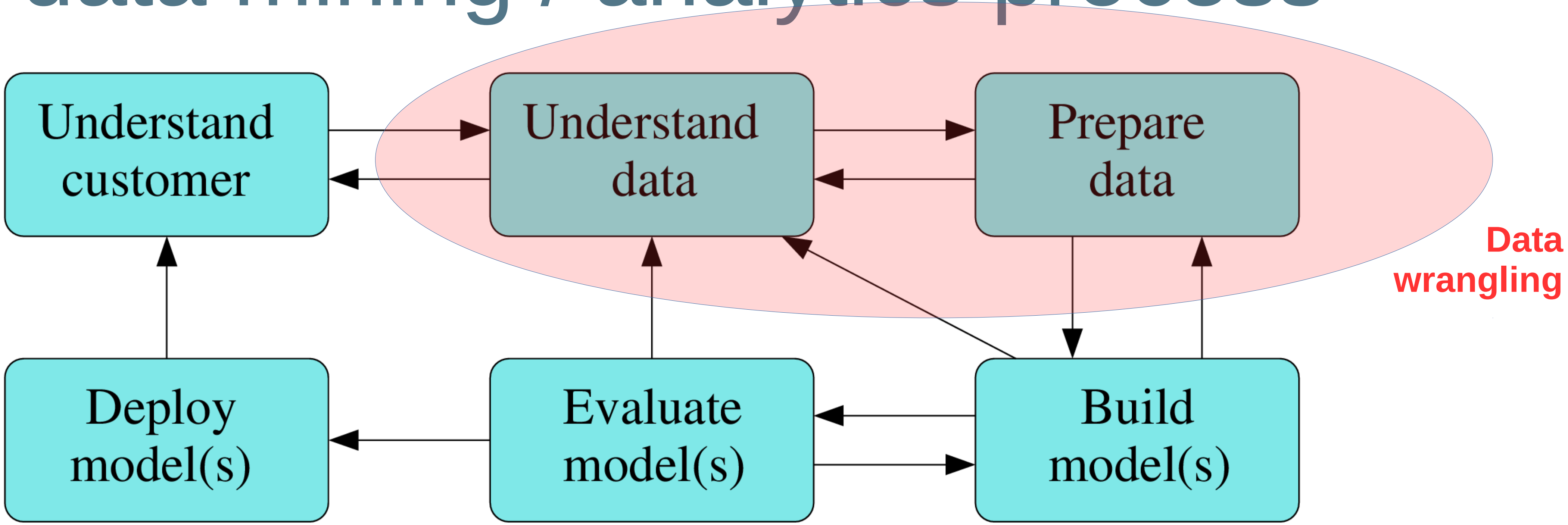
# COMP3430 / COMP8430 Data wrangling

## Lecture 2: The data wrangling process and understanding data (Lecturer: Peter Christen)

# Lecture outline

- The data wrangling process / pipeline / tasks
- Understanding data: sources, types, and formats
- Example data wrangling tools and resources

