COMP 499 Introduction to Data Analytics

Lecture 5 — Data Wrangling

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

Overview of Lecture

1. Data Wrangling Overview
   - Discover
   - Structure
   - Cleanse
   - Enrich
   - Validate
   - Publish
2. Data Cleaning — Professor Skiena lecture

Data Wrangling — Discovery

Discover what data is available

Extract step of ETL

Data Wrangling — Structure

Organize data into suitable format

Transform step of ETL

Data Wrangling — Cleanse

Clean the data

Iterative step with basic data analysis

Data Wrangling — Enrich

Discover and include related data

Integrate new data sets and data types

add more data fields

Data Wrangling — Validate

Check data is consistent and complete

Consistency

Does your data fit into expected values for it?
Do field values match the data type for the column?
Are values within acceptable ranges?
Are rows unique? Duplicated?

Completeness

Are all expected values included in your data?
Are some fields missing values?
Are there expected values that are not present in the dataset?

Test routines for your data wrangling process

Data Wrangling — Publish

Make available for analysis

Load step of ETL

into data warehouse in traditional business intelligence setting