant variable until reaching the stopping rule or running out of variables
Model with 2 variables
X_2
X_5
X_3
Quantifyinghealth.com

https://quantifyinghealth.com/stepwise-selection/

Classification examples

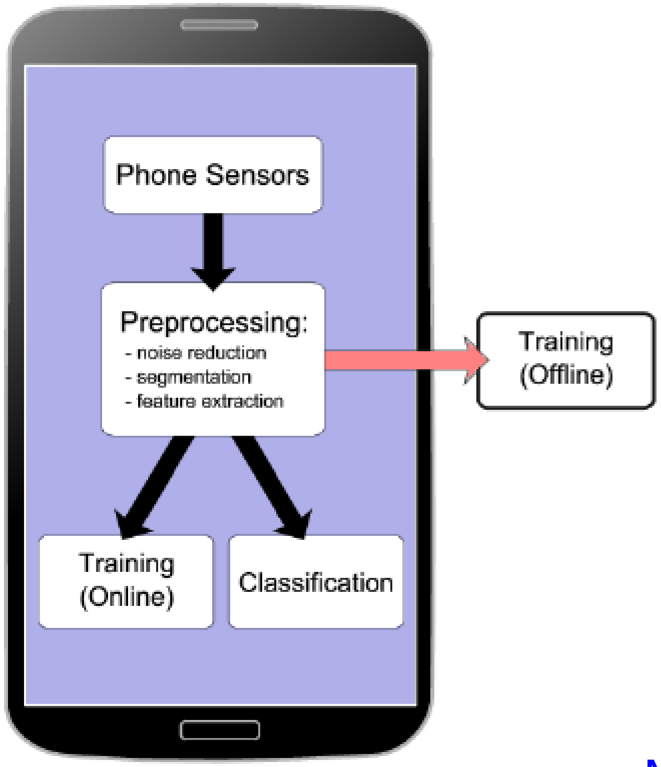
- Heart disease
- Human activity recognition (Sitting, Standing, Walking, Running)
- Text classification (Email Spam or not Spam)
- Image classification

