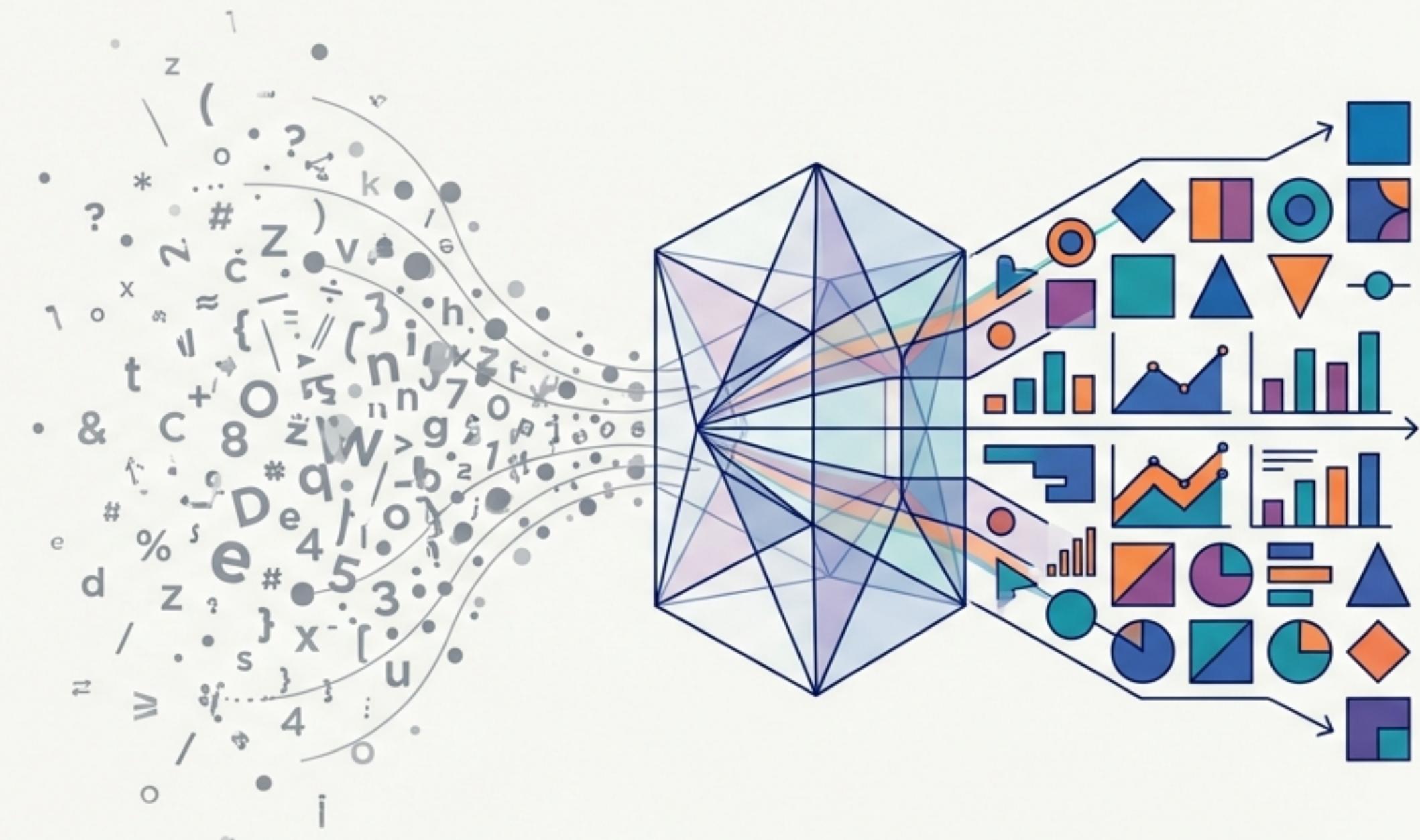


Turning Raw Data into Actionable Wisdom

An Introduction to the Data Analytics Value Chain



Introduction to Data Analytics • Unit I

The Science of Examining Data

Data Analytics is the process of collecting, organizing, and studying raw data to uncover trends, patterns, and insights. It differs from Data Analysis (which focuses solely on finding meaning) by also including the creation of tools, systems, and predictions.



Decision Making

Moving from gut feeling to clear facts.



Problem Solving

Identifying root causes of inefficiencies.



