

Data &
Data Science? $x = (s_l, s_w, p_l, p_w)$ x_1

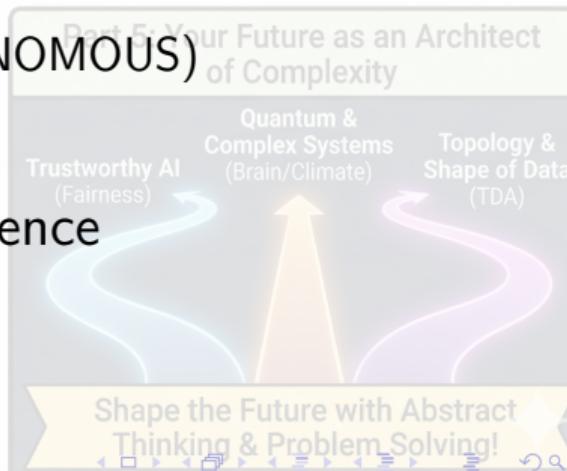
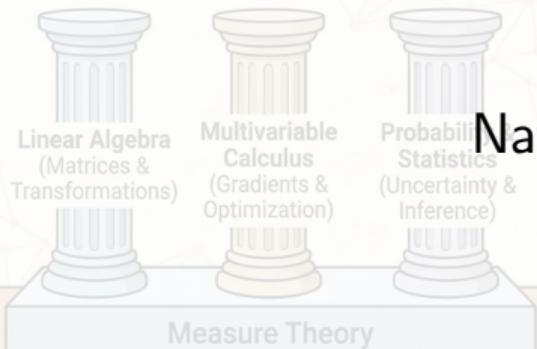
Applied Mathematics in the Age of AI

Bridging Theory and Computation

(Optimization, L_B)

Prof. Siju K. S

Saintgits College of Engineering (AUTONOMOUS)
Kottayam.



