

Feature Engineering

FE

Natural Language
Processing

NLP

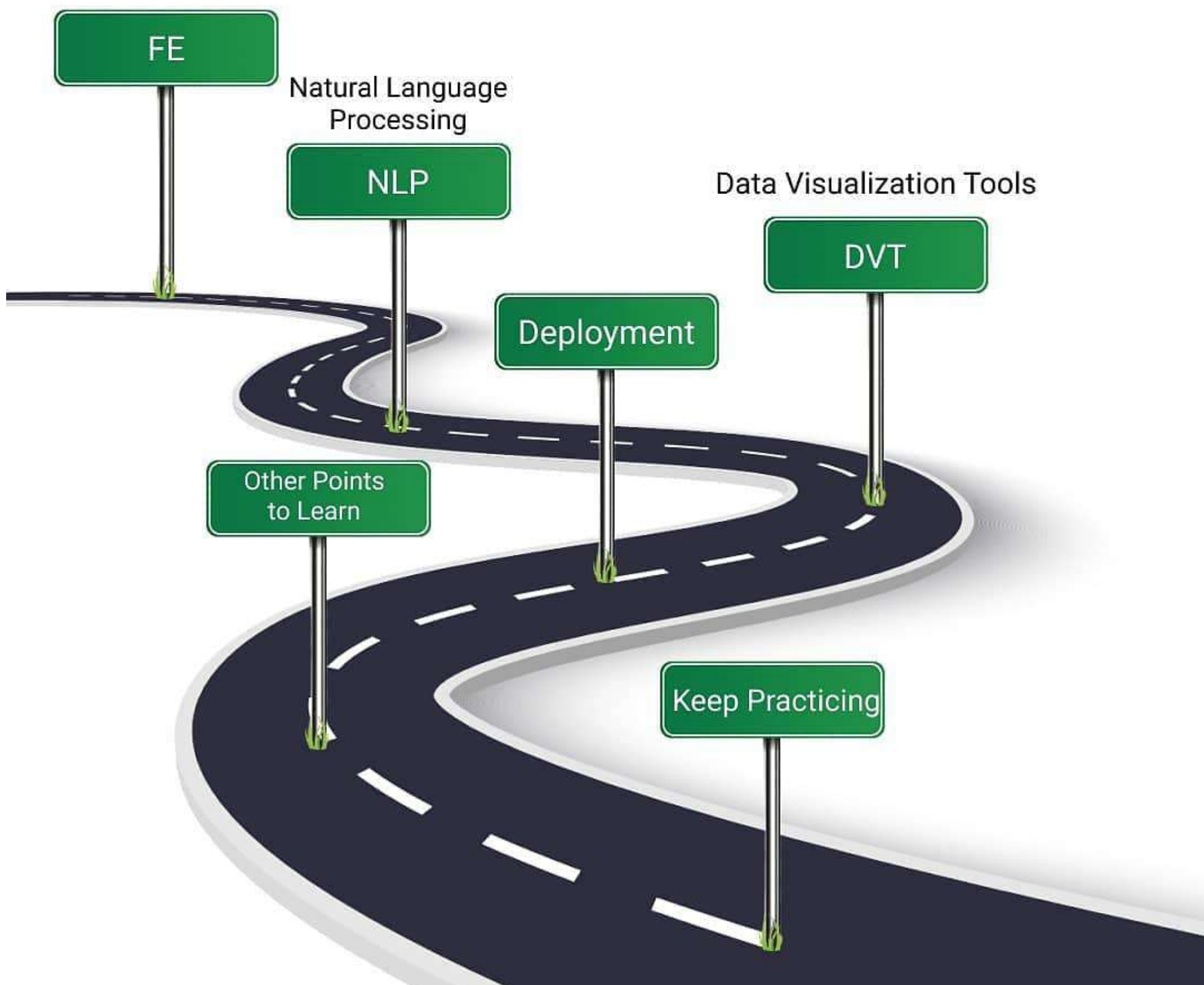
Data Visualization Tools

DVT

Deployment

Other Points
to Learn

Keep Practicing



MATHEMATICS

- **PART 1**

- LINEAR ALGEBRA
- ANALYTIC GEOMETRY
- MATRIX
- VECTOR CALCULUS
- OPTIMIZATION

- **PART 2**

- REGRESSION
- DIMENSIONALITY REDUCTION
- DENSITY ESTIMATION
- CLASSIFICATION

PROBABILITY

- INTRODUCTION TO PROBABILITY
- 1D RANDOM VARIABLE
- THE FUNCTION OF ONE RANDOM VARIABLE
- JOINT PROBABILITY DISTRIBUTION
- DISCRETE DISTRIBUTION
 - BINOMIAL (PYTHON | R)
 - BERNOULLI
 - GEOMETRIC ETC
- CONTINUOUS DISTRIBUTION
 - UNIFORM
 - EXPONENTIAL
 - GAMMA
- NORMAL DISTRIBUTION (PYTHON | R)

STATISTICS

- INTRODUCTION TO STATISTICS
- DATA DESCRIPTION
- RANDOM SAMPLES
- SAMPLING DISTRIBUTION
- PARAMETER ESTIMATION
- HYPOTHESES TESTING (PYTHON | R)
- ANOVA (PYTHON | R)
- RELIABILITY ENGINEERING
- STOCHASTIC PROCESS
- COMPUTER SIMULATION
- DESIGN OF EXPERIMENTS
- SIMPLE LINEAR REGRESSION
- CORRELATION
- MULTIPLE REGRESSION (PYTHON | R)
- NONPARAMETRIC STATISTICS
 - SIGN TEST
 - THE WILCOXON SIGNED-RANK TEST (R)
 - THE WILCOXON RANK SUM TEST
 - THE KRUSKAL-WALLIS TEST (R)
- STATISTICAL QUALITY CONTROL
- BASICS OF GRAPHS
- 4) PROGRAMMING

