KENT STATE UNIVERSITY

Data Processing and Analysis in Construction

Ruoxin Xiong, Ph.D.

Assistant Professor of Construction Management

rxiong3@kent.edu

Spring 2026

Learning Objectives

- Understand common data types in construction
- Explore advanced data collection, cleaning, and preprocessing techniques
- Discuss analytical and machine learning methodologies
- Apply analytics for decision-making in construction projects

Why Data Processing Matters in Construction?

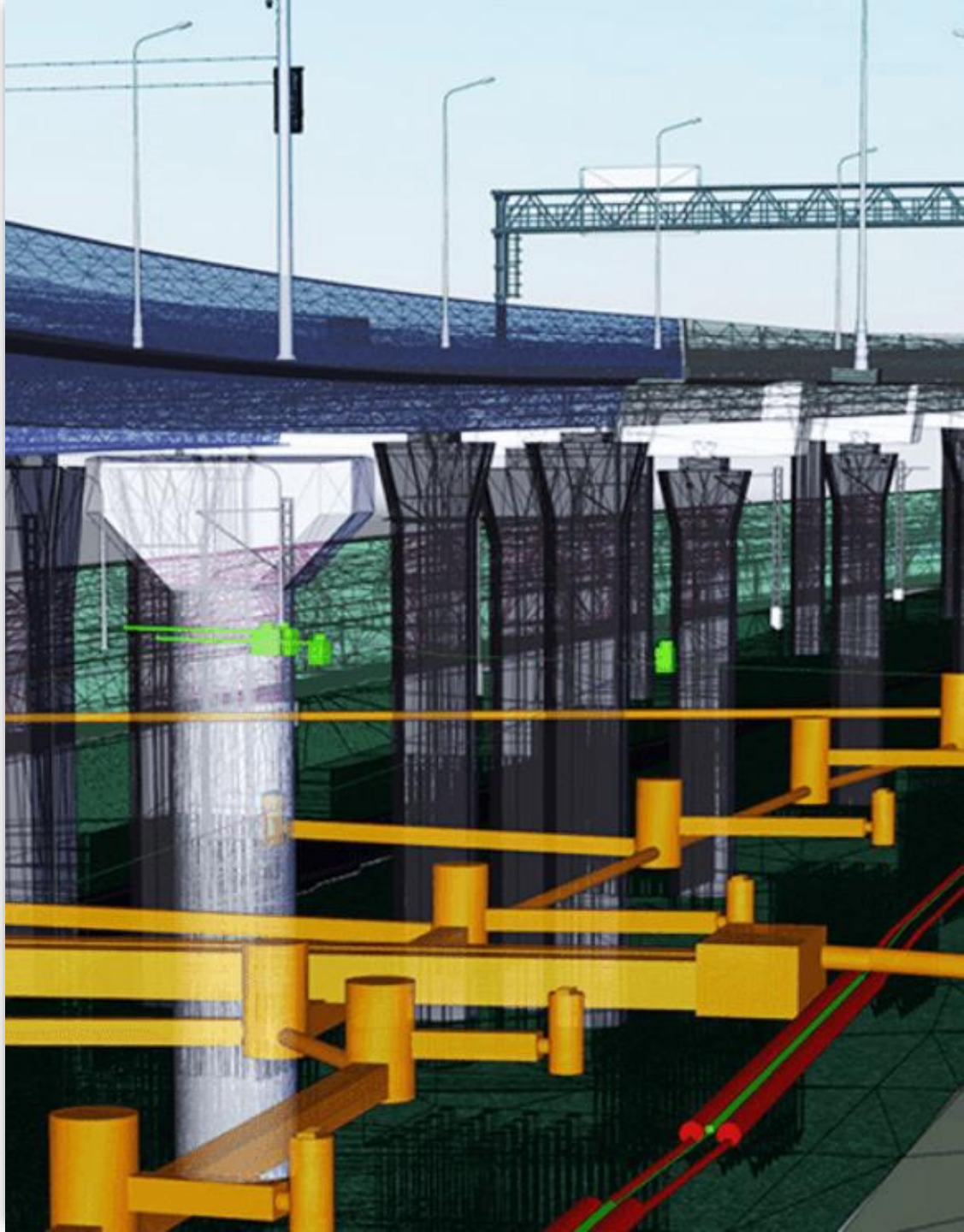
- **Increasing complexity of projects** → data as a critical asset for decision-making
- **Link between effective data management and cost/time savings.**
- **Role of technology adoption** (e.g., Building Information Modeling [BIM], IoT sensors, wearables)
- **High-level benefits**: risk mitigation, predictive maintenance, resource optimization

Types of Data in AEC

- **Geometric & Design Data**: CAD drawings, BIM models
- **Sensor & IoT Data**: IoT devices, drones, wearables, environmental sensors
- **Project Management Data**: Schedules (e.g., Gantt charts), cost data, resource logs
- **Historical & Contextual Data**: Weather records, soil tests, market benchmarks, regulatory information
- **Quality and Safety Data**: Inspection logs, incident reports

Geometric & Design Data

- **BIM (Building Information Modeling)**: 3D models with embedded metadata (e.g., material properties, supplier info).
- **CAD drawings:** 2D or 3D drawings referencing architectural, structural, MEP details.
- **Example:** Using Autodesk Revit or Bentley Systems to maintain a digital twin with layer-by-layer detail on a commercial building’s design.



Sensor & IoT Data

- **Equipment Telemetry**: GPS location, operational hours, fuel consumption, temperature
- **Drones & UAVs**: High-resolution imagery, LiDAR point clouds, thermal scans
- **Wearables**: Worker heart rate, location, activity tracking
- **Example**: A project manager using IoT sensors on a tower crane to measure loads, vibrations, and wind speeds in real time to forecast potential breakdowns or safety issues.