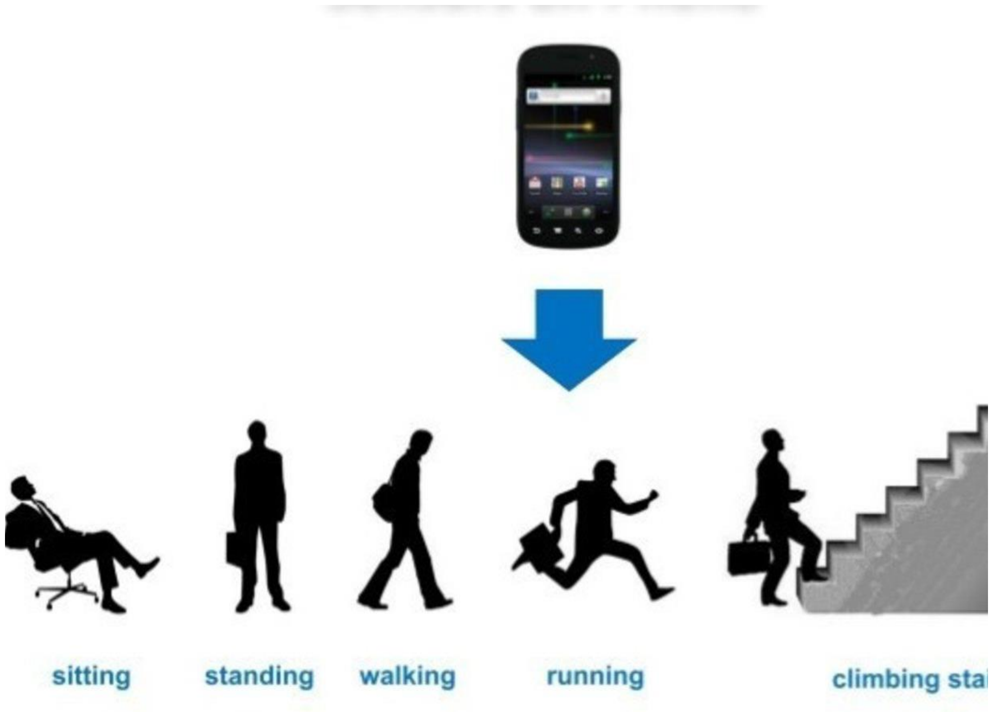


DOI: 10.3126/ajms.v10i1.21294

Heart disease data set source

Human Activity Recognition



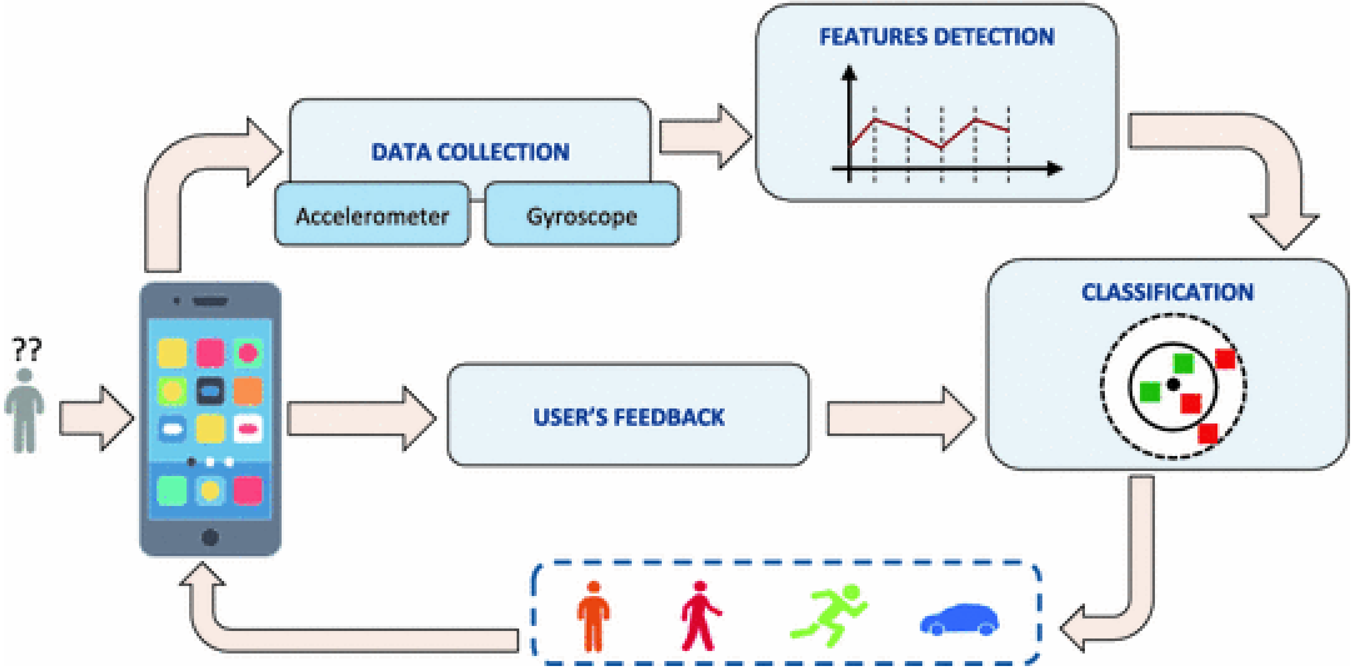


doi:10.3390/s150102059

[MATLAB video tutorial](#)

