

Business Data Analysis

MGM509

Lecture 1

From Business Challenges to Data
Science Solutions



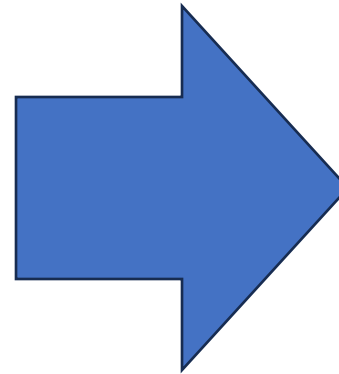
Lecture
1

What we will cover

1. Understand the types of business problems data science can solve
2. Learn about common data mining tasks and when to use them
3. Understand the full data mining process life cycle
4. Differentiate supervised vs unsupervised methods with examples

Lecture
1Why Data Science Needs a
Structured Process?

- **Business Problems Are Not Always Well-Defined**
- Real-world business challenges are often ambiguous, multi-faceted, and data-poor.
- Jumping directly into modeling risks:
 - Building irrelevant models
 - Wasting time on the wrong problem
 - Producing untrustworthy results



- **Data Mining Bridges the Gap**
- Converts business objectives into analytical questions
- Uses a structured, iterative process to guide success

Lecture
1

The Role of Data Science in Business

Turning Business Problems into Analytical Tasks

Effective data science is goal-directed:

- Understand what decisions need to be improved
- Identify what data can support these decisions

Aims of Data Science in Business

- Improve decisions with data
- Reduce guesswork by making evidence-based predictions
- Align technical outputs (models, insights) with actionable business needs

Lecture
1

Examples of Business Problems Solved with Data Science



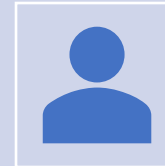
Customer retention and
churn management



Fraud and anomaly
detection



Forecasting and risk
analysis



Recommendation
systems and
personalization

Lecture
1

From Business Problems to Data Task

The Core Challenge

- Business problems are often **vague or high-level**.
- Data science reframes them into **precise analytical tasks**.
- Key question:
“What do we need to predict, classify, group, or discover?”

Begin with a clear
business question



Identify the data
needed



Match the question
to a data mining
task (e.g.,
classification,
clustering)

Lecture
1

8 Canonical Data Mining Tasks (1/2)

Classification

- Assign labels to predefined categories (e.g., spam vs. not spam)

Regression

- Predict numeric outcomes (e.g., revenue, prices)

Clustering

- Group similar records (e.g., customer segmentation)

Similarity Matching

- Finding items/entities similar to a target (e.g., recommender systems)

Lecture
1

8 Canonical Data Mining Tasks (2/2)

**Co-occurrence
Grouping**

- Discover patterns in frequently co-occurring items (e.g., market basket analysis)

Profiling

- Identify typical characteristics or behaviors of groups (e.g., loyal customers)

Link Prediction

- Anticipate connections between entities (e.g., friend recommendations)

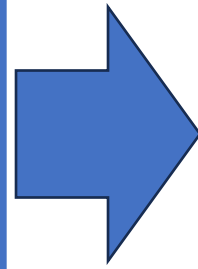
Data Reduction

- Simplify datasets to essential variables (e.g., dimensionality reduction)

Lecture
1

Example – Predicting Churn

Business Problem
“Which customers
are likely to leave
when their
contracts expire?”

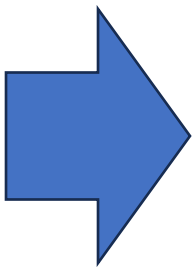


- Translated to Data Science Task
 - Task Type: Classification
 - Target Variable: Will churn (Yes/No)
 - Output: Probability or binary prediction

Lecture
1

Example – Customer Value Forecasting

Business Problem
“Which
customers are
likely to generate
the most
revenue?”