Sensor & IoT Data

Project Management & Administrative Data

- **Scheduling/Planning**: Gantt charts, resource allocations in software like MS Project
- **Costs & Procurement**: Vendor invoices, purchase orders, budget forecasts, time sheets
- **Example:** Tracking daily labor hours from timesheets to correlate labor usage with project milestones, identifying potential bottlenecks

Historical & Contextual Data

- **Local Codes & Regulations**: Permitting data, building codes, safety regulations.
- **Historical Project Archives**: Past performance records, cost overruns, risk event logs.
- **External Data**: Weather patterns, traffic conditions for urban projects, macroeconomic indicators (steel/concrete price indices).
- **Example**: A city-level open data portal providing 10 years of building permit records to identify average time from application to approval.

Question:

What are the common construction data categories based on their formats?

Construction Data by Format

- **Image Data:** Site Photographs, Drone Imagery
- **Video Data:** Surveillance Cameras, Drones & UAVs, Wearable Cameras
- **Point Cloud Data:** Laser Scanning (LiDAR), Photogrammetry
- **Audio Data**: On-Site Recordings, Acoustic Sensors
- **Textual / Document Data**: Project Documents, Regulatory & Legal Docs
- **Tabular / Numeric Data**: Schedules, Cost Logs, Sensor Data, Financial & Accounting Records

Question:

What potential data errors or anomalies were observed, and what underlying factors may have contributed to them?

Common Data Issues

- **Duplicates:** Repeated rows in your dataset
- **Missing Values**: Gaps in time-series, where certain timestamps lack valid readings
- **Outliers**: Unusually large (or small) values that don’t fit expected trends

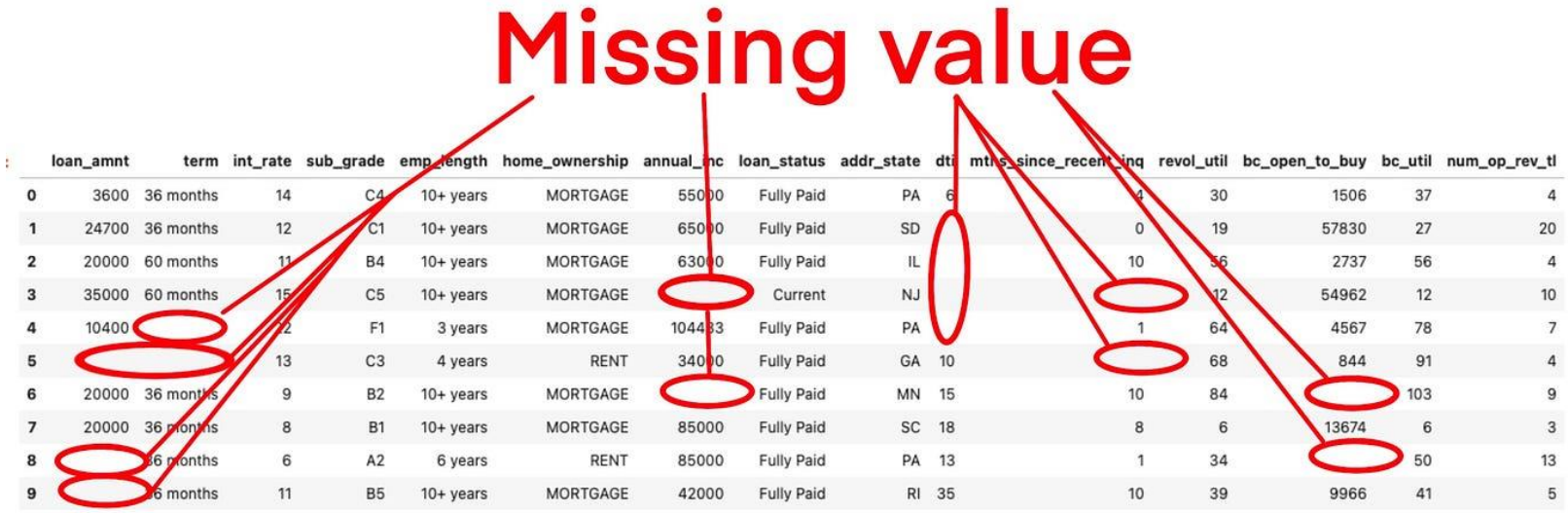
Why We Have Such Potential Errors?

- **Duplicates** often stem from software or transmission retries, repeated sensor readings, or buffering issues
  - Some sensors (e.g., a DHT22) only update data every 2 seconds or so. If the logging script reads faster than the sensor updates, you could repeatedly capture identical values.

S.N°	First Name	Last Name	Title	Company
1	Mary	Sue	Senior Marketing Manager	ABC Ltd.
2	Janet	Martin	Marketing Executive	ABC Ltd.
3	Bryan	Oscar	SEO Manager	ABC Ltd.
4	Jude	Taylor	Marketing Manager	ABC Ltd.
5	Mary S	Sue	Senior Marketing Manager	ABC Ltd.

Why We Have Such Potential Errors?

