# Machine Learning in Healthcare

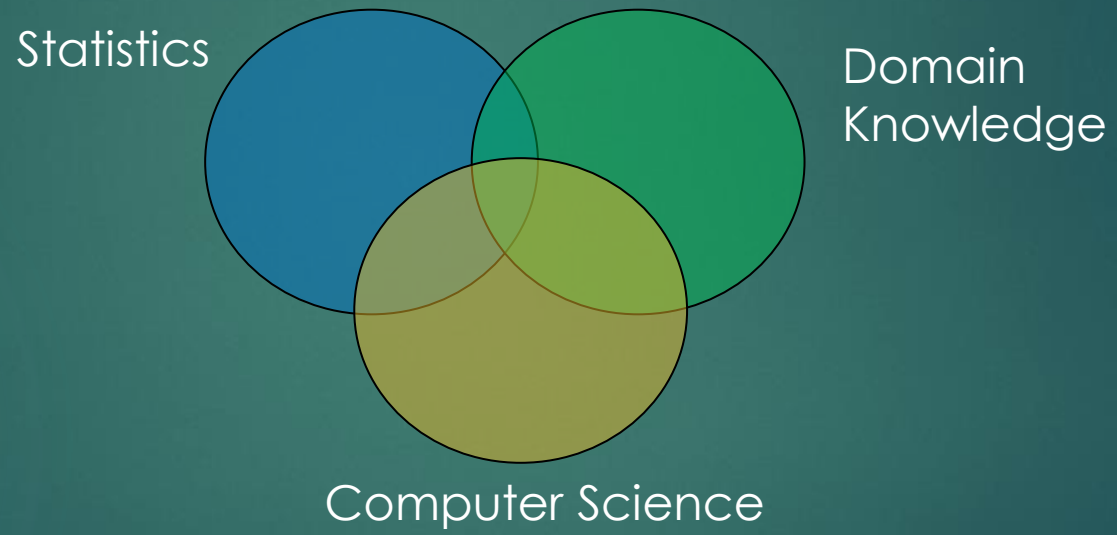
Deepak Narayan

# Introduction

- AIML for predicting disease progression
- Identifying patterns and trends in patient data
- Machine Learning Algorithms :
  - Linear Regression
  - Logistic Regression
  - Tree Models
- Stages of medical science :
  - Diagnosis (detection)
  - Prognosis (prediction)
  - Treatment (course of action)

# Data Science

- Convergence of



# Machine Learning

- Ability of machines to learn without being explicitly programmed.(Arthur Samuel, 1950s)

# Definitions and Terminologies

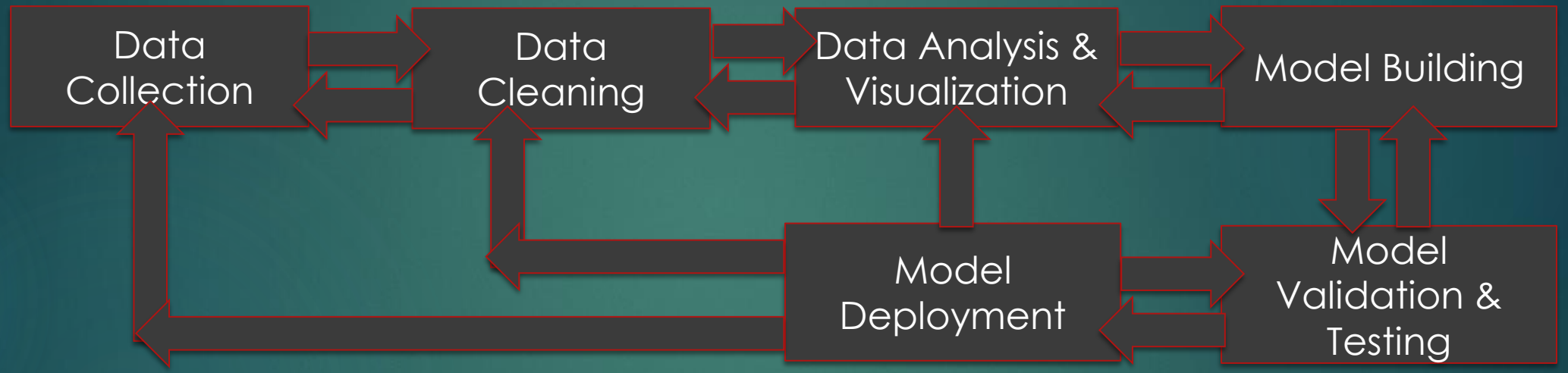
- Iris Data set (R.A.Fisher, statistician and biologist):

- Target (species of Iris plant) : output variable
- PL, PW, SL, SW : input variables
- Output variable or response variable : our primary focus of interest
- Input variables : Features, Predictors, Inputs

# Definitions and Terminologies

- 2 kinds of problems are seen :
  - Supervised Learning :
  - Unsupervised Learning :

# CRISP-DM




Disclaimer : In a real life scenario, a lot of this happens parallelly, although above figure shows it in a serial/sequential fashion.
One important thing that is missing from my above figure, is the problem statement. What is it that we are trying to solve, or achieve?

*– Deepak.*

# Medical prognosis

- Using medical data to predict risks and to predict disease progression and mortality (survival models) based on data available on patient
- Predicting the risk of a future event.
- Prognosis is also useful for guiding treatment.

# Medical prognosis

CHADSVASc score :

- Atrial fibrillation is a abnormal heart rhythm that puts the patient at a risk of stroke.

- https://en.wikipedia.org/wiki/CHA2DS2%E2%80%93VASc_score

| | $CHA_2DS_2$-VASc risk factor | Points |
|---|---|---|
| **C** | Congestive heart failure | +1 |
| **H** | Hypertension | +1 |
| **$A_2$** | Age 75 years or older | +2 |
| **D** | Diabetes mellitus | +1 |
| **$S_2$** | Previous stroke, transient ischaemic attack or thromboembolism | +2 |
| **V** | Vascular disease | +1 |
| **A** | Age 65–74 years | +1 |
| **Sc** | Sex category (female) | +1 |

# Medical prognosis

- Coming to concordance-index.

- The basic idea behind evaluating a prognostic model is to see how well it performs on pairs of patients.

# Medical prognosis

- Calculation of c-index is shown.
- Example of how to calculate c-index using the table provided.

$$\text{C-index} = \frac{\#\text{ concordant pairs} + 0.5 \times \#\text{ risk ties}}{\#\text{ permissible pairs}}$$

| Patient | Event | Risk |
|---|---|---|
| A | Yes | 0.8 |
| B | No | 0.43 |
| C | Yes | 0.62 |
| D | Yes | 0.58 |
| E | No | 0.62 |

# Decision Trees

- Can be used for classification and regression problems.
- Non parametric
- Splits the data into quadrants
- High overfitting based on data, so high variance, low bias.

# Decision Trees

- Bagging
  - Sampling with replacement : Single data point can be sampled multiple times and used to train models.
  - Aggregation : aggregating all these models together and getting an average
  - Parameters : Number of samples to consider, number of trees to create
  - **Reduces variance**

# Decision Trees

- Boosting
  - Incremental Learning
  - Second model learns from first, third model from the second, and so on
  - The errors in previous model is weighted and given in the next model so that it learns from it more and so on.
  - **Reduces bias**

# Random Forests

- Ensemble method
- Compared to bagging, reduces the correlation in the models
- m '**random**' predictors possible at each level = sqrt(p) or p/3