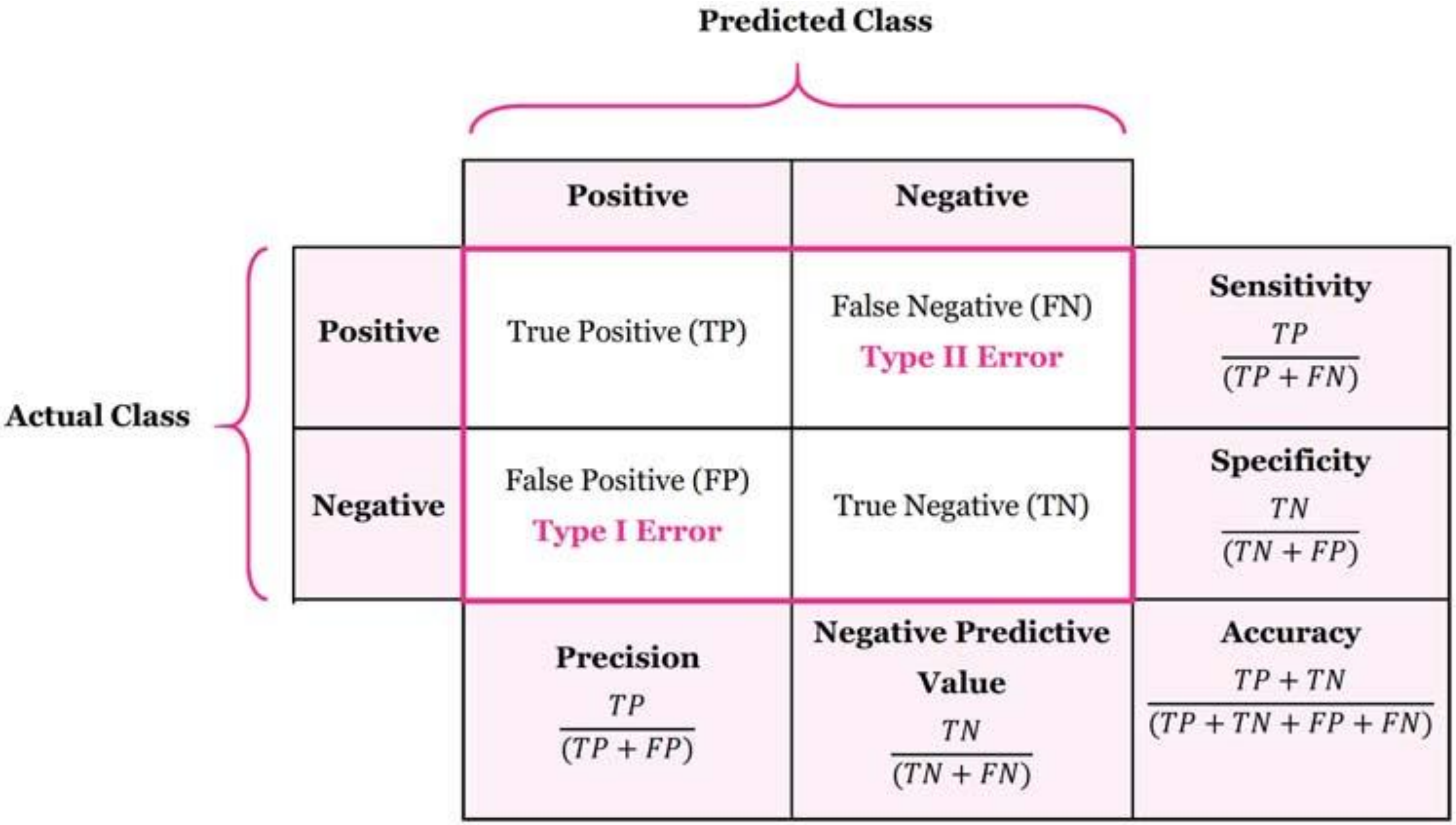
[Data source](#)

