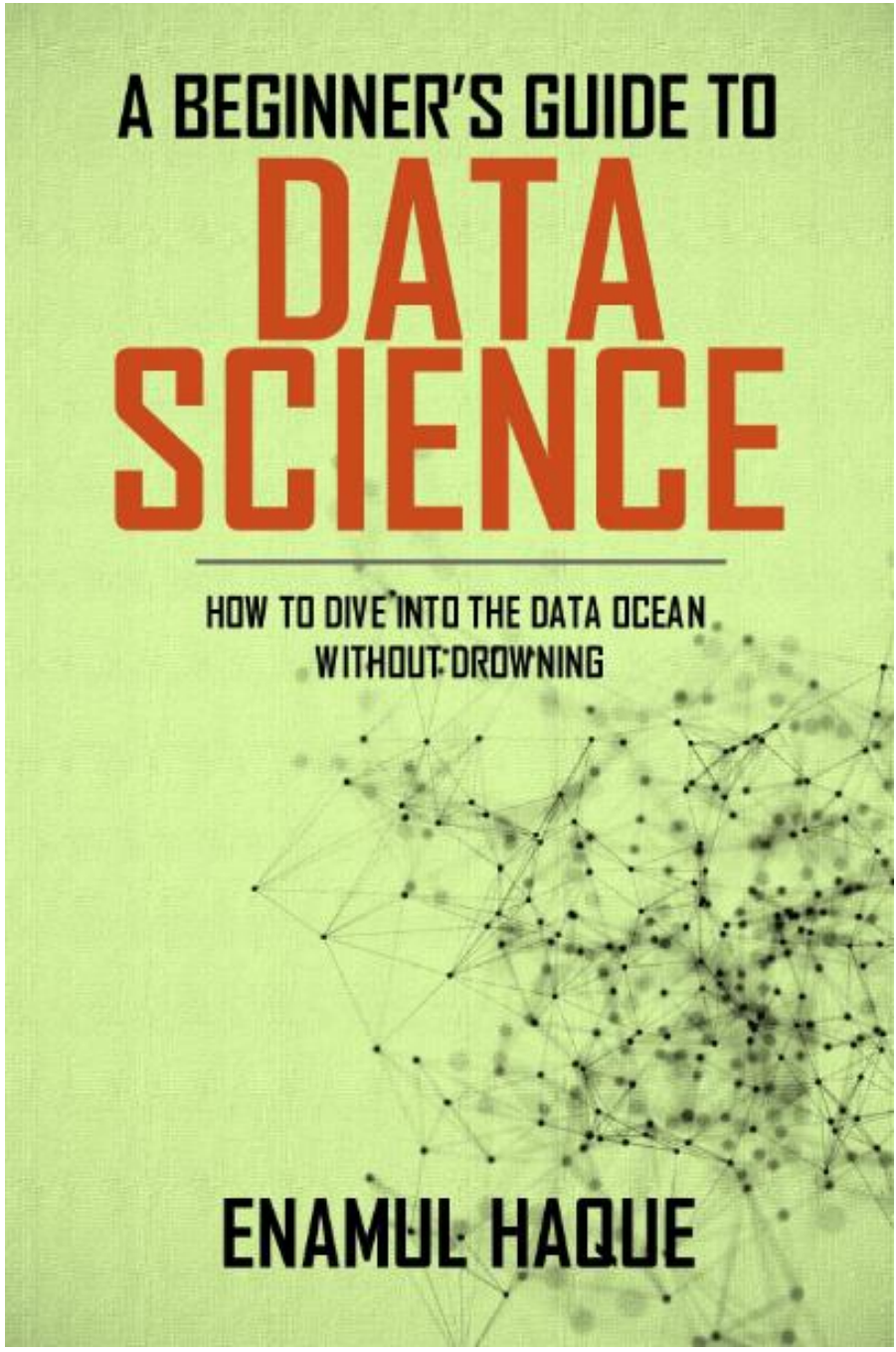

A BEGINNER'S GUIDE TO
DATA SCIENCE
HOW TO DIVE INTO THE DATA OCEAN WITHOUT DROWNING
ENAMUL HAQUE

# A Beginners Guide To DATA SCIENCE

*How to dive into the data ocean without drowning*

ENAMUL HAQUE


**Enel Publications**
**London, UK**
**Amazon Kindle Direct Publishing**

**First Printing Edition, April 2021**

ISBN 9798731261074

# CHAPTER FIVE:
# DATA SCIENCE DISCIPLINES

*"Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world."* ... – Atul Butte, Stanford University

# Core Disciplines of Data Science

Many core branches of learning contribute to the data science discipline. It suggests generic questions that data scientist should ask as they work through solving problems.

## Data engineering

As the name implies, data engineering is concerned with data, namely its delivery, storage and processing. Accordingly, the main task of engineers is to provide a reliable infrastructure for data. With the advent of big data, the area of responsibility has changed dramatically. Previously, these experts wrote large SQL queries and distilled data using tools such as Informatica ETL, Pentaho ETL, Talend, but now the requirements for data engineers have increased. A data engineer understands programming better than any data scientist, but the opposite is true when it comes to statistics.

## Scientific method

The Scientific Method is the science part of data science. According to Wikipedia, the Scientific Method is a process for acquiring new knowledge by applying the principles of reasoning on empirical evidence derived from testing hypotheses through repeatable experiments. When a scientist hears someone assertion about a fact, they naturally want to know both the evidence and the standard of acceptance for that evidence.

## Mathematics

Mathematics (along with statistics) is the cerebral part of Data Science. We will look into this separately.

## Statistics

Statistics is the study of the collection, organisation, analysis, and interpretation of data. It involves exploring data, discovering patterns and relationships, creating models, and making inferences about the future. Statistics is the discipline that has the straightest-line pedigree to data science. The statistician is responsible for understanding the analysis that will be done on the data to be collected and organised appropriately.

## Advanced computing

Advanced computing is the heavy lifting of data science. According to Wikipedia, computer programming (often shortened to programming or coding) is designing, writing, testing, debugging, and maintaining computer programs' source code. This source code is written in one or more programming languages. The purpose of programming is to create a set of instructions that computers use to perform specific operations or to exhibit desired behaviours. Writing source code often requires expertise in many different subjects, including knowledge of the application domain, specialised algorithms and formal logic.

## Visualisation

Data visualisation helps you turn all that granular data into easily understood, visually compelling—and valuable—business information. By tapping into external data sources, today’s data visualisation tools don’t simply let you see your KPIs more; clearly, they unify data and apply AI-driven analytics to reveal relationships between your KPIs, the market, and the world.

## Hacker mindset

A typical scientific mindset is building models, training, plot graphs, and analysing the different attributes to come up with a solution. The mindset of a hacker is very different from that of a scientist. They focus more on finding Solutions using simple methods. While the data scientists use so many various components to a problem, the hacker works at eliminating complexity to come up with a solution.[1] Therefore, the hacker mindset is more complimentary because the scientific mind's confines do not bound them.

## Domaine expertise

Domain expertise is the glue that holds data science together. According to Wikipedia, subject matter or domain expertise is proficiency, special knowledge or skills, in a particular area or topic. Any domain of knowledge can be subject to a data science inquiry, including but not limited to medicine, politics, the physical and biological sciences, marketing, information security, demographics, and even literature. Every data science team must include at least one person who is a subject matter expert on the problem being solved.

# Mathematics in Data Science

Mathematics is the foundation for any modern scientific discipline. And it's not a secret to anyone that almost all modern data science methods (including machine learning) are based on some kind of mathematical calculations. Sometimes, as a data scientist (or even a junior analyst), you need to know basic mathematics in order to correctly apply its methods. For other purposes, you can use the API or a ready-made algorithm.

But, at the same time, good possession of Nanami math on which to build your algorithm to create recommendations for the use of the product never hurt. This will give you an edge over your competitors and help you maintain confidence in your knowledge. It's always good to know what's under the hood rather than just sitting behind the wheel without knowing anything about the car.

Of course, you will need other knowledge, programming skills, a little business acumen, unique analytical thinking and curiosity about data, which are so necessary for a leading data scientist. In this article, I have tried to collect the most important mathematical concepts to help you in this endeavour.

Knowledge of mathematics basics is essential for professionals who want to move into this area from another specialisation (hardware development, trade, chemical industry, medicine and healthcare, business management, etc.).

And while they may think they've worked with spreadsheets, numeric calculations, and predictions for a long time, the math skill requirements are vastly different from their current job in data science.

Here are some handpicked suggestions of topics that will come in handy to stay at the top of the data science game.

## Functions, variables, equations, graphs

Let's start with basic things like linear equations and end with Newton's binomial and its properties.

- Logarithm, exponential, polynomial functions, rational numbers.
- Foundations of geometry and basic theorems, trigonometric identities.
- Real and complex numbers, their basic properties.
- Series, Sums and Inequalities.
- Plotting, rectangular and polar coordinate systems, tapered sections.

If you want to figure out how to quickly find something in a database with a million sorted items, then you need the concept of binary search. To understand dynamics, you first need to understand logarithms and recurrence equations. Or, if you want to analyse time series, use concepts like periodic functions and exponential law.

## Statistics

This is what you need to know to grow as a data scientist. The importance of a clear understanding of the basic concepts of statistics and probability cannot be overstated in discussions about data science. Many practitioners in the field refer to classical machine learning (not a neural network) as nothing more than statistical learning. The topic is vast and endless, and therefore focused planning is essential to cover as many core concepts as possible.

- Summary and descriptive statistics, mean, variance, covariance, correlation.
- Fundamentals of probability theory: basic ideas, expectation, calculus of probability, Bayes' theorem, conditional probability.

- Probability distribution functions - uniform, normal, binomial, chi-square, Student's t distribution, central limit theorem.
- Sampling, measurement, error, random number generator.
- Hypothesis testing, A/B testing, confidence interval, P-value.
- ANOVA, t-test.
- Linear regression, regularisation.

Where you can use them? During interviews. Trust me. As a forward-looking data scientist, you can quickly make a good impression on your future employer by mastering all of the above concepts. While working, you will often have to deal with the need to use certain concepts.

## Linear algebra

Facebook friends recommendation, Spotify song recommendation, Salvador Dali-style effect of photography using deep neural network transfer learning. What do they all have in common? Matrices and matrix algebra are used everywhere. Matrix algebra is an important aspect of mathematics that helps you understand how most machine learning algorithms function in a data stream. The following are the most important topics to explore:

- A matrix and vectors' main properties are dot product, linear transformation, transposition, conjugation, rank, determinant.
- Inner and outer product, matrix multiplication rule and various algorithms, inverse matrix.
- Spatial matrices - square, unit, triangular, sparse, dense, symmetric, Hermitian, anti-Hermitian and unitary matrices, unit vector.

- The concept of matrix decomposition / LU-decomposition, Gauss / Gauss-Jordan method, solution of systems of linear algebraic equations of the form Ax = b.
- Vector space, basis, hull, orthogonality, linear least squares.
- Matrix eigenvalue, eigenvector, diagonalization, singular value decomposition (SVD).

Where you can use them? If you are using principal component analysis (PCA) for dimensionality reduction, you will most likely use singular value decomposition for a more compact data dimension with fewer parameters. All neural network algorithms use linear algebra techniques to represent and process network structures and learning operations.

## Mathematical analysis

Whether you liked it at university or not, we encounter calculus in many aspects of data science and machine learning. It is hidden behind a seemingly simple analytical solution to a common problem with the least value of a quadratic function in linear regression. It is also embedded in every backpropagation method generated by the neural network for training. Knowledge of mathematical analysis will prove to be very valuable for your work. The following are topics to explore:

- Single variable function, limit, continuity and differentiability.
- The formula of finite increments, disclosure of uncertainties, L'Hôpital's theorem.
- Maximum and minimum.
- Rules for the product and differentiation of a complex function.
- Taylor series, infinite series summation/integration concept.
- The main theorem and formula for finite increments of integral calculus, calculation of definite and improper integrals.

- Beta and Gamma Functions.
- Functions of a set of variables, limit, continuity and partial derivatives.
- Fundamentals of ordinary differential equations and partial differential equations (not the most difficult).

Where to use them? You've probably wondered how the logistic regression algorithm is used. To find the minimum loss function, the gradient descent method is very often used. To understand how this works, it is necessary to use mathematical analysis concepts: gradient, derivatives, limits, differentiation of a complex function.

## Discrete math

Discrete mathematics is rarely touched upon when discussing a topic such as "mathematics in data science." Nevertheless, modern data science is built with the help of computing systems in which discrete mathematics is a key element. Discrete mathematics courses will help you master important concepts for the daily use of algorithms and data structures when working on analytical projects. Below are some of the topics to explore:

- Set, subset, boolean.
- Counting functions, combinatorics, countability.
- The main methods of proof are induction, proof by contradiction.
- Foundations of inductive, deductive and propositional logic.
- The main data structures are stacks, queues, graphs, arrays, hash tables, trees.
- Graph invariants: connected components, vertex degree, Ford - Fulkerson theorem, graph colouring.
- Recurrent formulas (equations, relations).
- Function growth, "O" notation is large.