Session Overview

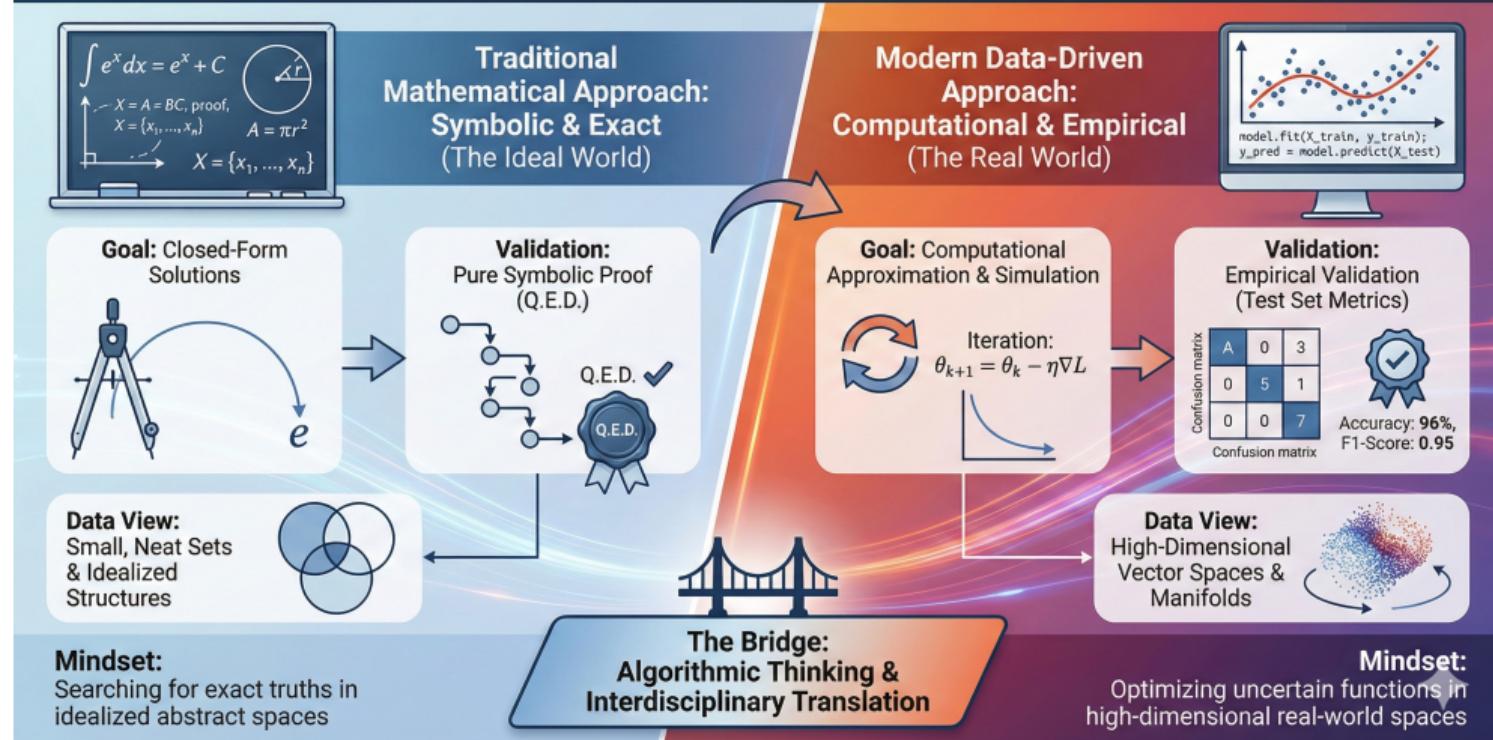
1 Data Foundations & EDA

2 Mathematics of Machine Learning

3 Future Outlook

The Modern Mathematician

Perspective Shift: The Modern Mathematician in the Data Age



Mathematical Formalism of Data

Set-Theoretic Definition: Data can be considered as elements of a set:

$$X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d$$

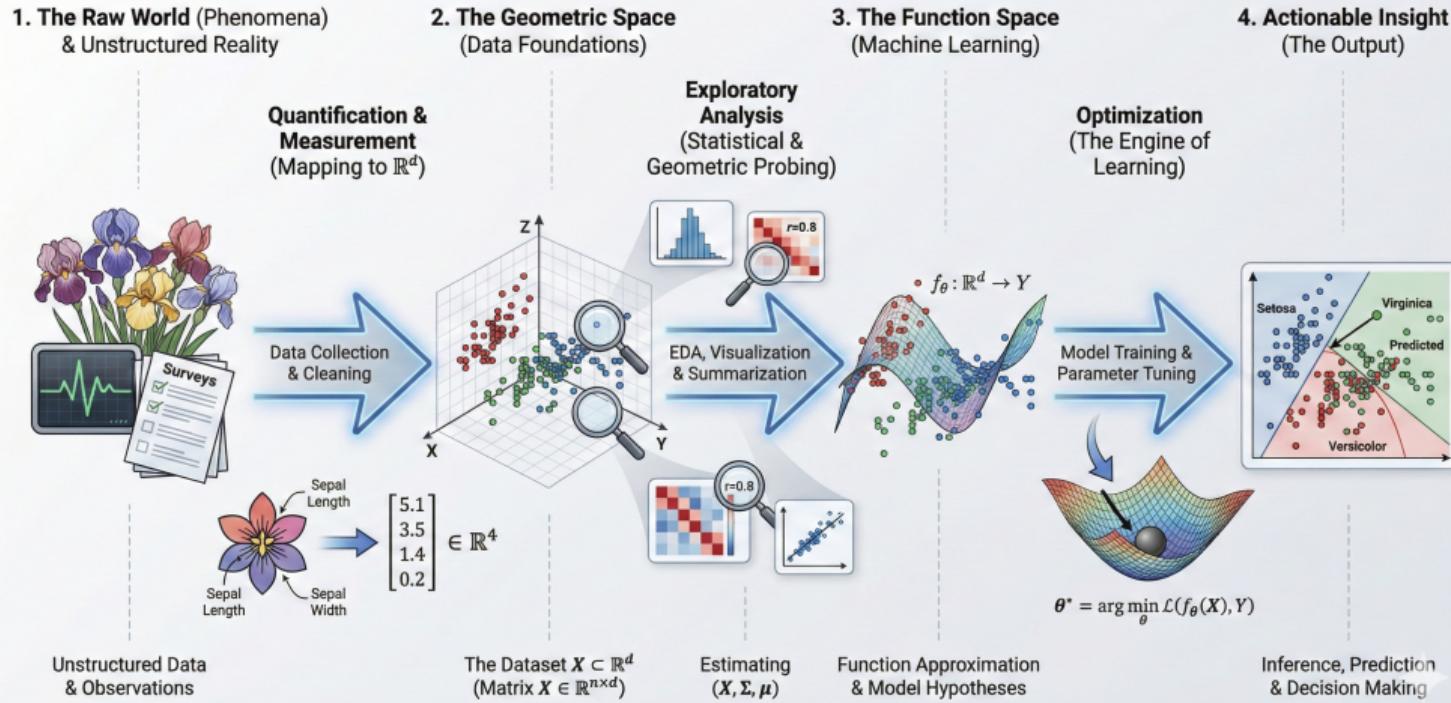
where each x_i is a d -dimensional vector representing measurable features.