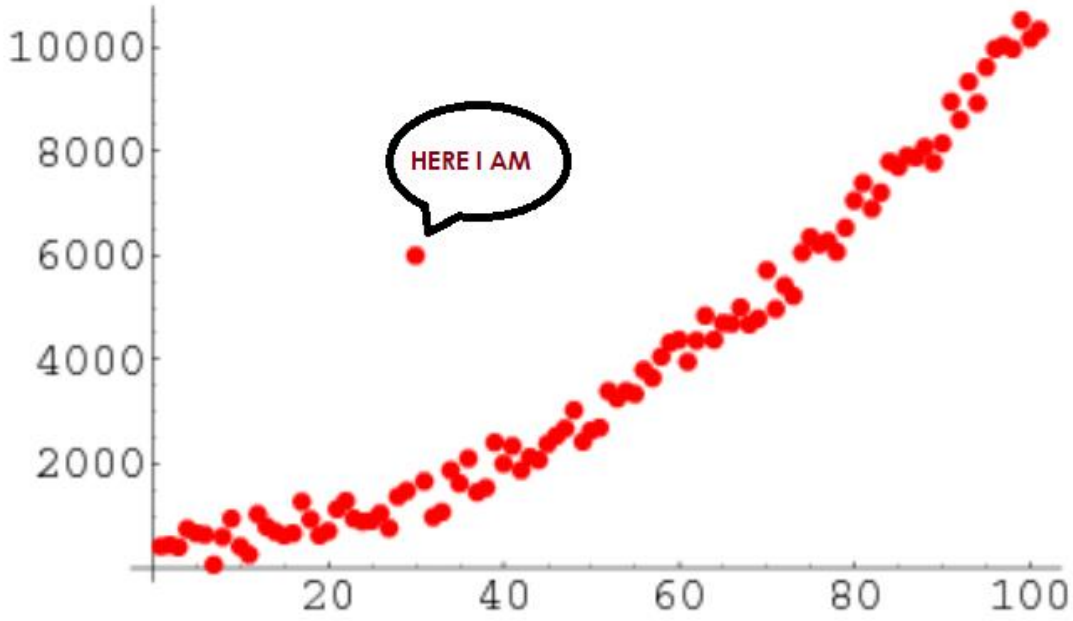
- **Missing Values** occur when sensors fail to respond or the logging process is interrupted, causing gaps in the dataset
  - A temperature/humidity sensor may occasionally fail to respond (due to hardware glitch or electrical noise)
  - Unstable Power or Connection
  - Script/Software Bugs



	loan_amnt	term	int_rate	sub_grade	emp_length	home_ownership	annual_inc	loan_status	addr_state	dti	mths_since_recent_inq	revol_util	bc_open_to_buy	bc_util	num_op_rev_tl
0	3600	36 months	14	C4	10+ years	MORTGAGE	55000	Fully Paid	PA	6	4	30	1506	37	4
1	24700	36 months	12	C1	10+ years	MORTGAGE	65000	Fully Paid	SD		0	19	57830	27	20
2	20000	60 months	11	B4	10+ years	MORTGAGE	63000	Fully Paid	IL		10	56	2737	56	4
3	35000	60 months	15	C5	10+ years	MORTGAGE		Current	NJ			12	54962	12	10
4	10400		12	F1	3 years	MORTGAGE	104433	Fully Paid	PA		1	64	4567	78	7
5			13	C3	4 years	RENT	34000	Fully Paid	GA	10		68	844	91	4
6	20000	36 months	9	B2	10+ years	MORTGAGE		Fully Paid	MN	15	10	84		103	9
7	20000	36 months	8	B1	10+ years	MORTGAGE	85000	Fully Paid	SC	18	8	6	13674	6	3
8		36 months	6	A2	6 years	RENT	85000	Fully Paid	PA	13	1	34		50	13
9		36 months	11	B5	10+ years	MORTGAGE	42000	Fully Paid	RI	35	10	39	9966	41	5

Why We Have Such Potential Errors?

- **Outliers** can arise from true environmental changes, sensor malfunctions, interference, or human errors
  - Measurement Noise & Interference
  - Hardware or Calibration Issues: Over time, some sensors drift out of calibration
  - Human errors: setup issues



Why We Have Such Potential Errors?

- Duplicates
- Missing Values
- Outliers

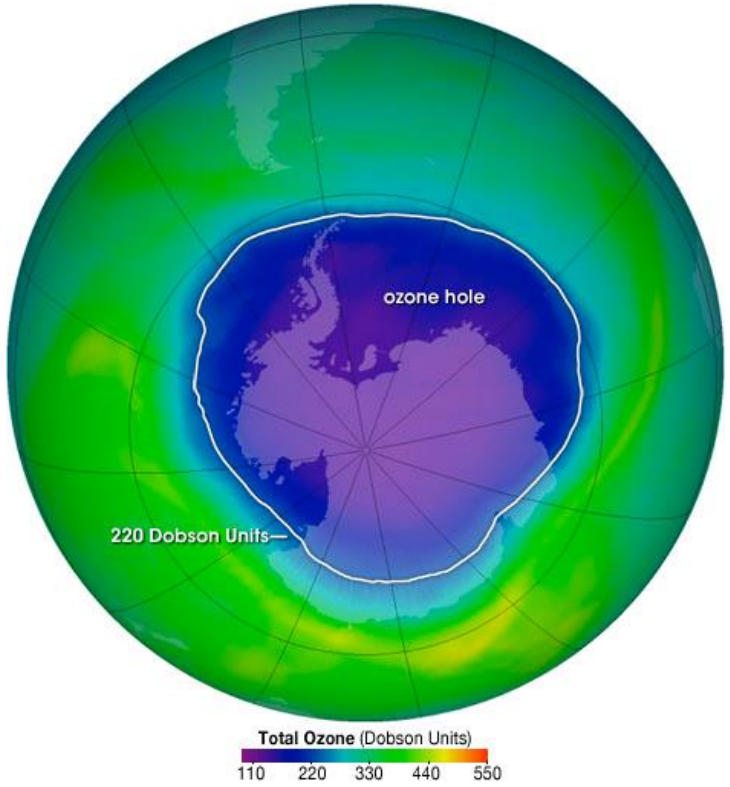
Why We Have Such Potential Errors?

