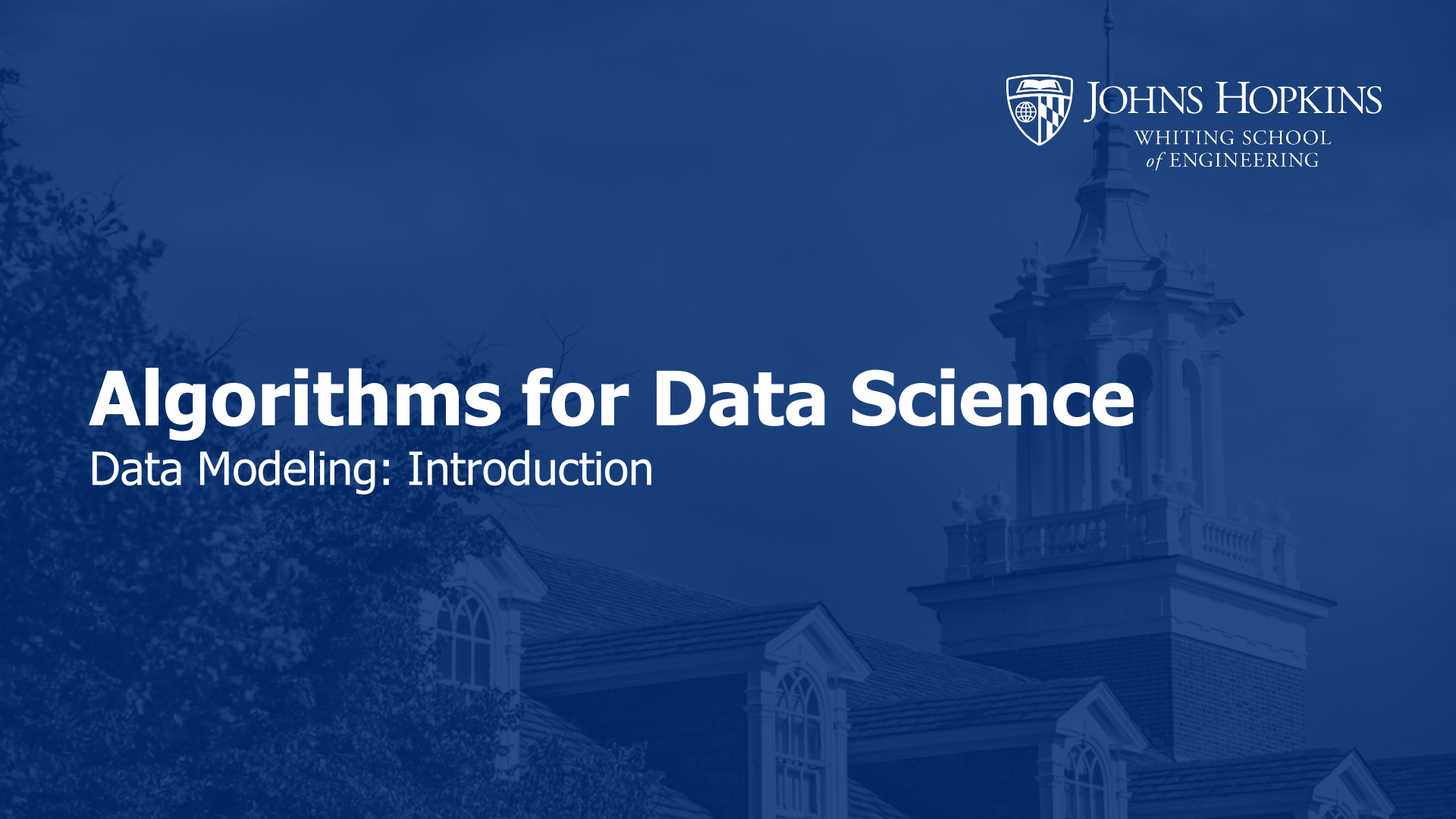


Algorithms for Data Science

Data Modeling: Introduction

Module Learning Objectives

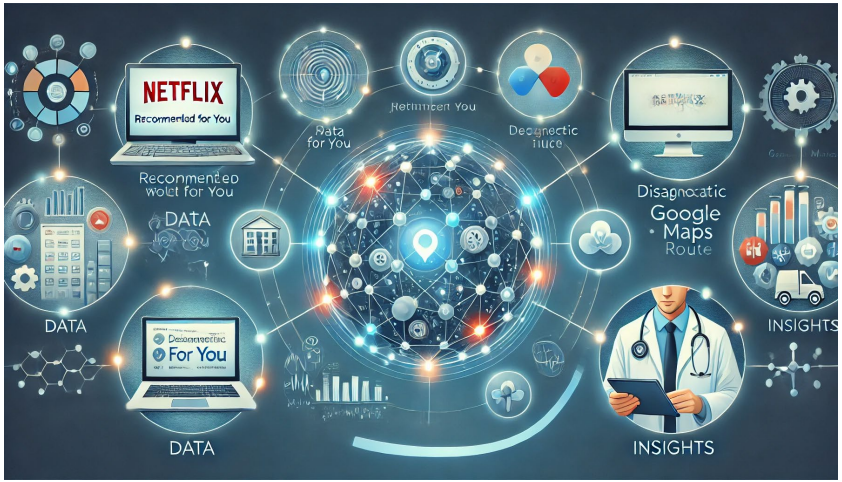
1. Discuss how patterns, tendencies, and correlations are uncovered during the training phase.

