

# Data Collection and Label- ing

CS 203: Software Tools and Techniques for AI

# Module Overview

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

## Four Core Components

# Part 1: Data Collection

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |



# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |

# Outline

|                                   |    |
|-----------------------------------|----|
| Four Core Components .....        | 3  |
| Why Data Collection Matters ..... | 5  |
| Common Data Sources .....         | 6  |
| Instrumentation .....             | 7  |
| Basic Instrumentation .....       | 8  |
| Web Analytics Tools .....         | 9  |
| The Cost of Bad Data .....        | 10 |
| Data Quality Issues .....         | 11 |
| Validation with Pydantic .....    | 12 |
| What is Data Labeling? .....      | 13 |
| Types of Labeling Tasks .....     | 14 |
| Label Studio .....                | 15 |
| Inter-Annotator Agreement .....   | 16 |