



Business School
UNIVERSITY OF COLORADO **DENVER**

Information Systems Program

Unit 4

Data Integration Concepts, Processes, and Techniques

Lesson 3: Data Cleaning Tasks



Lesson Objectives

- Explain the three types of data cleaning tasks
- Provide examples depicting data cleaning tasks
- Reflect on the tedious nature of data cleaning



Parsing

- Locates and separates individual data elements in text
- Studied in computer science for decades
- Regular expressions for pattern specification
- Natural language processing for context-sensitive parsing



Parsing Example

Raw input in source file

Aimee Christina Parker, Prod. Mgr.
Microsoft
One Microsoft Way
Redmond, WA



Parsed data in target file

First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Prod. Mgr.
Firm	Microsoft
Street	One Microsoft Way
City	Redmond
State	WA



Correcting Values

- Missing values
 - Default value for inapplicable values
 - Typical value: average, median, or mode
 - Complex processing for predicting values using relationships to other fields
- Conflicting values
 - More recent value
 - More credible source



Correction Example

Parsed data

First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Prod. Mgr.
Firm	Microsoft
Street	One Microsoft Way
City	Redmond
State	WA



Corrected data

First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Prod. Mgr.
Firm	Microsoft
Street	15580 NE 31st St.
City	Redmond
State	WA
Postal Code	98052
Country	USA

6



Standardization

- Applies conversion routines to transform data into preferred formats
- Uses both standard and custom business rules
- Common standardizations:
 - Unit of measure transformations
 - Standard abbreviations (state names, titles, street types)



Standardization Example

Corrected data

First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Prod. Mgr.
Firm	Microsoft
Street	15580 NE 31st St.
City	Redmond
State	Washington
Postal Code	98052
Country	USA



Standardized data

First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Product Manager
Firm	Microsoft Corporation
Street	15580 NE 31st Street
City	Redmond
State	WA
Postal Code	98052
Country	USA

Summary

- Common data cleaning tasks
- Many approaches for missing values
- Organization specific standardization rules