Opportunity ID

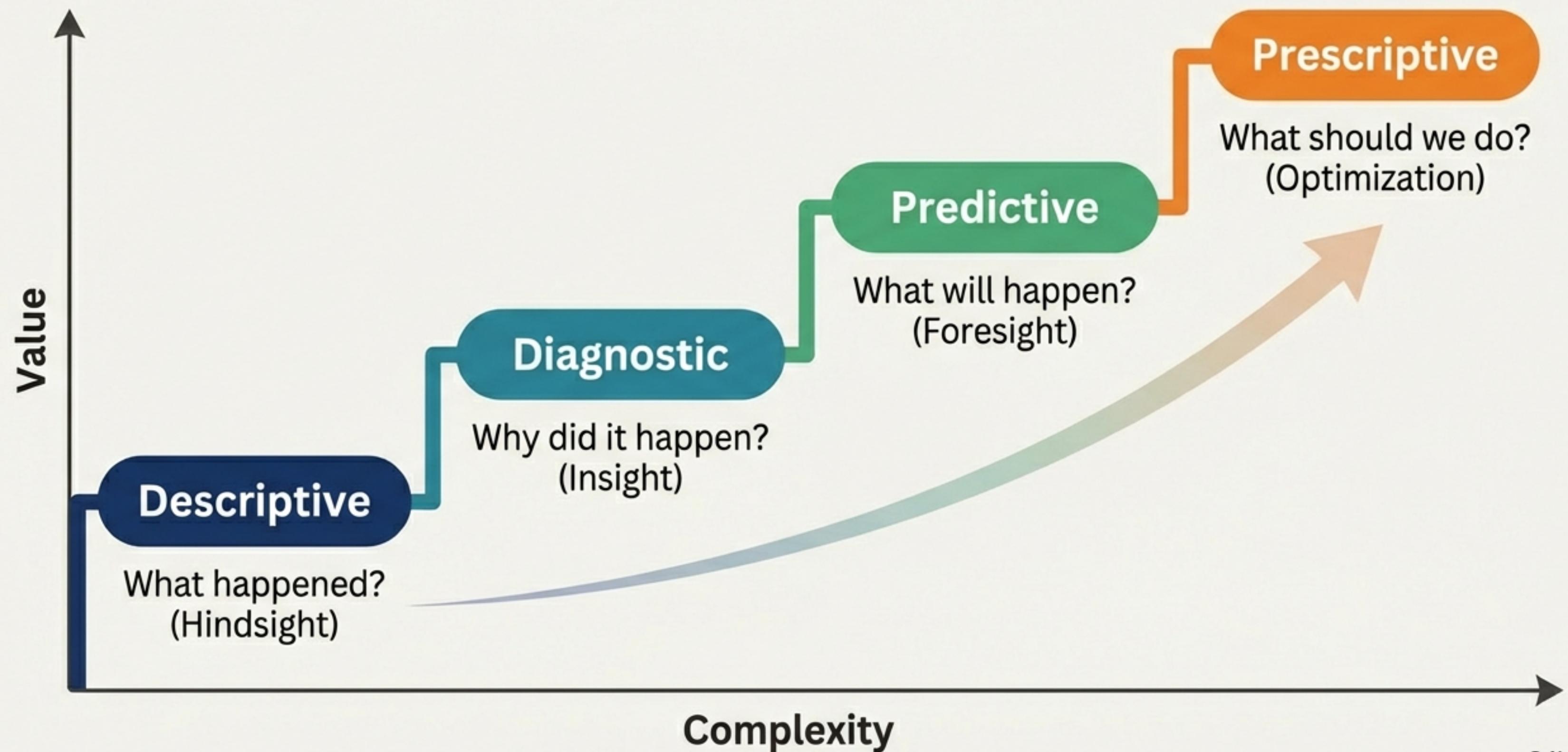
Spotting non-obvious growth trends.



