COMP 499 Introduction to Data Analytics

Lecture 1 — Context

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. Context
   - Data to Application Data
   - Data Levels of Interpretation
   - Lexical to Pragmatics
   - Approaches to Data Analytics

Context — Data to Knowledge

Data

raw, calibrated, normalized, validated
derived, aggregated, interpreted
Metadata describes source and properties of the data

Information

newsworthy
actionable
Claude Shannon's information theory

Knowledge

applicable wisdom, organized information
concepts, relations, constraints, taxonomy/ontology
axioms, rules, plans

Application — aka Knowledge Translation

Context — Data Level of Interpretation

Raw Data

raw values obtained directly from the measurement device

Calibrated Data

raw physical values, corrected with calibration operators

Validated Data

calibrated data that has been filtered through quality assurance procedures
(most commonly used data for scientific purposes)

Derived Data

frequently aggregated data, such as gridded or averaged data

Interpreted Data

derived data that is related to other data sets, or to the literature of the field

Context — Syntax to Pragmatics

Lexical = atomic units

Defined by regular expressions
Represented as enum's

Syntax = structure

Defined by grammars
Represented as Abstract Syntax Trees (AST)

Semantics = meaning

Defined by interpretation mappings
Represented as actions (procedural) in compiling

Pragmatics = goals

Context — Approaches to Data Analysis

Scripting

Unix tools, eg

text files, csv files for inputs, outputs, intermediate steps

stepwise development of analysis

script captures steps, parameters

easy to replay

Notebooks

Jupyter, eg

interactive scripting with "literate programming"

keep track of thought processes during analysis

work with files to replay analysis

"Spreadsheet" Environments

OpenRefine, eg

lots of tools, little guidance

need macros, histories, to capture/replay work

often proprietary