- Translated to Data Science Task
 - Task Type: Regression
 - Target Variable: Predicted customer lifetime value (CLV)

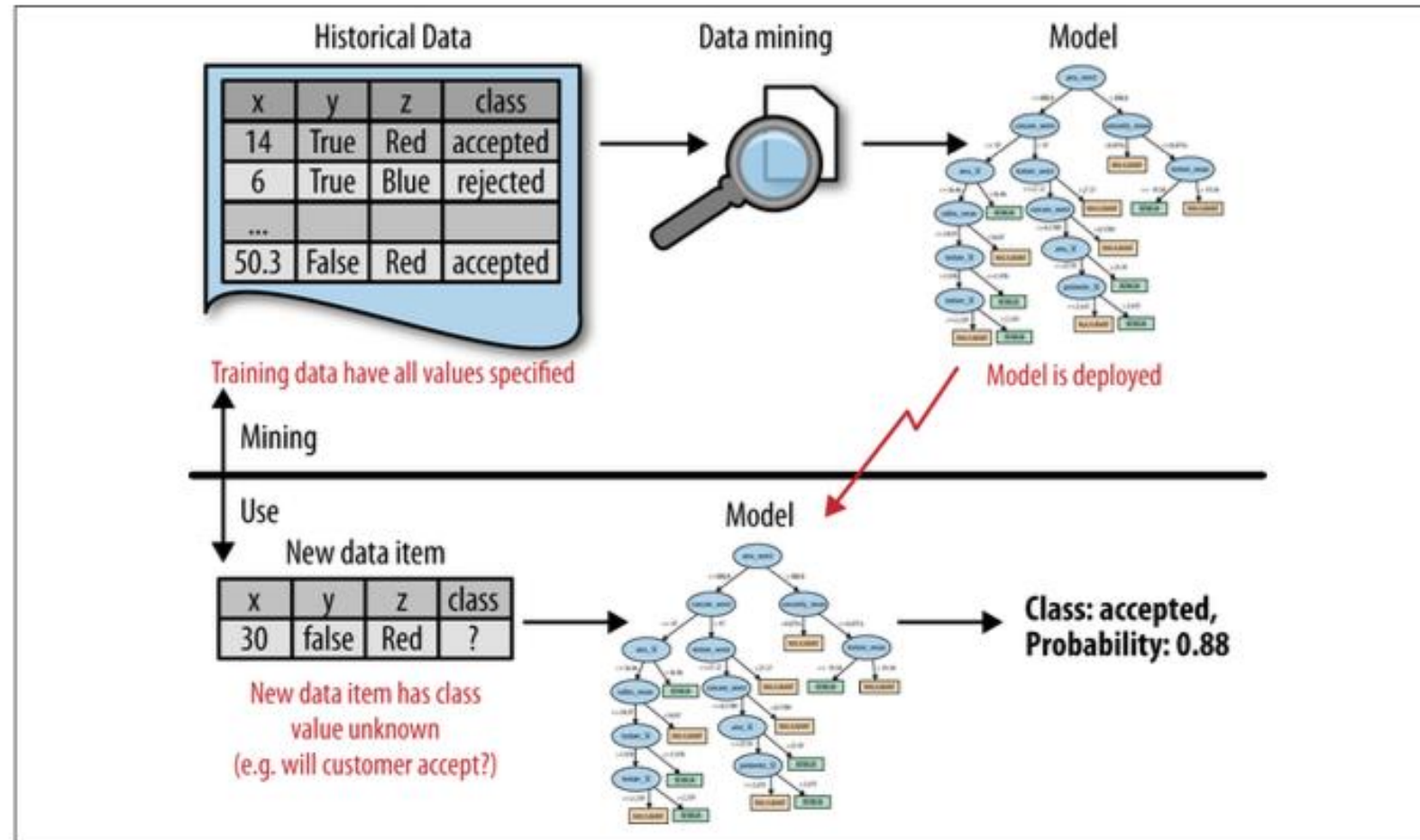
Lecture
1

Matching Problem Types to Tasks

Business Need	Data Mining Task
Predict churn	Classification
Segment Customers	Clustering
Estimate Product Demand	Regression
Recommend Products	Similarity Matching
Identify abnormal transactions	Anomaly Detection

Lecture 1

Data Mining Vs. Using Data Mining Results



Lecture
1

Data Mining Results

Result $\neq$ Just a Model

Not Just Models but Insights and Actions

- It could be:
 - A set of rules
 - A ranking
 - A cluster definition
 - A probability estimate

The form of the result should match the business goal

Lecture
1

Examples of Types of Results

Prediction Scores

- E.g., “Customer A has a 78% chance of churning.”