Where they can be used? Graph invariants and fast algorithms are essential when analysing any social networks. With any algorithm, you

need to understand the temporal and spatial complexity using the big O notation. This is necessary, for example, when determining how the run time and the required size increase with the increase in the amount of input data.

## Optimisation, operations research topics

These topics are not much different from the traditional discourse of applied mathematics since they are mainly important and most used in specialised fields of study: in theoretical computer science, control theory, operations research. But a general understanding of these effective methods can be instrumental in the field of machine learning. Almost every machine learning algorithm/method aims to minimize some sort of estimation error, given various constraints. This is the goal of optimisation. Study topics:

- Optimisation basics - how to formulate a problem.
- Maximum, minimum, convex function, global solution.
- Linear programming, simplex method.
- Integer programming.
- Constraint programming, knapsack problem.
- Randomised optimisation methods - search by ascent to the top, simulated annealing algorithm, genetic algorithm.

Where they can be used? Simple linear regression problems, as opposed to logistic ones, using the least-squares loss function, often have an exact analytical solution. To understand the reason, you need to know about such a concept as convexity in optimisation. It will also explain why we should have enough "rough" solutions for many machine learning problems. An optimisation is a powerful tool worth exploring in detail.

# Mathematical Analysis

A good data analyst without basic mathematics is nowhere (and the data researcher is even more so). So, let's understand areas to be at the interest of data science.

## The basics of mathematical analysis

- Functions and their properties.
- Function limit (basic views).
- Derivative function (its geometric and mechanical meaning).
- Derivative of a complex function.
- Extremes feature. The bulge function.
- Private derivatives and gradient.
- The gradient in optimization tasks.
- Derivative in the direction.
- Touching plane and linear approximation.

## The basics of linear algebra

- Vector space.
- Linear independence.
- Norm and scalar work of vectors.
- Determining the matrix. Operations on the matrix.
- Rank and determiner of the matrix.
- Line equation systems.
- Matrix types.
- Own vectors and own values.

- Matrix decompositions (spectral, singular).
- Approaching the matrix of the lower rank.
- Singular decomposition and low-burn approximation.

## Optimisation methods

- Optimising non-smooth functions (the problem of local lows).
- The method of imitation of the ignition.
- Genetic algorithms. Algorithm of differential evolution.
- Ndder-Mead Method.

## Probability theory and mathematical statistics

- Determining probability. Probability properties.
- Conditional probabilities. The formula of full probability. Formula Bayes.
- Discreet random values.
- Continuous random values.
- Sample distribution estimate. Statistics.
- Distribution characteristics.
- Important statistics (selective average, median, fashion, variance, interquartile swing).
- The central limit of the theorem.
- Confidence intervals.

# Statistical Modelling

A statistical model is a mathematical model that embodies statistical assumptions concerning sample data generation (and similar data from a larger population). A statistical model represents, often in considerably idealised form, the data-generating process. A statistical model is usually specified as a mathematical relationship between one or more random variables and other non-random variables. Some of the useful statistical modelling methods are described below:

## Spatial models

Spatial dependency is the co-variation of properties within geographic space: characteristics at proximal locations appear to be correlated, either positively or negatively. Spatial dependency leads to the spatial auto-correlation problem in statistics since, like temporal auto-correlation, this violates standard statistical techniques that assume independence among observations[2]

## Time series

Methods for time series analyses may be divided into two classes: frequency-domain methods and time-domain methods. The former include spectral analysis and recently wavelet analysis; the latter include auto-correlation and cross-correlation analysis. In the time domain, correlation analyses can be made in a filter-like manner using scaled correlation, thereby mitigating the need to operate in the frequency domain.

Additionally, time series analysis techniques may be divided into parametric and non-parametric methods. The parametric approaches assume that the underlying stationary stochastic process has a particular structure that can be described using a small number of parameters (for example, using an autoregressive or moving average model). In these approaches, the task is to estimate the model's parameters that describe the stochastic process. By contrast, non-parametric approaches explicitly estimate the covariance or the spectrum of the process without assuming that the process has any particular structure. Methods of time series analysis may also be divided into linear and non-linear, and univariate and multivariate.

## Survival analysis

Survival analysis is a branch of statistics for analysing the expected duration of time until one or more events happen, such as a death in biological organisms and failure in mechanical systems. This topic is called reliability theory or reliability analysis in engineering, duration analysis or duration modelling in economics, and event history analysis in sociology. Survival analysis attempts to answer questions such as: what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival? Survival models are used by actuaries and statisticians, and marketers designing churn and user retention models.[3]

Survival models are also used to predict time-to-event (time from becoming radicalised to turning into a terrorist or when a gun is purchased and used in a murder) or to model and predict decay.

## Market segmentation

Market segmentation, also called customer profiling, is a marketing strategy that involves dividing a broad target market into subsets of consumers, businesses, or countries that have or are perceived to have common needs, interests, and priorities, and then designing and implementing strategies to target them. Market segmentation strategies are generally used to identify and further define the target customers and provide supporting data for marketing plan elements such as positioning to achieve certain marketing plan objectives. Businesses may develop product differentiation strategies or an undifferentiated approach involving specific products or product lines depending on the target segment's specific demand and attributes.

## Recommendation systems

Recommender systems or recommendation systems (sometimes replacing “system” with a synonym such as a platform or an engine) are a subclass of information filtering system that seeks to predict the ‘rating’ or ‘preference’ that a user would give an item.

## Association rule learning

Association rule learning is a method for discovering interesting relations between variables in large databases. For example, the rule { onions, potatoes } ==> { burger } found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat. In fraud detection, association rules are used to detect patterns associated with the fraud. Linkage analysis is performed to identify additional fraud cases: if a credit card transaction from user A was used to make a fraudulent purchase at store B, by analysing all transactions from store B, we might find another user C with fraudulent activity.

## Attribution modelling

An attribution model is the rule or set of rules determining how credit for sales and conversions is assigned to touchpoints in conversion paths. For example, the Last Interaction model in Google Analytics assigns 100% credit to the final touchpoints (i.e., clicks) that immediately precede sales or conversions. Macro-economic models use long-term, aggregated historical data to assign an attribution weight to a number of channels for each sale or conversion. These models are also used for advertising mix optimisation.

## Scoring

The scoring model is a special kind of predictive models. Predictive models can predict defaulting on loan payments, risk of accident, client churn or attrition, or chance of buying a good. Scoring models typically use a logarithmic scale (each additional 50 points in your score, reducing the risk of defaulting by 50%). They are based on logistic regression and decision trees or a combination of multiple algorithms. Scoring technology is typically applied to transactional data, sometimes in real-time (credit card fraud detection, click fraud).

## Predictive Modelling

Predictive modelling leverages statistics to predict outcomes. Most often, the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred. For example, predictive models are often used to detect crimes and identify suspects after the crime has taken place. They may also be used for weather forecasting, to predict stock market prices, or to predict sales, incorporating time series or spatial models. Neural networks, linear regression, decision trees and naive Bayes are techniques used for predictive modelling. They are associated with creating a training set, cross-validation, and model fitting and selection.

## Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is the main task of exploratory data mining and a common statistical data analysis technique used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Unlike supervised classification (below), clustering does not use training sets. Though there are some hybrid implementations called semi-supervised learning.

## Supervised classification

Supervised classification, also called supervised learning, is the machine learning task of inferring a function from labelled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and the desired output value (also called label, class or category). A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances.

## Extreme value theory

Extreme value theory or extreme value analysis (EVA) is a branch of statistics dealing with the extreme deviations from the median of probability distributions. It seeks to assess, from a given ordered sample of a given random variable, the probability of more extreme events than any previously observed. For instance, floods occur once every 10, 100, or 500 years. These models have been performing poorly recently to predict catastrophic events, resulting in massive losses for insurance companies.

## Simulations

Monte-Carlo simulations are used in many contexts: to produce high-quality pseudo-random numbers in complex settings such as multi-layer Spatio-temporal hierarchical Bayesian models, to estimate parameters to compute statistics associated with infrequent events, or even to generate a large amount of data (for instance, cross and auto-correlated time series) to test and compare various algorithms, especially for stock trading or in engineering.

## Churn analysis

Customer churn analysis helps you identify and focus on higher-value customers, determine what actions typically precede a lost customer or sale, and better understand what factors influence customer retention. Statistical techniques involved include survival analysis as well as Markov chains with four states: brand new customer, returning customer, inactive (lost) customer, and re-acquired customer, along with path analysis (including root cause analysis) to understand how customers move from one state to another, to maximise profit. Related topics: customer lifetime value, cost of user acquisition, user retention.

## Inventory management

Inventory management overseeing and controlling the ordering, storage, and use of components that a company will use to produce the items it will sell and oversee and control quantities of finished products for sale. Inventory management is an operations research technique leveraging analytics (time series, seasonality, regression), especially for sales forecasting and optimum pricing — broken down per product category, market segment, and geography. It is strongly related to pricing optimisation. This is not just for brick and mortar operations: inventory could mean the amount of available banner ad slots on a publisher website in the next 60 days, with estimates of how much traffic (and conversions)

each banner ad slot is expected to deliver to the potential advertiser. You don't want to over-sell or under-sell this virtual inventory. Thus you need good statistical models to predict the web traffic and conversions (to pre-sell the inventory) for each advertiser category.

## Optimum bidding

This is an example of an automated, black-box, machine-to-machine communication system, sometimes working in real-time via various API's. It is backed by statistical models. Applications include detecting and purchasing the right keywords at the right price on Google AdWords, based on expected conversion rates for millions of keywords, most of them having no historical data; keywords are categorized using an indexation algorithm (see item #18 in this article) and aggregated into buckets (categories) to get some historical data with statistical significance, at the bucket level. This is a real problem for companies such as Amazon or eBay. Or it could be used as the core algorithm for automated high-frequency stock trading.

## Optimum pricing

While at first glance, it sounds like an econometric problem handled with efficiency curves or even a pure business problem, it is highly statistical in nature. Optimum pricing considers available and predicted inventory, production costs, prices from competitors, and profit margins. Price elasticity models are often used to determine how high prices can be boosted before reaching strong resistance. Modern systems offer prices-on-demand, in real-time, for instance, when booking a flight or a hotel room. User-dependent pricing — a way to further optimise pricing, offering different prices based on user segment — is a controversial issue. It is accepted in the insurance industry: bad car drivers paying more than good ones for the same coverage, or smokers/women / old people paying

a different fee for healthcare insurance (this is the only price discrimination allowed by Obamacare).

## Indexation

Any system based on taxonomies uses an indexation algorithm created to build and maintain the taxonomy. For instance, product reviews (both products and reviewers must be categorised using an indexation algorithm, then mapped onto each other), scoring algorithms to detect the top people to follow in a specific domain, digital content management, and of course, search engine technology. Indexation is a very efficient clustering algorithm, and the time used to massive index amounts of content grows linearly — that is very fast — with the size of your dataset. Basically, it relies on a few hundred categories manually selected after parsing tons of documents, extracting billions of keywords, filtering them, producing a keyword frequency table, and focusing on top keywords.

Finally, an indexation algorithm can be used to automatically create an index for any document — report, article, blog, website, data repository, metadata, catalogue, or book. Indeed, that's the origin of the word indexation. Surprisingly, publishers still pay people today for indexing jobs: you can find these jobs listed on the American Society for Indexing website. This is an opportunity for data scientist entrepreneurs: offering publishers software that does this job automatically, at a fraction of the cost.

## Search engines

Good search engine technology relies heavily on statistical modelling. Enterprise search engines help companies — for instance, Amazon — sell their products by providing users with an easy way to find them. The core algorithm used in any search engine is an indexation or automated tagging system. Google search could be improved as follows:

- Eliminate page rank — this algorithm has been fooled by cheaters developing link farms and other web spam,
- Add new content more frequently in your index to make search results less static, less frozen in time,
- Show more relevant articles using better user/search keyword/landing page matching algorithms which ultimately means better indexation systems, and
- Use better attribution models to show the article's source, not copies published on LinkedIn or elsewhere. (this could be as simple as putting more weights on small publishers and identifying the first occurrence of an article: timestamp detection and management).

## Cross-Selling

Usually, based on collaborative filtering algorithms, the idea is to find — especially in retail — which products to sell to a client based on recent purchases or interests. For instance, trying to sell engine oil to a customer buying gasoline. In banking, a company might want to sell several services: a checking account first, then a saving account, then a business account, then a loan and so on, to a specific customer segment. The challenge is to identify the correct order in which products must be promoted, the right customer segments, and the optimum time lag between the various promotions. Cross-selling is different from up-selling.

