LM11 · QUANTITATIVE METHODS · CFA LEVEL I

Introduction to Big Data Techniques

Fintech, Big Data, AI, machine learning, investment applications

Why This Matters

The investment industry increasingly relies on **large datasets**, **alternative data**, and **AI/ML tools**. Understanding how they fit together is essential for modern analysis.

Learning Outcome Statements

LOS a Describe aspects of "fintech" directly relevant to gathering and analyzing financial data

LOS b Describe Big Data, artificial intelligence, and machine learning

LOS c Describe applications of Big Data and data science to investment management

---

LOS A · FINTECH

Fintech

Fintech = Finance + Technology. Refers to the **technological innovation in the design and delivery of financial services and products**.

Major Drivers

- Rapid growth in data (volume and velocity)
- Technological advances (compute, storage, networking)

Major Applications

- **Analysis of large datasets** — traditional data + alternative data (social media, sensor networks)
- **Analytical tools** for extremely large datasets → AI, especially useful for complex *non-linear* relationships

---

LOS B · CHARACTERISTICS OF BIG DATA

Big Data: The Four V's

Big Data = the vast amount of data generated by industry, governments, individuals, and electronic devices from traditional *and* alternative sources.

CHARACTERISTIC	DEFINITION
Volume	Amount of data — millions to billions of points. Measured in **terabytes (1,000 GB)** and **petabytes (1,000 TB)**.
Velocity	Speed of recording/transmitting/generating. Real-time = *low latency*; lagged = *high latency*.
Variety	Text, image, speech, video. Structured (tables), semi-structured (web pages, photos with tags), and unstructured (raw video).
Veracity	Credibility / reliability of different data sources — quality vs quantity.

Structured vs Unstructured

- **Structured:** organized in tables, database entries (spreadsheets, relational DB)
- **Semi-structured:** web pages, photos with metadata
- **Unstructured:** voice, images, video (cannot be represented in tabular format — all data must be structured before use)

LOS B · TRADITIONAL & ALTERNATIVE SOURCES

Sources of Data

Traditional Sources

- **Financial markets:** price and volume
- **Businesses:** regulatory filings, press releases
- **Governments:** economic data

Three Main Sources of Alternative Data

- **Individuals:** text, photos, audio, website clicks and engagement, transactional data, web footprints
- **Business processes:** sales, credit card data, corporate exhaust (supply-chain information, bank records, retail scanner data)
- **Sensor data:** smartphones, RFID chips, PoS scanner data; also **Internet of Things (IoT)** — an array of physical devices (smart buildings, appliances, traffic sensors)

Challenges

- **Quality:** selection bias, outliers, missing data
- **Volume:** data cleansing / organizing; appropriateness
- **Ethical / Legal:** use of personal info, web scraping rights

CHECK YOUR UNDERSTANDING · QUIZ 1

Quiz 1

Select the best answer. Feedback appears after each click.

Q1. Which of the following is NOT one of the 4 V's of Big Data?

Q2. Real-time data are said to have:

Q3. Which of the following is an example of ALTERNATIVE data?

Q4. Unstructured data (e.g., voice, video):

LOS B · AI AND MACHINE LEARNING

Artificial Intelligence & Machine Learning

Artificial Intelligence (AI)

Computer systems that exhibit **cognitive and decision-making ability** comparable or superior to human beings. Ranges from expert systems (if-then rules) to **neural networks** (modeled on how brains learn and process information).

Machine Learning (ML)

Seeks to **extract knowledge from large amounts of data** without assumptions about the data's underlying probability distribution. Learns from known examples to determine underlying structure.

Requires **large training datasets**: training data → validation data → test data (final check of predictive ability).

TYPE	DESCRIPTION
Supervised	Input and output data are *labeled*. Machine learns to model outputs from inputs, then is given new inputs.

Unsupervised	Input data are *not labeled*. Machine describes the structure of the data (clusters, dimensions).
Deep Learning	Uses layers of neural networks to identify patterns; supervised or unsupervised. Image and speech recognition are applications.

Overfitting vs Underfitting

- **Overfitting:** model too complex — learns input/output too exactly and identifies spurious patterns.
- **Underfitting:** model not complex enough — fails to identify actual patterns and relationships.

LOS C · APPLICATIONS

Fintech Applications in Investment Management

Text Analytics

Analysis of **unstructured data in text or voice form**. Can partially automate tasks like evaluating company regulatory filings or extracting sentiment from transcripts.

Natural Language Processing (NLP)

Use of computers and AI to **interpret human language**. Applications: checking regulatory compliance in employee communications; evaluating large volumes of research reports to detect *subtle changes in sentiment*.

Risk Governance

Understanding a firm's exposure to a wide variety of risks. ML and Big Data are useful in modeling and testing risk — especially with **real-time data** to monitor risk exposures.

Algorithmic Trading

Computerized securities trading based on predetermined rules. Uses: optimizing execution instructions, dividing large orders across exchanges, and **high-frequency trading** that exploits intraday mispricings.

Data Science Process

Data science = methods for processing and visualizing data.

- Processing: **capture, curation, storage, search, transfer**
- Visualization: charts/graphs for structured data; word clouds / mind maps for less structured data

CHECK YOUR UNDERSTANDING · QUIZ 2

Quiz 2

Select the best answer. Feedback appears after each click.

Q1. Machine learning that uses LABELED input and output data is called:

Q2. An ML model that learns the training data TOO precisely and identifies spurious patterns has:

Q3. Which fintech application uses AI to interpret human language?

Q4. An example of ALTERNATIVE data from 'Individuals' would be:

Q5. High-frequency algorithmic trading MOST directly benefits from:

CHECK YOUR UNDERSTANDING · QUIZ 3

Quiz 3

Select the best answer. Feedback appears after each click.

Q1. Veracity in Big Data refers to:

Q2. The most granular layer of AI that uses multi-layer neural networks for image and speech recognition is called:

Q3. Data processing steps described in the curriculum include capture, curation, storage, search, and:

MODULE COMPLETE

LM11 · Introduction to Big Data Techniques Summary

Final Score

0 / 0

Cheat Sheet

$$Fintech = Finance + Technology$$

$$4V's: Volume, Velocity, Variety, Veracity$$

$$Datatypes: Structured \cdot Semi-structured \cdot Unstructured$$

$$Altdatasources: Individuals \cdot Business \cdot Sensors/IoT$$

$$ML: supervised(labeled), unsupervised(unlabeled), deep(layers)$$

$$Applications: Text, NLP, RiskGov, AlgoTrading$$

Last-Minute Exam Checklist

Low latency: real-time data

Unstructured: must be structured before use

Overfitting: too complex, captures noise

Underfitting: too simple, misses patterns

Alt data categories: Individuals / Business / Sensors

Data processing: capture, curation, storage, search, transfer

Step 1 of 10