- Not All Anomalies Are Errors
- Be Honest in Data Recording

Why We Have Such Potential Errors?

- Not All Anomalies Are Errors
- Be Honest in Data Recording

- The Hole in the Ozone Layer (1985)



Why We Have Such Potential Errors?

- **Not All Anomalies Are Errors**
- **Be Honest in Data Recording**

- **The Hole in the Ozone Layer (1985)**
  - Satellite data showed **unusually low ozone concentrations over Antarctica**. Initially, scientists suspected **sensor errors or instrument malfunction**.

  - Further checks confirmed a real phenomenon—an ozone hole largely caused by CFCs (chlorofluorocarbons). This discovery spurred global environmental policy changes (e.g., Montreal Protocol).

Data Quality: “garbage in, garbage out”

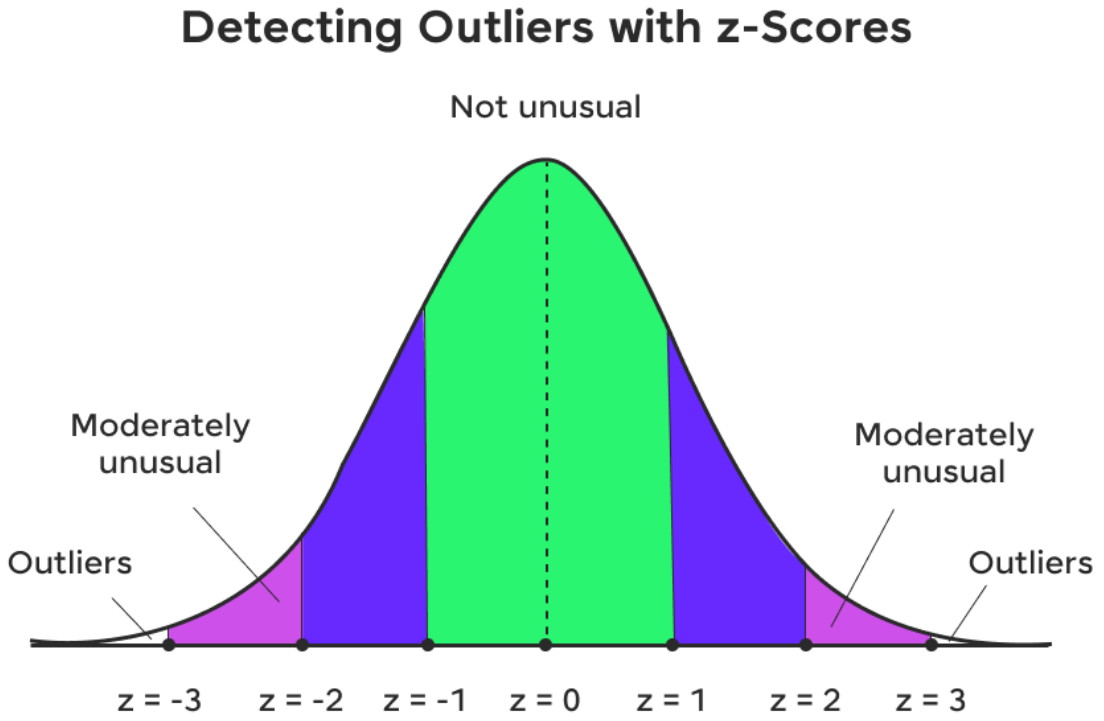
- How would you deal with potential data errors (e.g., missing values, outliers, duplicates)?

Some Data Preprocessing Methods

- **Data Cleaning**
  - **Handling Missing Values**
    - **Dropping Rows/Columns**: Remove instances or features with excessive missing data when they are not critical or exceed a defined threshold.
    - **Imputation**: Replace missing values with statistical estimates (mean, median, mode) or more advanced imputation (e.g., k-Nearest Neighbors, regression-based).
    - **Domain-Specific Rules**: Fill in missing data using domain knowledge (e.g., if “weather data” is missing, estimate from nearest-day records).
  - **Removing or Correcting Duplicates**
    - Decide whether duplicates are true copies (accidental re-logs) or valid repeated measurements.

Some Data Preprocessing Methods

- **Outlier Detection & Treatment (optional)**
  - **Statistical Methods**
    - Z-score
    - Interquartile Range (IQR)
  - **Advanced Outlier Detection**
    - Isolation Forest, Local Outlier Factor for complex, high-dimensional data
  - **Domain Knowledge**
    - Not all “outliers” are errors—some may be genuine but extreme events.

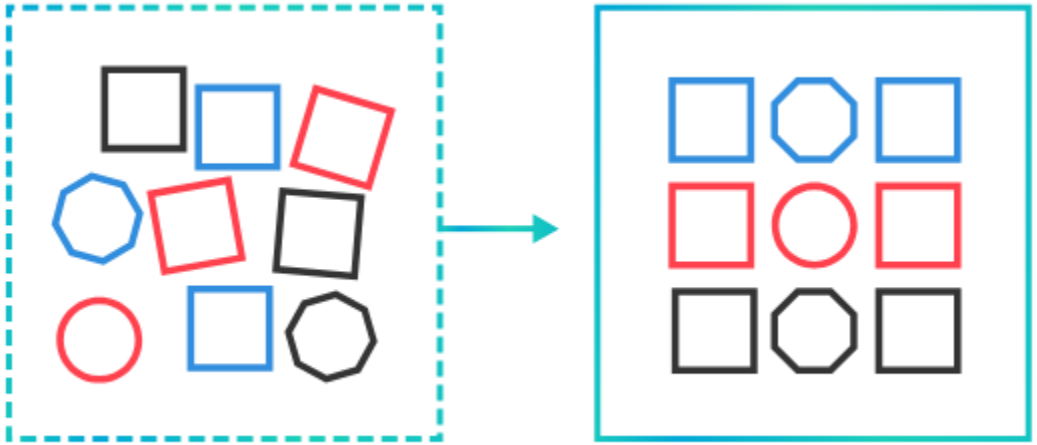


Some Data Preprocessing Methods

- **How do you deal with different data units, formats, scales?**
  - Temperature: 20^{o}C
  - Humidity: 30%
  - Comfort level: hot, cold, neutral

