Quantitative Methods

Introduction to Big Data Techniques

Intro and Exam Focus

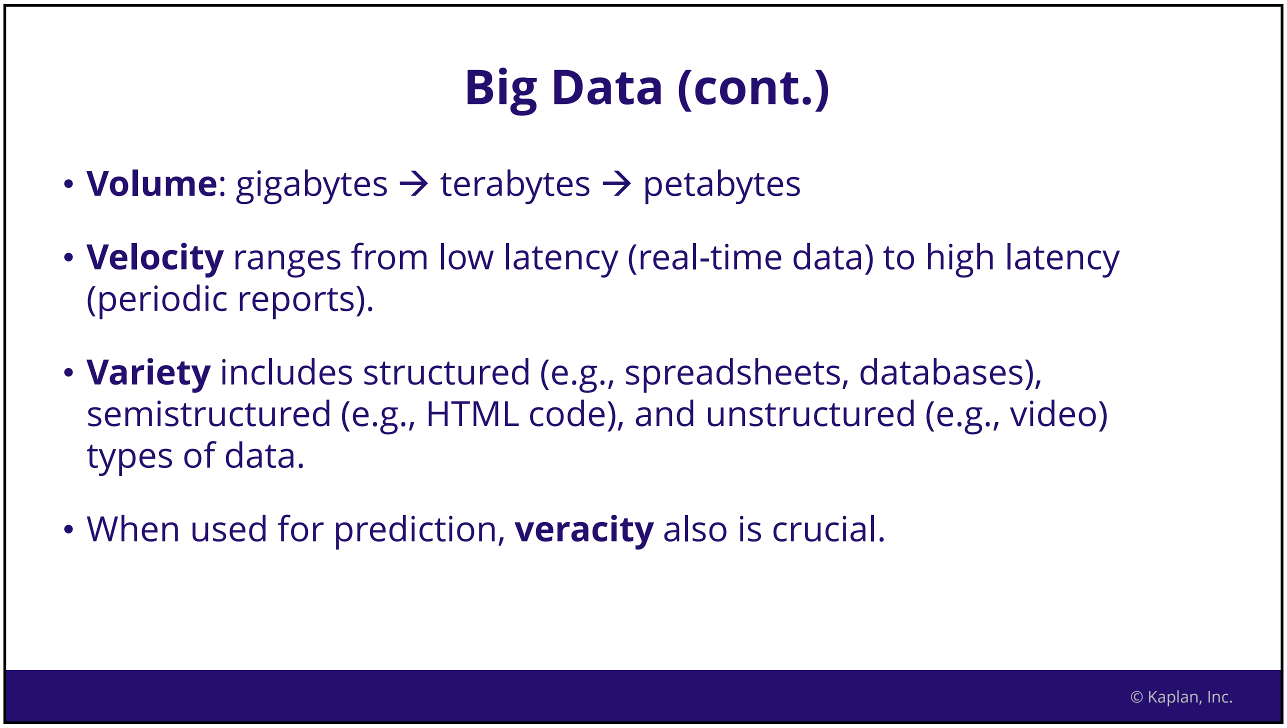
- Big data definitions and characteristics
- Machine learning vs. artificial intelligence
  - Supervised vs. unsupervised learning
  - Underfitting vs. overfitting
- Data science applications

© Kaplan, Inc.

Big Data

- **Big Data:** extremely large and complex datasets
  - **Traditional sources**: securities markets, financial statements, regulatory filings, economic statistics
  - **Nontraditional sources** (“alternative data”)
    - Individuals (social media, online reviews, website visits, emails)
    - Business processes (“corporate exhaust”) (e.g., retail scanner data)
    - Internet of Things (e.g., smart buildings/vehicles)