Improved Efficiency

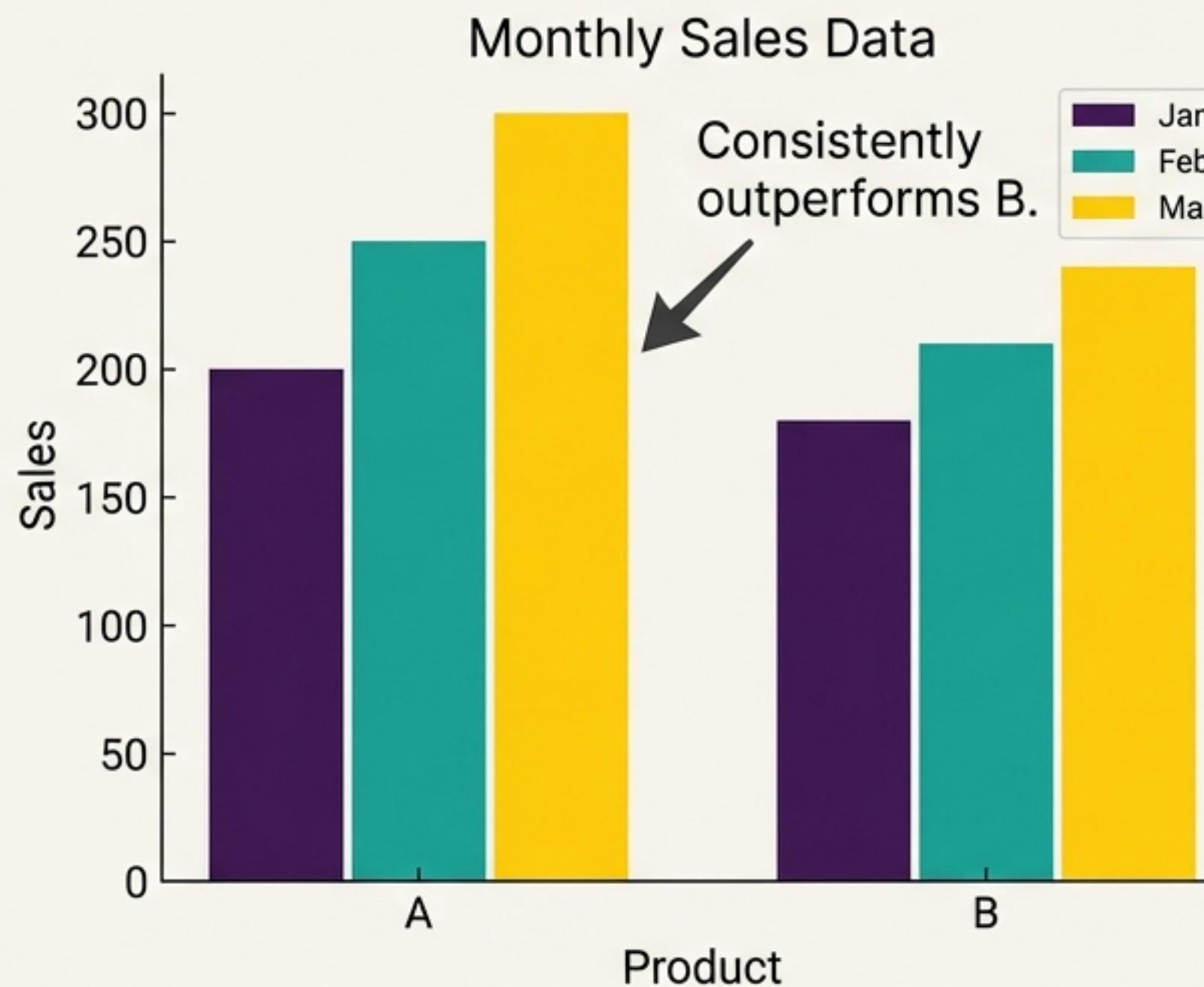
Reducing waste and optimizing workflows.