Some Data Preprocessing Methods

- **Data Transformation:**
  - **Feature Scaling:**
    - **Standardization (Z-score):** Centers data around zero with a unit standard deviation, useful for many ML algorithms
  - **Encoding Categorical Variables: e.g., *hot, neutral, cold***
    - **Assigns integer codes to categories: +1, 0, -1**



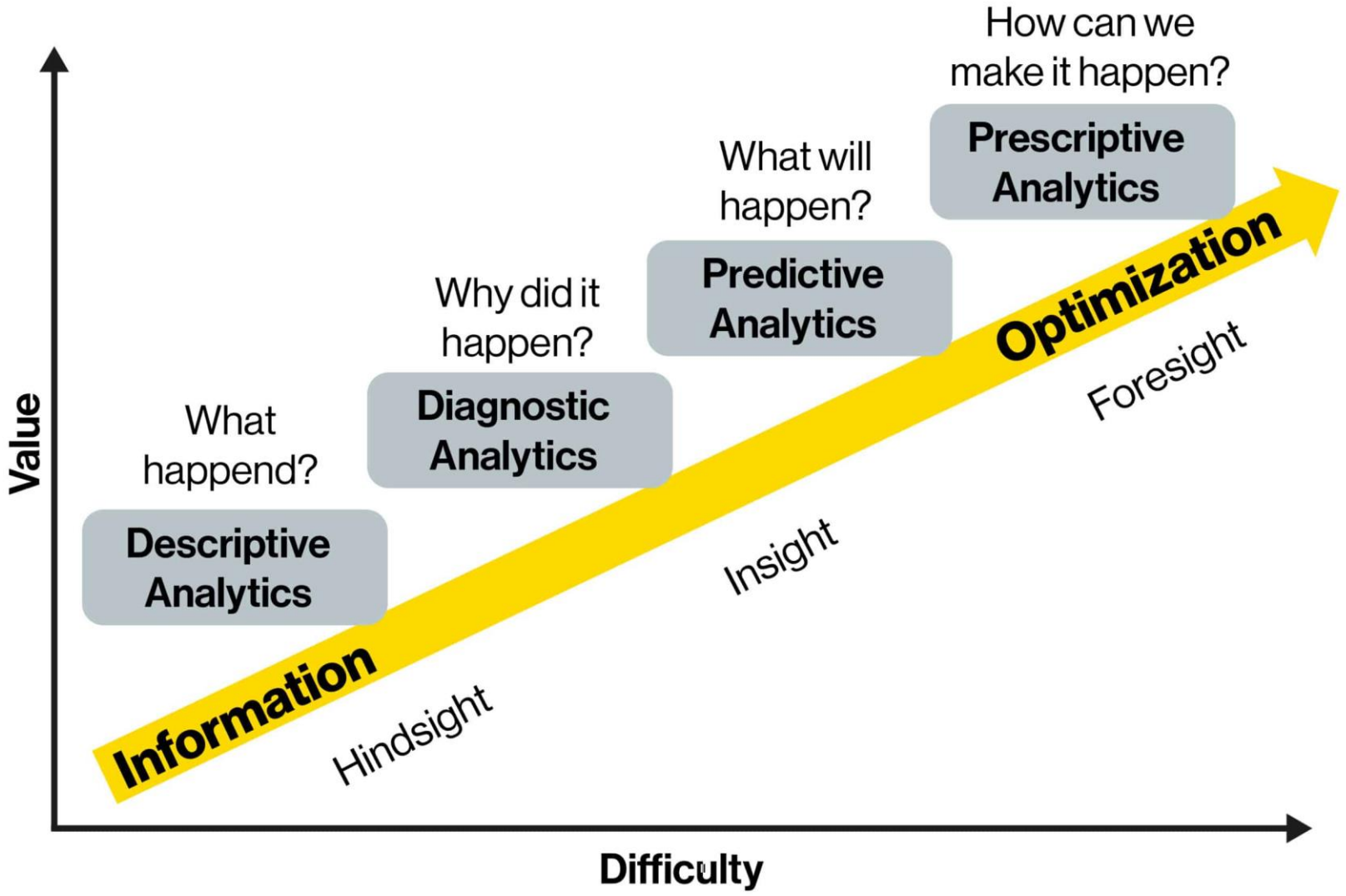
Putting It All Together

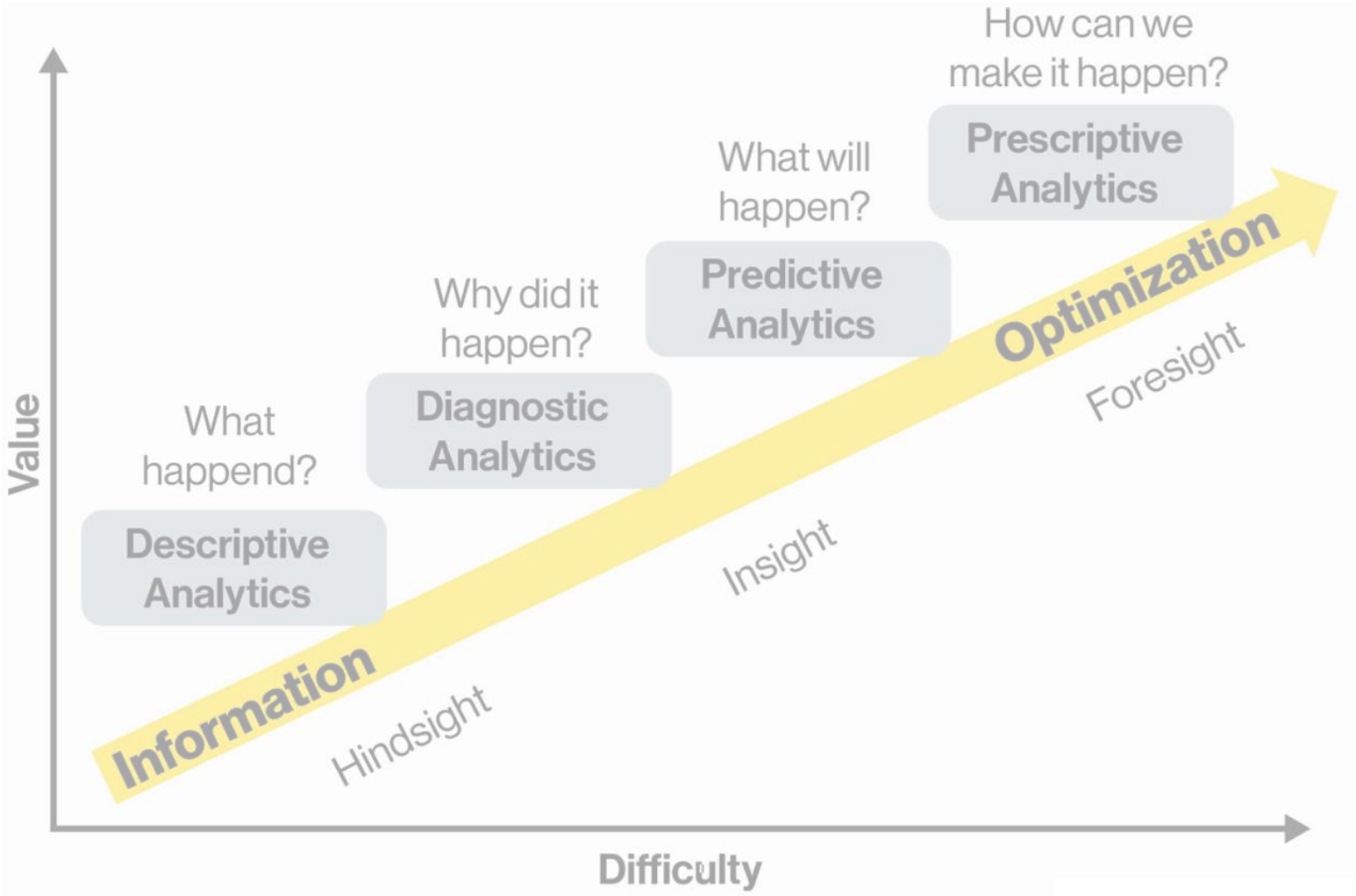
- **Typical Workflow:**
  - **Raw Data → Cleaning**: Remove duplicates, handle missing or invalid entries.