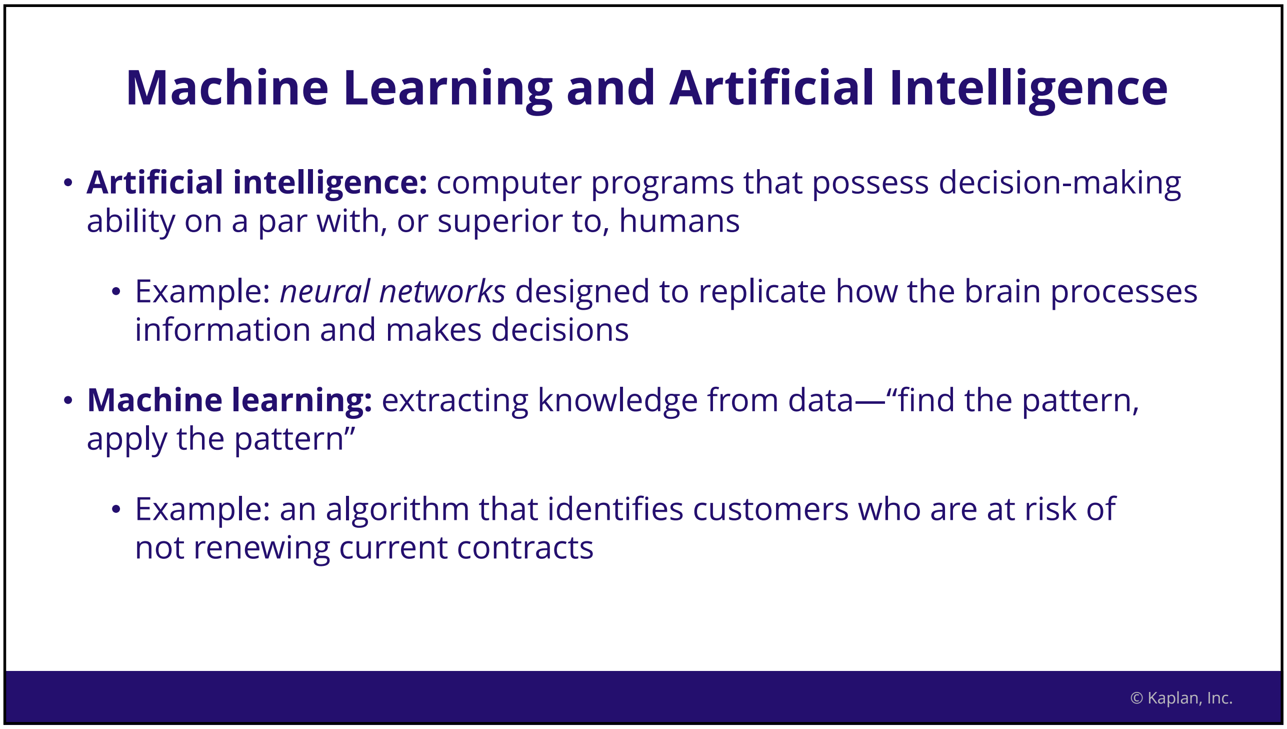
© Kaplan, Inc.



Big Data (cont.)

- **Volume**: gigabytes → terabytes → petabytes
- **Velocity** ranges from low latency (real-time data) to high latency (periodic reports).
- **Variety** includes structured (e.g., spreadsheets, databases), semistructured (e.g., HTML code), and unstructured (e.g., video) types of data.
- When used for prediction, **veracity** also is crucial.

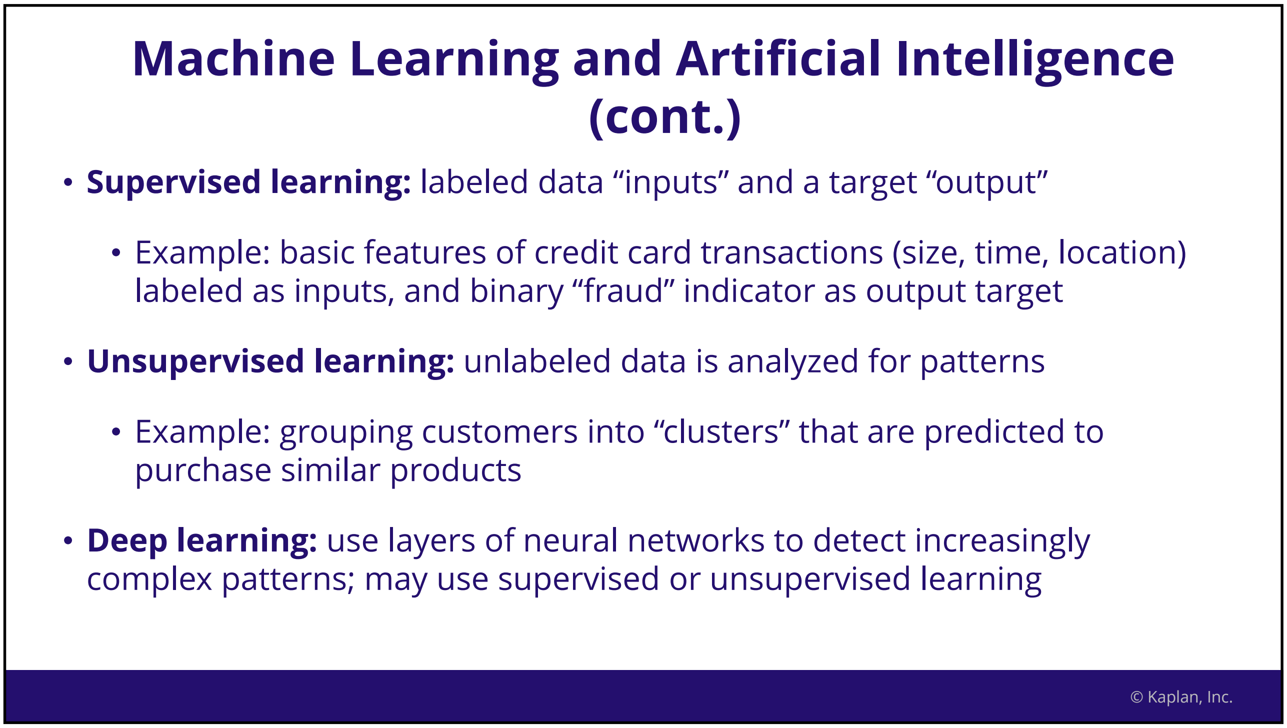
© Kaplan, Inc.