## Clinical trials

Clinical trials are experiments done in clinical research, usually involving small data. Such prospective biomedical or behavioural research studies on human participants are designed to answer specific biomedical or behavioural interventions, including new treatments and known interventions that warrant further research and comparison. Clinical trials generate data on safety and efficacy. Primary concerns include how to

test patients are sampled (mainly if they are compensated), conflict of interests in these studies, and the lack of reproducibility.


*Figure 1 - The field of statistics affects all areas of life*

## Multivariate testing

Multivariate testing is a technique for testing a hypothesis in which multiple variables are modified. The goal is to determine which combinations of variations perform the best out of all possible combinations. Websites and mobile apps are made of combinations of changeable elements that are optimised using multivariate testing. This involves careful design-of-experiment, and the tiny, temporary difference (in yield or web traffic) between two versions of a webpage might not have statistical significance. While ANOVA[4] and tests of hypotheses are used by industrial or healthcare statisticians for multivariate testing, we have developed model-free, data-driven systems based on data binning and model-free confidence intervals. Stopping a multivariate testing experiment (they

usually last 14 days for web page optimisation) as soon as the winning combination is identified helps save a lot of money. Note that external events — for instance, a holiday or some server outage — can impact multivariate testing results and need to be addressed.

## Queuing systems

A queue management system is used to control queues. Queues of people form in various situations and locations in a queue area, for instance, in a call centre. The process of queue formation and propagation is defined as queuing theory. People's arrival in a queue is typically modelled using a Poisson method to serve a client modelled using an exponential distribution. While being a statistical problem, it is considered to be part of operations research.

## Supply chain optimisation

Supply chain optimisation is applying processes and tools to ensure a manufacturing and distribution supply chain's optimal operation. This includes the optimal placement of inventory within the supply chain, minimising operating costs (including manufacturing costs, transportation costs, and distribution costs). This often involves applying mathematical modelling techniques such as graph theory to find optimum delivery routes (and optimum locations of warehouses), the simplex algorithm, and Monte Carlo simulations.

# পঞ্চম অধ্যায়: ডেটা সায়েন্স ডিসিপ্লিন

**(please note, that this is machine translation)**

# ডেটা সায়েন্সের কোর ডিসিপ্লিন

শেখার অনেক মূল শাখা তথ্য বিজ্ঞান শৃঙ্খলায় অবদান রাখে। এটি সাধারণ প্রশ্নগুলি প্রস্তাব করে যা ডেটা বিজ্ঞানীদের জিজ্ঞাসা করা উচিত কারণ তারা সমস্যা সমাধানের মাধ্যমে কাজ করে।

## ডেটা ইঞ্জিনিয়ারিং

নাম থেকে বোঝা যায়, ডেটা ইঞ্জিনিয়ারিং ডেটার সাথে সম্পর্কিত, যথা এর ডেলিভারি, স্টোরেজ এবং প্রসেসিং। তদনুসারে, প্রকৌশলীদের প্রধান কাজ হল তথ্যের জন্য একটি নির্ভরযোগ্য অবকাঠামো প্রদান করা। বড় তথ্যের আবির্ভাবের সাথে, দায়িত্বের ক্ষেত্রটি নাটকীয়ভাবে পরিবর্তিত হয়েছে। পূর্বে, এই বিশেষজ্ঞরা ইনফরম্যাটিকা ইটিএল, পেন্টাহো ইটিএল, ট্যালেন্ডের মতো সরঞ্জাম ব্যবহার করে বড় এসকিউএল প্রশ্ন এবং পাতিত ডেটা লিখেছিলেন, কিন্তু এখন ডেটা ইঞ্জিনিয়ারদের প্রয়োজনীয়তা বেড়েছে। একজন ডেটা ইঞ্জিনিয়ার যেকোন ডেটা সায়েন্টিস্টের চেয়ে প্রোগ্রামিংকে ভালো বোঝেন, কিন্তু পরিসংখ্যানের ক্ষেত্রে এর বিপরীত সত্য।

## বৈজ্ঞানিক পদ্ধতি

বৈজ্ঞানিক পদ্ধতি হল ডেটা সায়েন্সের বিজ্ঞান অংশ। উইকিপিডিয়ার মতে, বৈজ্ঞানিক পদ্ধতি হল পুনরাবৃত্তিযোগ্য পরীক্ষার মাধ্যমে অনুমান পরীক্ষা থেকে প্রাপ্ত অভিজ্ঞতাগত প্রমাণের উপর যুক্তির নীতি প্রয়োগ করে নতুন জ্ঞান অর্জনের একটি প্রক্রিয়া। যখন একজন বিজ্ঞানী কোন সত্য সম্পর্কে কারো বক্তব্য শুনেন, তখন তারা স্বাভাবিকভাবেই প্রমাণ এবং গ্রহণযোগ্যতার মান উভয়ই জানতে চায়।

## গণিত

গণিত (পরিসংখ্যান সহ) ডেটা সায়েন্সের সেরিব্রাল অংশ। আমরা এটি আলাদাভাবে দেখব।

## পরিসংখ্যান

পরিসংখ্যান হলো তথ্য সংগ্রহ, সংগঠন, বিশ্লেষণ এবং ব্যাখ্যা অধ্যয়ন। এতে ডেটা অন্বেষণ, নিদর্শন এবং সম্পর্ক আবিষ্কার, মডেল তৈরি করা এবং ভবিষ্যত সম্পর্কে অনুমান করা জড়িত। পরিসংখ্যান হচ্ছে এমন একটি শৃঙ্খলা যার মধ্যে ডেটা সায়েন্সের জন্য সরলরেখার বংশধর রয়েছে। পরিসংখ্যানবিদ বিশ্লেষণ বোঝার জন্য দায়ী যা তথ্য সংগ্রহ করা হবে এবং যথাযথভাবে সংগঠিত হবে।

## উন্নত কম্পিউটিং

উন্নত কম্পিউটিং হচ্ছে ডেটা সায়েন্সের ভারী উত্তোলন। উইকিপিডিয়ার মতে, কম্পিউটার প্রোগ্রামিং (প্রায়শই প্রোগ্রামিং বা কোডিংয়ের জন্য সংক্ষিপ্ত করা হয়) হল কম্পিউটার প্রোগ্রামগুলির সোর্স কোড ডিজাইন করা, লেখা, পরীক্ষা করা, ডিবাগ করা এবং বজায় রাখা। এই সোর্স কোডটি এক বা একাধিক প্রোগ্রামিং ভাষায় লেখা। প্রোগ্রামিং এর উদ্দেশ্য হল কম্পিউটারগুলি নির্দিষ্ট অপারেশন করতে বা পছন্দসই আচরণ প্রদর্শন করার জন্য নির্দেশাবলী তৈরি করে। সোর্স কোড লেখার জন্য প্রায়শই অ্যাপ্লিকেশন ডোমেনের জ্ঞান, বিশেষ অ্যালগরিদম এবং আনুষ্ঠানিক যুক্তি সহ বিভিন্ন বিষয়ে দক্ষতার প্রয়োজন হয়।

## ভিজ্যুয়ালাইজেশন

ডেটা ভিজ্যুয়ালাইজেশন আপনাকে সেই সমস্ত দানাদার ডেটা সহজেই বোঝা যায়, চাক্ষুষভাবে আকর্ষণীয় — এবং মূল্যবান -ব্যবসায়িক তথ্যে পরিণত করে। বাহ্যিক ডেটা উত্সগুলিতে ট্যাপ করে, আজকের ডেটা ভিজ্যুয়ালাইজেশন সরঞ্জামগুলি আপনাকে কেবল আপনার কেপিআই দেখতে দেয় না; স্পষ্টতই, তারা আপনার কেপিআই, বাজার এবং বিশ্বের মধ্যে সম্পর্ক প্রকাশ করতে ডেটা একত্রিত করে এবং এআই-চালিত বিশ্লেষণ প্রয়োগ করে।

## হ্যাকারের মানসিকতা

একটি সাধারণ বৈজ্ঞানিক মানসিকতা হল মডেল তৈরি করা, প্রশিক্ষণ দেওয়া, প্লট গ্রাফ তৈরি করা এবং বিভিন্ন গুণাবলী বিশ্লেষণ করে সমাধান করা। একজন হ্যাকারের মানসিকতা একজন বিজ্ঞানীর থেকে অনেক আলাদা। তারা সহজ পদ্ধতি ব্যবহার করে সমাধান খোঁজার দিকে বেশি মনোনিবেশ করে। যদিও ডেটা বিজ্ঞানীরা একটি সমস্যার জন্য এতগুলি বিভিন্ন উপাদান ব্যবহার করেন, হ্যাকার একটি সমাধান নিয়ে আসতে জটিলতা দূর করে।[১] অতএব, হ্যাকার মানসিকতা আরো প্রশংসনীয় কারণ বৈজ্ঞানিক মনের সীমাবদ্ধতা তাদের আবদ্ধ করে না।

## ডোমেইন দক্ষতা

ডোমেইন দক্ষতা হল সেই আঠালো যা ডাটা সায়েন্সকে একসাথে ধরে রাখে। উইকিপিডিয়ার মতে, বিষয়বস্তু বা ডোমেইন দক্ষতা হল একটি বিশেষ ক্ষেত্র বা বিষয়ে দক্ষতা, বিশেষ জ্ঞান বা দক্ষতা। জ্ঞানের যে কোন ক্ষেত্র একটি ডাটা সায়েন্স অনুসন্ধানের বিষয় হতে পারে, যার মধ্যে medicineষধ, রাজনীতি, ভৌত ও জৈবিক বিজ্ঞান, বিপণন, তথ্য নিরাপত্তা, জনসংখ্যাতাত্ত্বিক, এমনকি সাহিত্যও সীমাবদ্ধ নয়। প্রতিটি ডেটা সায়েন্স টিমে কমপক্ষে একজনকে অন্তর্ভুক্ত করতে হবে যিনি সমস্যার সমাধানের বিষয়ে একজন বিষয় বিশেষজ্ঞ।

# ডেটা সায়েন্সে গণিত

গণিত যেকোনো আধুনিক বৈজ্ঞানিক অনুশাসনের ভিত্তি। এবং এটি কারও কাছে গোপন নয় যে প্রায় সমস্ত আধুনিক ডেটা সায়েন্স পদ্ধতি (মেশিন লার্নিং সহ) কিছু ধরণের গাণিতিক গণনার উপর ভিত্তি করে। কখনও কখনও, একজন তথ্য বিজ্ঞানী (অথবা এমনকি একজন জুনিয়র বিশ্লেষক) হিসাবে, আপনার পদ্ধতিগুলি সঠিকভাবে প্রয়োগ করার জন্য আপনাকে মৌলিক গণিত জানতে হবে। অন্যান্য উদ্দেশ্যে, আপনি API বা একটি প্রস্তুত অ্যালগরিদম ব্যবহার করতে পারেন।

কিন্তু, একই সময়ে, নানামি গণিতের ভাল দখল যার উপর আপনার অ্যালগরিদম তৈরি করতে হবে যাতে পণ্য ব্যবহারের জন্য সুপারিশ তৈরি করা যায়। এটি আপনাকে আপনার প্রতিযোগীদের উপর একটি প্রান্ত দেবে এবং আপনাকে আপনার জ্ঞানের উপর আস্থা বজায় রাখতে সহায়তা করবে। গাড়ির বিষয়ে কিছু না জেনে শুধু চাকার পিছনে বসে থাকার চেয়ে হুডের নীচে কী আছে তা জানা সবসময় ভাল।

অবশ্যই, আপনার অন্যান্য জ্ঞান, প্রোগ্রামিং দক্ষতা, সামান্য ব্যবসায়িক দক্ষতা, অনন্য বিশ্লেষণাত্মক চিন্তাভাবনা এবং ডেটা সম্পর্কে কৌতূহল প্রয়োজন হবে, যা একজন শীর্ষস্থানীয় ডেটা বিজ্ঞানীর জন্য খুব প্রয়োজনীয়। এই প্রবন্ধে, আমি এই প্রচেষ্টায় আপনাকে সাহায্য করার জন্য সবচেয়ে গুরুত্বপূর্ণ গাণিতিক ধারণাগুলি সংগ্রহ করার চেষ্টা করেছি।

গণিতের মূল বিষয়গুলির জ্ঞান এমন পেশাদারদের জন্য অপরিহার্য যারা এই অঞ্চলে অন্য বিশেষায়ণ (হার্ডওয়্যার ডেভেলপমেন্ট, ট্রেড, কেমিক্যাল ইন্ডাস্ট্রি, মেডিসিন অ্যান্ড হেলথকেয়ার, বিজনেস ম্যানেজমেন্ট ইত্যাদি) থেকে এই এলাকায় যেতে চান।

এবং যখন তারা মনে করতে পারে যে তারা দীর্ঘদিন ধরে স্প্রেডশীট, সংখ্যাসূচক গণনা এবং ভবিষ্যদ্বাণী নিয়ে কাজ করেছে, গণিতের দক্ষতার প্রয়োজনীয়তা তাদের ডেটা সায়েন্সে বর্তমান চাকরির থেকে একেবারেই আলাদা।