Smartphone data for Human Activity Recognition



Source

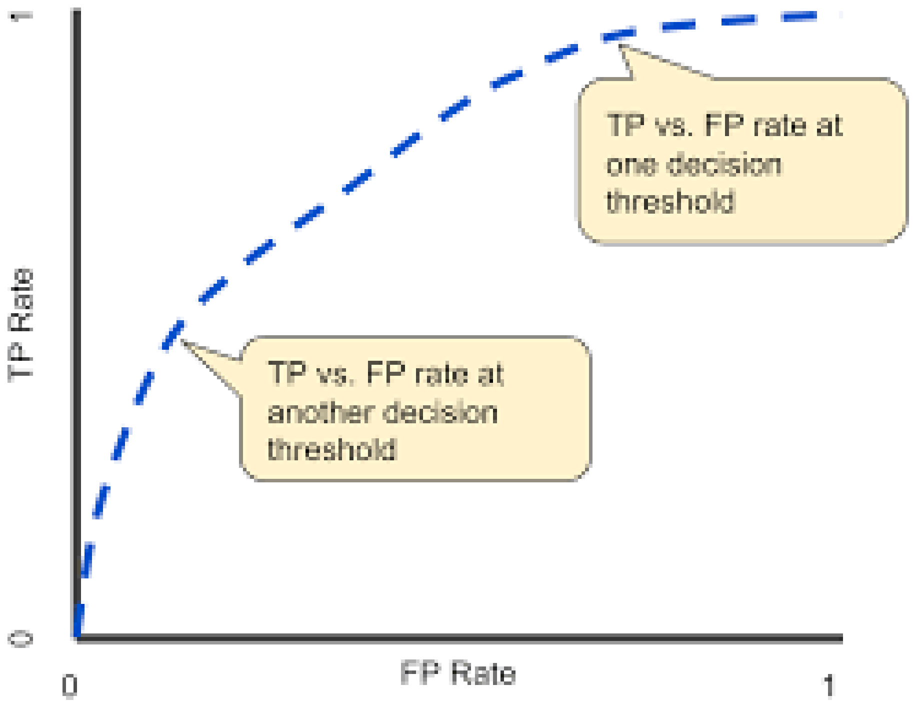
Confusion Matrix



		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

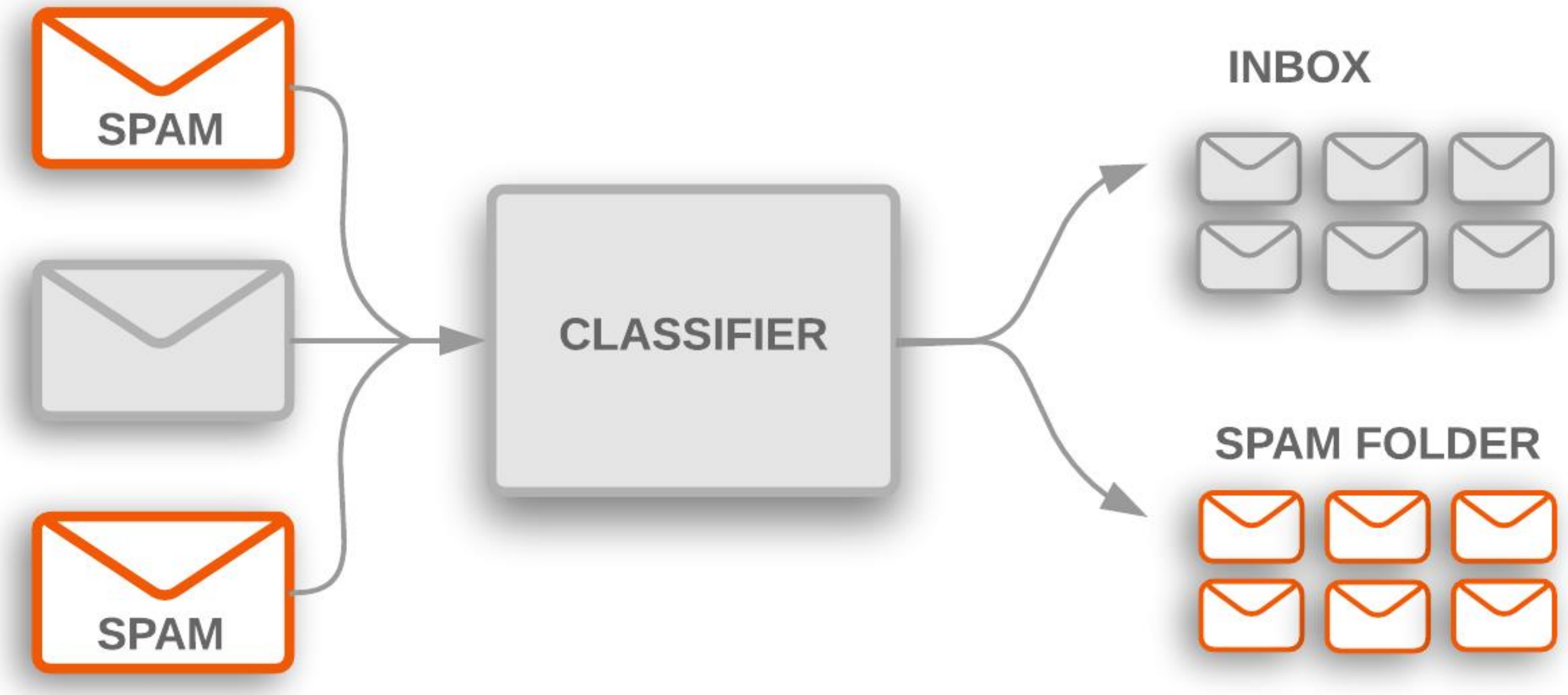
Source: https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html

