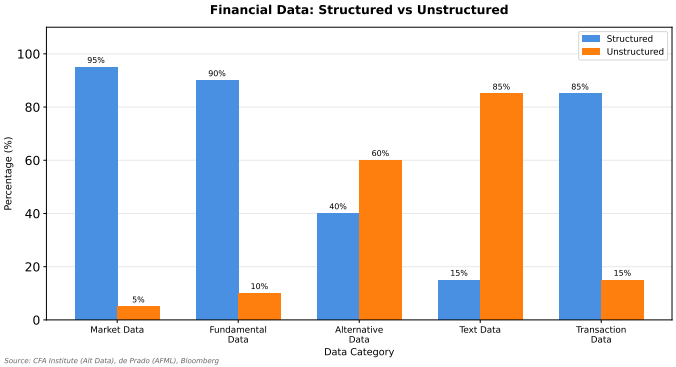
Lesson 26: Financial Data

Mini-Lecture Version (30 min)

Digital Finance

Learning Objectives: Distinguish between structured and unstructured financial data — Identify major sources and vendors of financial data — Understand the alternative data revolution and its applications — Recognize data quality issues and preprocessing requirements

Structured vs. Unstructured Data



Modern ML: Combines both types (e.g., sentiment scores from text as features in tabular models).

This concept is fundamental to understanding Financial Data.

Traditional Financial Data Sources

Market Data:

- Stock prices (exchanges)
- Bond yields (TRACE, Bloomberg)
- Derivatives (CME, Eurex)
- FX rates (interbank, EBS)
- (See full lecture for details)

Fundamental Data:

- Financial statements (EDGAR, SEDAR)
- Company events (earnings, M&A)
- Economic indicators (BLS, Fed, ECB)
- Industry metrics (PMI, CPI)

Credit/Risk Data:

- Credit bureaus (Experian, Equifax, TransUnion)
- Ratings (Moody's, S&P, Fitch)
- Loan performance data
- Default histories

Major Vendors:

- $28,320/year multi-terminal, $31,980 single terminal (2025)
- Refinitiv (formerly Thomson Reuters)
- FactSet
- S&P Capital IQ
- Morningstar

Trend: Declining costs for basic data (Yahoo Finance free), but premium data remains expensive.

This concept is fundamental to understanding Financial Data.

The Alternative Data Revolution

What is Alternative Data?

- Non-traditional data sources
- Often unstructured or semi-structured
- Provides early signals
- Competitive edge (information advantage)

Categories:

1. **Web-scraped:** Prices, reviews, job postings
2. **Sensor/IoT:** Satellite, credit cards, mobile location
3. **Social:** Twitter sentiment, Reddit mentions
4. **Business:** Email receipts, app usage

Example Use Cases:

- Satellite images: Count cars in parking lots (retail sales proxy)
- Credit card data: Real-time consumer spending
- Job postings: Company growth indicators
- App downloads: User engagement trends
- (See full lecture for details)

Market Size:

- $1.7B in 2020
- Projected $17B by 2027
- Hedge funds are largest buyers

Challenges: Quality control, legal/ethical concerns, data decay (alpha decay).

Historical context helps explain current Financial Data landscape.

Alternative Data: Examples in Detail

Satellite Imagery:

- Providers: Orbital Insight, RS Metrics
- Use: Count oil tanks, construction activity
- Example: China steel production estimates
- Frequency: Daily to weekly
- (See full lecture for details)

Credit Card Transactions:

- Providers: Facteus, Second Measure
- Use: Real-time revenue tracking
- Example: Restaurant chain performance
- Privacy: Aggregated, anonymized

Social Media Sentiment:

- Providers: RavenPack, Bloomberg sentiment
- Use: Market mood, event detection
- Example: Tweet volume predicting volatility
- Challenges: Noise, manipulation

Web Traffic:

- Providers: SimilarWeb, Alexa (discontinued)
- Use: Company engagement metrics
- Example: E-commerce site visits
- Limitation: Sample-based estimates

Key Question: Does alternative data provide genuine alpha or just noise? Evidence: Mixed, diminishing returns as adoption increases (alpha decay).

Real-world examples demonstrate Financial Data applications.

