Data Collection and Labeling

CS 203: Software Tools and Techniques for AI

Prof. Nipun Batra

Module Overview

Four Core Components

1. **Data Collection** - Tools and techniques for gathering data from diverse sources

Key Insight

Quality data is the foundation of successful AI systems. Garbage in, garbage out!

Module Overview

Four Core Components

1. **Data Collection** - Tools and techniques for gathering data from diverse sources
2. **Data Validation** - Ensuring data quality and reliability

Key Insight

Quality data is the foundation of successful AI systems. Garbage in, garbage out!

Module Overview

Four Core Components

1. **Data Collection** - Tools and techniques for gathering data from diverse sources
2. **Data Validation** - Ensuring data quality and reliability
3. **Data Labeling** - Annotating datasets with ground truth

Key Insight

Quality data is the foundation of successful AI systems. Garbage in, garbage out!

Module Overview

Four Core Components

1. **Data Collection** - Tools and techniques for gathering data from diverse sources
2. **Data Validation** - Ensuring data quality and reliability
3. **Data Labeling** - Annotating datasets with ground truth
4. **Data Augmentation** - Expanding datasets strategically

Key Insight

Quality data is the foundation of successful AI systems. Garbage in, garbage out!

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time → models need fresh data

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time → models need fresh data
- **Production systems** require automated collection pipelines

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 1: Data Collection

Instrumenting and Logging Data Sources

Why Data Collection Matters

- **Real-world AI systems** depend on continuous data flow
- **User behavior** changes over time $\rightarrow$ models need fresh data
- **Production systems** require automated collection pipelines
- **Debugging** often requires understanding what data was seen

Common Data Sources

Digital Sources

- Web applications
- Mobile apps
- IoT devices
- APIs and databases

Physical Sources

- Sensors
- Cameras
- Microphones
- Manual entry

Part 2: Data Validation

Ensuring Data Quality and Reliability

Why Data Validation Matters

The Cost of Bad Data

- **Garbage In, Garbage Out**
- **Silent Failures**
- **Expensive Debugging**
- **Lost Trust**

Real-World Impact

E-commerce: Missing IDs → 30% drop in CTR **Medical:** Wrong units → Misdiagnoses **Fraud Detection:** Duplicates → 45% false positives

Types of Data Quality Issues

1. **Completeness** - Missing or null values
2. **Accuracy** - Incorrect or outdated values
3. **Consistency** - Contradictory data
4. **Timeliness** - Stale or outdated data

Part 3: Data Labeling

Annotating Datasets with Ground Truth

What is Data Labeling?

Data Labeling: Adding meaningful tags or annotations to raw data

Why It's Critical

- **Supervised Learning** requires labeled examples

Types of Labeling Tasks

Classification

Sequence Labeling

Part 3: Data Labeling

Annotating Datasets with Ground Truth

What is Data Labeling?

Data Labeling: Adding meaningful tags or annotations to raw data

Why It's Critical

- **Supervised Learning** requires labeled examples
- **Quality labels** $\rightarrow$ Better models

Types of Labeling Tasks

Classification

Sequence Labeling

Part 3: Data Labeling

Annotating Datasets with Ground Truth

What is Data Labeling?

Data Labeling: Adding meaningful tags or annotations to raw data

Why It's Critical

- **Supervised Learning** requires labeled examples
- **Quality labels** $\rightarrow$ Better models
- **Consistent labels** $\rightarrow$ Reliable evaluation

Types of Labeling Tasks

Classification

Sequence Labeling

Part 3: Data Labeling

Annotating Datasets with Ground Truth

What is Data Labeling?

Data Labeling: Adding meaningful tags or annotations to raw data

Why It's Critical

- **Supervised Learning** requires labeled examples
- **Quality labels** $\rightarrow$ Better models
- **Consistent labels** $\rightarrow$ Reliable evaluation
- **Cost:** Often 60-80% of ML project time and budget

Types of Labeling Tasks

Classification

Sequence Labeling

Part 4: Data Augmentation

Expanding Training Datasets Strategically

Why Data Augmentation?

The Problem

- Limited labeled data is expensive
- Class imbalance leads to biased models
- Overfitting on small datasets
- Rare events underrepresented

The Solution

Data Augmentation: Creating new training examples by applying transformations
Benefits: Increase dataset size, Improve generalization, Reduce overfitting, Balance classes

Image Data Augmentation

Summary

Key Takeaways

Data Collection

- Instrument systems for automatic capture
- Use appropriate tools (analytics, APM, CDC)
- Implement buffering, batching, error handling
- Respect privacy and comply with regulations

Data Validation

- Validate early and often
- Use schema validation (Pydantic, Pandera)
- Monitor data quality metrics
- Set up alerts for anomalies

Key Takeaways (cont.)

Questions?

Thank You!

Next: Reproducibility & Versioning

Contact: nipun.batra@iitgn.ac.in