  - **Feature Engineering**: Create or transform features, encode categorical data.
  - **Scaling/Normalization**: If needed for the chosen model.
  - **Outlier Checks**: Decide how to handle extreme values.
  - **Final Dataset**: Use this for model training, validation, and deployment.

Common Data Analytical Types

- How would you deal with data and gain insights from the data?

Common Data Analytical Types





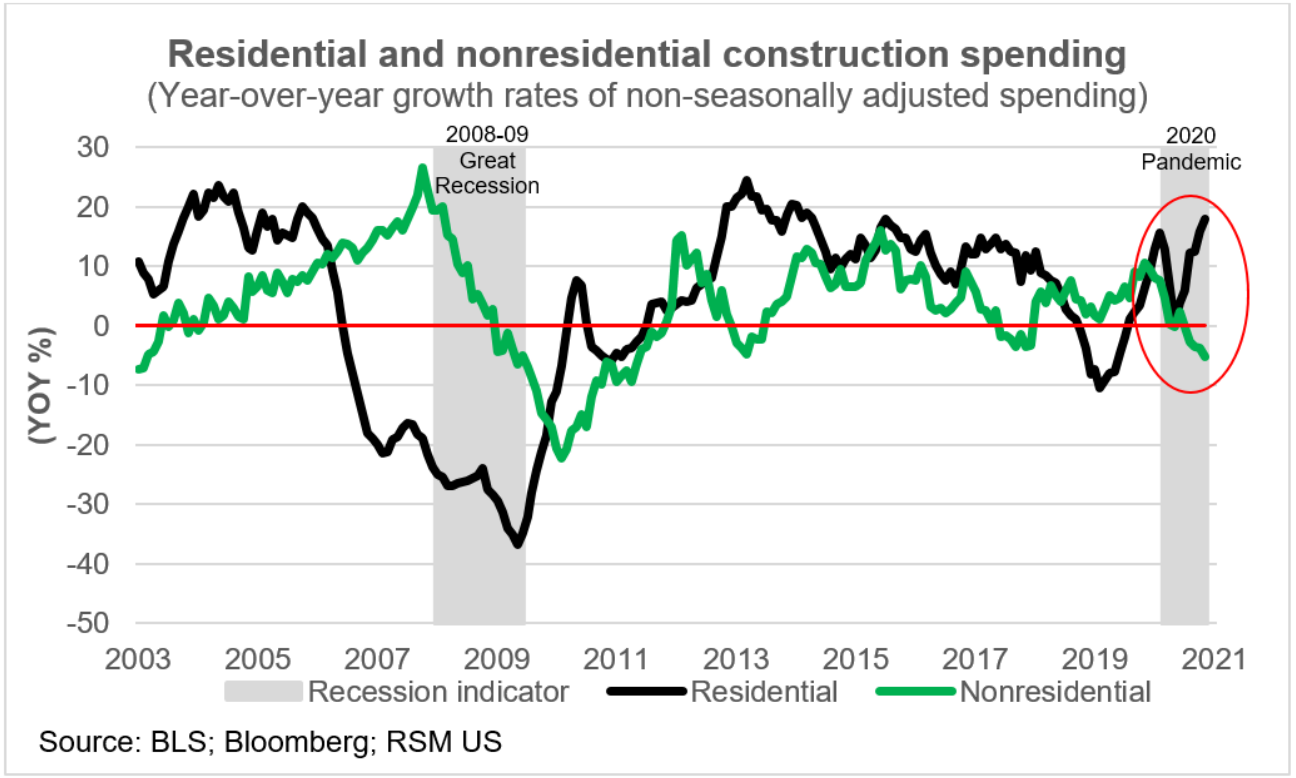
Can you give any real-world examples using these methods?