Data Quality: The GIGO Principle

Garbage In, Garbage Out:

- ML models amplify data quality issues
- No algorithm fixes bad data
- Quality ¿ Quantity (usually)

Common Data Problems:

- **Missing values:** Deletions, NaN, nulls
- **Outliers:** Errors vs. true extremes
- **Inconsistencies:** Units, formats, definitions
- **Duplicates:** Same record multiple times
- **Errors:** Typos, wrong values

Finance-Specific Issues:

- **Survivorship bias:** Only successful firms remain
- **Look-ahead bias:** Using future information
- **Corporate actions:** Splits, dividends, mergers
- **Restatements:** Accounting changes, revisions
- **Stale data:** Delayed or infrequent updates

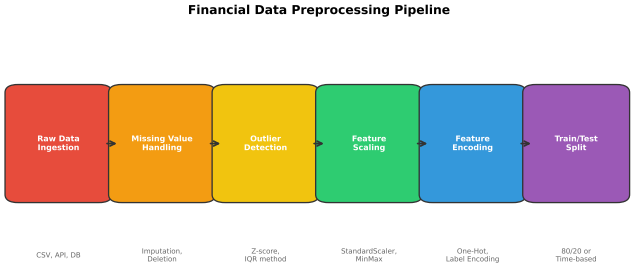
Impact on ML:

- Biased predictions
- Overfitting to noise
- Poor generalization
- Misleading performance metrics

Best Practice: Spend 50-80% of project time on data cleaning and validation.

This concept is fundamental to understanding Financial Data.

Data Preprocessing Pipeline



Source: scikit-learn.org, pandas.pydata.org, Géron (Hands-On ML)

Automation: Modern ML pipelines use tools like Apache Airflow, Prefect for orchestration.

This concept is fundamental to understanding Financial Data.

Handling Missing Data

Why Data is Missing:

1. **MCAR** (Missing Completely At Random): Pure chance, no pattern
2. **MAR** (Missing At Random): Related to observed data
3. **MNAR** (Missing Not At Random): Related to unobserved value itself

Finance Example:

- MCAR: Random system glitch
- MAR: Small firms don't report segment data
- MNAR: Firms hide bad performance

Strategies:

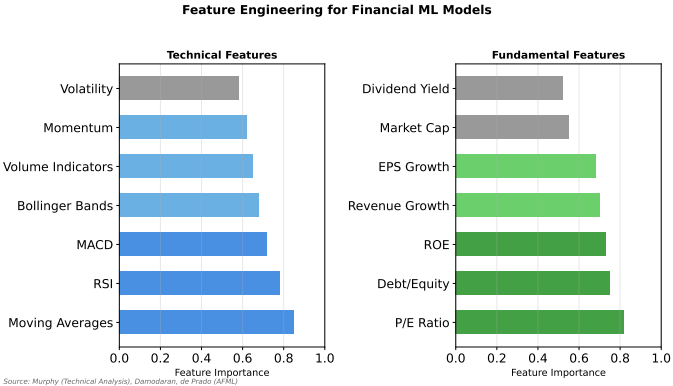
- **Deletion:** Drop rows/columns (only if ¡ 5% missing, MCAR)
- **Mean/Median imputation:** Replace with average (simple, biased)
- **Forward/backward fill:** Time series (assumes persistence)
- **Model-based:** Predict missing values (KNN, regression)
- **Indicator variable:** Flag missingness as feature

Best Practice:

- Understand WHY data is missing
- Test sensitivity to imputation method
- Document assumptions

This concept is fundamental to understanding Financial Data.

Feature Engineering Fundamentals



Art + Science: Combines domain expertise with systematic experimentation to create predictive features.

This concept is fundamental to understanding Financial Data.

Key Takeaways

1. Distinguish between structured and unstructured financial data
2. Identify major sources and vendors of financial data
3. Understand the alternative data revolution and its applications
4. Recognize data quality issues and preprocessing requirements

Bottom Line: Financial Data is transforming how financial services operate and compete.

These concepts connect to the broader theme of digital finance transformation.

Financial Data in Visual Perspective

Technology view

Application view

Future view

Visual representations help reinforce key concepts of financial data.

Concrete Examples: Making It Real

Technical Examples

- Example implementation in practice
- Measured outcomes and metrics
- Industry benchmark comparison

Case Study

- Real-world deployment scenario
- Quantifiable results achieved

Industry Leaders

- Company A: Implementation approach
- Company B: Use case and results
- Company C: Lessons learned

Market Data

- Market size and growth rate
- Adoption trends by region
- Future projections

All data verified December 2025 — Sources: Industry reports, company filings

Quiz Questions (1–5)

Q1. What is the primary purpose of financial data?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above

Quiz Questions (1–5)

Q1. What is the primary purpose of financial data?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above
Answer: D – All these factors contribute to the value proposition.

Q2. Which technology is most commonly associated with financial data?
A) APIs B) Blockchain C) Machine Learning D) Cloud Computing

Quiz Questions (1–5)

Q1. What is the primary purpose of financial data?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above
Answer: D – All these factors contribute to the value proposition.

Q2. Which technology is most commonly associated with financial data?
A) APIs B) Blockchain C) Machine Learning D) Cloud Computing
Answer: A – APIs enable integration and interoperability.

Q3. What is a key regulatory consideration for financial data?
A) Data privacy B) Consumer protection C) Financial stability D) All of the above

Quiz Questions (1–5)

Q1. What is the primary purpose of financial data?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above
Answer: D – All these factors contribute to the value proposition.