2. Discuss how algorithms supply the steps necessary to carry out operations or computations.

3. Discuss how algorithms are used to solve problems in a variety of ways, from simple to complex.

What Powers the Decisions Around Us?

- How does Netflix recommend the perfect show for you?
- What helps doctors diagnose diseases accurately?
- Why does Google Maps always find the fastest route?

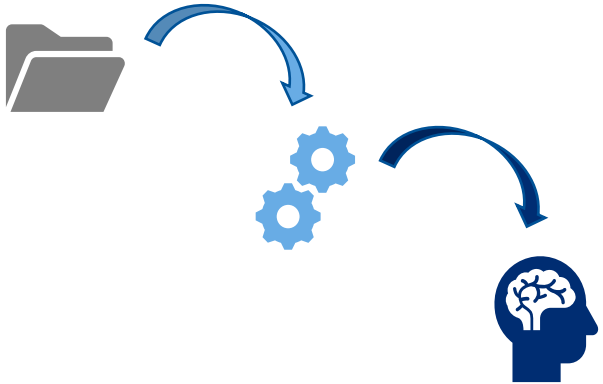


Data modeling is at the heart of decision-making systems that drive the modern world.

What is Data Modeling?

Data modeling is the process of creating a mathematical representation of real-world processes or systems to generate insights or make predictions.

- Simplifies complex systems by capturing essential patterns.
- Turns raw data into actionable insights.
- Forms the foundation for AI, ML, and DS solutions.



"All models are wrong, but some are useful" – George Box

What is Data Modeling?

Data modeling is the process of creating a mathematical representation of real-world processes or systems to generate insights or make predictions.

- A model can be viewed as a function *f* that maps input data X to outputs *y*:

$$y = f(X; \theta)$$

- Where θ represents the parameters of the model that are learned from the data.

What is Data Modeling?

Data modeling is the process of creating a mathematical representation of real-world processes or systems to generate insights or make predictions.

- Models are often represented as optimization problems, where the goal is to minimize a loss function L over the parameters θ

$$\theta^* = \arg\min L(\theta; X, y)$$

- Where θ represents the parameters of the model that are learned from the data.

What is Data Modeling?

Data modeling is the process of creating a mathematical representation of real-world processes or systems to generate insights or make predictions.

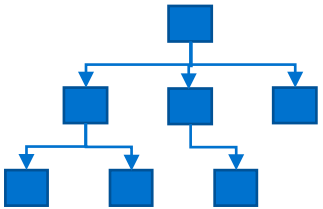
Linear Regression

- Used to predict a continuous outcome variable based on one or more predictors.

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

Decision Trees

- Non-parametric models that partition data into subsets based on values of input features.



Neural Networks

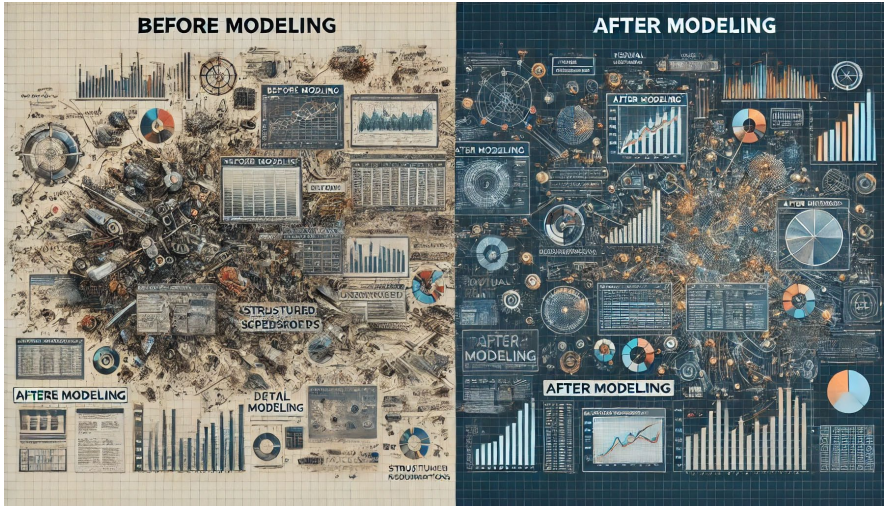
- Complex models inspired by human brains capable of capturing non-linear relationships.

$$a^{(l)} = \sigma\left(W^{(l)} a^{(l-1)} + b^{(l)}\right)$$

Why Data Modeling?

Without data models, raw data is like a library with unorganized books, valuable but inaccessible.

- Transforms raw data into meaningful insights.
- Drives decision-making in critical systems.
- Reduces complexity, makes sense of large datasets.
- Enables predictive and prescriptive analytics for proactive solutions.



Algorithms vs. Models: What's the Difference?

Algorithms

- A step-by-step **procedure for calculations**.
- E.g. Linear regression, decision trees, gradient descent.

Recipe

Models

- A **trained artifact derived** from applying an algorithm to data.
- E.g. Regression line predicting housing prices.

Dish created using the recipe

Algorithms vs. Models

Mathematically, an algorithm can be considered as a function that maps input data D to a model M:

$$\text{Algorithm: } D \rightarrow M$$

Mathematically, a model is a function f parametrized by θ that maps input features X to output predictions y:

$$f_\theta : X \rightarrow y$$