Australian National University

COMP3430 / COMP8430
Data wrangling

Lecture 25: Wrangling dynamic and spatial data

(Lecturer: Peter Christen)

Lecture outline

- Big Data and its characteristics
- Data wrangling dynamic data and data streams
- Data wrangling location (spatial) data

Big Data characteristics (1)

- We now live in the era of Big Data
  - Massive amounts of data are being collected by many organisations (both in the private and public sectors), as well as by individuals
  - Many companies and services move online, generating constant streams of (transactional) data
  - The Internet of Things (IoT) is resulting in billions of sensor being connected, producing vasts amount of data
- Temporal, dynamic and streaming data pose novel challenges for both data wrangling (as well as data mining)
  - Not much is published about these topics (ongoing research in some areas), so the following slides are mostly discussion points

Big Data characteristics (2)

- The four V's that characterise Big Data
  - **Volume** (the size of data): Data are too large to be stored for later processing or analysis
  - **Variety** (different forms of data): Different and novel processing, exploration, cleaning, and integration approaches are required
  - **Velocity** (streaming data): Data are being generated on a constant basis (data streams), and require ongoing (near) real-time processing, wrangling and analysis
  - **Veracity** (uncertainty of data): Data quality can change over time, will be of different quality from different sources, will be difficult to ascertain quality of all data due to huge volume

Wrangling dynamic data (1)

- Dynamic data are when values of known entities change over time
- Dynamic data occur in many domains
  - Name and address changes of people
  - Price changes for consumer products
- New records (with changed attribute values) for a known entity should be verified for consistency
  - Gender and/or age of a person to be consistent with previous values of the same person
  - Updated price of a product is sensible (for example, same camera doesn't suddenly cost 10 times as much)
- Overall, databases should be consistent
  - We cannot suddenly have 5,000 people living at the same address

Wrangling dynamic data (2)

- The temporal dimension makes wrangling of dynamic challenging
- For data exploration, distributions and ranges might change, new values might occur
  - Such as new postcodes for new suburbs previously not seen
  - Changes in name popularity, or number of people living in a postcode area
- Difficult to decide if a new value is an outlier and maybe wrong, or a new valid value
- Difficult to decide what the valid minimum and maximum values should be
  - Changes in external factors (new policies or regulations) will affect data types and/or values

Wrangling dynamic data (3)

- Data cleaning is also challenged by dynamically changing data
  - What was normal once might become unusual later, and might need to be cleaned (such as addresses of residencies replaced by businesses)
  - Difficult to apply transformation techniques such as binning and histograms, as ranges and distributions change over time
  - Difficult to normalise data if min-max values change (for 0-1 normalisation), or mean and standard deviation change (for z-score normalisation)
  - Normalised data in the past might become skewed if distributions change over time
  - Missingness of data likely changes as data collection processes change (new Web input forms, new automatic collection equipment, etc.)

Wrangling dynamic data (4)

- Due to challenge of data cleaning, data integration will also be challenged
- Schema matching: Database schemas can change over time
  - New or refined attributes, split attributes, new database tables, etc.
- Record linkage: Attribute values (and maybe their types and structures) likely change over time
  - Such as names, addresses and contact numbers of people or businesses
- Requirements of how data are to be fused will change over time
  - Depending upon record linkage results, as well as of the applications that require the fused data