Class Labels

- E.g., “This transaction is classified as fraudulent.”

Segments or Clusters

- E.g., “These 3 customer types differ in spending behavior.”

Action Rules

- E.g., “Offer Plan B if usage drops below 500 mins.”

Lecture
1

Why Result Format Matters?

- Tailored to Business Needs
- Always ask:
“How will this output be used in real-world decision-making?”
- A good result should:

Support a
decision or
action

Be easy to
implement or
integrate

Reflect
operational
constraints

**Lecture
1****Supervised vs. Unsupervised Learning**

- Two approaches define how the model learns from data.

Supervised Learning

Learns from labeled examples

Unsupervised Learning

Finds structure without labeled outcomes

Lecture
1

Supervised vs. Unsupervised Learning

	Supervised Learning	Unsupervised Learning
Characteristics	• Training data includes known outcomes (labels) • Goal: Learn a function that maps input features to an output	• No labeled outcomes provided • Goal: Discover patterns, groupings, or structures in data
Common Tasks	• Classification (e.g., fraud/no fraud) • Regression (e.g., sales forecast)	• Clustering (e.g., customer segmentation) • Association rule mining (e.g., market basket analysis) • Anomaly detection (e.g., finding outliers)

Lecture
1Supervised vs. Unsupervised
Learning Examples**Supervised Learning**

- Predict customer churn using historical activity
- Detect fraud in banking transactions
- Score customer credit risk based on demographics

Unsupervised Learning

- Group customers based on purchase behavior
- Identify underlying topics in documents
- Reduce variables to core factors

Lecture
1

Why a Process Framework?

- Data Mining Is Iterative, Not Linear
- Successful projects follow a structured yet flexible process
- Helps teams:
 - Align technical work with business needs
 - Avoid common pitfalls
 - Plan for evaluation and deployment
- **Solution:** Use a proven process model like CRISP-DM (Cross Industry Standard Process for Data Mining)

Lecture
1

Data Mining Process Overview (CRISP-DM)

- Cross Industry Standard Process for Data Mining
- Not strictly sequential; often involves **loops and iteration**