The Spectrum of Analytical Insights

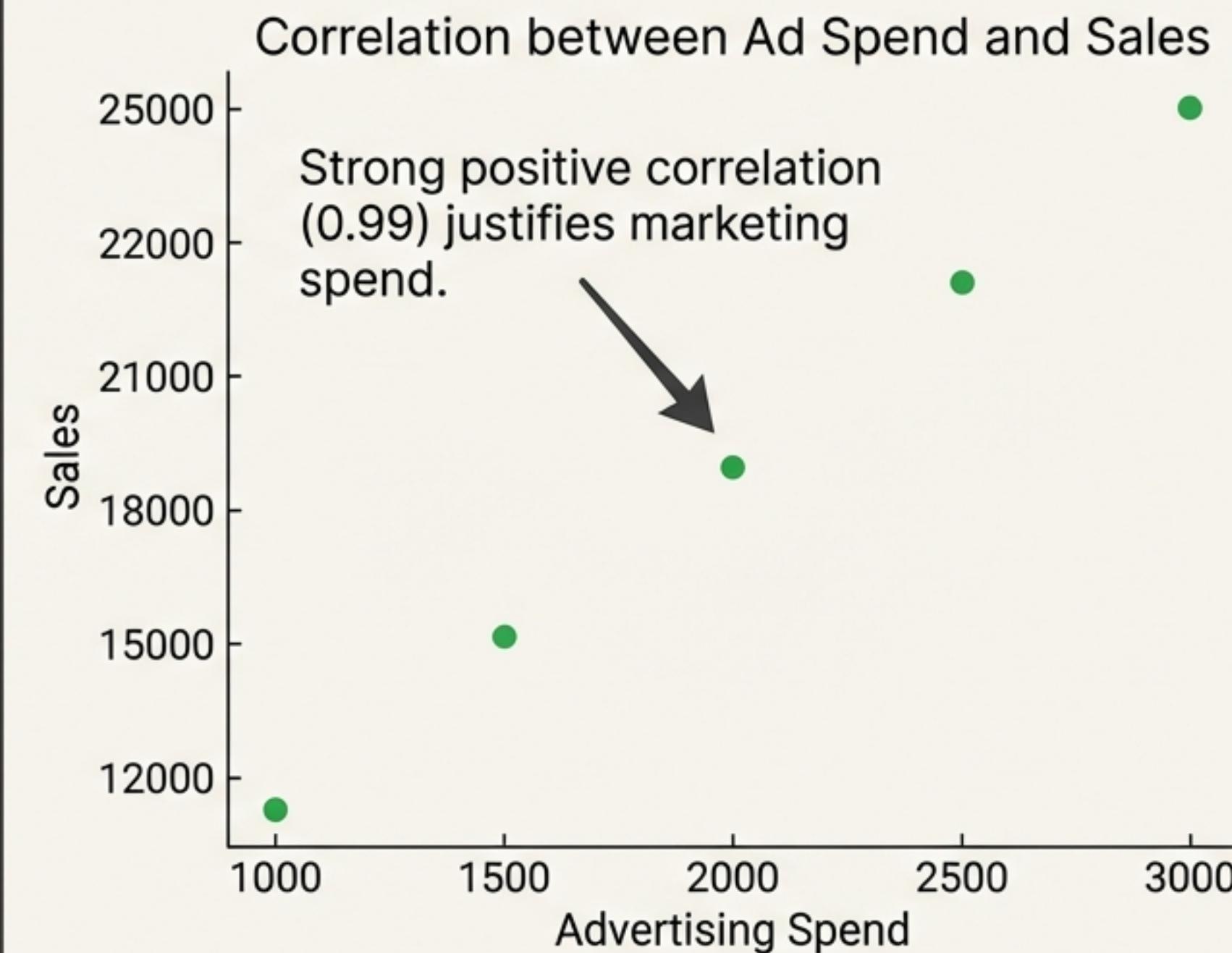


Understanding the Past and Present

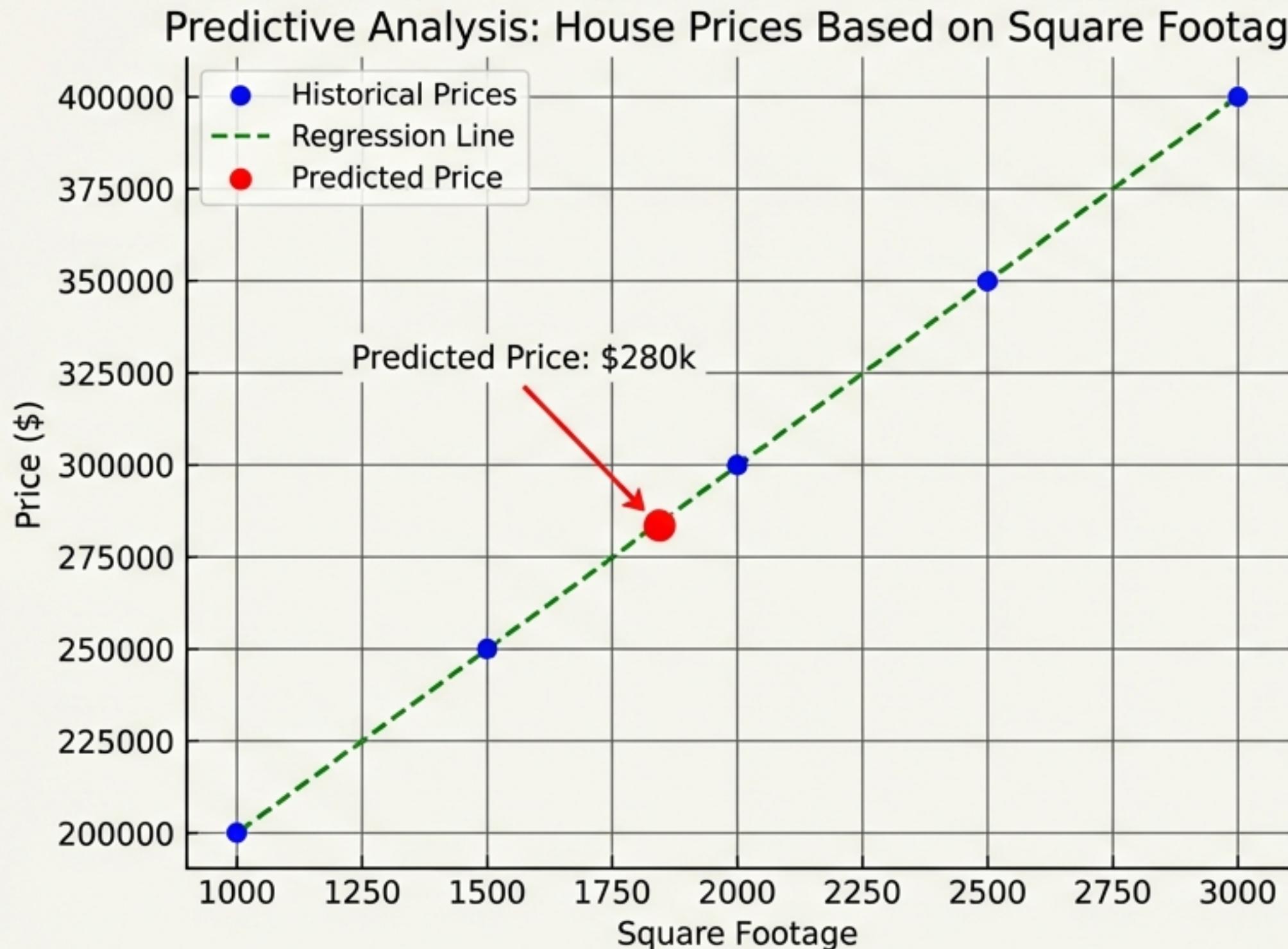
Descriptive: Summarizing History



Diagnostic: Investigating Causality



Shaping the Future with Prediction



The Logic: Linear Regression

```
model = LinearRegression()  
model.fit(X, y)  
predicted_price = model.predict([[1800]])
```

Prescriptive Analytics takes this further by using optimization models to suggest the best course of action based on these predictions.

The Data Analytics Lifecycle



Methods of Inquiry

Qualitative Analytics



- **Source:** Words, images, symbols, interviews.
- **Methods:** Content Analysis, Grounded Theory, Narrative Analytics.
- **Goal:** Deriving meaning and context from non-statistical inputs.

Quantitative Analytics



- **Source:** Numerical data and structured records.
- **Methods:** Hypothesis testing, Regression, Mean/Average calculations.
- **Goal:** Measuring variables and testing theories statistically.