Measure Theoretic View: We view the data space as a triple (X, Σ, μ) , where:

- X : The sample space.
- Σ : The σ -algebra of measurable subsets.
- μ : The measure assigning probability/size.

The Data Science Pipeline

The Data Science Pipeline: A Vector Space Perspective



Exploratory Data Analysis (EDA)

Module 1: Data Foundations and Exploratory Data Analysis – A Mathematical Perspective

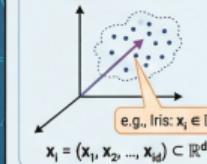
1. MATHEMATICAL FORMALISM & FOUNDATIONS



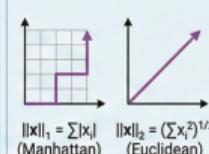
Rigorous Abstraction

Data as a Set in Vector Space

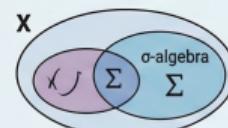
$$X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d$$



Norm-based Measurement



Measure Theoretic View



Measure μ
 (Size/Probability)



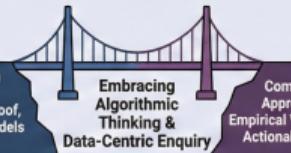
Role of Math & Stats:

- Optimization, Linear Algebra, Probability, Information Theory

2. THE PERSPECTIVE SHIFT & EDA

FROM:

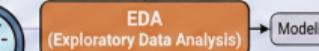
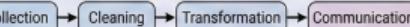
Closed-form Solutions,
 Symbolic Proof,
 Abstract Models



TO:

Computational Approximation,
 Empirical Validation,
 Actionable Insight

Key Steps in Data Analysis



Discover patterns, test assumptions,
 reveal hidden story.

Statistical Tools



Mean, Variance,
 Correlation, Normality
 Tests

Visualization Tools



Histogram Scatter Plot Heatmap
 Heatmap Pairplot Boxplot

EDA Purpose: Understand structure, identify anomalies,
 guide modelling, prevent misleading conclusions.

3. CASE STUDY: THE IRIS DATASET



Historical Context: Sir R.A. Fisher (1936)
 - "The Use of Multiple Measurements in Taxonomic Problems"

Dataset Structure

| Sepal L | Sepal W | Petal L | Petal W | Species |
|---------|---------|---------|---------|---------|
| ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... |

$X \subset \mathbb{R}^4$, 150 samples, 3 classes



EDA in Action

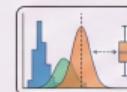
Pairplot
 Visualizing Relationships & Class Separation



Heatmap
 Identifying Feature Correlations



Histogram & Boxplot
 Analyzing Feature Distributions & Outliers



Significance: A benchmark linking statistical theory to modern machine learning classification.

Metrics and Norms in Data Analysis

To analyze similarity (clustering) or error (regression), we require a metric space structure.