ROC curve



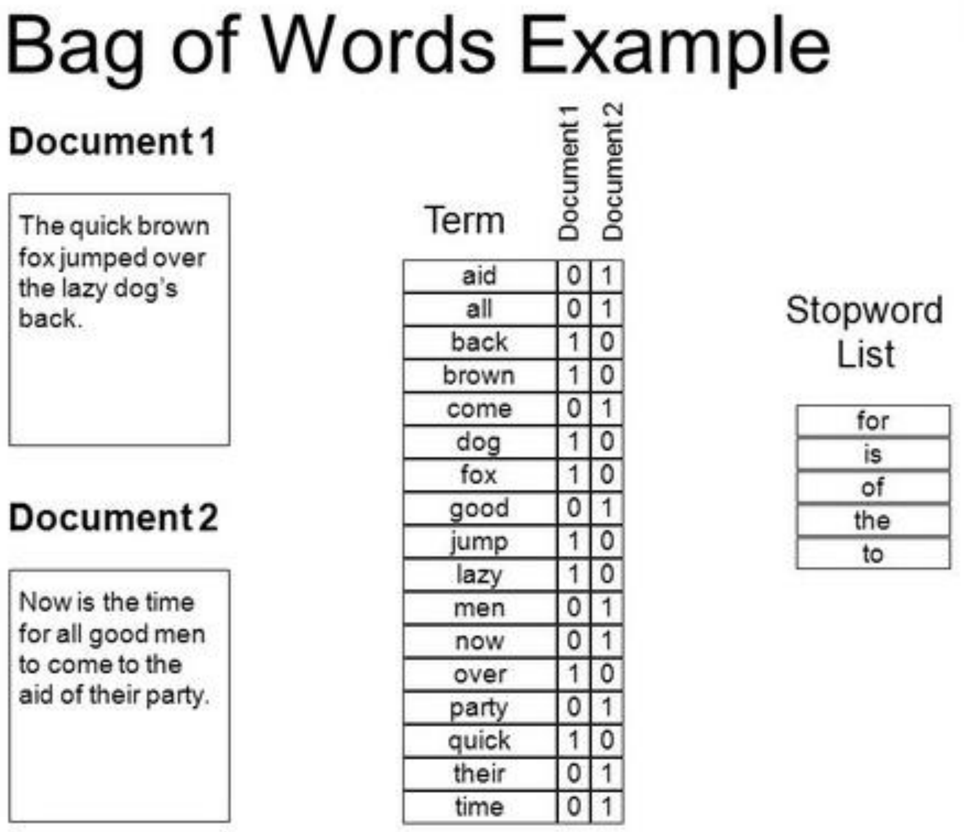
Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Text classification



Source

Text Classification



Bag of Words Example

Document 1

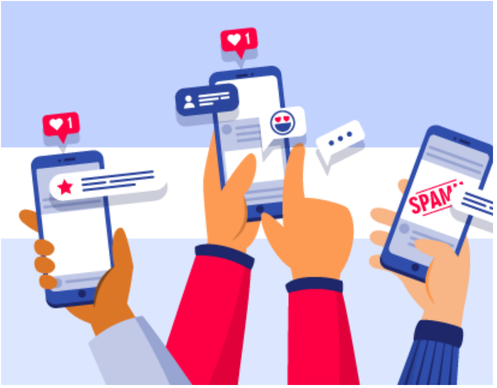
The quick brown fox jumped over the lazy dog's back.

