Australian National University

COMP3430 / 8430
Data wrangling

Lecture 1: Overview and course introduction
(Lecturer: Peter Christen)

Lecture outline

- What is data wrangling
- Data wrangling tasks and outcomes
- Examples of data wrangling

- Course topics overview

- Example data wrangling using Rattle

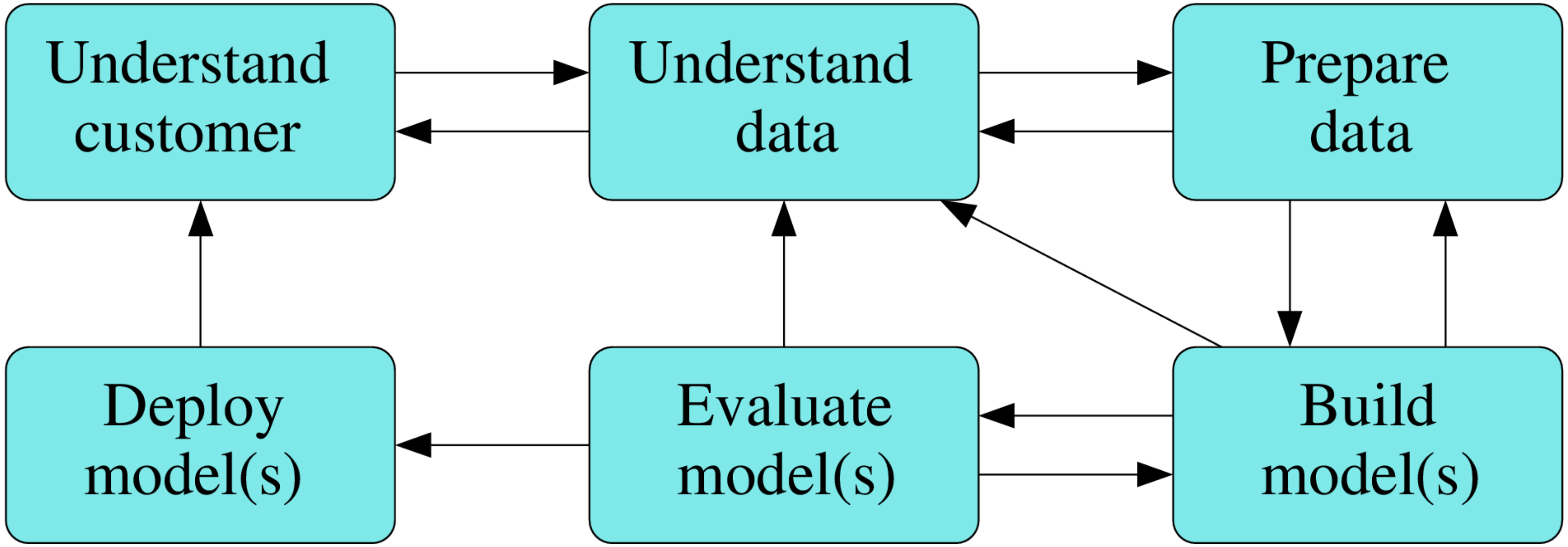
What is data wrangling?

In a recent article, the New York Times (2014) writes: “*Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in the mundane labor of collecting and preparing data, before it can be explored for useful information.*”

What is data wrangling?

- It is the (manual) process of converting raw data from various sources into another format to allow more convenient further processing by using a variety of tools and techniques
- Also known as data munching, data cleansing, data preparation, data cleaning, data janitoring, extraction / transformation / loading (within data warehousing), etc.

The data mining / analytics process



Typically up to 90% of time and effort are spent in the first three steps!

- (based on: CRISP-DM, the *CRoss Industry Standard Process for Data Mining*)

Main data wrangling tasks

- Data extraction
- Data quality assessment
- Data profiling, exploration, summarisation, and visualisation
- Data cleaning (transformation, reshaping, aggregation, reduction, imputation, parsing, standardisation)
- Data integration (schema matching and mapping, data matching, record linkage, deduplication, data fusion)

Examples of data wrangling tasks (1)

Given a dataset containing health care treatment outcomes:

- Calculate appropriate summary statistics and measures to assess the quality of the data set, and the types of questions it is suitable for answering
- Identify anomalies and problems with the data that require attention, as well as possible solutions
- Generate appropriate visualisations to convey the key aspects and limitations of the data set

Examples of data wrangling tasks (2)

- Say we are interested in the effect of student health on education outcomes
- Given eight separate data sets on education outcomes (one from each state and territory), as well as the health data from example 1:

1) Integrate the separate education data sets in different formats and standards into a single cohesive data set that is suitable for analysis
2) Detect individuals who move between states and make sure they don't show up twice (or more)
3) Combine this data with the health data from example (1), even though most of the individuals are too young to have a unique identifier (either with Medicare, ATO or Centrelink)

Outcomes of data wrangling tasks

- Understanding and characterisation of the quality aspects of a given data set or database
- Cleaned, standardised, consistent and integrated data in a format that is suitable for the further processing and/or analytics tasks at hand
- Documentation of the data quality assessment, profiling, exploration and cleaning conducted

Course topics overview

- The data wrangling process
- Data quality: assessment and dimensions
- Data exploration and profiling
- Data cleaning and pre-processing
- Data integration: schema matching and mapping; record linkage (data matching); and data fusion
- Advanced topics: Wrangling dynamic data and streams

Next

- Recording 2: An example of a practical data wrangling exercise using the Rattle tool

- Administrative course issues will be discussed in the first in-person lecture