Norm-based Measurements:

The L_1 Norm (Manhattan - useful for sparse data):

$$\|x\|_1 = \sum_{i=1}^d |x_i|$$

The L_2 Norm (Euclidean - standard geometric distance):

$$\|x\|_2 = \left(\sum_{i=1}^d x_i^2 \right)^{1/2}$$

Machine Learning: An Optimization Problem

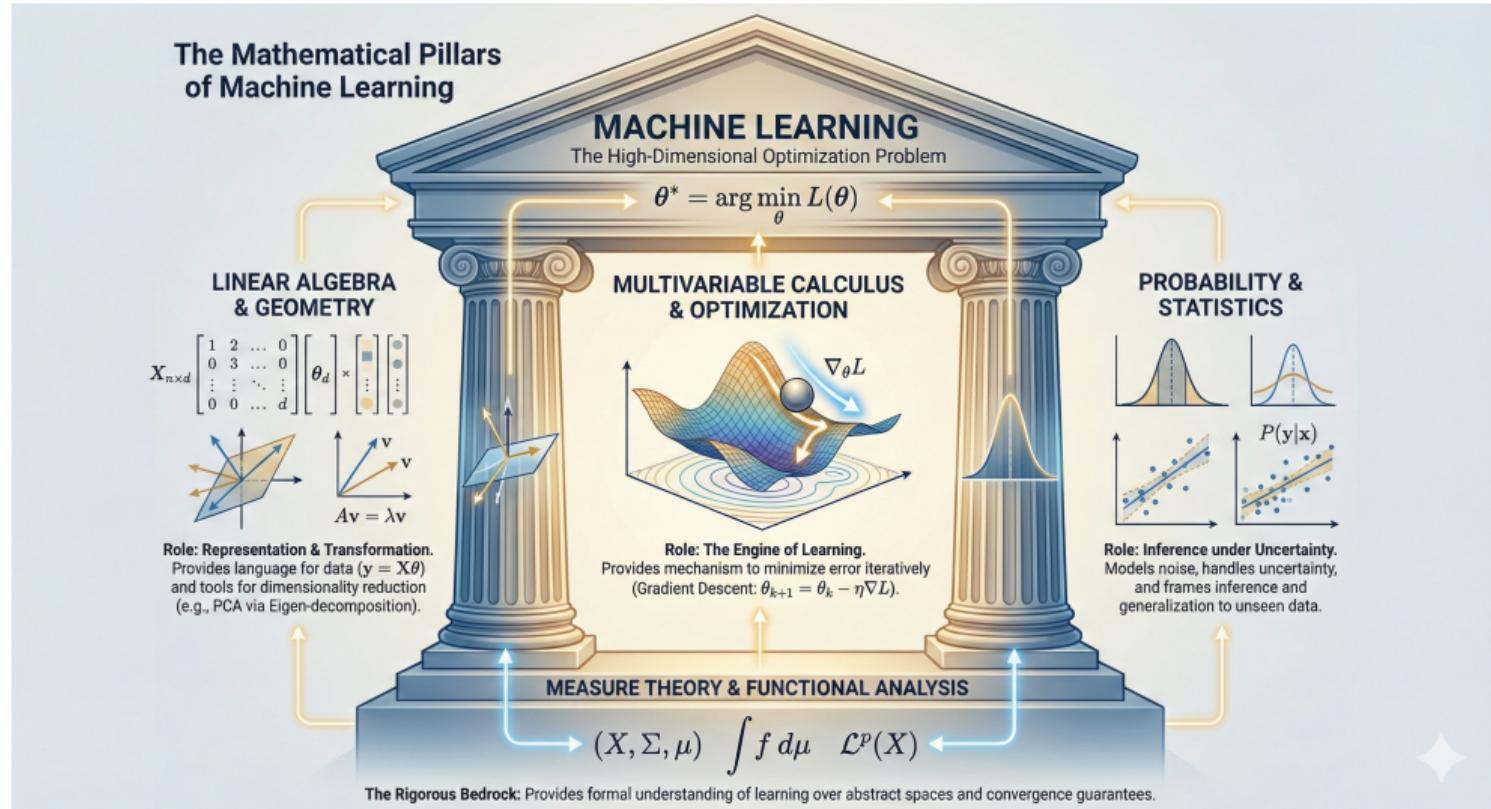
Given inputs X and outputs Y , we seek a function $f_\theta : \mathbb{R}^d \rightarrow Y$.
The goal is to find parameters θ^* that minimize a Loss Function L :

$$\theta^* = \arg \min_{\theta} L(f_\theta(X), Y)$$

Key Insight

ML is fundamentally an optimization problem in a high-dimensional vector space, constrained by computational resources and statistical uncertainty.