My Example:

- **Descriptive**: Project managers view a dashboard showing 70% of tasks completed, but cost usage is at 80% of budget.
- **Diagnostic**: They discover that rework in the foundation stage caused extra labor hours. A deeper correlation analysis links it to subpar soil reports.
- **Predictive**: A forecasting model suggests the project might go 2 weeks over schedule unless additional resources are allocated.
- **Prescriptive**: An optimization tool recommends hiring an extra concrete crew and slightly adjusting subcontractor schedules to minimize the delay to under 1 week.

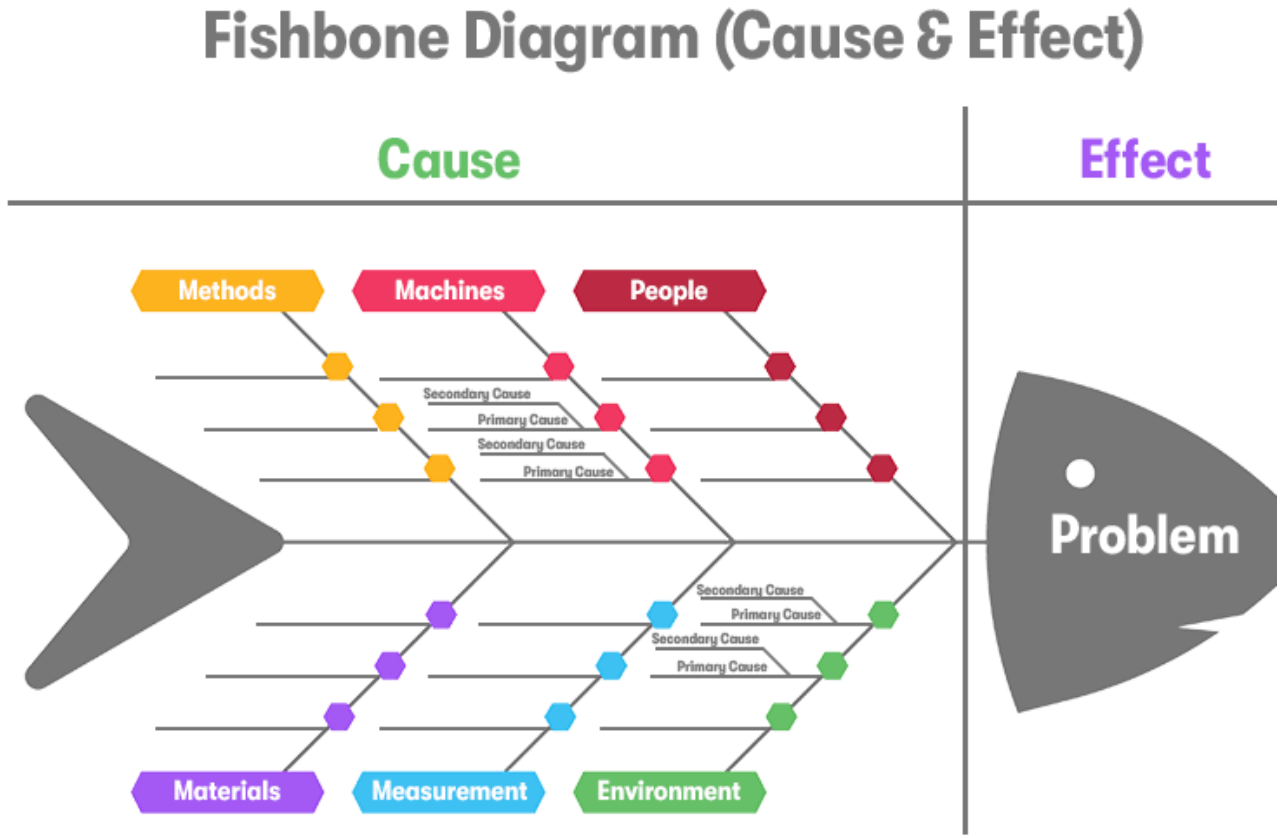
Common Data Analytical Methods

- **Descriptive Analytics**
  - **What** has happened or is currently happening?
    - ❖Data Summaries & Aggregations
    - ❖Dashboards & Visualization



Common Data Analytical Methods

- **Diagnostic Analytics**
  - **Why** it happened?
    - ❖Root Cause Analysis (RCA): Fishbone (Ishikawa) diagrams
    - ❖Correlation & Trend Analysis: (e.g., Pearson, Spearman correlations)