Q2. Which technology is most commonly associated with financial data?
A) APIs B) Blockchain C) Machine Learning D) Cloud Computing
Answer: A – APIs enable integration and interoperability.

Q3. What is a key regulatory consideration for financial data?
A) Data privacy B) Consumer protection C) Financial stability D) All of the above
Answer: D – All regulatory aspects must be considered.

Q4. Which industry sector benefits most from financial data?
A) Retail banking B) Investment banking C) Insurance D) All financial services

Quiz Questions (1–5)

Q1. What is the primary purpose of financial data?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above
Answer: D – All these factors contribute to the value proposition.

Q2. Which technology is most commonly associated with financial data?
A) APIs B) Blockchain C) Machine Learning D) Cloud Computing
Answer: A – APIs enable integration and interoperability.

Q3. What is a key regulatory consideration for financial data?
A) Data privacy B) Consumer protection C) Financial stability D) All of the above
Answer: D – All regulatory aspects must be considered.

Q4. Which industry sector benefits most from financial data?
A) Retail banking B) Investment banking C) Insurance D) All financial services
Answer: D – Benefits span across all financial services.

Q5. What is the main challenge in implementing financial data?
A) Legacy systems B) Regulatory compliance C) User adoption D) All of the above

Quiz Questions (1–5)

Q1. What is the primary purpose of financial data?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above
Answer: D – All these factors contribute to the value proposition.

Q2. Which technology is most commonly associated with financial data?
A) APIs B) Blockchain C) Machine Learning D) Cloud Computing
Answer: A – APIs enable integration and interoperability.

Q3. What is a key regulatory consideration for financial data?
A) Data privacy B) Consumer protection C) Financial stability D) All of the above
Answer: D – All regulatory aspects must be considered.

Q4. Which industry sector benefits most from financial data?
A) Retail banking B) Investment banking C) Insurance D) All financial services
Answer: D – Benefits span across all financial services.

Q5. What is the main challenge in implementing financial data?
A) Legacy systems B) Regulatory compliance C) User adoption D) All of the above
Answer: D – Multiple challenges must be addressed.

Quiz Questions (6–10)

Q6. How has financial data evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above

Quiz Questions (6–10)

Q6. How has financial data evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above
Answer: D – The evolution has involved multiple trends.

Q7. What metric best measures success in financial data?
A) User adoption B) Revenue growth C) Cost reduction D) All can be relevant

Quiz Questions (6–10)

Q6. How has financial data evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above
Answer: D – The evolution has involved multiple trends.

Q7. What metric best measures success in financial data?
A) User adoption B) Revenue growth C) Cost reduction D) All can be relevant
Answer: D – Success metrics depend on specific goals.

Q8. Which region leads in financial data adoption?
A) North America B) Europe C) Asia-Pacific D) Varies by segment

Quiz Questions (6–10)

Q6. How has financial data evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above
Answer: D – The evolution has involved multiple trends.

Q7. What metric best measures success in financial data?
A) User adoption B) Revenue growth C) Cost reduction D) All can be relevant
Answer: D – Success metrics depend on specific goals.

Q8. Which region leads in financial data adoption?
A) North America B) Europe C) Asia-Pacific D) Varies by segment
Answer: D – Leadership varies by specific market segment.

Q9. What is the future outlook for financial data?
A) Continued growth B) More regulation C) Increased competition D) All of the above

Quiz Questions (6–10)

Q6. How has financial data evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above
Answer: D – The evolution has involved multiple trends.

Q7. What metric best measures success in financial data?
A) User adoption B) Revenue growth C) Cost reduction D) All can be relevant
Answer: D – Success metrics depend on specific goals.

Q8. Which region leads in financial data adoption?
A) North America B) Europe C) Asia-Pacific D) Varies by segment
Answer: D – Leadership varies by specific market segment.

Q9. What is the future outlook for financial data?
A) Continued growth B) More regulation C) Increased competition D) All of the above
Answer: D – Multiple trends will shape the future.

Q10. What is a key takeaway about financial data?
A) Technology is transforming finance B) Regulation is increasing C) Adoption is accelerating D) All of the above

Quiz Questions (6–10)

Q6. How has financial data evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above
Answer: D – The evolution has involved multiple trends.

Q7. What metric best measures success in financial data?
A) User adoption B) Revenue growth C) Cost reduction D) All can be relevant
Answer: D – Success metrics depend on specific goals.

Q8. Which region leads in financial data adoption?
A) North America B) Europe C) Asia-Pacific D) Varies by segment
Answer: D – Leadership varies by specific market segment.

Q9. What is the future outlook for financial data?
A) Continued growth B) More regulation C) Increased competition D) All of the above
Answer: D – Multiple trends will shape the future.

Q10. What is a key takeaway about financial data?
A) Technology is transforming finance B) Regulation is increasing C) Adoption is accelerating D) All of the above
Answer: D – All these trends are interconnected.