এখানে বিষয়গুলির কিছু হ্যান্ডপিকড পরামর্শ দেওয়া হয়েছে যা ডেটা সায়েন্স গেমের শীর্ষে থাকার জন্য কাজে আসবে।

## ফাংশন, ভেরিয়েবল, সমীকরণ, গ্রাফ

আসুন রৈখিক সমীকরণের মতো মৌলিক বিষয়গুলি দিয়ে শুরু করি এবং নিউটনের দ্বিপদ এবং এর বৈশিষ্ট্যগুলির সাথে শেষ করি।

- লগারিদম, সূচকীয়, বহুপদী ফাংশন, যুক্তিসঙ্গত সংখ্যা।
- জ্যামিতি এবং মৌলিক তত্ত্বের ভিত্তি, ত্রিকোণমিতিক পরিচয়।
- বাস্তব এবং জটিল সংখ্যা, তাদের মৌলিক বৈশিষ্ট্য।
- সিরিজ, যোগফল এবং অসমতা।
- প্লটিং, আয়তক্ষেত্রাকার এবং মেরু সমন্বয় ব্যবস্থা, টেপারড বিভাগ।

আপনি যদি লক্ষ লক্ষ সাজানো আইটেমের সাথে একটি ডাটাবেসে কীভাবে দ্রুত কিছু খুঁজে বের করতে চান তা জানতে চান তবে আপনার বাইনারি অনুসন্ধানের ধারণাটি প্রয়োজন। গতিবিদ্যা বোঝার জন্য, আপনাকে প্রথমে লগারিদম এবং পুনরাবৃত্তি সমীকরণ বুঝতে হবে। অথবা, যদি আপনি সময় সিরিজ বিশ্লেষণ করতে চান, পর্যায়ক্রমিক ফাংশন এবং সূচকীয় আইনের মত ধারণা ব্যবহার করুন।

## পরিসংখ্যান

ডেটা সায়েন্টিস্ট হিসেবে বেড়ে ওঠার জন্য আপনার এটাই জানা দরকার। পরিসংখ্যান এবং সম্ভাবনার মৌলিক ধারণাগুলির একটি পরিষ্কার বোঝার গুরুত্ব তথ্য বিজ্ঞান সম্পর্কে আলোচনায় বাড়াবাড়ি করা যাবে না। ক্ষেত্রের অনেক অনুশীলনকারীরা ক্লাসিক্যাল মেশিন লার্নিংকে (নিউরাল নেটওয়ার্ক নয়) পরিসংখ্যানগত শিক্ষার চেয়ে বেশি কিছু বলে না। বিষয়টি বিস্তৃত এবং অবিরাম, এবং তাই যতটা সম্ভব মূল ধারণাগুলি কভার করার জন্য নিবদ্ধ পরিকল্পনা অপরিহার্য।

- সংক্ষিপ্তসার এবং বর্ণনামূলক পরিসংখ্যান, গড়, বৈচিত্র্য, সহবাস, পারস্পরিক সম্পর্ক।
- সম্ভাব্য তত্ত্বের মৌলিক বিষয়: মৌলিক ধারণা, প্রত্যাশা, সম্ভাবনার ক্যালকুলাস, বায়েসের উপপাদ্য, শর্তাধীন সম্ভাবনা।

- সম্ভাব্যতা বিতরণ ফাংশন - অভিন্ন, স্বাভাবিক, দ্বিপদ, চি -বর্গ, ছাত্রদের t বিতরণ, কেন্দ্রীয় সীমা উপপাদ্য।
- নমুনা, পরিমাপ, ত্রুটি, এলোমেলো সংখ্যা জেনারেটর।
- হাইপোথিসিস টেস্টিং, এ/বি টেস্টিং, কনফিডেন্স ব্যবধান, পি-ভ্যালু।
- আনোভা, টি-টেস্ট।
- লিনিয়ার রিগ্রেশন, রেগুলারাইজেশন।

আপনি তাদের কোথায় ব্যবহার করতে পারেন? সাক্ষাৎকারের সময়। আমাকে বিশ্বাস কর. একজন দূরদর্শী ডেটা বিজ্ঞানী হিসাবে, আপনি উপরের সমস্ত ধারণাগুলি আয়ত্ত করে দ্রুত আপনার ভবিষ্যতের নিয়োগকর্তার উপর একটি ভাল ছাপ ফেলতে পারেন। কাজ করার সময়, আপনাকে প্রায়শই কিছু ধারণা ব্যবহার করার প্রয়োজন মোকাবেলা করতে হবে।

## রৈখিক বীজগণিত

ফেসবুক বন্ধুদের সুপারিশ, স্পটিফাই গানের সুপারিশ, সালভাদর ডালি-স্টাইলের প্রভাব গভীর স্নায়ু নেটওয়ার্ক ট্রান্সফার লার্নিং ব্যবহার করে ফটোগ্রাফির। তাদের সবার কি মিল আছে? ম্যাট্রিক্স এবং ম্যাট্রিক্স বীজগণিত সর্বত্র ব্যবহৃত হয়। ম্যাট্রিক্স বীজগণিত হল গণিতের একটি গুরুত্বপূর্ণ দিক যা আপনাকে বুঝতে সাহায্য করে কিভাবে একটি যন্ত্র প্রবাহে অধিকাংশ মেশিন লার্নিং অ্যালগরিদম কাজ করে। অন্বেষণের জন্য সবচেয়ে গুরুত্বপূর্ণ বিষয়গুলি হল:

- একটি ম্যাট্রিক্স এবং ভেক্টরের প্রধান বৈশিষ্ট্য হল ডট প্রোডাক্ট, লিনিয়ার ট্রান্সফরমেশন, ট্রান্সপোজিশন, কনজুগেশন, র‍্যাঙ্ক, নির্ধারক।
- অভ্যন্তরীণ এবং বাইরের পণ্য, ম্যাট্রিক্স গুণের নিয়ম এবং বিভিন্ন অ্যালগরিদম, বিপরীত ম্যাট্রিক্স।
- স্থানিক ম্যাট্রিক্স - বর্গ, একক, ত্রিভুজাকার, স্পার্স, ঘন, প্রতিসম, হার্মিশিয়ান, অ্যান্টি -হারমিটিয়ান এবং একক ম্যাট্রিক্স, ইউনিট ভেক্টর।
- ম্যাট্রিক্স পচনের ধারণা
- ভেক্টর স্পেস, বেসিস, হুল, অরথগোনালিটি, লিনিয়ার ন্যূনতম স্কোয়ার।
- ম্যাট্রিক্স eigenvalue, eigenvector, diagonalisation, singular value decomposition (SVD)।

আপনি তাদের কোথায় ব্যবহার করতে পারেন? যদি আপনি মাত্রিকতা হ্রাসের জন্য প্রধান উপাদান বিশ্লেষণ (PCA) ব্যবহার করেন, আপনি সম্ভবত কম প্যারামিটার সহ আরও কমপ্যাক্ট ডেটা

মাত্রার জন্য একবচন মান পচন ব্যবহার করবেন। সমস্ত নিউরাল নেটওয়ার্ক অ্যালগরিদম নেটওয়ার্ক স্ট্রাকচার এবং লার্নিং অপারেশনগুলিকে প্রতিনিধিত্ব এবং প্রক্রিয়া করার জন্য রৈখিক বীজগণিত কৌশল ব্যবহার করে।

## গাণিতিক বিশ্লেষণ

আপনি বিশ্ববিদ্যালয়ে এটি পছন্দ করেন বা না করেন, আমরা ডেটা সায়েন্স এবং মেশিন লার্নিংয়ের অনেক ক্ষেত্রে ক্যালকুলাসের মুখোমুখি হই। এটি একটি সাধারণ সমস্যাটির একটি আপাতদৃষ্টিতে সহজ বিশ্লেষণাত্মক সমাধানের পিছনে লুকিয়ে আছে যা লিনিয়ার রিগ্রেশনে একটি চতুর্ভুজ ফাংশনের ন্যূনতম মান দিয়ে থাকে। এটি প্রশিক্ষণের জন্য নিউরাল নেটওয়ার্ক দ্বারা উত্পন্ন প্রতিটি ব্যাকপ্রোপ্যাগেশন পদ্ধতিতেও অন্তর্ভুক্ত। গাণিতিক বিশ্লেষণের জ্ঞান আপনার কাজের জন্য খুবই মূল্যবান প্রমাণিত হবে। অন্বেষণ করার জন্য নিম্নলিখিত বিষয়গুলি রয়েছে:

- একক পরিবর্তনশীল ফাংশন, সীমা, ধারাবাহিকতা এবং ভিন্নতা।
- সীমাবদ্ধ বৃদ্ধির সূত্র, অনিশ্চয়তা প্রকাশ, L'Hôpital এর উপপাদ্য।
- সর্বোচ্চ এবং সর্বনিম্ন।
- একটি জটিল ফাংশনের পণ্য এবং ভিন্নতার নিয়ম।
- টেলর সিরিজ, অসীম সিরিজ সংক্ষেপণ/ইন্টিগ্রেশন ধারণা।
- অবিচ্ছেদ্য ক্যালকুলাসের সীমাবদ্ধ বৃদ্ধির মূল উপপাদ্য এবং সূত্র, নির্দিষ্ট এবং অনুপযুক্ত ইন্টিগ্রালের গণনা।
- বিটা এবং গামা ফাংশন।
- ভেরিয়েবলের একটি সেটের কাজ, সীমা, ধারাবাহিকতা এবং আংশিক ডেরিভেটিভস।
- সাধারণ ডিফারেনশিয়াল সমীকরণ এবং আংশিক ডিফারেনশিয়াল সমীকরণের মৌলিক বিষয়গুলি (সবচেয়ে কঠিন নয়)।

এগুলি কোথায় ব্যবহার করবেন? আপনি সম্ভবত ভাবছেন কিভাবে লজিস্টিক রিগ্রেশন অ্যালগরিদম ব্যবহার করা হয়। সর্বনিম্ন ক্ষতি ফাংশন খুঁজে পেতে, গ্রেডিয়েন্ট বংশধর পদ্ধতিটি প্রায়শই ব্যবহৃত হয়। এটি কীভাবে কাজ করে তা বোঝার জন্য, গাণিতিক বিশ্লেষণ ধারণাগুলি ব্যবহার করা প্রয়োজন: গ্রেডিয়েন্ট, ডেরিভেটিভস, সীমা, একটি জটিল ফাংশনের পার্থক্য।

## আলাদা গণিত

"ডেটা সায়েন্সে গণিত" এর মতো একটি বিষয় নিয়ে আলোচনা করার সময় আলাদা গণিতকে খুব কমই স্পর্শ করা হয়। তবুও, আধুনিক ডেটা সায়েন্স তৈরি করা হয় কম্পিউটিং সিস্টেমের সাহায্যে যেখানে আলাদা গণিত একটি মূল উপাদান। পৃথক গণিত কোর্স আপনাকে বিশ্লেষণাত্মক প্রকল্পগুলিতে কাজ করার সময় অ্যালগরিদম এবং ডেটা স্ট্রাকচারের দৈনন্দিন ব্যবহারের জন্য গুরুত্বপূর্ণ ধারণাগুলি আয়ত্ত করতে সহায়তা করবে। অন্বেষণ করার জন্য কিছু বিষয় নিচে দেওয়া হল:

- সেট, উপসেট, বুলিয়ান।
- গণনা ফাংশন, combinatorics, countability।
- প্রমাণের প্রধান পদ্ধতি হল আনয়ন, দ্বন্দ্ব দ্বারা প্রমাণ।
- প্রবর্তনমূলক, বিয়োগমূলক এবং প্রস্তাবিত যুক্তির ভিত্তি।
- প্রধান ডেটা স্ট্রাকচার হল স্ট্যাক, কিউ, গ্রাফ, অ্যারে, হ্যাশ টেবিল, ট্রি।
- গ্রাফ ইনভারিয়েন্টস: সংযুক্ত উপাদান, ভারটেক্স ডিগ্রী, ফোর্ড - ফুলকারসন তত্ত্ব, গ্রাফ কালারিং।
- পুনরাবৃত্ত সূত্র (সমীকরণ, সম্পর্ক)।
- ফাংশন বৃদ্ধি, "O" স্বরলিপি বড়।

যেকোনো সামাজিক নেটওয়ার্ক বিশ্লেষণ করার সময় গ্রাফ ইনভারিয়েন্টস এবং দ্রুত অ্যালগরিদম অপরিহার্য। এগুলো কোথায় ব্যবহার করা যাবে? যেকোনো অ্যালগরিদমের সাথে, আপনাকে বড় O স্বরলিপি ব্যবহার করে সাময়িক এবং স্থানিক জটিলতা বুঝতে হবে। এটি প্রয়োজনীয়, উদাহরণস্বরূপ, ইনপুট ডেটার পরিমাণ বৃদ্ধির সাথে সাথে রান সময় এবং প্রয়োজনীয় আকার কীভাবে বৃদ্ধি করে তা নির্ধারণ করার সময়।