Machine Learning and Artificial Intelligence

- **Artificial intelligence:** computer programs that possess decision-making ability on a par with, or superior to, humans
  - Example: *neural networks* designed to replicate how the brain processes information and makes decisions
- **Machine learning:** extracting knowledge from data—“find the pattern, apply the pattern”
  - Example: an algorithm that identifies customers who are at risk of not renewing current contracts

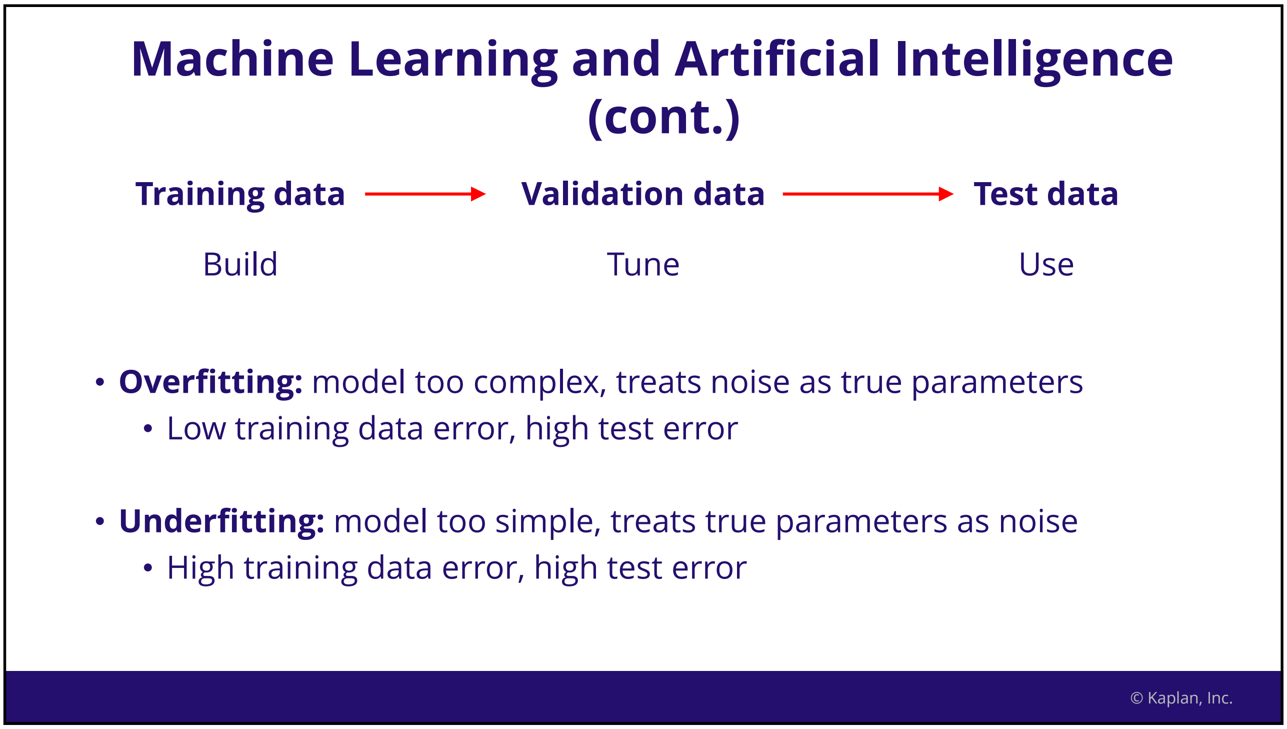
© Kaplan, Inc.



