Make Datasets

```
library(tidyverse)
library(lubridate)
```

Define where to get the data from and where to put the resulting data sets:

```
data_source_dir = "../data_raw/"
data_destination_dir = "../datasets/"
```

Define helper functions that get information from a filename:

```
get_provider_by_filename = function(filename) {
  split_result = str_split(filename, "_")
  return(split_result[[1]][2])
}

# example:
get_provider_by_filename("1544519617_vodafone_ul.txt")
```

```
## [1] "vodafone"
```

```
get_datatype_by_filename = function(filename) {
  split_result = str_split(filename, "_")
  ending = split_result[[1]][3]
  split_result_ending = str_split(ending, "\\.")
  return(split_result_ending[[1]][1])
}

# example:
get_datatype_by_filename("1544519617_vodafone_ul.txt")
```

```
## [1] "ul"
```

This function creates a dataset for a given `datatype` like “ul”, “dl”, “context” or cells. The provider is added as an extra column as well as the scenario.

```
make_dataset = function(data_location, datatype) {

  # read all datasets and store them in a list to combine them later
  datasets = list()

  scenarios = c("urban", "suburban", "campus", "highway")
  for (cur_scenario in scenarios) {

    # get the current path where the files are located and list the files
    cur_path = str_c(data_location, "/", cur_scenario)
    data_files = list.files(cur_path)

    # now read each file
    for (cur_filename in data_files) {
```

```
      # only read when the datatype matches the one we want
      cur_datatype = get_datatype_by_filename(cur_filename)
      if (cur_datatype != datatype) {
        next
      }

      # read the file and add a column for the provider and for the scenario
      cur_provider = get_provider_by_filename(cur_filename)
      cur_dataset = read_csv(str_c(cur_path, "/", cur_filename), col_type=cols()) %>%
        mutate(scenario=cur_scenario, provider=cur_provider)

      # attach it to the list
      datasets[[length(datasets)+1]] = cur_dataset
    }
  }

  # build the final dataset, convert the seconds to proper dates and sort by time
  final_dataset = bind_rows(datasets) %>%
    mutate(timestamp=as_datetime(timestamp_ms)) %>%
    arrange(timestamp)
  return(final_dataset)
}
```