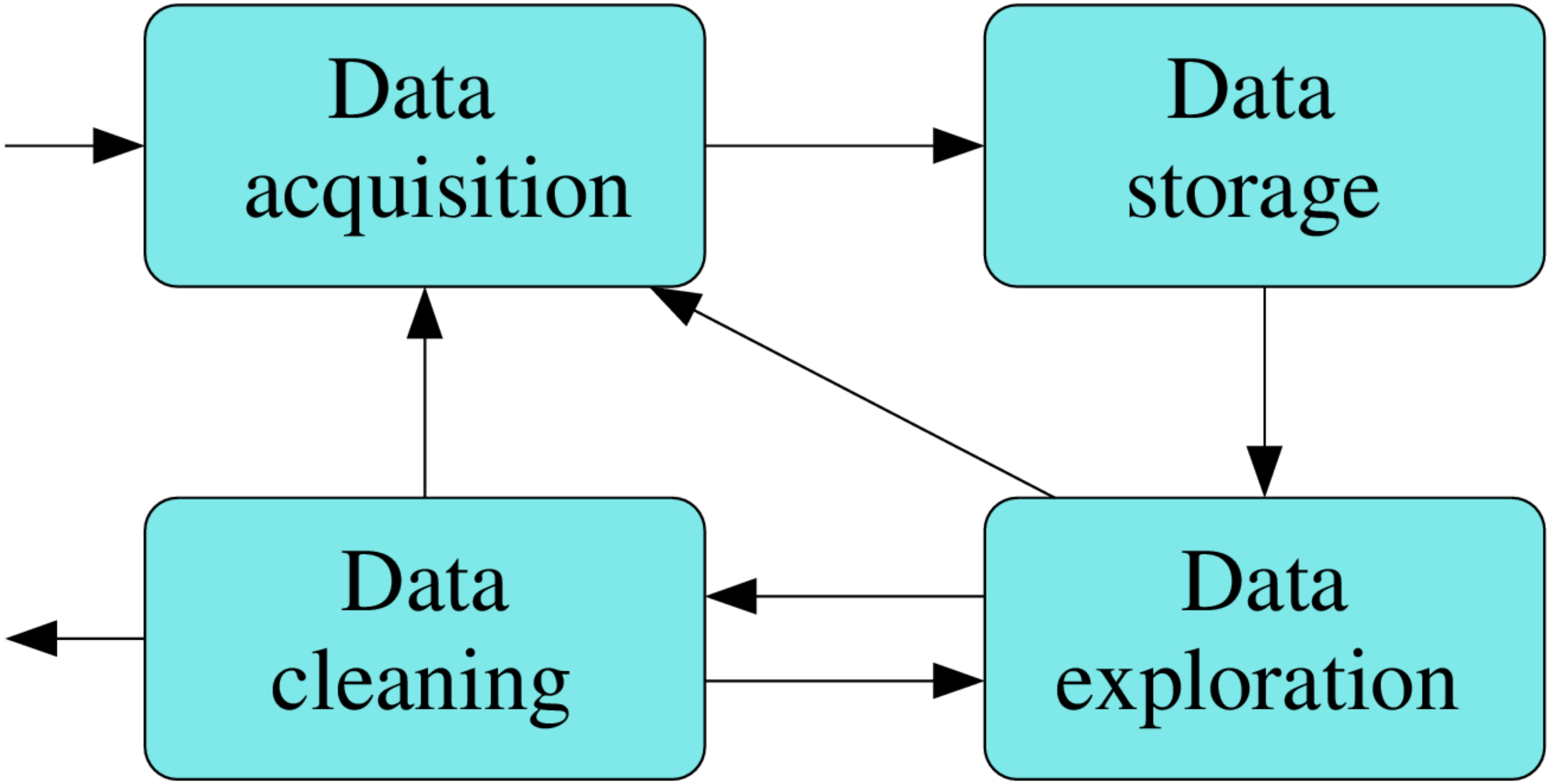
# The data mining / analytics process



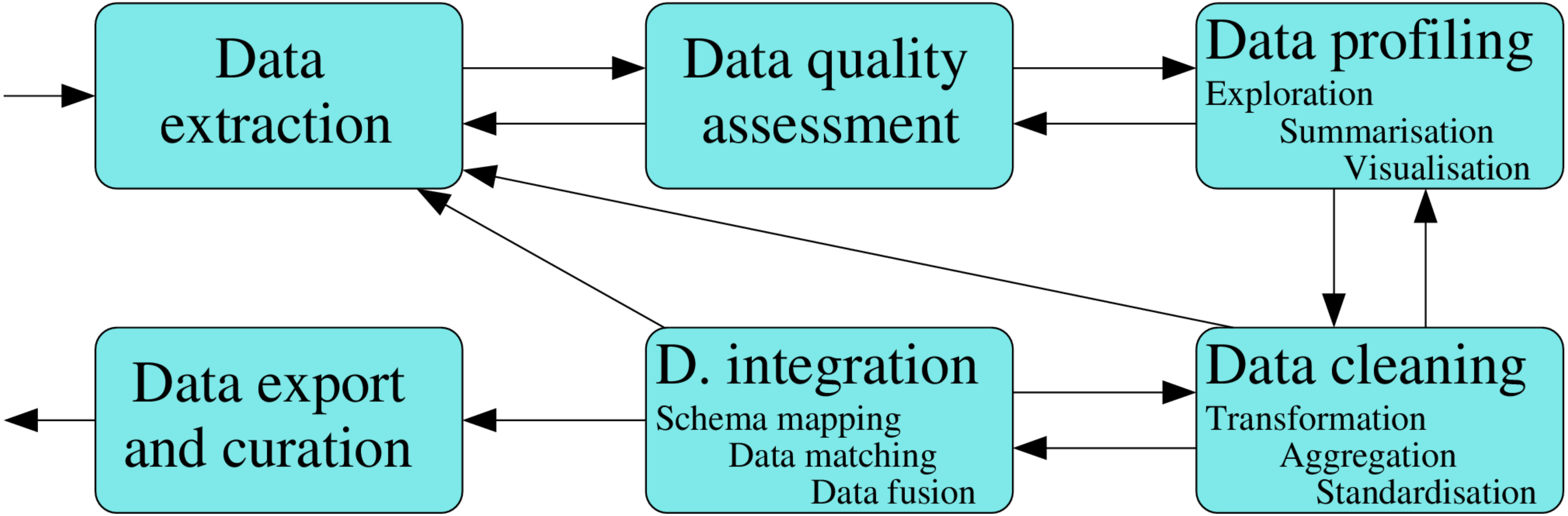
Typically up to 90% of time and effort are spent in the first three steps!