## অপ্টিমাইজেশন, অপারেশন গবেষণা বিষয়

এই বিষয়গুলি প্রযোজ্য গণিতের প্রচলিত বক্তৃতা থেকে খুব বেশি আলাদা নয় কারণ এগুলি প্রধানত গুরুত্বপূর্ণ এবং অধ্যয়নের বিশেষ ক্ষেত্রে সর্বাধিক ব্যবহৃত হয়: তাত্ত্বিক কম্পিউটার বিজ্ঞান, নিয়ন্ত্রণ তত্ত্ব, অপারেশন গবেষণায়। কিন্তু এই কার্যকরী পদ্ধতির একটি সাধারণ উপলব্ধি মেশিন লার্নিং এর ক্ষেত্রে সহায়ক হতে পারে। প্রায় প্রতিটি মেশিন লার্নিং অ্যালগরিদম/পদ্ধতির লক্ষ্য বিভিন্ন সীমাবদ্ধতার কারণে অনুমানের ত্রুটিকে কিছুটা কমানো। এটি অপ্টিমাইজেশনের লক্ষ্য। অধ্যয়নের বিষয়:

- অপ্টিমাইজেশনের মূল বিষয়গুলি - কীভাবে একটি সমস্যা তৈরি করা যায়।

- সর্বোচ্চ, সর্বনিম্ন, উত্তল ফাংশন, বৈশ্বিক সমাধান।
- লিনিয়ার প্রোগ্রামিং, সিমপ্লেক্স পদ্ধতি।
- ইন্টিজার প্রোগ্রামিং।
- সীমাবদ্ধ প্রোগ্রামিং, ন্যাপস্যাক সমস্যা।
- র্যান্ডমাইজড অপটিমাইজেশন পদ্ধতি - উপরের দিকে আরোহণ দ্বারা অনুসন্ধান, সিমুলেটেড অ্যানিলিং অ্যালগরিদম, জেনেটিক অ্যালগরিদম।

এগুলো কোথায় ব্যবহার করা যাবে? লজিস্টিকগুলির বিপরীতে, ন্যূনতম-বর্গ ক্ষতির ফাংশন ব্যবহার করে সহজ রৈখিক প্রতিক্রিয়া সমস্যা, প্রায়ই একটি সঠিক বিশ্লেষণাত্মক সমাধান থাকে। কারণটি বুঝতে, আপনাকে অপ্টিমাইজেশনে উত্তলতার মতো ধারণা সম্পর্কে জানতে হবে। এটি আরও ব্যাখ্যা করবে কেন আমাদের অনেক মেশিন লার্নিং সমস্যার জন্য পর্যাপ্ত "রুক্ষ" সমাধান থাকা উচিত। একটি অপ্টিমাইজেশান একটি শক্তিশালী হাতিয়ার যা বিস্তারিতভাবে অন্বেষণ করা যায়।

# গাণিতিক বিশ্লেষণ

মৌলিক গণিত ছাড়া একজন ভাল ডেটা বিশ্লেষক কোথাও নেই (এবং ডেটা গবেষক আরও বেশি)। সুতরাং, আসুন ডেটা সায়েন্সের স্বার্থে ক্ষেত্রগুলি বুঝতে পারি।

## গাণিতিক বিশ্লেষণের মূল বিষয়

- ফাংশন এবং তাদের বৈশিষ্ট্য।
- ফাংশন সীমা (মৌলিক মতামত)।
- ডেরিভেটিভ ফাংশন (এর জ্যামিতিক এবং যান্ত্রিক অর্থ)।
- একটি জটিল ফাংশনের ডেরিভেটিভ।
- চরম বৈশিষ্ট্য। বুল ফাংশন।
- ব্যক্তিগত ডেরিভেটিভস এবং গ্রেডিয়েন্ট।
- অপ্টিমাইজেশান কাজের মধ্যে গ্রেডিয়েন্ট।
- দিক থেকে ডেরিভেটিভ।
- সমতল স্পর্শ এবং রৈখিক আনুমানিকতা।

## রৈখিক বীজগণিতের মূল বিষয়

- ভেক্টর স্পেস।
- রৈখিক স্বাধীনতা।
- ভেক্টরগুলির আদর্শ এবং স্কেলার কাজ।
- ম্যাট্রিক্স নির্ধারণ। ম্যাট্রিক্সে অপারেশন।
- র্যাংক এবং ম্যাট্রিক্সের নির্ধারক।

- লাইন সমীকরণ সিস্টেম।
- ম্যাট্রিক্স প্রকার।
- নিজস্ব ভেক্টর এবং নিজস্ব মান।
- ম্যাট্রিক্স পচন (বর্ণালী, একবচন)।
- নিম্ন র্যাঙ্কের ম্যাট্রিক্সের কাছে।
- একবচন পচন এবং কম বার্ন আনুমানিক।

## অপ্টিমাইজেশন পদ্ধতি

- নন-স্মুথ ফাংশন অপ্টিমাইজ করা (স্থানীয় নিম্নগতির সমস্যা)।
- ইগনিশন অনুকরণ পদ্ধতি।
- জেনেটিক আলগোরিদিম. ডিফারেনশিয়াল বিবর্তনের অ্যালগরিদম।
- Ndder-Mead পদ্ধতি।

## সম্ভাব্যতা তত্ত্ব এবং গাণিতিক পরিসংখ্যান

- সম্ভাব্যতা নির্ধারণ। সম্ভাব্য বৈশিষ্ট্য।
- শর্তাধীন সম্ভাবনা। পূর্ণ সম্ভাবনার সূত্র। ফর্মুলা বায়েস।
- বিচক্ষণ এলোমেলো মান।
- ক্রমাগত এলোমেলো মান।
- নমুনা বিতরণের অনুমান। পরিসংখ্যান।
- বিতরণের বৈশিষ্ট্য।
- গুরুত্বপূর্ণ পরিসংখ্যান (নির্বাচনী গড়, মধ্যমা, ফ্যাশন, বৈচিত্র্য, অন্তর্বর্তী সুইং)।
- উপপাদ্যের কেন্দ্রীয় সীমা।
- আস্থা অন্তর.

# পরিসংখ্যানগত মডেলিং

একটি পরিসংখ্যানগত মডেল হল একটি গাণিতিক মডেল যা নমুনা ডেটা প্রজন্ম (এবং বৃহত্তর জনসংখ্যার অনুরূপ ডেটা) সম্পর্কিত পরিসংখ্যানগত অনুমানকে ধারণ করে। একটি পরিসংখ্যানগত মডেল প্রতিনিধিত্ব করে, প্রায়শই যথেষ্ট আদর্শ আকারে, ডেটা তৈরি করার প্রক্রিয়া। একটি পরিসংখ্যানগত মডেল সাধারণত এক বা একাধিক এলোমেলো ভেরিয়েবল এবং অন্যান্য নন-র্যান্ডম ভেরিয়েবলের মধ্যে গাণিতিক সম্পর্ক হিসেবে নির্দিষ্ট করা হয়। কিছু উপকারী পরিসংখ্যানগত মডেলিং পদ্ধতি নীচে বর্ণিত হয়েছে:

## স্থানিক মডেল

স্থানিক নির্ভরতা হল ভৌগলিক স্থানের মধ্যে বৈশিষ্ট্যের সহ-বৈচিত্র্য: প্রক্সিমাল লোকেশনের বৈশিষ্ট্যগুলি ইতিবাচক বা নেতিবাচকভাবে সম্পর্কযুক্ত বলে মনে হয়। স্থানিক নির্ভরতা পরিসংখ্যানগুলিতে স্থানিক অটো-পারস্পরিক সম্পর্কের সমস্যার দিকে পরিচালিত করে, যেমন সাময়িক অটো-পারস্পরিক সম্পর্কের মতো, এটি মানসম্মত পরিসংখ্যান কৌশলগুলিকে লঙ্ঘন করে যা পর্যবেক্ষণের মধ্যে স্বাধীনতা অনুমান করে[৩]

## সময় সিরিজ

সময় সিরিজ বিশ্লেষণের পদ্ধতি দুটি শ্রেণীতে বিভক্ত করা যেতে পারে: ফ্রিকোয়েন্সি-ডোমেন পদ্ধতি এবং সময়-ডোমেন পদ্ধতি। প্রাক্তন বর্ণালী বিশ্লেষণ এবং সম্প্রতি তরঙ্গাকৃতি বিশ্লেষণ

অন্তর্ভুক্ত; পরেরটির মধ্যে রয়েছে অটো-পারস্পরিক সম্পর্ক এবং ক্রস-পারস্পরিক সম্পর্ক বিশ্লেষণ। টাইম ডোমেইনে, পারস্পরিক সম্পর্ক বিশ্লেষণগুলি স্কেলযুক্ত পারস্পরিক সম্পর্ক ব্যবহার করে ফিল্টারের মতো পদ্ধতিতে তৈরি করা যেতে পারে, যার ফলে ফ্রিকোয়েন্সি ডোমেনে কাজ করার প্রয়োজনীয়তা হ্রাস পায়।

উপরন্তু, সময় সিরিজ বিশ্লেষণ কৌশলগুলি প্যারামেট্রিক এবং নন-প্যারামেট্রিক পদ্ধতিতে বিভক্ত করা যেতে পারে। প্যারামেট্রিক পদ্ধতিগুলি অনুমান করে যে অন্তর্নিহিত স্থির স্টোকাস্টিক প্রক্রিয়াটির একটি নির্দিষ্ট কাঠামো রয়েছে যা অল্প সংখ্যক পরামিতি ব্যবহার করে বর্ণনা করা যেতে পারে (উদাহরণস্বরূপ, একটি অটোরগ্রেসিভ বা মুভিং এভারেজ মডেল ব্যবহার করে)। এই পদ্ধতির মধ্যে, কাজটি স্টোকাস্টিক প্রক্রিয়া বর্ণনা করে এমন মডেলের পরামিতিগুলি অনুমান করা। বিপরীতে, অ-প্যারামেট্রিক পদ্ধতিগুলি স্পষ্টভাবে অনুমান করে যে প্রক্রিয়াটির কোন বিশেষ কাঠামো আছে তা না ধরে কোভারিয়েন্স বা প্রক্রিয়ার বর্ণালী অনুমান করে। সময় সিরিজ বিশ্লেষণের পদ্ধতিগুলিও লিনিয়ার এবং নন-লিনিয়ার, এবং ইউনিভারিয়েট এবং মাল্টিভারিয়েটে বিভক্ত হতে পারে।

## বেঁচে থাকার বিশ্লেষণ

বেঁচে থাকার বিশ্লেষণ পরিসংখ্যানের একটি শাখা যা এক বা একাধিক ঘটনা না হওয়া পর্যন্ত সময়ের প্রত্যাশিত সময়কাল বিশ্লেষণ করে, যেমন জৈবিক প্রাণীর মৃত্যু এবং যান্ত্রিক ব্যবস্থায় ব্যর্থতা। এই বিষয়টিকে বলা হয় নির্ভরযোগ্যতা তত্ত্ব বা ইঞ্জিনিয়ারিংয়ে নির্ভরযোগ্যতা বিশ্লেষণ, অর্থনীতিতে সময়কাল বিশ্লেষণ বা সময়কাল মডেলিং এবং সমাজবিজ্ঞানে ইভেন্ট ইতিহাস বিশ্লেষণ। বেঁচে থাকার বিশ্লেষণ প্রশ্নের উত্তর দেওয়ার চেষ্টা করে যেমন: একটি নির্দিষ্ট সময়ের অতীত বেঁচে থাকা জনসংখ্যার অনুপাত কত? যারা বেঁচে আছে, তারা কোন হারে মারা যাবে বা ব্যর্থ হবে? মৃত্যু বা ব্যর্থতার একাধিক কারণ কি বিবেচনায় নেওয়া যেতে পারে? নির্দিষ্ট পরিস্থিতি বা বৈশিষ্ট্যগুলি কীভাবে বেঁচে থাকার সম্ভাবনা বাড়ায় বা হ্রাস করে? বেঁচে থাকার মডেলগুলি অ্যাকচুয়ারী এবং পরিসংখ্যানবিদরা ব্যবহার করেন এবং বিপণনকারীরা মন্থন এবং ব্যবহারকারী ধরে রাখার মডেলগুলি ডিজাইন করেন।[7]

বেঁচে থাকার মডেলগুলি সময়-থেকে-ইভেন্টের পূর্বাভাস দেওয়ার জন্যও ব্যবহার করা হয় (মৌলবাদী হওয়া থেকে সন্ত্রাসী হয়ে ওঠার সময় বা যখন বন্দুক কেনা হয় এবং হত্যায় ব্যবহৃত হয়)