Document 2

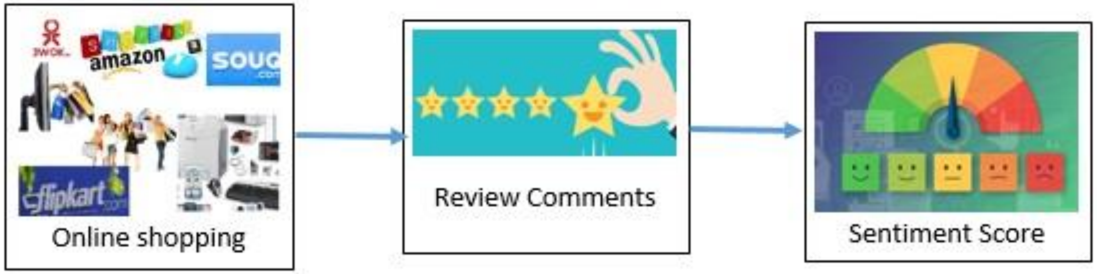
Now is the time for all good men to come to the aid of their party.

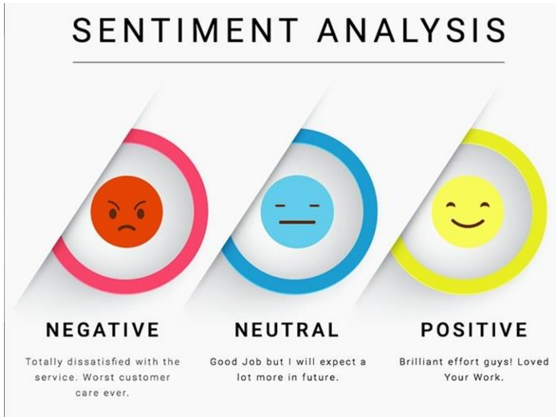
Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

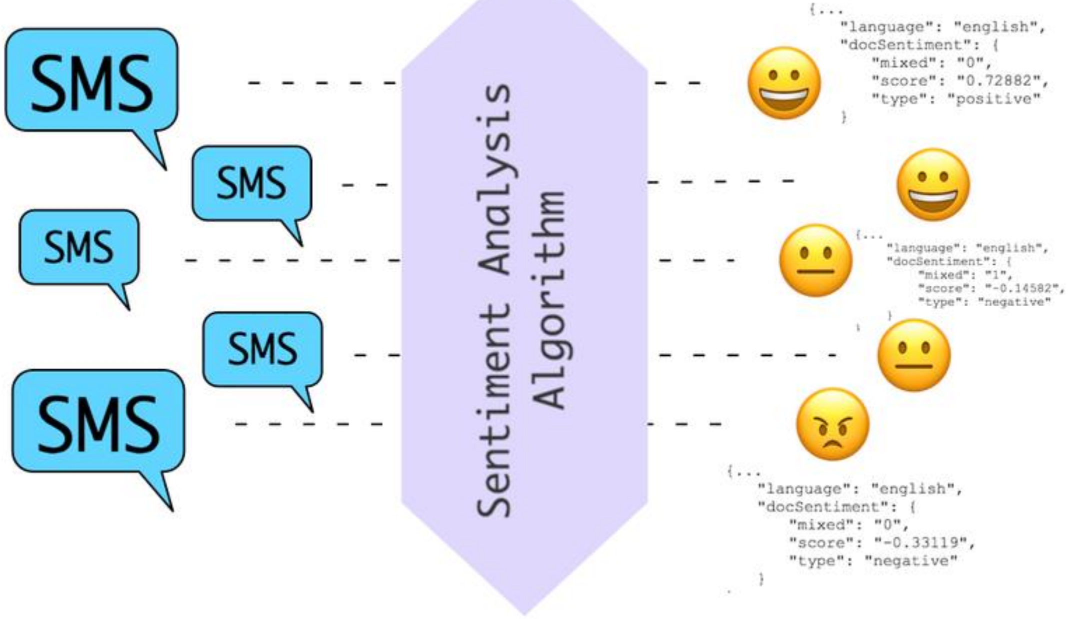
Stopword List
for
is
of
the
to



- SMS spam or not dataset source
- Do not disturb messages
- MATLAB script
- Bag of words
- Governments taking suggestions from citizens with lakhs of responses