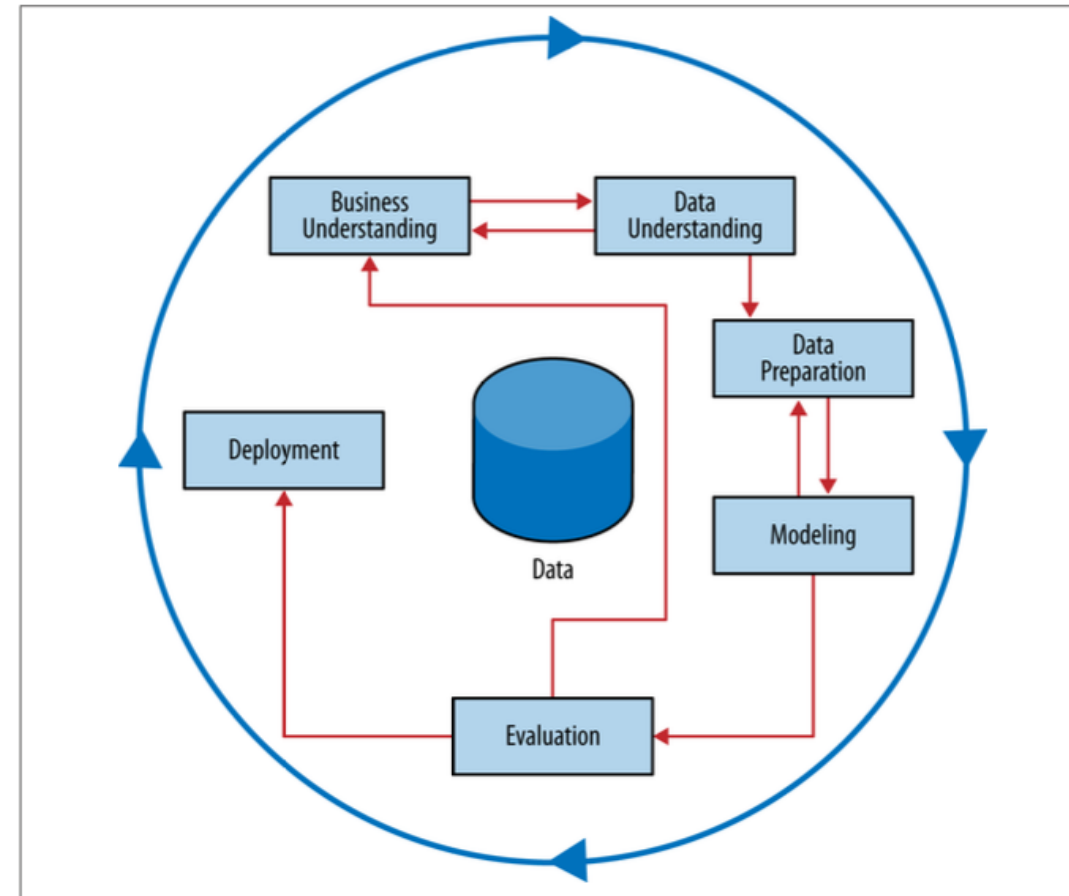


Figure 2-2. The CRISP data mining process.

Lecture
1

Phase 1: Business Understanding

- Clarify What the Business Needs
- Define project objectives and success criteria
- Understand:
 - What decision needs support
 - Constraints and expectations
 - How results will be used



Technical success is meaningless without business value

Lecture
1

Phase 2: Data Understanding

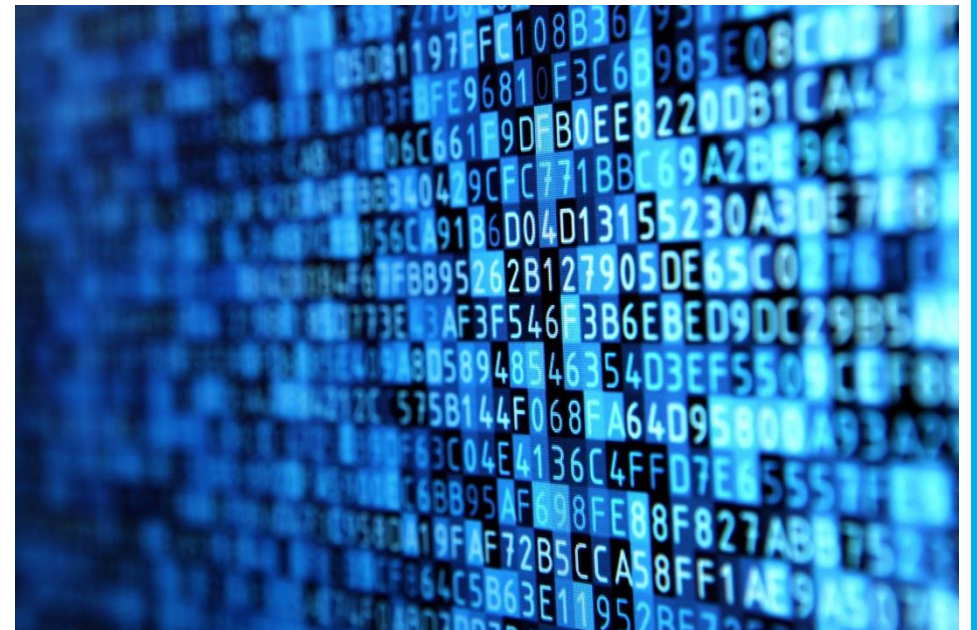
- **Get Familiar with the Data**
- Identify relevant data sources
- Examine:
 - Data quality
 - Completeness
 - Unexpected patterns
- Example: Are “churn” labels complete and timely?