## বাজার বিভাজন

মার্কেট সেগমেন্টেশন, যাকে কাস্টমার প্রোফাইলিংও বলা হয়, একটি মার্কেটিং স্ট্র্যাটেজি যা একটি বিস্তৃত টার্গেট মার্কেটকে ভোক্তা, ব্যবসা, বা দেশগুলির উপসেটগুলিতে বিভক্ত করে থাকে যাদের সাধারণ চাহিদা, আগ্রহ এবং অগ্রাধিকার আছে বা অনুভূত হয় এবং তারপর লক্ষ্য নির্ধারণের কৌশলগুলি ডিজাইন এবং বাস্তবায়ন করে তাদের মার্কেট সেগমেন্টেশন কৌশলগুলি সাধারণত টার্গেট গ্রাহকদের চিহ্নিত করতে এবং আরও সংজ্ঞায়িত করতে এবং বিপণন পরিকল্পনার উপাদানগুলির জন্য সহায়ক ডেটা সরবরাহ করতে ব্যবহৃত হয় যেমন নির্দিষ্ট বিপণন পরিকল্পনার লক্ষ্য অর্জনের জন্য অবস্থান। টার্গেট সেগমেন্টের সুনির্দিষ্ট চাহিদা এবং বৈশিষ্ট্যের উপর নির্ভর করে ব্যবসাগুলি পণ্য বিভাজন কৌশল বা নির্দিষ্ট পণ্য বা পণ্যের রেখার সাথে জড়িত একটি অভিন্ন পদ্ধতির বিকাশ করতে পারে।

## সুপারিশ সিস্টেম

রিকমেন্ডার সিস্টেম বা সুপারিশ সিস্টেম (কখনও কখনও প্ল্যাটফর্ম বা ইঞ্জিনের মতো প্রতিশব্দ দিয়ে "সিস্টেম" প্রতিস্থাপন করা হয়) হল তথ্য ফিল্টারিং সিস্টেমের একটি উপশ্রেণী যা ব্যবহারকারী একটি আইটেমকে 'রেটিং' বা 'পছন্দ' সম্পর্কে ভবিষ্যদ্বাণী করতে চায়।

## সমিতির নিয়ম শেখা

অ্যাসোসিয়েশন রুল লার্নিং বড় ডাটাবেসে ভেরিয়েবলের মধ্যে আকর্ষণীয় সম্পর্ক আবিষ্কারের একটি পদ্ধতি। জালিয়াতি শনাক্তকরণের ক্ষেত্রে, জালিয়াতির সাথে যুক্ত নিদর্শন সনাক্ত করতে সমিতির নিয়ম ব্যবহার করা হয়। উদাহরণস্বরূপ, একটি সুপার মার্কেটের বিক্রয় তথ্যে পাওয়া {{পেঁয়াজ, আলু} ==> {বার্গার} নিয়মটি ইঙ্গিত করবে যে যদি কোন গ্রাহক পেঁয়াজ এবং আলু একসাথে কিনে থাকে তবে তারা হ্যামবার্গার মাংসও কিনতে পারে। অতিরিক্ত জালিয়াতির ঘটনা শনাক্ত করার জন্য লিংকেজ বিশ্লেষণ করা হয়: যদি স্টোর বি থেকে সমস্ত লেনদেন বিশ্লেষণ করে স্টোর বি -তে জালিয়াতিমূলক কেনাকাটা করার জন্য ব্যবহারকারীর ক্রেডিট কার্ড লেনদেন ব্যবহার করা হয়, তাহলে আমরা জালিয়াতিমূলক কার্যকলাপের সাথে অন্য ব্যবহারকারী সি খুঁজে পেতে পারি।

## অ্যাট্রিবিউশন মডেলিং

একটি অ্যাট্রিবিউশন মডেল হল নিয়ম বা নিয়মের সেট যা নির্ধারণ করে কিভাবে বিক্রয় এবং রূপান্তরের জন্য ক্রেডিট রূপান্তর পথের টাচপয়েন্টগুলিতে বরাদ্দ করা হয়। উদাহরণস্বরূপ, গুগল

অ্যানালিটিক্সের লাস্ট ইন্টারঅ্যাকশন মডেলটি চূড়ান্ত টাচপয়েন্টগুলিতে (অর্থাৎ, ক্লিক) 100% ক্রেডিট বরাদ্দ করে যা অবিলম্বে বিক্রয় বা রূপান্তরের আগে। ম্যাক্রো-ইকোনমিক মডেলগুলি দীর্ঘমেয়াদী, সমষ্টিগত dataতিহাসিক তথ্য ব্যবহার করে প্রতিটি বিক্রয় বা রূপান্তরের জন্য বেশ কয়েকটি চ্যানেলে একটি অ্যাট্রিবিউশন ওজন নির্ধারণ করে। এই মডেলগুলি বিজ্ঞাপন মিশ্রণ অপ্টিমাইজেশনের জন্যও ব্যবহৃত হয়।

## স্কোরিং

স্কোরিং মডেল একটি বিশেষ ধরনের ভবিষ্যদ্বাণীমূলক মডেল। ভবিষ্যদ্বাণীমূলক মডেলগুলি loanণ পরিশোধ, দুর্ঘটনার ঝুঁকি, ক্লায়েন্ট মন্থন বা অবনতি, বা ভাল জিনিস কেনার সম্ভাবনা সম্পর্কে ভবিষ্যদ্বাণী করতে পারে। স্কোরিং মডেলগুলি সাধারণত লগারিদমিক স্কেল ব্যবহার করে (আপনার স্কোরের প্রতিটি অতিরিক্ত 50 পয়েন্ট, ডিফল্ট হওয়ার ঝুঁকি 50%কমিয়ে দেয়)। এগুলি লজিস্টিক রিগ্রেশন এবং ডিসিশন ট্রি বা একাধিক অ্যালগরিদমের সংমিশ্রণের উপর ভিত্তি করে। স্কোরিং টেকনোলজি সাধারণত লেনদেনের ডেটাতে প্রয়োগ করা হয়, কখনও কখনও রিয়েল-টাইমে (ক্রেডিট কার্ড জালিয়াতি সনাক্তকরণ, ক্লিক জালিয়াতি)।

## ভবিষ্যদ্বাণীপূর্ণ মডেলিং

ভবিষ্যদ্বাণীমূলক মডেলিং ফলাফলের পূর্বাভাস দেওয়ার জন্য পরিসংখ্যান ব্যবহার করে। প্রায়শই, যে ঘটনাটি ভবিষ্যদ্বাণী করতে চায় তা ভবিষ্যতে হয়, কিন্তু ভবিষ্যদ্বাণীমূলক মডেলিং যে কোন ধরনের অজানা ইভেন্টে প্রয়োগ করা যেতে পারে, তা কখনই ঘটে না কেন। উদাহরণস্বরূপ, ভবিষ্যদ্বাণীমূলক মডেলগুলি প্রায়ই অপরাধ সনাক্ত করতে এবং অপরাধ সংঘটিত হওয়ার পর সন্দেহভাজনদের চিহ্নিত করতে ব্যবহৃত হয়। এগুলি আবহাওয়ার পূর্বাভাস, স্টক মার্কেটের দামের পূর্বাভাস দেওয়ার জন্য, বা বিক্রির পূর্বাভাস দিতে, সময় সিরিজ বা স্থানিক মডেল অন্তর্ভুক্ত করার জন্যও ব্যবহার করা যেতে পারে। নিউরাল নেটওয়ার্ক, লিনিয়ার রিগ্রেশন, ডিসিশন ট্রি এবং সাদাসিধা Bayes হল ভবিষ্যদ্বাণীমূলক মডেলিংয়ের জন্য ব্যবহৃত কৌশল। তারা একটি প্রশিক্ষণ সেট, ক্রস-বৈধতা, এবং মডেল ফিটিং এবং নির্বাচন তৈরির সাথে যুক্ত।

## ক্লাস্টারিং

ক্লাস্টার বিশ্লেষণ বা ক্লাস্টারিং হল বস্তুর একটি সেটকে এমনভাবে গোষ্ঠীভুক্ত করার কাজ যে একই গোষ্ঠীর বস্তু (যাকে ক্লাস্টার বলা হয়) অন্য গোষ্ঠীর (ক্লাস্টার) তুলনায় একে অপরের সাথে

অনেক বেশি অনুরূপ (কিছু অর্থে বা অন্যভাবে) । এটি অনুসন্ধানমূলক ডেটা মাইনিং এবং মেশিন লার্নিং, প্যাটার্ন রিকগনিশন, ইমেজ এনালাইসিস, ইনফরমেশন রিক্রিভালেশন এবং বায়োইনফরম্যাটিক্স সহ অনেক ক্ষেত্রে ব্যবহৃত একটি সাধারণ পরিসংখ্যানগত ডেটা বিশ্লেষণ কৌশল।

তত্ত্বাবধানে শ্রেণীবিন্যাসের বিপরীতে (নীচে), ক্লাস্টারিং প্রশিক্ষণ সেট ব্যবহার করে না। যাইহোক, কিছু সংকর বাস্তবায়ন আছে যাকে বলা হয় আধা-তত্ত্বাবধানে শেখা।

## তত্ত্বাবধানে শ্রেণীবিভাগ

তত্ত্বাবধানে শ্রেণীবিন্যাস, যাকে তত্ত্বাবধানে শেখাও বলা হয়, মেশিন লার্নিং কাজটি লেবেলযুক্ত প্রশিক্ষণ ডেটা থেকে একটি ফাংশন বের করা। প্রশিক্ষণ তথ্য প্রশিক্ষণ উদাহরণ একটি সেট গঠিত। তত্ত্বাবধানে শেখার ক্ষেত্রে, প্রতিটি উদাহরণ একটি ইনপুট বস্তু (সাধারণত একটি ভেক্টর) এবং পছন্দসই আউটপুট মান (যাকে লেবেল, ক্লাস বা বিভাগও বলা হয়) সমন্বিত একটি জোড়া। একটি তত্ত্বাবধানে শেখার অ্যালগরিদম প্রশিক্ষণ ডেটা বিশ্লেষণ করে এবং একটি অনুমিত ফাংশন তৈরি করে, যা নতুন উদাহরণ ম্যাপিংয়ের জন্য ব্যবহার করা যেতে পারে। একটি অনুকূল দৃশ্যকল্প অ্যালগরিদমকে অদেখা দৃষ্টান্তের জন্য ক্লাস লেবেলগুলি সঠিকভাবে নির্ধারণ করার অনুমতি দেবে।

## চরম মূল্য তত্ত্ব

চরম মূল্য তত্ত্ব বা চরম মূল্য বিশ্লেষণ (ইভিএ) পরিসংখ্যানের একটি শাখা যা সম্ভাব্যতা বিতরণের মধ্যমা থেকে চরম বিচ্যুতি মোকাবেলা করে। এটি একটি প্রদত্ত র্যান্ডম ভেরিয়েবলের প্রদত্ত অর্ডারকৃত নমুনা থেকে মূল্যায়ন করতে চায়, যা পূর্বে পর্যবেক্ষণের চেয়ে বেশি চরম ঘটনার সম্ভাবনা। উদাহরণস্বরূপ, প্রতি 10, 100 বা 500 বছরে একবার বন্যা হয়। এই মডেলগুলি সম্প্রতি বিপর্যয়কর ঘটনাগুলির পূর্বাভাস দেওয়ার জন্য খারাপভাবে কাজ করছে, যার ফলে বীমা কোম্পানিগুলির ব্যাপক ক্ষতি হয়েছে।

## সিমুলেশন

মন্টে-কার্লো সিমুলেশনগুলি অনেক প্রসঙ্গে ব্যবহৃত হয়: বহু-স্তরের স্প্যাটিও-টেম্পোরাল হায়ারার্কিক্যাল বেইসিয়ান মডেলের মতো জটিল সেটিংসে উচ্চমানের ছদ্ম-র্যান্ডম সংখ্যা তৈরি করতে, বিরল ঘটনাগুলির সাথে সম্পর্কিত পরিসংখ্যান গণনা করার জন্য পরামিতিগুলি অনুমান করতে, অথবা এমনকি একটি তৈরি করতে বিশেষ করে স্টক ট্রেডিং বা ইঞ্জিনিয়ারিংয়ের জন্য বিভিন্ন

অ্যালগরিদম পরীক্ষা এবং তুলনা করার জন্য প্রচুর পরিমাণে ডেটা (উদাহরণস্বরূপ, ক্রস এবং অটো-সম্পর্কযুক্ত সময় সিরিজ)।