The Mathematical Pillars of Machine Learning



Machine Learning Core

Module 2: Machine Learning Core – A Mathematical & Computational Perspective

1. THE MATHEMATICAL GOAL: FUNCTION APPROXIMATION & OPTIMIZATION

Data Space: $X \subset \mathbb{R}^d$ (From observed data (X, Y) , learn a function...) → Function Space & Loss Minimization: $f_\theta: \mathbb{R}^d \rightarrow Y$ (Objective: $\theta^* = \arg \min_{\theta} L(f_\theta(X), Y)$) (Minimize the Loss Function L (discrepancy between prediction and truth))

2. THE MATHEMATICAL FOUNDATION (THE PILLARS)

MACHINE LEARNING

- MULTIVARIABLE CALCULUS: Optimization Engine: Gradients guide updates
- LINEAR ALGEBRA: $y = X\theta$, $Av = \lambda v$ (Structure & Transformation: Vectorization, PCA)
- PROBABILITY & STATISTICS: Inference & Uncertainty: Modelling noise & distributions
- OPTIMIZATION THEORY & MEASURE THEORY: Rigor & Convergence: Guarantees & formal understanding ((X, Σ, μ))

ML problems are fundamentally optimization problems in high-dimensional spaces.

3. GRADIENT DESCENT: THE CORE ENGINE

Iterative update of loss landscape: $\theta_k \rightarrow \theta_k - \eta \nabla_{\theta} L(\theta_k)$

LOSS FUNCTIONS & REGULARIZATION

- MSE: $L(\theta) = \frac{1}{n} \sum_i (y_i - f_\theta(x_i))^2$ Measures average squared difference
- L2 Regularization ($+ \lambda \|\theta\|_2^2$): Penalties complex models to prevent overfitting
- L1 Regularization ($+ \lambda \|\theta\|_1$): Penalties complex models to prevent overfitting

4. LINEAR ALGEBRA: THE BACKBONE OF COMPUTATION

VECTORIZED OPERATIONS: $\begin{matrix} d \\ n \end{matrix} \times \begin{matrix} d \\ n \end{matrix} = \begin{matrix} n \\ n \end{matrix}$ ($y = X\theta$. Efficient batch processing for n samples & d features.)

DIMENSIONALITY REDUCTION: Eigenvalue Decomposition ($Av = \lambda v$). Basis for PCA and feature extraction.

5. THE ML WORKFLOW: FROM DATA TO MODEL (IRIS EXAMPLE)

DATA (Iris): Sepal Length, Width, Petal Length, Width → PREPARATION: Split (Train/Test) → Train Set, Test Set → MODEL (Logistic Regression): $\text{model.fit}(X_{\text{train}}, y_{\text{train}})$ → PERFORMANCE: Tune Hyperparameters, Evaluation, Assess model accuracy and errors. (accuracy_scores, confusion_matrix)

The Engine: Gradient Descent

To solve $\min_{\theta} L(\theta)$, we use iterative updates based on Multivariable Calculus.

Update Rule:

$$\theta_{k+1} = \theta_k - \eta \nabla_{\theta} L(\theta_k)$$

Where:

- η : Learning rate (step size).
- $\nabla_{\theta} L$: The gradient vector (direction of steepest ascent).

Regularization (L_2 - Ridge): Prevents overfitting by penalizing large norms:

$$L(\theta) = \frac{1}{n} \sum (y_i - f_{\theta}(x_i))^2 + \lambda \|\theta\|_2^2$$

Linear Algebra: The Computational Backbone

Vectorization:

$$y = X\theta$$

$$X \in \mathbb{R}^{n \times d}, \quad \theta \in \mathbb{R}^d$$

Dimensionality Reduction:

Eigenvalue Decomposition helps us find principal components:

$$Av = \lambda v$$

The Future: Mathematics in 2030

The Mathematician of 2030: Architect of Intelligence & Complexity

A Infographic concept for an expanding intellectual landscape. Building upon, the foundations shown on previous infographics.

FRONTIER 1: TRUSTWORTHY AI & MATHEMATICAL GUARANTEES



Scope: Developing math for Explainability (XAI), Causality, Fairness, & Robustness.

Role: The "Safety Engineer" of Intelligent Systems.

