Study Guide: Data Collection and Governance

DSC 208R - Data Management for Analytics

Overview

This study guide covers the sourcing stage of the data science lifecycle, focusing on data acquisition, reorganization, preparation, and labeling. It also addresses principles of data governance and privacy laws that every data scientist should understand.

1 Learning Objectives

By the end of this module, you should be able to:

- Understand the complete lifecycle of real-world data science projects
- Identify the time allocation challenges in data science work
- Explain the sourcing stage and its importance in the data science pipeline
- Describe the four main activities in the sourcing stage
- Recognize the challenges in data sourcing and preparation
- Understand the principles of data-centric AI
- Apply best practices in data governance and privacy

2 The Data Science Lifecycle

2.1 Key Components

1. Data acquisition

2. Data preparation
3. Data cleaning
4. Feature engineering
5. Model selection
6. Training & inference
7. Serving
8. Monitoring

2.2 Time Allocation in Data Science

Research Findings

Multiple industry surveys (CrowdFlower 2016, Kaggle 2018, IDC-Alteryx 2019) consistently show that data scientists spend the majority of their time on:

- Data collection
- Data cleaning
- Data organization

Rather than on model building and algorithm development.

3 The Sourcing Stage

3.1 Definition

The sourcing stage is where raw datasets are transformed into ”analytics/ML-ready” datasets. This stage ends when data is prepared for:

- SQL analytics for Business Intelligence
- Feature engineering for ML/AI analytics

3.2 Challenges in Data Sourcing

1. **Heterogeneity**: Diverse data modalities, file formats, and sources
2. **Access constraints**: Limited availability or permissions
3. **Application diversity**: Various prediction applications with different requirements
4. **Data volatility**: Unpredictable and continual edits to datasets
5. **Data quality issues**: Messy, incomplete, ambiguous, or erroneous data
6. **Scale**: Managing large volumes of data
7. **Governance**: Poor data management practices in organizations

4 Four Key Activities in the Sourcing Stage

Sourcing Process Flow

Raw Data Sources → Acquiring → Reorganizing → Cleaning → Data/Feature Engineering → Analytics Results
Note: Labeling & Amplification may be required in some cases

4.1 1. Data Acquisition

- Methods for obtaining data from various sources
- Understanding data access protocols and permissions
- Techniques for data extraction and collection

4.2 2. Data Reorganization

- Transforming data into usable formats
- Structuring unstructured or semi-structured data
- Normalizing data representations

4.3 3. Data Cleaning

- Identifying and handling missing values
- Detecting and correcting errors
- Removing duplicates and outliers
- Standardizing formats and units

4.4 4. Data/Feature Engineering

- Creating new features from existing data
- Transforming variables for better model performance
- Dimensionality reduction techniques
- Feature selection methods

5 Data-Centric AI

Data-Centric Approach

The Data-Centric AI movement emphasizes improving data quality rather than just model architecture. This approach recognizes that high-quality, well-prepared data is often more important than sophisticated algorithms.

5.1 Principles

- Focus on systematic data improvement
- Iterative data refinement
- Consistent data labeling
- Comprehensive data documentation

6 Data Governance and Privacy

6.1 Data Governance

- Policies for data management
- Data quality standards
- Data lifecycle management
- Roles and responsibilities

6.2 Privacy Considerations

- Relevant privacy laws and regulations
- Anonymization and pseudonymization techniques
- Consent management
- Data minimization principles

7 Study Questions

1. Why do data scientists spend more time on data preparation than on model building?

   Solution

   Data scientists spend more time on data preparation because real-world data is typically messy, incomplete, and not immediately usable for analysis. Industry surveys consistently show that 70-80% of a data scientist's time is spent on data collection, cleaning, and organization. This is necessary because high-quality data is fundamental to accurate models and insights. Even the most sophisticated algorithms will produce poor results with low-quality data, making thorough preparation an essential investment.

2. What are the main challenges in the sourcing stage of the data science lifecycle?

Solution

The main challenges include: (1) Data heterogeneity - dealing with diverse formats, structures, and sources; (2) Access constraints - limited permissions or availability; (3) Application diversity - different use cases requiring different data preparations; (4) Data volatility - constantly changing data; (5) Quality issues - missing values, errors, and inconsistencies; (6) Scale - managing large volumes of data efficiently; and (7) Governance issues - navigating organizational data management practices and policies.