Lecture
1

Phase 3: Data Preparation

- **Prepping Raw Data for Modeling**
- Most time-consuming phase
- Crucial to prevent data leakage or biased models
- Tasks may include:
 - Data cleaning (e.g., missing values)
 - Feature selection or engineering
 - Data transformation and integration



Lecture
1

Data Preparation Pitfalls & Iteration

- **Avoid "Data Leakage"**
- **Leakage:** when future data unintentionally influences training
- Example: Including a churn-related event that happens after the prediction point
- Clean, accurate, and time-respectful data is critical to trustworthy models
- **Iteration is the Norm**
- Not a Waterfall Process
- You may revisit:
 - Business understanding after initial modeling
 - Data preparation after evaluation
- Key: Stay flexible but methodical
- "Most data mining is 80% data prep, 20% modeling — and 100% learning."

Lecture
1

Phase 4: Modeling

- **Building Predictive or Descriptive Models**
- Select modeling techniques:
 - Decision trees, logistic regression, clustering, etc.
- Iterate to improve:
 - Feature choices
 - Model parameters

Models should be aligned with the business question and data structure

**Lecture
1****Modeling Techniques**

- Choosing the Right Tool for the Task

Classification	Decision Trees, Random Forests, Logistic Regression
Regression	Linear/Nonlinear models, Gradient Boosting
Clustering	K-Means, Hierarchical clustering
Association	Apriori, FP-Growth

Technique depends on task type, data type, and business context

Lecture
1

Modeling with a Purpose

- Always Link Models to Decisions
- Ask: What will this model enable us to do better?
- Consider:
 - Interpretability vs. accuracy
 - Deployment complexity
- Risk of error

A model is only useful if it improves real-world decisions

Lecture
1

Phase 5: Evaluation

- **Is the Model Really Useful?**
- Technical accuracy $\neq$ business success
- Evaluate with:
 - Holdout/test datasets
 - Business impact metrics (e.g., ROI, lift, cost savings)
- Ask:
 - Does the model solve the right problem?
 - Is the performance sufficient for action?
 - Practical constraints (e.g., speed, interpretability)



Don't skip evaluation—poor models can harm decisions

Lecture
1

Common Pitfalls in Evaluation

- Avoid Misleading Results

Overfitting

- Model performs well on training data but fails on new data

Leakage

- Future data mistakenly used in training

Biased metrics

- Accuracy can be misleading on imbalanced data

Always validate with realistic test conditions and business logic

Lecture
1

Phase 6: Deployment

- **Making the Model Operational**
- Examples:
 - Integrate into CRM system
 - Generate daily customer rankings
 - Create dashboards or alert systems
- Doesn't require full automation:
 - Sometimes simple “rules on paper” are enough



The real value of a model is realized only when deployed

Lecture
1

Putting Models to Work

- **Turning Insight Into Action**
- Deployment options:
 - Manual: score lists, reports
 - Automated: embedded into apps, websites, CRM
- Example:
 - Use churn model to trigger retention offers
 - Feed product scores into a recommendation engine

A model that sits unused has ZERO business value

Lecture
1

Deployment Considerations

- Operational Success Factors

Monitoring	Does the model still work overtime?
Maintenance	Is retraining needed as data changes?
Integration	How easily can it plug into systems or workflows?

Deployment is not the end— it's the beginning of real-world testing

Lecture
1

Beyond Data Mining– Supporting Tools

- **Many Tools Work Alongside Data Science**
- Data science is powerful, but not isolated
- Other techniques often play supporting roles:

Statistics

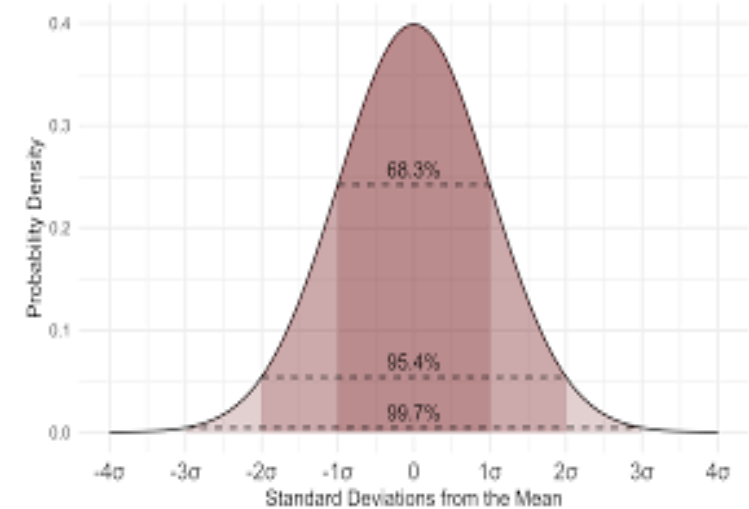
Database
queryingData
warehousingMachine
learningRegression
analysis