Wrangling data streams

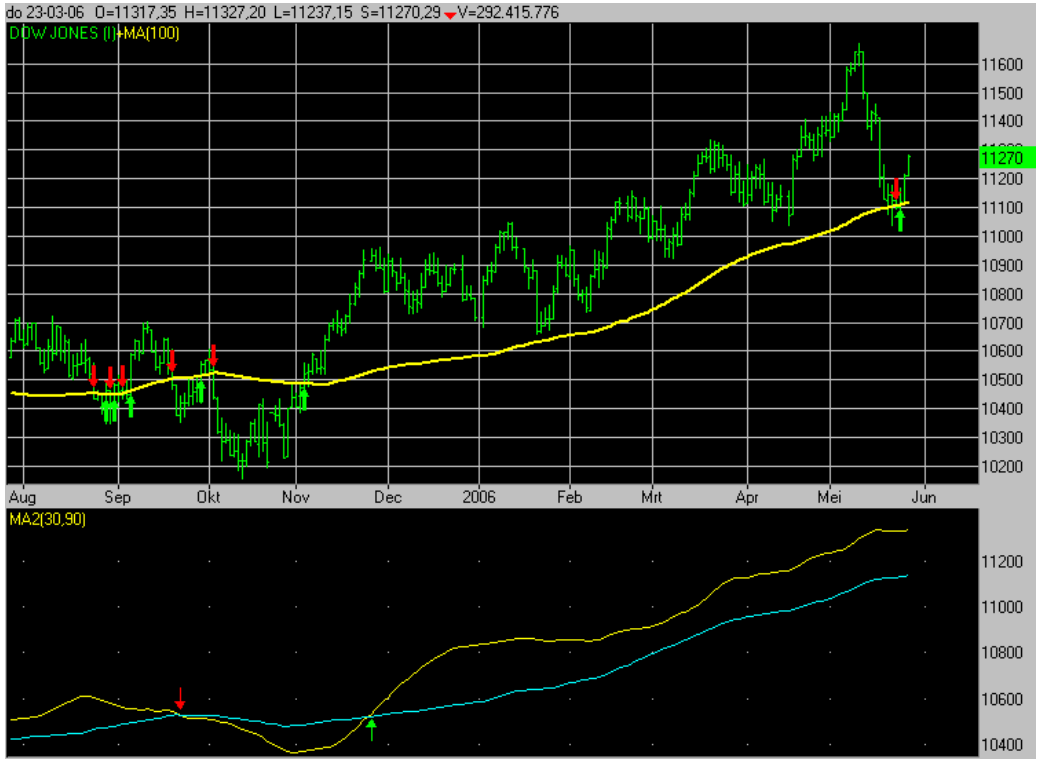
- Data streams are characterised by rapid arrival of individual (transactional) records, but unpredictable arrival rate
- Data volume is often too large to store and process off-line
  - Applications often require (near) real-time processing and analysis
  - Each record can be accessed (inspected and modified) only once
  - Examples are online financial transactions such as credit card purchases
- Data exploration and cleaning need to be conducted in real-time
  - So all the challenges of dynamic data, plus the additional requirement of processing in real-time

Example technique: Sliding window (1)

- Given a data stream with unpredictable arrival rate
  - Unknown in advance how many records arrive in a certain time interval
- Assume we want to monitor:
  - Minimum, averages, medians, and maximums of numerical attributes
  - Number of unique values and percentage of missing values for categorical attributes
- We keep in memory the last *x* records, and calculate the above measures on these *x* records
  - When a new record arrives, the oldest record will be removed
  - We can visualise these calculated values as moving (dynamic) graphs

Example technique: Sliding window (2)

- Such an approach smoothes the input data (“moving averages”)
  - Different weights can be given to more recent compared to older data
  - Irregular changes (such as outliers and missing or inaccurate values) are smoothed out
  - Trends are clearly visible



Data wrangling location (spatial) data (1)

- Location data are increasingly being generated and used in various application domains
  - Location data are automatically generated by smartphones, cars, etc.
  - Generally contain location and time when the location was recorded (as individual location points or trajectories)
  - Applications range from mapping apps, spatial data analysis, location specific advertisements, all the way to fitness and dating applications
- Allow all kinds of useful applications, but challenging due to privacy concerns
  - Once location data are integrated with other geographical or personal data
  - For example, identify where celebrities or rich individuals live by tracking routes from prestigious night clubs to end points of taxi / Uber rides

Data wrangling location (spatial) data (2)

- Location data is dynamic and often requires real-time processing and analysis
  - As we have seen before, this is challenging
  - Location data might be incomplete and/or inaccurate (GPS turned off, people inside buildings, vehicles in tunnels, etc.)
  - It might not be possible to integrate location data legally
  - By themselves, location data is of limited use (i.e. if context is not known), only traffic densities are available (such as number of people using public transport at a certain station at what time)
- Anonymising location data is challenging
  - Only a few data points make a taxi trip unique, once integrated with other data re-identification is often possible

Data wrangling location (spatial) data (3)

- Location data provide new ways of exploring and cleaning data
  - Location data can be visualised on maps, allowing for interactive human inspection (which can provide immediate feedback on data quality – for example too many individuals located at a remote train station points to either an event or equipment malfunction)
  - Automatic verification of validity of locations and trajectories is possible based on domain knowledge (such as maximum speed of a taxi in a city will highlight trajectories that are not possible; or a smartphone cannot be in two locations at the same time)
  - Missing locations can potentially be inferred (taxi trajectory from A to B must go via either C or D – if for example A and B are on different sides of a river and C and D are bridges)