## মন্থন বিশ্লেষণ

গ্রাহক মন্থন বিশ্লেষণ আপনাকে উচ্চ মূল্যের গ্রাহকদের সনাক্ত করতে এবং তাদের উপর মনোযোগ কেন্দ্রীভূত করতে, সাধারণত হারানো গ্রাহক বা বিক্রির পূর্বে কোন কাজগুলি নির্ধারণ করতে সাহায্য করে এবং কোন বিষয়গুলি গ্রাহক ধারণকে প্রভাবিত করে তা আরও ভালভাবে বুঝতে সাহায্য করে। পরিসংখ্যানগত কৌশলগুলির মধ্যে রয়েছে বেঁচে থাকার বিশ্লেষণের পাশাপাশি চারটি রাজ্যের মার্কভ চেইন: একেবারে নতুন গ্রাহক, ফেরত আসা গ্রাহক, নিষ্ক্রিয় (হারিয়ে যাওয়া) গ্রাহক এবং পুনরায় অধিগ্রহণ করা গ্রাহক, পথ বিশ্লেষণ সহ (মূল কারণ বিশ্লেষণ সহ) গ্রাহকরা কীভাবে চলে যান তা বোঝার জন্য। এক রাজ্য থেকে অন্য রাজ্যে, মুনাফা বাড়ানোর জন্য। সম্পর্কিত বিষয়: গ্রাহকের আজীবন মূল্য, ব্যবহারকারী অধিগ্রহণের খরচ, ব্যবহারকারী ধরে রাখা।

## ইনভেন্টরি ম্যানেজমেন্ট

ইনভেন্টরি ম্যানেজমেন্ট একটি কোম্পানি যে আইটেমগুলি বিক্রি করবে তা বিক্রয় করার জন্য ব্যবহার করা সামগ্রীর অর্ডার, স্টোরেজ এবং ব্যবহার তত্ত্বাবধান এবং নিয়ন্ত্রণ করবে এবং বিক্রয়ের জন্য সমাপ্ত পণ্যগুলির তত্ত্বাবধান এবং নিয়ন্ত্রণ করবে। ইনভেন্টরি ম্যানেজমেন্ট হল একটি অপারেশন রিসার্চ টেকনিক যা অ্যানালিটিক্স (টাইম সিরিজ, সিজনালিটি, রিগ্রেশন) ব্যবহার করে, বিশেষ করে বিক্রয় পূর্বাভাস এবং সর্বোত্তম মূল্য নির্ধারণের জন্য - প্রতি প্রোডাক্ট ক্যাটাগরি, মার্কেট সেগমেন্ট এবং ভূগোল ভেঙ্গে। এটি প্রাইস অপ্টিমাইজেশনের সাথে দৃ strongly়ভাবে সম্পর্কিত। এটি শুধুমাত্র ইট-মর্টার অপারেশনের জন্য নয়: ইনভেন্টরির অর্থ হতে পারে আগামী 60 দিনের মধ্যে একটি প্রকাশক ওয়েবসাইটে উপলব্ধ ব্যানার বিজ্ঞাপনের স্লটগুলির পরিমাণ, প্রতিটি ব্যানার বিজ্ঞাপনের স্লটটি কতটা ট্রাফিক (এবং রূপান্তর) সরবরাহ করবে তা অনুমান করে সম্ভাব্য বিজ্ঞাপনদাতার কাছে। আপনি এই ভার্চুয়াল ইনভেন্টরির অতিরিক্ত বিক্রি বা কম বিক্রি করতে চান না। এভাবে,

## সর্বোত্তম বিডিং

এটি একটি স্বয়ংক্রিয়, ব্ল্যাক-বক্স, মেশিন-টু-মেশিন যোগাযোগ ব্যবস্থার একটি উদাহরণ, কখনও কখনও বিভিন্ন API এর মাধ্যমে রিয়েল-টাইমে কাজ করে। পরিসংখ্যানগত মডেলগুলি এটিকে সমর্থন করে। অ্যাপ্লিকেশনগুলির মধ্যে রয়েছে লক্ষ লক্ষ কীওয়ার্ডের প্রত্যাশিত রূপান্তর হারের

ভিত্তিতে গুগল অ্যাডওয়ার্ডে সঠিক মূল্যে সঠিক কীওয়ার্ডগুলি সনাক্ত করা এবং কেনা; কীওয়ার্ডগুলিকে একটি সূচীকরণ অ্যালগরিদম ব্যবহার করে শ্রেণীবদ্ধ করা হয় (এই নিবন্ধে আইটেম #18 দেখুন) এবং বালতি স্তরে পরিসংখ্যানগত তাত্পর্য সহ কিছু historicalতিহাসিক তথ্য পেতে বালতিতে (বিভাগ) একত্রিত করা হয়। এটি অ্যামাজন বা ইবে এর মতো সংস্থার জন্য একটি বাস্তব সমস্যা। অথবা এটি স্বয়ংক্রিয় উচ্চ-ফ্রিকোয়েন্সি স্টক ট্রেডিংয়ের মূল অ্যালগরিদম হিসাবে ব্যবহার করা যেতে পারে।

## সর্বোত্তম মূল্য

প্রথম নজরে দেখে মনে হচ্ছে এটি একটি অর্থনৈতিক সমস্যা যা দক্ষতা বক্রতা বা এমনকি একটি বিশুদ্ধ ব্যবসায়িক সমস্যা দ্বারা পরিচালিত হয়, এটি প্রকৃতির অত্যন্ত পরিসংখ্যানগত। সর্বোত্তম মূল্য উপলব্ধ এবং পূর্বাভাসকৃত ইনভেন্টরি, উৎপাদন খরচ, প্রতিযোগীদের কাছ থেকে মূল্য এবং মুনাফা মার্জিন বিবেচনা করে। দামের স্থিতিস্থাপকতা মডেলগুলি প্রায়ই শক্তিশালী প্রতিরোধে পৌঁছানোর আগে উচ্চ মূল্য কীভাবে বাড়ানো যায় তা নির্ধারণ করতে ব্যবহৃত হয়। আধুনিক সিস্টেমগুলি রিয়েল-টাইমে চাহিদা অনুযায়ী দাম দেয়, উদাহরণস্বরূপ, ফ্লাইট বা হোটেল রুম বুক করার সময়। ব্যবহারকারী-নির্ভর মূল্য-মূল্যকে আরও অপ্টিমাইজ করার একটি উপায়, ব্যবহারকারী অংশের উপর ভিত্তি করে বিভিন্ন মূল্য প্রদান-একটি বিতর্কিত সমস্যা। এটি বীমা শিল্পে গৃহীত হয়: খারাপ গাড়ি চালকরা একই কভারেজের জন্য ভালদের চেয়ে বেশি অর্থ প্রদান করে, অথবা ধূমপায়ী/মহিলা/বয়স্ক ব্যক্তিরা স্বাস্থ্যসেবা বীমার জন্য আলাদা ফি প্রদান করে।

## ইনডেক্সেশন

শ্রেণীবিন্যাসের উপর ভিত্তি করে যে কোনও সিস্টেম শ্রেণীবিন্যাস তৈরি এবং বজায় রাখার জন্য তৈরি একটি সূচীকরণ অ্যালগরিদম ব্যবহার করে। উদাহরণস্বরূপ, প্রোডাক্ট রিভিউ (প্রোডাক্ট এবং রিভিউয়ার উভয়কেই একটি ইনডেকশন অ্যালগরিদম ব্যবহার করে শ্রেণীভুক্ত করতে হবে, তারপর একে অপরের সাথে ম্যাপ করা হবে), একটি নির্দিষ্ট ডোমেইন, ডিজিটাল কন্টেন্ট ম্যানেজমেন্ট এবং অবশ্যই সার্চ ইঞ্জিন টেকনোলজিতে অনুসরণ করার জন্য শীর্ষ ব্যক্তিদের সনাক্ত করতে অ্যালগরিদম স্কোর করা। সূচীকরণ একটি খুব দক্ষ ক্লাস্টারিং অ্যালগরিদম, এবং সময়সীমার ব্যাপক সূচকে ব্যবহৃত সময়গুলি রৈখিকভাবে বৃদ্ধি পায় - এটি খুব দ্রুত - আপনার ডেটাসেটের আকারের সাথে। মূলত, এটি টন ডকুমেন্ট বিশ্লেষণ, বিলিয়ন বিলিয়ন কীওয়ার্ড বের করা, ফিল্টার করা, একটি কীওয়ার্ড ফ্রিকোয়েন্সি টেবিল তৈরি এবং শীর্ষ কীওয়ার্ডগুলিতে ফোকাস করার পরে ম্যানুয়ালি নির্বাচিত কয়েকশো বিভাগের উপর নির্ভর করে।

অবশেষে, একটি সূচীকরণ অ্যালগরিদম স্বয়ংক্রিয়ভাবে যে কোনও নথির জন্য একটি সূচক তৈরি করতে ব্যবহার করা যেতে পারে - প্রতিবেদন, নিবন্ধ, ব্লগ, ওয়েবসাইট, ডেটা সংগ্রহস্থল, মেটাডেটা, ক্যাটালগ বা বই। প্রকৃতপক্ষে, এটি সূচক শব্দটির উৎপত্তি। আশ্চর্যজনকভাবে, প্রকাশকরা এখনও চাকরির সূচকের জন্য মানুষকে অর্থ প্রদান করে: আপনি আমেরিকান সোসাইটি ফর ইনডেক্সিং ওয়েবসাইটে তালিকাভুক্ত এই কাজগুলি খুঁজে পেতে পারেন। ডেটা বিজ্ঞানী উদ্যোক্তাদের জন্য এটি একটি সুযোগ: প্রকাশকদের সফটওয়্যার দেওয়া যা এই কাজটি স্বয়ংক্রিয়ভাবে করে, খরচের একটি অংশে।

## সার্চ ইঞ্জিন

ভাল সার্চ ইঞ্জিন প্রযুক্তি পরিসংখ্যানগত মডেলিংয়ের উপর অনেক বেশি নির্ভর করে। এন্টারপ্রাইজ সার্চ ইঞ্জিন কোম্পানিকে সাহায্য করে - উদাহরণস্বরূপ, অ্যামাজন - ব্যবহারকারীদের তাদের খুঁজে বের করার সহজ উপায় দিয়ে তাদের পণ্য বিক্রি করে। যে কোন সার্চ ইঞ্জিনে ব্যবহৃত মূল অ্যালগরিদম হল একটি ইনডেক্সেশন বা স্বয়ংক্রিয় ট্যাগিং সিস্টেম। গুগল অনুসন্ধান নিম্নরূপ উন্নত করা যেতে পারে:

- পেজ র‍্যাঙ্ক নির্মূল করুন - এই অ্যালগরিদমটি প্রতারকদের দ্বারা বোকা বানানো হয়েছে লিঙ্ক খামার এবং অন্যান্য ওয়েব স্প্যাম,
- সার্চ রেজাল্ট কম স্ট্যাটিক, কম হিমায়িত করতে আপনার ইনডেক্সে ঘন ঘন নতুন কন্টেন্ট যোগ করুন,
- ভাল ব্যবহারকারী/অনুসন্ধান কীওয়ার্ড/ল্যান্ডিং পৃষ্ঠা মিলে অ্যালগরিদম ব্যবহার করে আরও প্রাসঙ্গিক নিবন্ধ দেখান যা শেষ পর্যন্ত আরও ভাল সূচীকরণ সিস্টেম এবং
- নিবন্ধের উৎস দেখানোর জন্য আরো ভালো অ্যাট্রিবিউশন মডেল ব্যবহার করুন, লিঙ্কডইন বা অন্য কোথাও প্রকাশিত কপি নয়। (এটি ছোট প্রকাশকদের উপর বেশি চাপ দেওয়া এবং একটি নিবন্ধের প্রথম ঘটনা চিহ্নিত করার মতো সহজ হতে পারে: টাইমস্ট্যাম্প সনাক্তকরণ এবং ব্যবস্থাপনা)।

## ক্রস সেলিং

ক্রস-সেলিং আপ-সেলিং থেকে আলাদা। সাধারণত, সহযোগী ফিল্টারিং অ্যালগরিদমের উপর ভিত্তি করে, ধারণাটি খুঁজে বের করা হয় - বিশেষত খুচরাতে - সাম্প্রতিক ক্রয় বা আগ্রহের ভিত্তিতে কোন পণ্য ক্লায়েন্টকে বিক্রি করতে হবে। উদাহরণস্বরূপ, পেট্রল কেনা গ্রাহকের কাছে ইঞ্জিন তেল

বিক্রির চেষ্টা করা। ব্যাঙ্কিং -এ, একটি কোম্পানি হয়তো বেশ কিছু পরিষেবা বিক্রি করতে চায়: প্রথমে একটি চেকিং অ্যাকাউন্ট, তারপর একটি সেভিং অ্যাকাউন্ট, তারপর একটি ব্যবসায়িক অ্যাকাউন্ট, তারপর একটি নির্দিষ্ট গ্রাহক বিভাগে loanণ ইত্যাদি। চ্যালেঞ্জ হল সঠিক ক্রম চিহ্নিত করা যাতে কোন পণ্যের প্রচার করা উচিত, সঠিক গ্রাহক বিভাগ এবং বিভিন্ন প্রচারের মধ্যে সর্বোত্তম সময় ব্যবধান।