The right combination depends on the problem and the context

Lecture
1

Statistics

- Foundation of Inference and Uncertainty
- Helps with:
 - Hypothesis testing
 - Confidence intervals
 - Probability estimation
- Often used before or after model building:
 - To explore data patterns
 - To interpret model outcomes



Statistics is complementary, not competitive, with data mining

Lecture
1

Database Querying

- Asking Questions Directly from the Data
- Uses structured query languages (e.g., SQL)
- Great for:
 - Exploratory analysis
 - Counting events or frequencies
 - Filtering or aggregating data

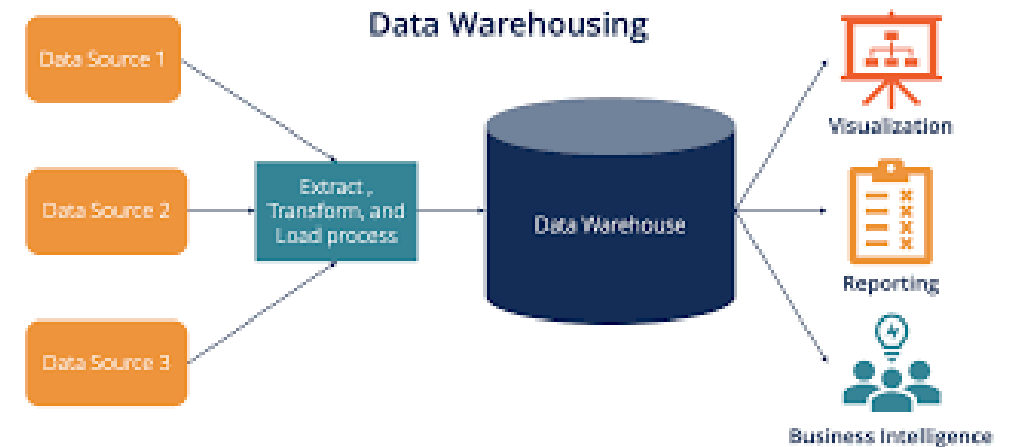


Querying is often the first step in building an understanding of the data

Lecture
1

Data Warehousing

- Centralized Storage for Analytics
- Supports:
 - Long-term data collection
 - Enterprise-wide access
 - Efficient querying and reporting
- Enables:
 - Consistency across departments
 - Reliable inputs to data science projects

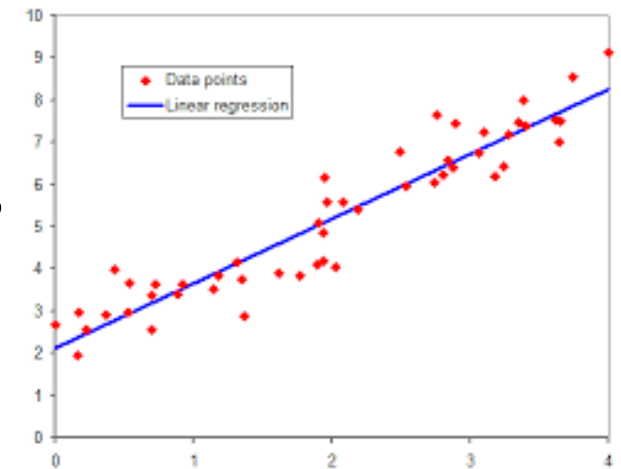


Think of it as the infrastructure layer beneath analytics

Lecture
1

Regression Analysis

- A Time-Tested Predictive Tool
- Models relationships between variables
- Types:
 - Linear regression: Predict numeric outcomes
 - Logistic regression: Predict binary classes
- Still widely used for its:
 - Interpretability
 - Speed
 - Simplicity
- A go-to method when transparency is as important as accuracy