PROGRAMMING

- PYTHON
- R
- DATABASE
- OTHER
 - DATA STRUCTURE
 - TIME COMPLEXITY
 - WEB SCRAPING (PYTHON | R)
 - LINUX
 - GIT

REFERENCES:

- PYTHON
- R
- SQL
- DATA STRUCTURE

MACHINE LEARNING

- **INTRODUCTION:**
 - HOW MODEL WORKS
 - BASIC DATA EXPLORATION
 - FIRST ML MODEL
 - MODEL VALIDATION
 - UNDERFITTING & OVERFITTING
 - RANDOM FORESTS (PYTHON | R)
 - SCIKIT-LEARN
- **INTRODUCTION:**
 - HANDLING MISSING VALUES
 - HANDLING CATEGORICAL VARIABLES
 - PIPELINES
 - CROSS-VALIDATION (R)
 - XGBOOST (PYTHON | R)
 - DATA LEAKAGE

DEEP LEARNING

- ARTIFICIAL NEURAL NETWORK
- CONVOLUTIONAL NEURAL NETWORK
- RECURRENT NEURAL NETWORK
- TENSORFLOW
- KERAS
- PYTORCH
- A SINGLE NEURON
- DEEP NEURAL NETWORK
- STOCHASTIC GRADIENT DESCENT
- OVERFITTING AND UNDERFITTING
- DROPOUT BATCH NORMALIZATION
- BINARY CLASSIFICATION

FEATURE ENGINEERING

- BASELINE MODEL
- CATEGORICAL ENCODINGS
- FEATURE GENERATION
- FEATURE SELECTION

NATURAL LANGUAGE PROCESSING

IN NLP DISTINGUISH YOURSELF BY LEARNING TO WORK WITH TEXT DATA.

- TEXT CLASSIFICATION
- WORD VECTORS

DATA VISUALIZATION TOOLS

- EXCEL VBA
- BI (BUSINESS INTELLIGENCE):
 - TABLEAU
 - POWER BI
 - QLIK VIEW
 - QLIK SENSE

DEPLOYMENT

THE LAST PART IS DOING THE DEPLOYMENT. DEFINITELY, WHETHER YOU ARE FRESHER OR 5+ YEARS OF EXPERIENCE, OR 10+ YEARS OF EXPERIENCE, DEPLOYMENT IS NECESSARY. BECAUSE DEPLOYMENT WILL DEFINITELY GIVE YOU A FACT IS THAT YOU WORKED A LOT.

- MICROSOFT AZURE
- HEROKU
- GOOGLE CLOUD PLATFORM
- FLASK
- DJANGO

Data Scientist

Roadmap

Mathematics

- Linear Algebra
- Analytics Geometry
- Matrix
- Vector Calculus
- Optimization
- Regression
- Dimensionality Reduction
- Density Estimation
- Classification

Probability

- Discrete Distribution
 - Binomial
 - Bernoulli
 - Geometric etc
- Continuous Distribution
 - Uniform
 - Exponential
 - Gamma
- Normal Distribution
- Introduction to Probability
- 1D Random Variable
- Function of One Random Variable
- Joint Probability Distribution

Statistics

- Introduction to Statistics
- Data Description
- Random Samples
- Sampling Distribution
- Parameter Estimation
- Hypotheses Testing
 - ANOVA
- Reliability Engineering
- Stochastic Process
- Computer Simulation
- Design of Experiments
- Simple Linear Regression
- Correlation
- Multiple Regression
- Nonparametric Statistics
 - Sign Test
 - The Wilcoxon Signed-Rank Test
 - The Wilcoxon Rank Sum Test
 - The Kruskal-Wallis Test
- Statistical Quality Control
- Basic of Graphs

Programming

- Python**
 - Python Basics
 - List
 - Set
 - Tuples
 - Dictionary
 - Function, etc.
 - NumPy
 - Pandas
 - Matplotlib/Seaborn, etc.
- DataBase**
 - SQL
 - MongoDB
- R**
 - R Basic
 - Vector
 - List
 - Data Frame
 - Matrix
 - Array, etc
 - dplyr
 - ggplot2
 - TidyR
 - Shiny, etc.
- Other**
 - Data Structure
 - Array, etc
 - Web Scraping
 - Linux
 - Git

Machine Learning

- Introduction**
 - How Model Works
 - Basic Data Exploration
 - First ML Model
 - Model Validation
 - Underfitting & Overfitting
 - Random Forests
 - scikit-learn
- Intermediate**
 - Handling Missing Values
 - Handling Categorical Variables
 - Pipelines
 - Cross-Validation
 - XGBoost
 - Data Leakage

Deep Learning

- Artificial Neural Network
- Convolutional Neural Network
- Recurrent Neural Network
- Keras
- PyTorch
- TensorFlow
- A Single Neuron
- Deep Neural Network
- Stochastic Gradient Descent
- Overfitting and Underfitting
- Dropout Batch Normalization
- Binary Classification

Feature Engineering

- Baseline Model
- Categorical Encodings
- Feature Generation
- Feature Selection

Natural language Processing

- Text Classification
- Word Vectors

Data Visualization Tools

- Excel VBA
- BI (Business Intelligence)
 - Tableau
 - Power BI
 - Qlik View
 - Qlik Sense

Deployment

- Microsoft Azure
- Heroku
- Google Cloud Platform
- Flask
- Django

Other Points

- Domain Knowledge
- Communication Skill
- Reinforcement Learning
- Case Studies
 - Data Science at Netflix
 - Data Science at Flipkart
 - Project on Credit Card Fraud Detection
 - Project on Movie Recommendation , etc.

Keep Practicing