Machine Learning and Artificial Intelligence (cont.)

- **Supervised learning:** labeled data “inputs” and a target “output”
  - Example: basic features of credit card transactions (size, time, location) labeled as inputs, and binary “fraud” indicator as output target
- **Unsupervised learning:** unlabeled data is analyzed for patterns
  - Example: grouping customers into “clusters” that are predicted to purchase similar products
- **Deep learning:** use layers of neural networks to detect increasingly complex patterns; may use supervised or unsupervised learning

© Kaplan, Inc.



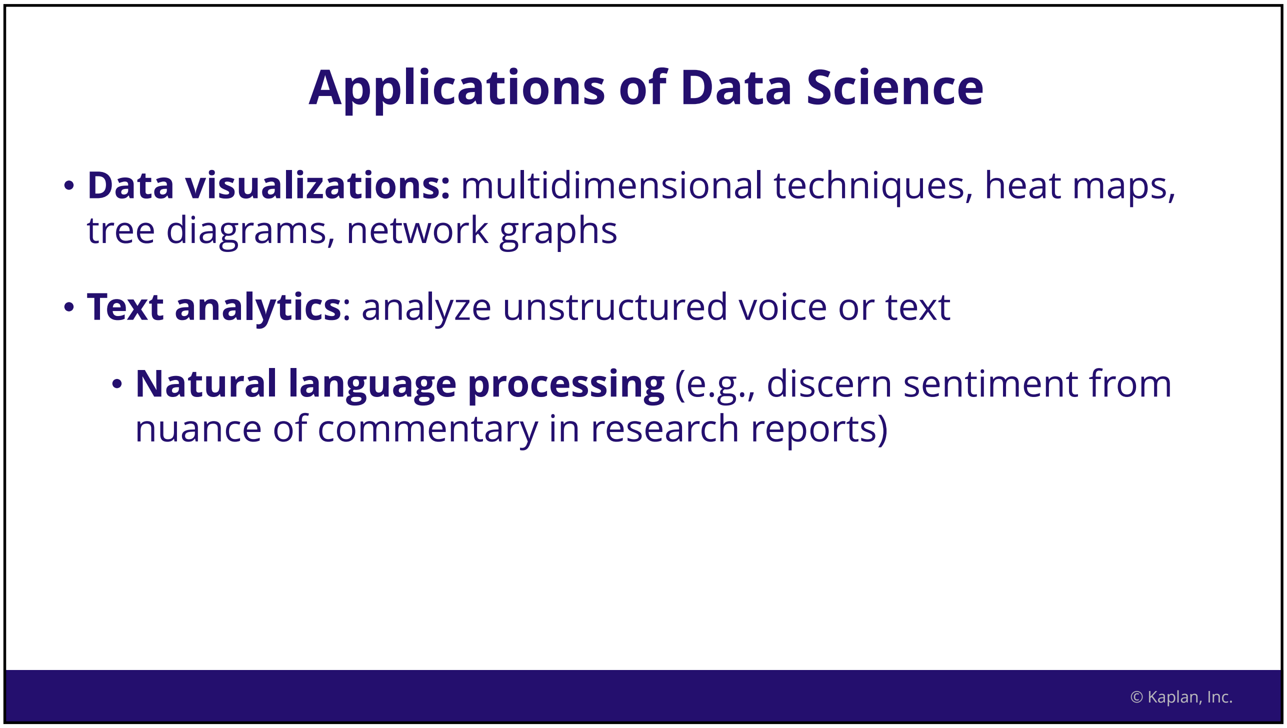
Machine Learning and Artificial Intelligence (cont.)

Training data → **Validation data** → **Test data**

Build Tune Use

- **Overfitting:** model too complex, treats noise as true parameters
  - Low training data error, high test error
- **Underfitting:** model too simple, treats true parameters as noise
  - High training data error, high test error

© Kaplan, Inc.



Applications of Data Science

- **Data visualizations:** multidimensional techniques, heat maps, tree diagrams, network graphs
- **Text analytics**: analyze unstructured voice or text
  - **Natural language processing** (e.g., discern sentiment from nuance of commentary in research reports)

© Kaplan, Inc.