Common Data Analytical Methods

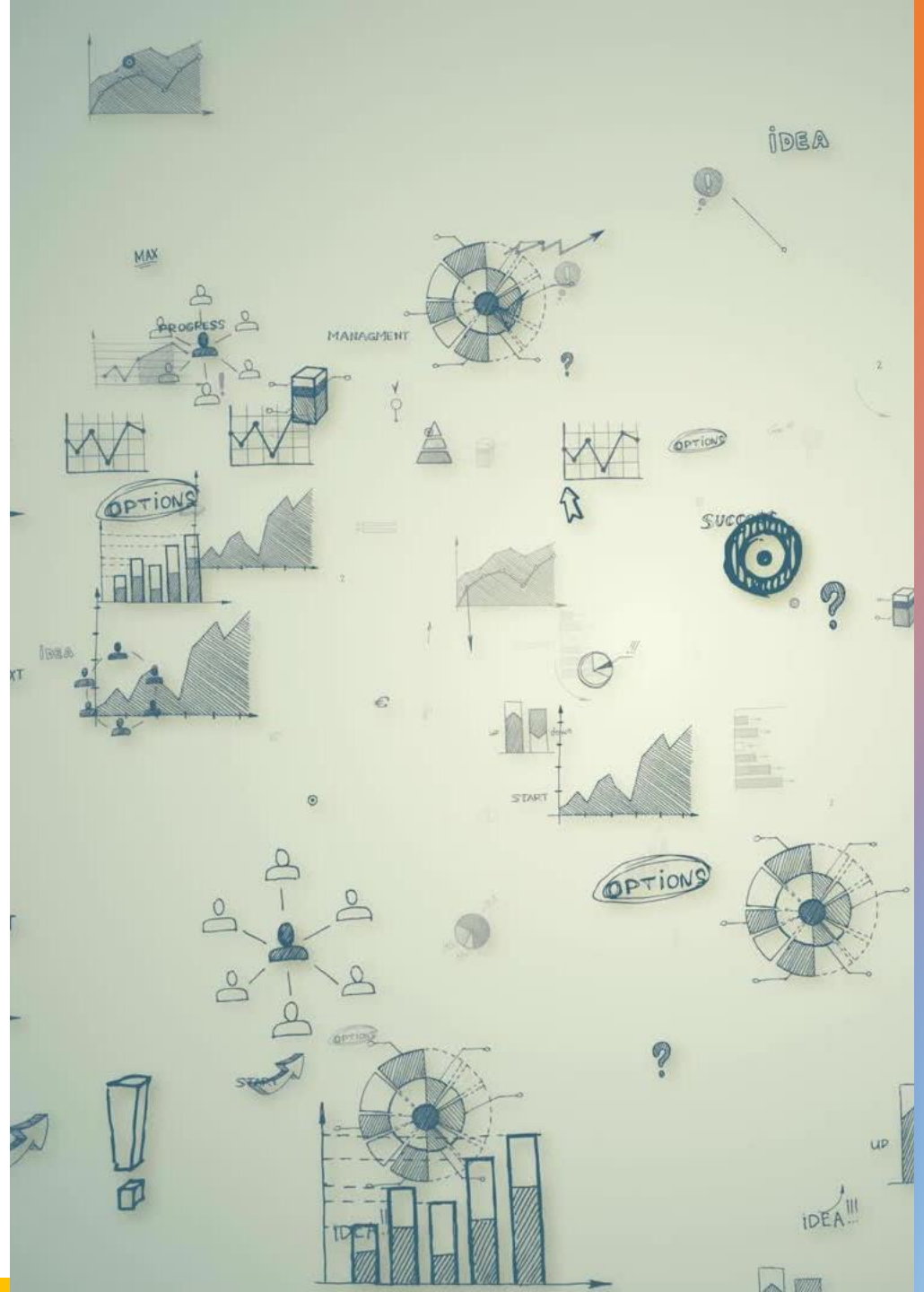
- **Predictive Analytics**
  - **What might happen** under current or similar conditions?
    - ❖Machine Learning to forecast future events or outcomes
    - ❖Time-Series Forecasting

Common Data Analytical Methods

- **Prescriptive Analytics**
  - **How** could we make it happened? (How do we achieve the best outcome?)
    - ❖ Optimization Models, e.g., *Linear/Integer Programming (LP/IP)* for resource allocation, scheduling, or budget optimization.
    - ❖ Multi-Objective Optimization
      - ❖ Balancing trade-offs (e.g., minimize cost while maximizing safety or minimizing project duration).

Data Visualization & Communication

- **Effective Dashboards:**
  - Key metrics for project managers, site engineers
- **Storytelling with Data:**
  - Tailor visual narratives to stakeholders (technical vs. executive)
- **Visualization Tools:**
  - Tableau, Python libraries (matplotlib, seaborn, Plotly)



Handling Data at Scale (Big Data Considerations)

- **Cloud services for large datasets** (e.g., AWS)
- **Distributed Processing Frameworks**: Spark, Hadoop

Handling Data at Scale (Big Data Considerations)

Challenges and Future Trends

- **Data Security & Privacy**: Protecting sensitive project and stakeholder data.
- **Integration with Digital Twins & BIM**: Real-time updates and simulation.
- **Emerging Tech**: AI-driven robotics, augmented reality for data overlay, advanced image processing.
- **Sustainability Considerations**: Data to optimize energy, material usage, and reduce waste.
- **Skills Gap & Training**: Need for data-savvy construction professionals

Summary & Next Steps

- Data is a strategic asset from planning to post-construction.
- Quality data processing → More accurate analytics and predictions.
- Machine learning and IoT-based approaches are transforming the industry.
- Future-forward skills in data science are critical for construction professionals.

Assignment 3

Thank You!