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

# The data wrangling process (1)



# The data wrangling process (2)



# Main data wrangling tasks

- **Data extraction**: From different sources, both internally and increasingly externally to an organisation
- **Data quality assessment**: Along a variety of dimensions
- **Data profiling**: Exploration, summarisation, and visualisation to better understand data
- **Data cleaning**: Transformation, reshaping, aggregation, reduction, imputation, parsing, standardisation
- **Data integration**: Schema matching and mapping, data matching, record linkage, deduplication, data fusion

# Understanding data

## What is data?

- Data is how we store observations in reusable form
- Observations are about entities and their attributes, as well as relationships between entities
- Sometimes (ideally) entities have unique identifiers (products have barcodes, most Australians have a Tax File Number (TFN) or a Medicare number, books have ISBNs, etc.)
- Unique entity identifiers should be **stable over time**, **accurate**, **complete**, and **robust** (like a checksum in an identifier number)

# Sources of data (1)

- Relational databases
  - Transactional data, mostly normalised into many tables, with keys between them, continuous and frequent updates on (single) records
- Data warehouses
  - Decision support data, processed and cleaned, historical data, aggregated, updated at certain intervals
- Internet
  - Click-stream data, log files, Web pages (HTML, XML), blogs, e-mails, posts, images, videos, audio, etc.

# Sources of data (2)

- Files
  - Portable text (like comma separated, tabulator, fixed column) or non-portable proprietary binary files
- Scientific instruments, experiments and simulations
  - Astronomy, genomics, seismology, physics, chemistry, etc.
- Sensors (often data streams)
  - Internet of Things (IoT)

# Data size and complexity

- “*We are drowning in data but starving of knowledge*”
  (Jiawei Han, author of the *Data Mining* text book)

- Automated data collection and mature database technology
  - Allows data to be stored efficiently, cheap, persistent
  - Using databases, data warehouses and other repositories
  - Data are increasingly stored distributed (storage area networks, grids, etc.)
- Large and massive data collections
  - Millions to billions of records
  - Tens to thousands of attributes (sometimes also called *variables*)
  - **Data are rarely collected for data analytics** (rather for online transaction processing, OLTP)

- A lot of data are *write only* (or *read once only*)

# Types and measurements of data (1)

- Numerical data
  - Integer, floating-point, binary, interval, ratio
  - Non-scalar (like velocity: speed and direction)
- Non-numerical data
  - Nominal data (just naming things, for example personal names)
  - Categorical data (grouping things, like postcodes, university course codes)
  - Ordinal data (ordering things, for example wine tasting, movie ratings)
- Series data
  - Ordering is an important feature (otherwise not series data)
  - One attribute must always be monotonic (increasing or decreasing)
  - Most common are *time series*

# Types and measurements of data (2)

- Multimedia data
  - Images, video, audio
  - Many standard formats used, binary, often compressed
- Different mappings and conversions between data types are possible and often needed
  - Some conversions are loss-less, others are lossy
- Different data wrangling (and data analytics) techniques can handle different types of data
  - Some are restricted to certain types of data, for example only numerical data

# Formats of data

- Structured data
  - Relational database tables, integrated data warehouses
  - Images, video, audio (can be compressed)
- Semi-structured data
  - XML, HTML, e-mails, SMS, log files
- Free-format data
  - Mainly free-format text - ASCII or Unicode

# Data wrangling tools and resources (1)

- Data wrangling books (mostly specific to a certain language or tool)
  - **Data Wrangling with Python**; Jacqueline Kazil and Katharine Jarmul, O'Reilly Media, 2016
  - **Python for Data Analysis**; Wes McKinney, O'Reilly Media, 2012 (second edition now available)
  - **Data Science from Scratch - First Principles with Python**; Joel Grus, O'Reilly Media, 2015
  - **Data Wrangling with R**; Bradley Boehmke, Springer, 2016
  - **R for Data Science - Import, Tidy, Transform, Visualize, and Model Data**; Garrett Grolemund and Hadley Wickham, O'Reilly Media, 2017
- Some of these can be found as PDF files for download

# Data wrangling tools and resources (2)

- Programming tools (mostly specific to a certain language or tool)
  - **Pandas** (Python): http://pandas.pydata.org/
    A library that allows efficient data structure and data manipulation and analysis tools, including visualisation (we will show Pandas examples throughout the course)
  - **Matplotlib** (Python) http://matplotlib.org
    A comprehensive 2D plotting library to produce high quality outputs as well as interactive environments
  - **Dplyr** ( R )
    https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html
    Summarise and transform data in rows and columns
- Many more modules / packages relevant to data wrangling

# Data wrangling tools and resources (3)

- Software
  - **Rattle** ( R ): http://rattle.togaware.com/
    A graphical user interface (GUI) on top of R, includes extensive data exploration, visualisation and transformation operations, developed by Graham Williams (previously Senior Data Miner at ATO), used in this course
  - **DataWrangler** (now TrifactaWrangler) https://www.trifacta.com/
    An interactive tool for data cleaning and transformation, developed by a Stanford/Berkeley Wrangler research project, now commercial
  - See also: https://blog.varonis.com/free-data-wrangling-tools/
- Many database and data warehouse systems do include some data wrangling functionalities