Process in Action: The Titanic Dataset

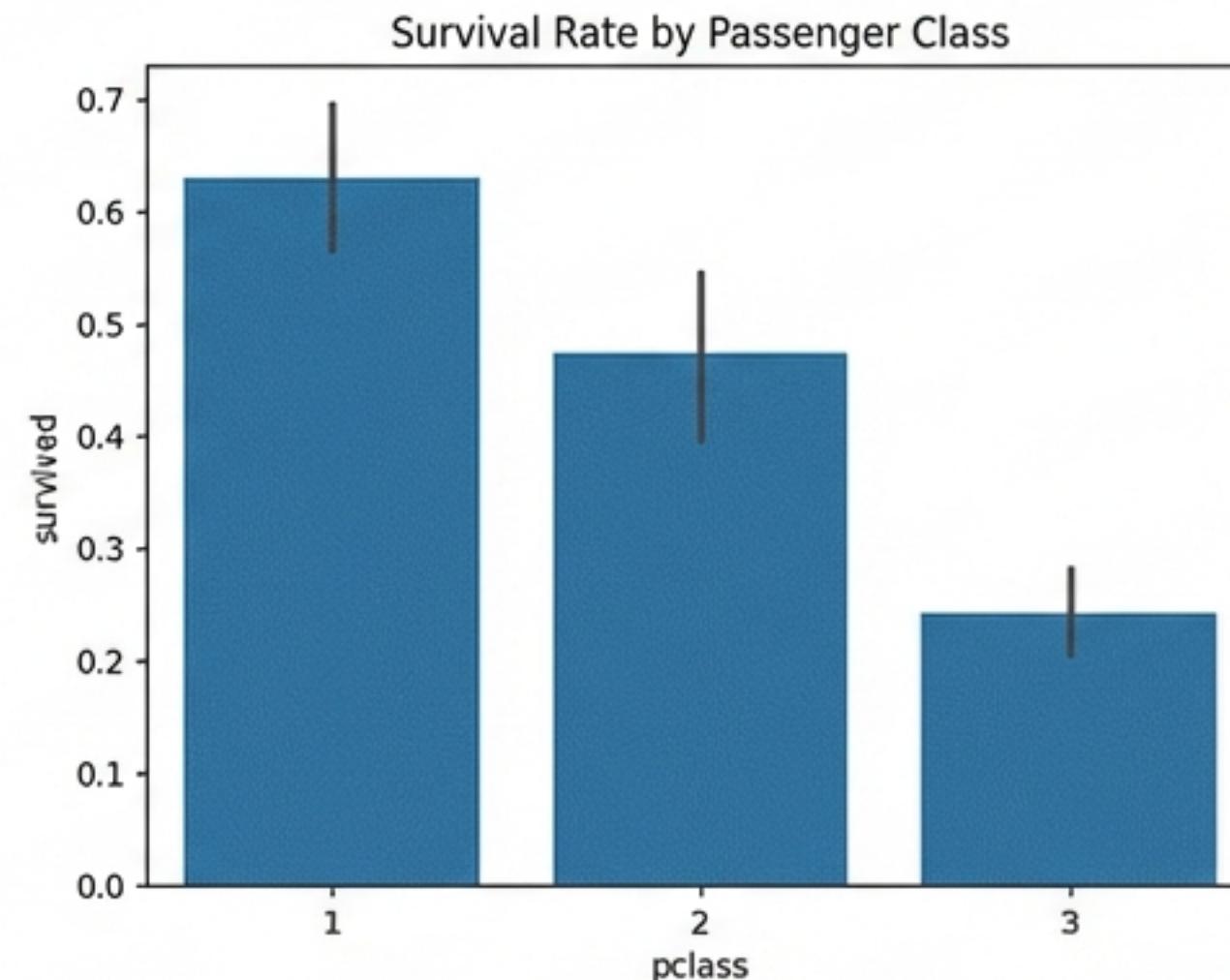
From Data Cleaning to Visual Insight

Input: Python/Pandas Cleaning

```
titanic['age'].fillna(titanic['age'].median(),  
inplace=True)  
titanic.drop(['deck', 'embark_town'], axis=1,  
inplace=True)  
titanic['sex'] = titanic['sex'].map({'male': 0,  
'female': 1})
```

Imputing missing age values and removing redundant columns.

Output: Survival Insight



The Foundation: Data Repositories

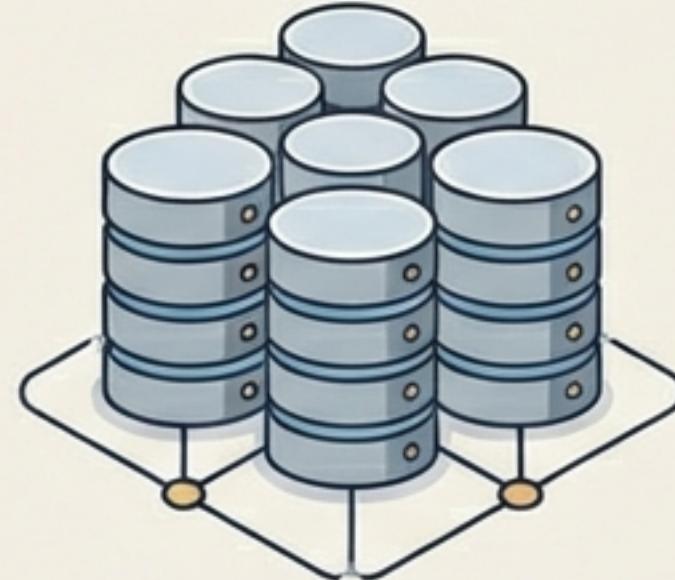
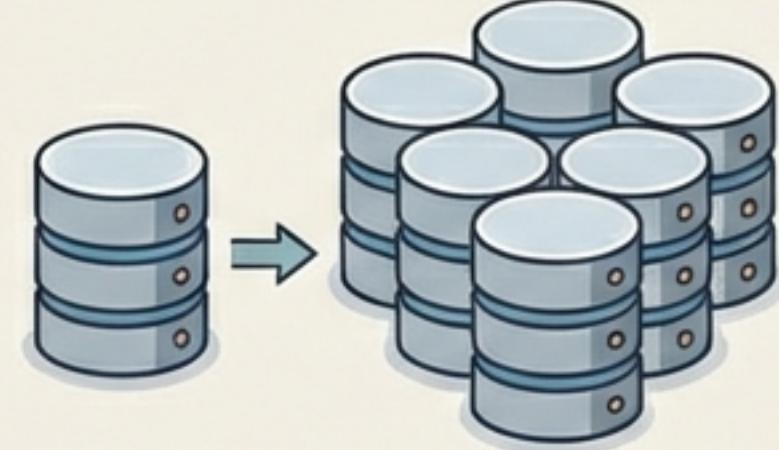
A Data Repository is a centralized system that serves as the “Single Version of the Truth,” allowing organizations to store, manage, and retrieve consolidated data for analysis.