## ক্লিনিকাল ট্রায়াল

ক্লিনিকাল ট্রায়াল হল ক্লিনিকাল গবেষণায় করা পরীক্ষা, সাধারণত ছোট তথ্য জড়িত। মানুষের অংশগ্রহণকারীদের উপর এই ধরনের সম্ভাব্য বায়োমেডিক্যাল বা আচরণগত গবেষণা অধ্যয়নগুলি নির্দিষ্ট বায়োমেডিক্যাল বা আচরণগত হস্তক্ষেপের উত্তর দেওয়ার জন্য ডিজাইন করা হয়েছে, যার মধ্যে রয়েছে নতুন চিকিত্সা এবং পরিচিত হস্তক্ষেপ যা আরও গবেষণা এবং তুলনার জন্য প্রয়োজনীয়। ক্লিনিকাল ট্রায়ালগুলি নিরাপত্তা এবং কার্যকারিতা সম্পর্কিত তথ্য তৈরি করে। প্রাথমিক উদ্বেগের মধ্যে রয়েছে কিভাবে রোগীদের নমুনা দেওয়া হয় (প্রধানত যদি তাদের ক্ষতিপূরণ দেওয়া হয়), এই গবেষণায় স্বার্থের দ্বন্দ্ব এবং পুনরুত্পাদনযোগ্যতার অভাব।


*চিত্র 2 - পরিসংখ্যানের ক্ষেত্র জীবনের সকল ক্ষেত্রকে প্রভাবিত করে*

## বহুবিধ পরীক্ষা

মাল্টিভেরিয়েট টেস্টিং একটি হাইপোথিসিস পরীক্ষা করার একটি কৌশল যেখানে একাধিক ভেরিয়েবল পরিবর্তন করা হয়। লক্ষ্য হল সম্ভাব্য সংমিশ্রণের মধ্যে কোন বৈচিত্রের সংমিশ্রণ সবচেয়ে ভালো করে তা নির্ধারণ করা। ওয়েবসাইট এবং মোবাইল অ্যাপগুলি পরিবর্তনশীল উপাদানগুলির সংমিশ্রণে গঠিত যা বহুবিধ পরীক্ষার মাধ্যমে অপ্টিমাইজ করা হয়। এর মধ্যে রয়েছে পরীক্ষার যত্নশীল নকশা, এবং একটি ওয়েবপেজের দুটি সংস্করণের মধ্যে ক্ষুদ্র, অস্থায়ী পার্থক্য (ফলন বা ওয়েব ট্রাফিকের মধ্যে) পরিসংখ্যানগত তাত্পর্য নাও থাকতে পারে। যখন ANOVA[8]এবং হাইপোথিসিসের পরীক্ষাগুলি বহুবিধ পরীক্ষার জন্য শিল্প বা স্বাস্থ্যসেবা পরিসংখ্যানবিদরা ব্যবহার করেন, আমরা ডেটা বিনিং এবং মডেল-মুক্ত আস্থা ব্যবধানের উপর ভিত্তি করে মডেল-মুক্ত, ডেটা-চালিত সিস্টেম তৈরি করেছি। একটি বহুমুখী পরীক্ষার পরীক্ষা বন্ধ করা (তারা সাধারণত ওয়েব পেজ অপ্টিমাইজেশনের জন্য 14 দিন স্থায়ী হয়) যত তাড়াতাড়ি বিজয়ী সংমিশ্রণটি চিহ্নিত করা হয় তা প্রচুর অর্থ সাশ্রয় করতে সহায়তা করে। মনে রাখবেন যে বহিরাগত ঘটনা - উদাহরণস্বরূপ, ছুটির দিন বা কিছু সার্ভার বিভ্রাট - বহুবিধ পরীক্ষার ফলাফলকে প্রভাবিত করতে পারে এবং এর সমাধান করা প্রয়োজন।

## সারিবদ্ধ ব্যবস্থা

সারি নিয়ন্ত্রণের জন্য একটি কিউ ম্যানেজমেন্ট সিস্টেম ব্যবহার করা হয়। একটি সারি এলাকায় বিভিন্ন পরিস্থিতিতে এবং অবস্থানে মানুষের সারি তৈরি হয়, উদাহরণস্বরূপ, একটি কল সেন্টারে। সারি গঠন ও বংশ বিস্তারের প্রক্রিয়াকে সারিবদ্ধ তত্ত্ব হিসেবে সংজ্ঞায়িত করা হয়। একটি সারিতে মানুষের আগমন সাধারণত একটি পয়েসন পদ্ধতি ব্যবহার করে একটি ক্লায়েন্টকে একটি সূচকীয় ডিস্ট্রিবিউশন ব্যবহার করে মডেল করা হয়। পরিসংখ্যানগত সমস্যা হওয়া সত্ত্বেও, এটি অপারেশন গবেষণার অংশ বলে মনে করা হয়।

## সাপ্লাই চেইন অপটিমাইজেশন

সাপ্লাই চেইন অপটিমাইজেশন একটি উৎপাদন ও বিতরণ সাপ্লাই চেইনের সর্বোত্তম ক্রিয়াকলাপ নিশ্চিত করার জন্য প্রক্রিয়া এবং সরঞ্জাম প্রয়োগ করে। এর মধ্যে রয়েছে সাপ্লাই চেইনের মধ্যে অনুকূল ইনভেন্টরি প্লেসমেন্ট, অপারেটিং খরচ কমানো (উৎপাদন খরচ, পরিবহন খরচ এবং বিতরণ খরচ সহ)। এটি প্রায়শই গাণিতিক মডেলিং কৌশল যেমন গ্রাফ তত্ত্ব প্রয়োগ করে অনুকূল ডেলিভারি রুট (এবং গুদামের সর্বোত্তম অবস্থান), সিমপ্লেক্স অ্যালগরিদম এবং মন্টে কার্লো সিমুলেশন ব্যবহার করে।

# *About The Author*

Enamul Haque is an author, researcher, and managing consultant best known for working with global companies such as Microsoft, Capgemini, Nokia, HCL Technologies, and the United Nations High Commissioner for Refugees (UNHCR) and International Telecommunication Union (ITU). He has over 26 years of rich experience in IT transformation and leading people for their professional growth and increase contribution to the organisation. Out of which, he treasured 13 years of experience in remote working and leading virtual teams.

As a consultant, Enamul worked with many of the world's best-known companies on their digital transformation and service integration

strategies for improving business performance and value creation, including Alstom, Bayer AG, Bombardier, Britvic, Cadent, Carphone, Chanel, Direct Line Group, Estee Lauder Companies, Heathrow Airport, Neste, Rockwell Automation, Rogers, Sandvik, Shell, SJ Johnson, Terex, True-Value, Unilever, Warner Brothers, among many others. He assists in re-skilling technical workforces to stay modern and ensure business continuity and compliance.

Enamul shares his industry knowledge among the MBA students as a guest lecturer at the University of Coventry, London campus. He worked very extensively as contributing writer for various newspapers, magazines, and other publications. Enamul is multilingual and lived and worked in many countries, including the USA, Switzerland, Finland, UAE, UK, India, and Germany.

Enamul Haque studied mathematics and analytics (*Cours de mathématiques spéciales*) at the Swiss Federal Institute of Technology (EPFL), Lausanne, and architecture and Technology of computer science (license en science Informatique) at the University of Geneva. He also has a diploma in Artificial Intelligence and Machine Learning from the University of Helsinki. He is currently pursuing a Harvard and Capgemini co-branded program on foundational behaviours of managerial success (proximity, performance, and perspective). The program is based on three key areas, such as understanding the importance of Managerial behaviours and the impact they have on teams (including virtual teams), the ability to demonstrate and apply new managerial practices in a changing environment and to be equipped to enable a cultural shift towards a more substantial employee experience and engagement.

AUTHOR OFFICIAL WEBSITE: https://www.enamulhaque.co.uk/
ALL BOOKS BY ENAMUL HAQUE: https://enamulhaque.co.uk/my-books
ENAMUL HAQUE BLOG: https://enamulhaque.co.uk/my-articles
GOODREADS AUTHOR PROFILE: https://www.goodreads.com/haquenam
AMAZON AUTHOR PROFILE: https://www.amazon.com/ author/enamulhaque
TWITTER HANDLE @HAQUENAM: https://twitter.com/haquenam
LINKEDIN PROFILE: https://www.linkedin.com/in/haquenam
YOUTUBE TUTORIAL: https://www.youtube.com/c/digitaldeepdive

FACEBOOK AUTHOR PAGE: https://www.facebook.com/authorenam

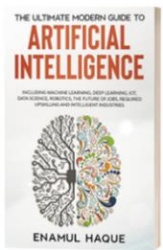
# *Other Books By the Author*

**THE ULTIMATE MODERN GUIDE TO CLOUD COMPUTING**
ISBN- 979-8666050637


*This book has the most simplified explanation of Cloud Computing, starting from understanding digital transformation, enabling technologies to define essential characteristics, service models, deployment models, etc., with a pragmatic approach. It provides the path to digital transformation through the adoption of Cloud Computing to help construct Intelligent Enterprises.*

**THE ULTIMATE MODERN GUIDE TO ARTIFICIAL INTELLIGENCE**
ISBN: 979-8691930768

*This book has the most simplified explanation of Cloud Computing, starting from understanding digital transformation, enabling technologies to define essential characteristics, service models, deployment models, etc., with a pragmatic approach. It provides the path to digital transformation through the adoption of Cloud Computing to help construct Intelligent Enterprises.*




**THE ULTIMATE MODERN GUIDE TO THE INTERNET OF THINGS (IOT)**
ISBN- 979-8691930768


*The Internet of Things explained: Simply and Non-Technically. IoT is a computing paradigm in which several technologies that connect various devices based on wireless Internet acquire environmental information through sensors and control. This book provides a rigorous understanding of the IoT framework, characteristics, architecture, applications, technologies etc. in plain English to improve your awareness A key objective of this book is to provide a systematic source of reference for all aspects of IoT.*

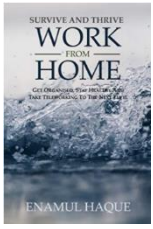
**THE ULTIMATE MODERN GUIDE TO DIGITAL TRANSFORMATION**

ISBN- 979-8702899572


*In this book, you'll learn how new technologies disrupt businesses and how to transform to survive with the convergence of cloud computing, big data, artificial intelligence, the internet of things, and many other emerging technologies and how they are changing how we operate the 21st century. This book will give you the digital practices needed to catapults your organisation into next-level success.*

**SURVIVE AND THRIVE WORK FROM HOME**

ISBN- 979-8580562872




*The impact of Pandemic and the new shifting trends for work-live balance. How we get there and it a success both for employees and employers. The fundamentals of remote working. Understanding the norms, teleworking history, benefits, challenges, and a very high-level overview of technology and culture's essential aspects. A collection of the best practices to do your work from home work for you. This includes the very best tips and tricks for working remotely by personality, job types etc. This has a selection of tops tools for remote use.*

**CLOUD SERVICE MANAGEMENT AND GOVERNANCE**

ISBN- 978-1716788352


*Once an organisation adopts cloud computing, it quickly becomes apparent that the traditional IT Service Management processes' traditional approaches will need to undergo drastic changes to integrate and run Bi-Modal IT Service Operations. This book is an alleyway to manage enterprise could services with a framework that consists of progressive Service Management practices to ensure practical, strategic, and modular methodology for the positive transformation of ITSM for cloud delivery models is followed.*

# *Notes and References*

[1] **Dan Radak** - Data Science Security Hacks - https://data-science-blog.com/blog/2020/06/04/data-science-security-hacks/

[2] **kirk86** - Statistical modeling summarization - https://kirk86.github.io/2017/11/stats-modeling/

[3] **Smriti Srivastava** - The 10 general applications of statistical models in data analytics - https://www.analyticsinsight.net/the-10-general-applications-of-statistical-models-in-data-analytics/

[4] **ANOVA** - Analysis of variance is a collection of statistical models and their associated estimation procedures used to analyse the differences among means. ANOVA was developed by the statistician Ronald Fisher.

[5] ড্যান রাদাক-ডেটা সায়েন্স সিকিউরিটি হ্যাকস-https://data-science-blog.com/blog/2020/06/04/data-science-security-hacks/

[6] kirk86 - পরিসংখ্যানগত মডেলিং সংক্ষিপ্তকরণ - https://kirk86.github.io/2017/11/stats-modeling/

[7] স্মৃতি শ্রীবাস্তব-তথ্য বিশ্লেষণে পরিসংখ্যানগত মডেলগুলির 10 টি সাধারণ অ্যাপ্লিকেশন-https://www.analyticsinsight.net/the-10-general-applications-of-statistical-models-in-data-analytics/

[8]ANOVA - বৈকল্পিক বিশ্লেষণ হল পরিসংখ্যানগত মডেলগুলির একটি সংগ্রহ এবং তাদের সাথে সম্পর্কিত অনুমান পদ্ধতি যা মাধ্যমের মধ্যে পার্থক্য বিশ্লেষণ করতে ব্যবহৃত হয়। ANOVA তৈরি করেছিলেন পরিসংখ্যানবিদ রোনাল্ড ফিশার।