3. How does heterogeneity of data sources affect the data preparation process?

Solution

Heterogeneity of data sources significantly complicates data preparation by requiring:

- Multiple data extraction methods for different source systems
- Various transformation techniques to standardize formats
- Complex integration processes to merge disparate data
- Additional validation steps to ensure consistency
- Custom handling for different data types (structured, semi-structured, unstructured)
- More extensive documentation to track data lineage

This heterogeneity increases the time and complexity of the preparation process and requires broader technical expertise.

4. Explain the relationship between data cleaning and feature engineering.

Solution

Data cleaning and feature engineering are sequential but interconnected processes:

- Data cleaning comes first and focuses on correcting errors, handling missing values, and ensuring data quality
- Feature engineering builds upon cleaned data to create meaningful variables for analysis
- Clean data is a prerequisite for effective feature engineering
- The boundary between them can blur, as some cleaning operations (like normalization) also serve as feature engineering
- Both processes require domain knowledge and understanding of the analytical goals
- Decisions made during cleaning (e.g., how to impute missing values) can impact subsequent feature engineering options

Together, they transform raw data into analysis-ready features that maximize model performance.

5. What is the data-centric AI movement, and why is it significant?

Solution

The data-centric AI movement, championed by Andrew Ng and others, shifts focus from model architecture to data quality. It's significant because:

- It recognizes that improving data quality often yields better results than refining algorithms
- It promotes systematic approaches to data improvement rather than ad-hoc fixes
- It encourages consistent labeling standards and documentation
- It addresses the reality that most AI projects fail due to data issues, not model limitations
- It provides a framework for iterative data refinement
- It helps organizations allocate resources more effectively by focusing on data quality

This movement represents a paradigm shift in how AI practitioners approach problem-solving, emphasizing the foundation (data) rather than just the structure built upon it (models).

6. How do data governance policies impact the work of data scientists?

Solution

Data governance policies impact data scientists by:

- Defining what data can be accessed and for what purposes
- Establishing protocols for data sharing and collaboration
- Setting standards for data quality and documentation
- Creating frameworks for data security and privacy compliance
- Determining data retention periods and archiving requirements
- Establishing roles and responsibilities in the data lifecycle
- Providing processes for resolving data-related issues

While these policies may initially seem to constrain data scientists, they ultimately enable more reliable, ethical, and sustainable data science work by ensuring data integrity and appropriate use.

7. What are the four main activities in the sourcing stage, and how do they relate to each other?

Solution

The four main activities in the sourcing stage are:

(a) **Data Acquisition:** Obtaining data from various sources

(b) **Data Reorganization:** Transforming data into usable formats

(c) **Data Cleaning:** Correcting errors and handling missing values

(d) **Data/Feature Engineering:** Creating meaningful features for analysis

These activities form a sequential but iterative process. Acquisition provides the raw material, reorganization structures it into a workable format, cleaning improves its quality, and feature engineering transforms it into analytically useful variables. Each stage depends on the previous one, but discoveries in later stages often necessitate revisiting earlier steps (e.g., finding data quality issues during cleaning might require returning to acquisition for additional data).

8. Why is data labeling sometimes necessary, and what challenges does it present?

Solution

Data labeling is necessary for supervised machine learning tasks where algorithms learn from labeled examples. It presents several challenges:

- **Resource intensity:** Labeling often requires significant human effort and time
- **Consistency issues:** Different labelers may interpret guidelines differently
- **Subjectivity:** Some tasks involve inherently subjective judgments
- **Expertise requirements:** Specialized domains may require expert knowledge
- **Scale limitations:** Large datasets may be impractical to label manually
- **Bias introduction:** Labelers may inadvertently introduce their biases
- **Evolving contexts:** Labels may need updates as real-world contexts change

Organizations address these challenges through clear guidelines, quality control processes, semi-automated approaches, and active learning techniques that prioritize which items to label.

8 Additional Resources

- Data-Centric AI Movement
- CrowdFlower Data Science Report 2016
- Kaggle State of ML and Data Science Survey 2018
- IDC-Alteryx State of Data Science and Analytics Report 2019

Key Takeaway

The sourcing stage is the foundation of successful data science projects. Mastering the skills of data acquisition, reorganization, cleaning, and engineering is essential for any data scientist, as these activities consume the majority of time in real-world projects.