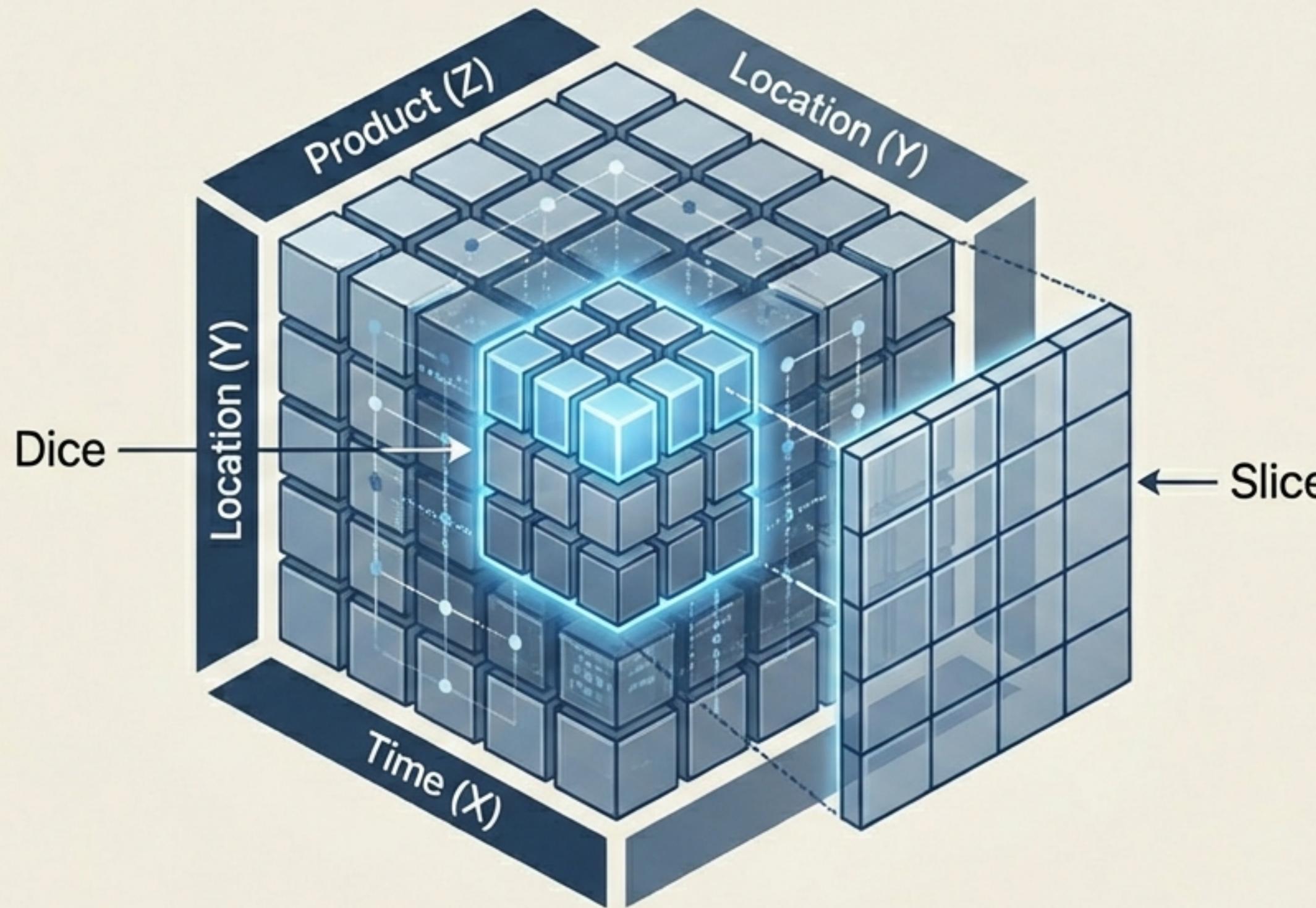
Key Challenges

- **Scalability**: Maintaining speed as volume expands.
- **Security**: Protecting centralized sensitive data.
- **Backup**: Critical redundancy to prevent data loss.

Storage Architectures: Warehouse vs. Lake vs. Mart

Data Warehouse	Data Lake	Data Mart
 <ul style="list-style-type: none">• Structured, historical data.• Subject-oriented.• Used for BI & Reporting.• Example: Snowflake, Redshift.	 <ul style="list-style-type: none">• Raw, unstructured & structured data.• 'Schema-on-read' flexibility.• Used for Machine Learning.• Example: Hadoop, AWS S3.	 <ul style="list-style-type: none">• Specialized subset of a warehouse.• Focused on one department (e.g., Sales, Marketing).• Faster access.

Multidimensional Analysis: The Data Cube



- **Dimensions:** Perspectives (Time, Region).
- **Measures:** Numerical facts (Revenue).
- **Operations:** Slice, Dice, Roll-up (Aggregate), Drill-down (Detail), Pivot (Rotate).