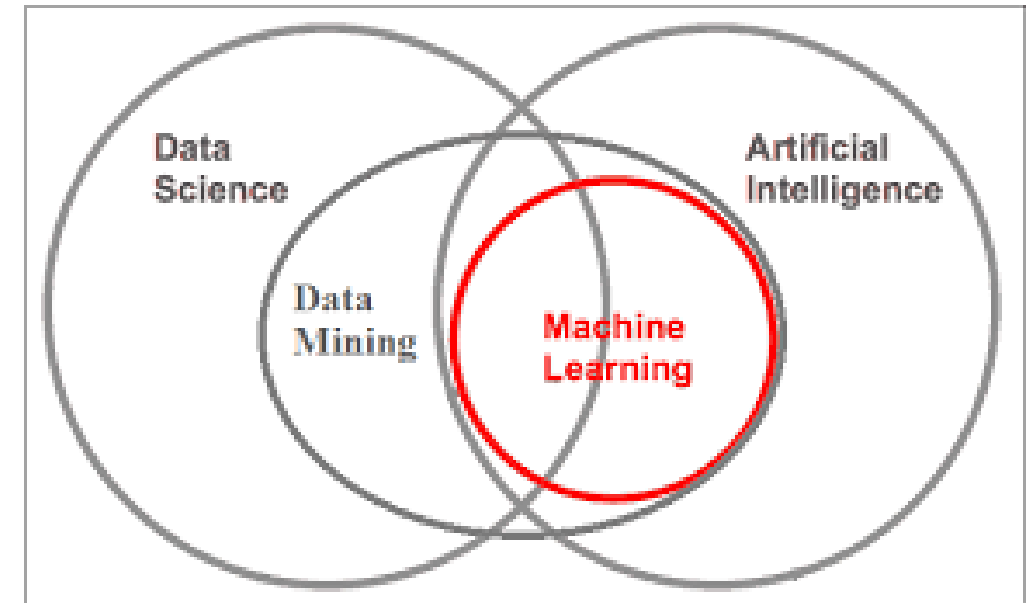


Lecture
1

Machine Learning and Data Mining

- **Similar Goals, Different Emphases**

- Machine Learning focuses on:
 - Generalization from data
 - Algorithm performance
- Data Mining focuses on:
 - Business context
 - Interpretability and deployment



ML powers the engine; data mining steers the car

Lecture
1

Managing Data Science Teams

- **Cross-Functional Collaboration Matters**
- Data science requires integration of:
 - Business domain expertise
 - Data engineering
 - Modeling and analytics skills
- Success depends on:
 - Open communication
 - Shared understanding of objectives
 - Access to relevant data



Encourage teams to work iteratively, not in silos

Lecture
1

Who Should be Involved (Skills & Roles)?



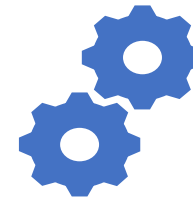
Business Stakeholder

defines the decision need



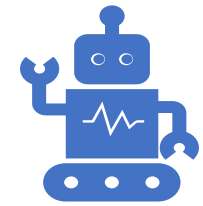
Data scientist/ analyst

leads modeling and evaluation



Data engineer

manages access and transformation



Deployment/ IT

implements model in production

Lecture
1

Success Factors for Data Science Projects

- A successful project depends on:

Clear business understanding

Proper task framing

Clean and relevant data

Iterative modeling and evaluation

Deployment that enables action

- Supported by tools like:
 - Statistics, querying, regression, machine learning

Technology alone isn't enough — process and people matter just as much

Lecture
1

Key Points

Business problems must be reframed as **data mining tasks**

Understand the distinction between:

- **Supervised** vs. **unsupervised** methods
- **Modeling** vs. **deployment**

CRISP-DM provides a proven, structured approach to projects

Effective data science = Alignment of data, technique, and decision-making

For Discussion Panel and Next Lecture:

- Use the open-source Telco Customer Churn dataset available on [Kaggle](#).
- Import the Dataset into Excel
- Familiarize yourself with the columns

Lecture 1



For Discussion Panel and Next Lecture



Thank You