SENTIMENT ANALYSIS
NEGATIVE
Totally dissatisfied with the service. Worst customer care ever.
NEUTRAL
Good Job but I will expect a lot more in future.
POSITIVE
Brilliant effort guys! Loved Your Work.



```
{...
    "language": "english",
    "docSentiment": {
        "mixed": "0",
        "score": "0.72882",
        "type": "positive"
    }
```

```
{...
    "language": "english",
    "docSentiment": {
        "mixed": "1",
        "score": "-0.14582",
        "type": "negative"
    }
```

```
{...
    "language": "english",
    "docSentiment": {
        "mixed": "0",
        "score": "-0.33119",
        "type": "negative"
    }
```

Sentiment Classification

- Sentiment classifier in MATLAB
- Social Media text mining
- US elections analysis through Twitter data – Deb Roy, MIT Paper

Image Classification

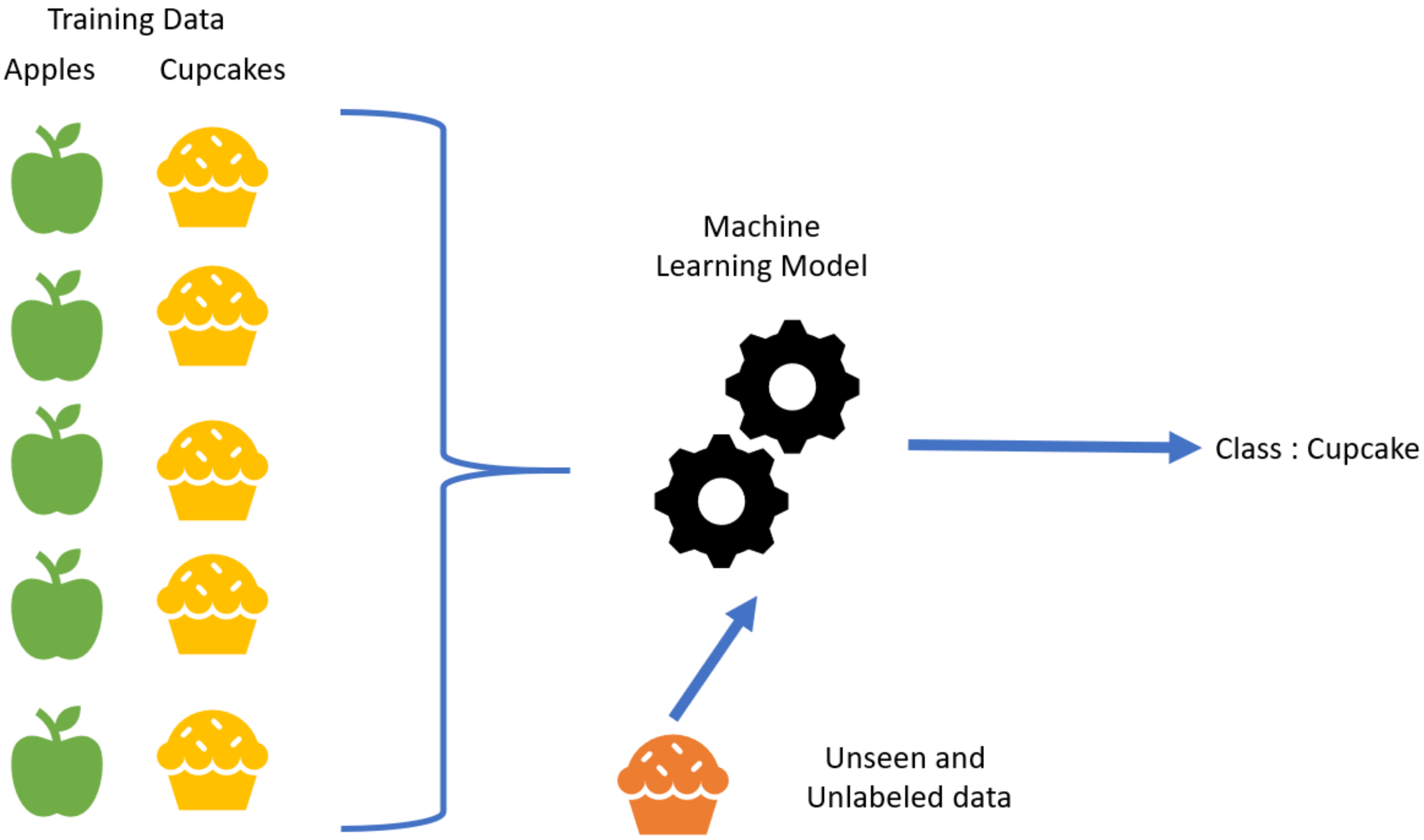
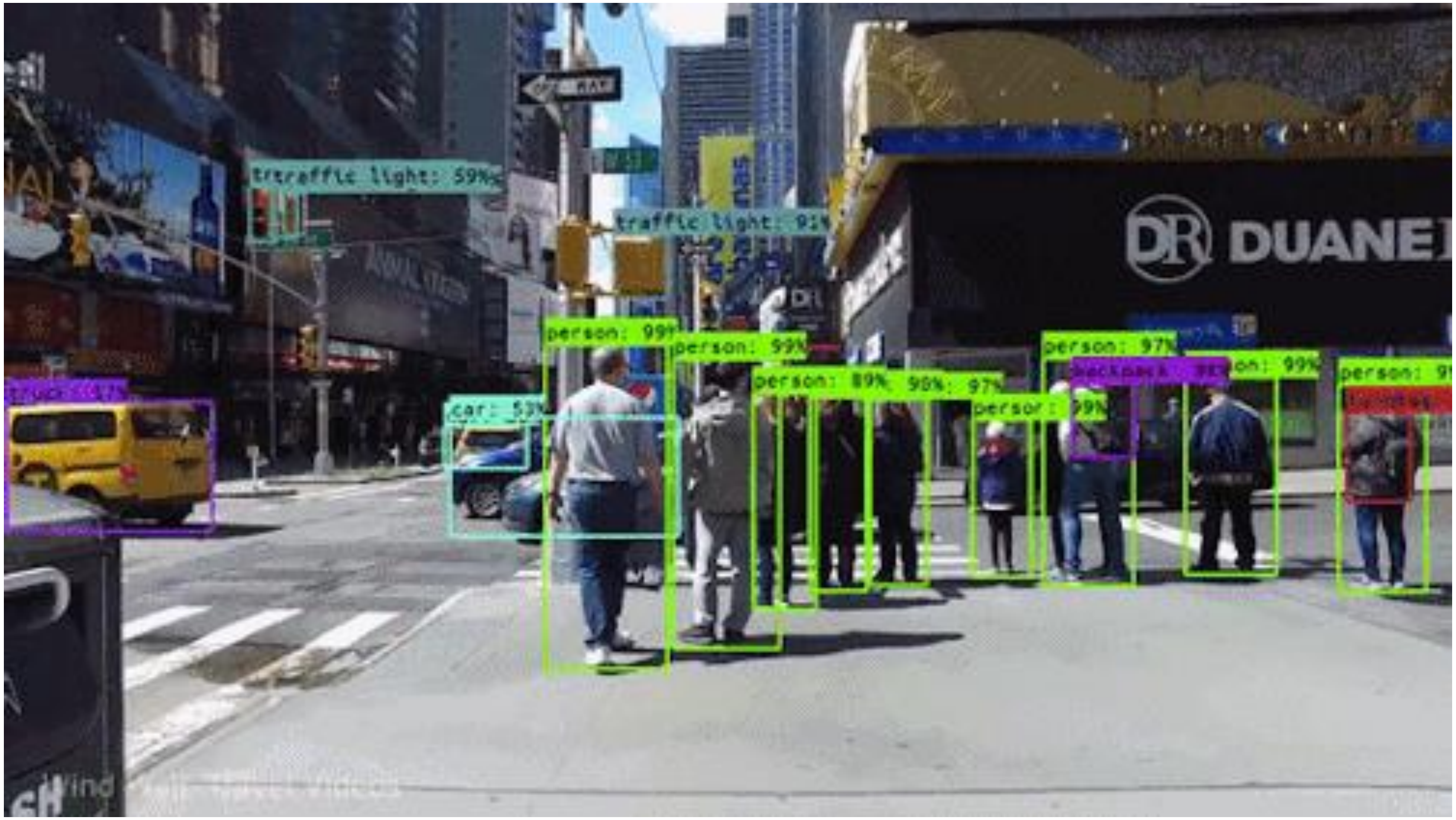


Image classification: Object Detection



How do self-driving cars see?