The Bridge: The ETL Pipeline

Extract

Extract • Transform • Load

Load



Heterogeneous
Sources

- Cleaning (Nulls/Dups)
- Mapping (Standardize)
- Enrichment

Target
Repository

Tools: Informatica, Talend, AWS Glue.

The Data Analyst: Role & Responsibilities

The bridge between technical data streams and business decision-makers.

Collection & Cleaning

Wrangling raw data via SQL/APIs.



Decision Support

Translating numbers into strategy.



Exploratory Analysis (EDA)

Identifying initial patterns.

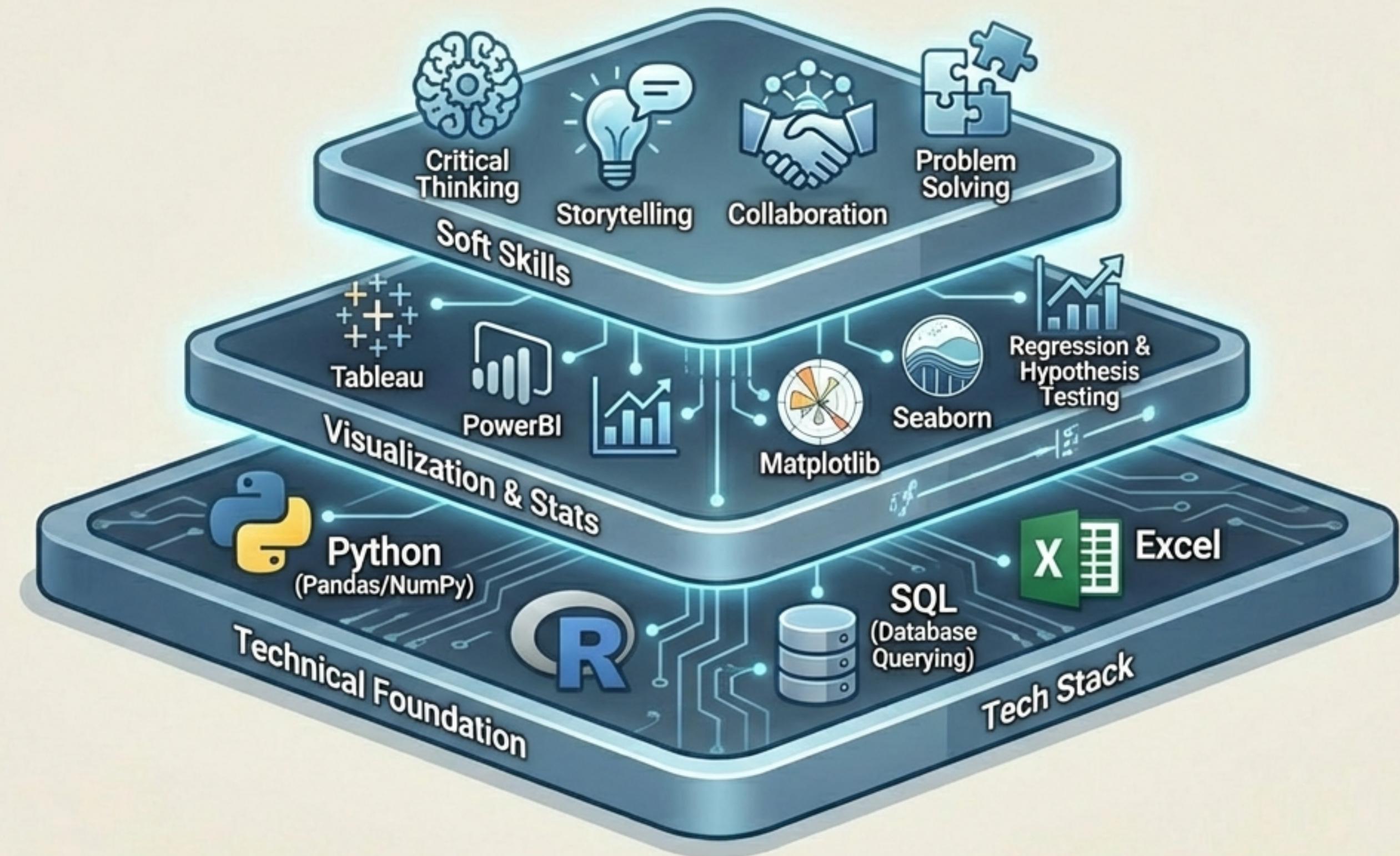


Visualization

Building dashboards in Tableau/PowerBI.



The Analyst's Toolkit



Universal Impact



The future belongs to those
who can interpret the data.