Key Math: High-Dim Probability, Information Geometry, Logic.

THE NEW METHODOLOGY: AI-AUGMENTED DISCOVERY

From Solver to Orchestrator

AI as Conjecture Generator & Proof assistant (e.g., Lean/Coq)



THE ENDURING CORE:
RIGOROUS ABSTRACTION & LOGIC

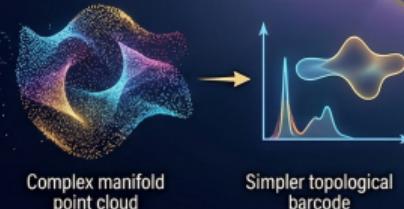
FRONTIER 2: QUANTUM & COMPLEX SYSTEMS MODELLING



Scope: designing Quantum Algorithms, Modelling Climate, Brain, & Social Networks.

Role: The "Systems Architect" of Reality.

Key Math: Linear Algebra over Complex Fields, Dynamical Systems, Network Theory.



3: THE GEOMETRY OF DATA & INFORMATION

Scope: Uncovering shape & structure in massive datasets beyond simple statistics.

Role: The "Cartographer" of the Information Age.

Key Math: Topological Data Analysis (TDA), Differential Geometry, Algebraic Topology.

2030 PROFESSIONAL HORIZON



The future mathematician is not just a calculator, but the essential bridge between abstract structure and real-world complexity.

Strategic Guidance for Indian Mathematics Academics

Strategic Guidance for Indian Mathematics Academics in the Data Age (Postgrads, Researchers, Faculty)

1. THE STRONG FOUNDATION (Embrace & Leverage)



Rigorous Theory:
Algebra,
Analysis,
Logic



Your Asset:
Deep abstract thinking &
problem-solving rigor.
Don't abandon it; build upon it.

2. BRIDGING THE GAP (Actionable Guidance & Skill Acquisition)



Computational Fluency



Learn programming
as a tool for
thought, not just
a skill.

Implement
algorithms, don't
just prove them.



Data-Centric Theory



Connect pure
math to ML
foundations:
Optimization,
High-Dim
Probability,
Matrix
Approximations.



Interdisciplinary Collaboration



Bio Eng Econ
Break silos.
Collaborate with
CS, Biology,
Economics, etc.,
on real-world
data problems.

3. THE EMPOWERED ACADEMIC (Impact & Opportunities in India)



Solving National
Challenges
(e.g., Healthcare,
Agriculture)

Modern Pedagogy



Curriculum Revamp:
Computation & Data
Mentoring Future
Data Scientists



Industry & Policy Linkages



Consulting & Collaboration
with Tech Hubs
Influencing Data
Policy & Ethics

From Abstract Masters to Real-World Architects: Shaping India's Intelligent Future through Mathematical Leadership.

Summary & Discussion

To contribute effectively, we must:

- ① Move from closed-form solutions to computational approximations.
- ② Accept empirical validation alongside symbolic proof.
- ③ Embrace interdisciplinary collaboration.

Thank You

Clear Vision

Introduction: The Curious Question

Data & Data Science?

Let's find out!

Part 1: Building Blocks of Data

$x = (s_i, s_w, p_i, p_w)$

$X = \begin{Bmatrix} x_1 \\ x_2 \\ \vdots \\ x_1, \dots, x_n \end{Bmatrix}$

Vector Space

L_1 L_2

Part 2: The Data Science Journey

Raw Data & Cleaning

Geometric Space (EDA)

Actionable Insight

Function Space (Optimization, f_θ)

Part 3: Mathematical Pillars

Linear Algebra (Matrices & Transformations)

Multivariable Calculus (Gradients & Optimization)

Probability & Statistics (Uncertainty & Inference)

Measure Theory

Part 4: The Modern Mathematician Shift

Traditional

$\hat{x} = \sum \frac{x_i}{n}$

$l \neq \sqrt{e} \Leftarrow$

$\sum x_i = \frac{1}{n} \approx \sum$

Modern

Computational & Empirical Bridge Builder

Part 5: Your Future as an Architect of Complexity

Quantum & Complex Systems (Brain/Climate)

Topology & Shape of Data (TDA)

Trustworthy AI (Fairness)

Shape the Future with Abstract Thinking & Problem Solving!