Now create the data sets:

```
dataset_ul = make_dataset(data_source_dir, datatype="ul")
glimpse(dataset_ul)
```

```
## Rows: 6,180
## Columns: 27
## $ time_s            <dbl> 10.49, 10.19, 10.60, 20.20, 21.71, 30.83, 30.20, ...
## $ timestamp_ms      <dbl> 1544432937, 1544432937, 1544432943, 1544432947, 1...
## $ distance_m        <dbl> 100.93, 96.30, 88.76, 216.95, 233.54, 316.60, 308...
## $ latitude          <dbl> 51.49052, 51.49055, 51.49053, 51.49071, 51.49070,...
## $ longitude         <dbl> 7.413815, 7.413966, 7.413821, 7.415693, 7.415705,...
## $ altitude          <dbl> 161.41, 157.63, 159.74, 152.47, 156.16, 156.66, 1...
## $ velocity_mps      <dbl> 11.80, 11.83, 11.70, 11.45, 11.49, 7.93, 8.15, 8....
## $ acceleration_mpss <dbl> 0.13, 0.03, 0.06, -0.32, -0.26, 0.23, 0.24, 0.32,...
## $ direction         <dbl> 79.23, 79.35, 78.93, 85.25, 85.64, 102.61, 101.39...
## $ isRegistered      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ rsrp_dbm          <dbl> -99, -85, -121, -84, -97, -96, -74, -108, -111, -...
## $ rsrq_db           <dbl> -9, -5, -15, -6, -12, -12, -5, -9, -13, -11, -6, ...
## $ rssnr_db          <dbl> -1, 22, -8, 11, -2, 5, 29, 2, 6, 11, 13, 16, -3, ...
## $ cqi               <dbl> 8, 10, 4, 13, 9, 5, 15, 2, 6, 15, 12, 9, 6, 11, 1...
## $ ss                <dbl> 36, 50, 12, 52, 42, 42, 62, 33, 20, 53, 44, 45, 3...
## $ ta                <dbl> 9, 7, 63, 4, 7, 7, 4, 21, 16, 7, 4, 4, 7, 16, 4, ...
## $ ci                <dbl> 13828122, 26385408, 13067274, 29391105, 13416987,...
## $ pci               <dbl> 452, 95, 289, 167, 62, 62, 167, 347, 385, 62, 167...
## $ id                <dbl> 0, 0, 0, 1, 1, 2, 2, 1, 2, 3, 3, 4, 4, 4, 5, 5, 5...
## $ payload_mb        <dbl> 1.0, 4.0, 6.0, 2.0, 6.0, 5.0, 4.0, 10.0, 0.1, 7.0...
## $ throughput_mbits  <dbl> 4.66, 24.52, 1.29, 14.86, 3.97, 6.52, 16.27, 3.18...
## $ rtt_ms            <dbl> 47, 35, 59, 51, 1493, 146, 57, 71, 544, 1348, 54,...
## $ txPower_dbm       <dbl> 20.56, 12.30, 21.16, 10.02, 21.12, 21.38, 4.34, 2...
## $ f_mhz             <dbl> 1750, 1720, 1770, 1720, 1750, 1750, 1720, 1770, 1...
## $ scenario          <chr> "campus", "campus", "campus", "campus", "campus",...
```

```
## $ provider          <chr> "o2", "tmobile", "vodafone", "tmobile", "o2", "o2...
## $ timestamp         <dttm> 2018-12-10 09:08:57, 2018-12-10 09:08:57, 2018-1...
```

```
dataset_dl = make_dataset(data_source_dir, datatype="dl")
glimpse(dataset_dl)
```

```
## Rows: 6,516
## Columns: 26
## $ time_s            <dbl> 10.38, 10.17, 10.53, 20.16, 28.46, 30.86, 30.21, ...
## $ timestamp_ms      <dbl> 1544432936, 1544432937, 1544432943, 1544432948, 1...
## $ distance_m        <dbl> 100.93, 96.30, 88.76, 216.95, 300.95, 316.60, 308...
## $ latitude          <dbl> 51.49052, 51.49055, 51.49053, 51.49071, 51.49063,...
## $ longitude         <dbl> 7.413815, 7.413966, 7.413821, 7.415693, 7.416688,...
## $ altitude          <dbl> 161.41, 157.63, 159.74, 152.47, 155.68, 156.66, 1...
## $ velocity_mps      <dbl> 11.80, 11.83, 11.70, 11.45, 8.02, 7.93, 8.15, 8.2...
## $ acceleration_mpss <dbl> 0.13, 0.03, 0.06, -0.32, 0.15, 0.23, 0.24, 0.32, ...
## $ direction         <dbl> 79.23, 79.35, 78.93, 85.25, 100.75, 102.61, 101.3...
## $ isRegistered      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ rsrp_dbm          <dbl> -99, -85, -121, -84, -96, -96, -74, -108, -111, -...
## $ rsrq_db           <dbl> -9, -5, -15, -6, -12, -12, -5, -9, -13, -6, -13, ...
## $ rssnr_db          <dbl> -1, 22, -8, 11, 5, 5, 29, 2, 6, 13, -1, 1, 16, -3...
## $ cqi               <dbl> 8, 10, 4, 13, 5, 5, 15, 2, 6, 12, 7, 6, 9, 5, 11,...
## $ ss                <dbl> 36, 50, 12, 52, 42, 42, 62, 33, 20, 44, 38, 16, 4...
## $ ta                <dbl> 9, 7, 63, 4, 7, 7, 4, 21, 16, 4, 7, 16, 4, 7, 16,...
## $ ci                <dbl> 13828122, 26385408, 13067274, 29391105, 13416987,...
## $ pci               <dbl> 452, 95, 289, 167, 62, 62, 167, 347, 385, 167, 62...
## $ id                <dbl> 0, 0, 0, 1, 1, 2, 2, 1, 2, 3, 3, 3, 4, 4, 4, 5, 5...
## $ payload_mb        <dbl> 6.0, 0.1, 0.1, 2.0, 10.0, 7.0, 2.0, 5.0, 1.0, 3.0...
## $ throughput_mbits  <dbl> 2.38, 6.84, 3.54, 9.71, 0.90, 1.09, 7.31, 18.57, ...
## $ rtt_ms            <dbl> 54, 41, 61, 58, 1573, 144, 57, 70, 548, 163, 1346...
## $ f_mhz             <dbl> 1845, 1815, 1865, 1815, 1845, 1845, 1815, 1865, 1...
## $ scenario          <chr> "campus", "campus", "campus", "campus", "campus",...
## $ provider          <chr> "o2", "tmobile", "vodafone", "tmobile", "o2", "o2...
## $ timestamp         <dttm> 2018-12-10 09:08:56, 2018-12-10 09:08:57, 2018-1...
```

```
dataset_context = make_dataset(data_source_dir, datatype="context")
glimpse(dataset_context)
```

```
## Rows: 71,027
## Columns: 21
## $ time_s            <dbl> 0.06, 1.07, 0.05, 2.07, 1.06, 3.07, 2.07, 4.07, 3...
## $ timestamp_ms      <dbl> 1544432926, 1544432927, 1544432927, 1544432928, 1...
## $ distance_m        <dbl> 0.00, 6.79, 0.00, 14.43, 3.37, 23.00, 7.22, 33.08...
## $ latitude          <dbl> 51.49033, 51.49036, 51.49035, 51.49038, 51.49038,...
## $ longitude         <dbl> 7.412292, 7.412422, 7.412391, 7.412551, 7.412549,...
## $ altitude          <dbl> 161.88, 162.54, 168.50, 162.67, 167.87, 162.96, 1...
## $ velocity_mps      <dbl> 6.76, 7.65, 3.35, 8.57, 3.81, 10.08, 9.01, 10.73,...
## $ acceleration_mpss <dbl> 0.00, 0.89, 0.00, 0.92, 0.45, 1.51, 5.15, 0.65, 1...
## $ direction         <dbl> 76.84, 77.47, 77.76, 76.36, 77.48, 77.02, 76.10, ...
## $ isRegistered      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ rsrp_dbm          <dbl> -98, -101, -91, -101, -91, -94, -88, -94, -88, -9...
## $ rsrq_db           <dbl> -10, -12, -6, -12, -6, -9, -6, -9, -6, -8, -6, -8...
## $ rssnr_db          <dbl> -1, -1, 12, -1, 12, 5, 18, 5, 18, 1, 18, 1, 20, 3...
## $ cqi               <dbl> 9, 6, 11, 6, 11, 12, 15, 12, 15, 10, 15, 10, 12, ...
## $ ss                <dbl> 37, 36, 45, 36, 45, 40, 47, 40, 47, 36, 47, 36, 4...
```

```
## $ ta                <dbl> 9, 9, 7, 9, 7, 9, 7, 9, 7, 9, 7, 9, 7, 24, 9, 7, ...
## $ ci                <dbl> 13828122, 13828122, 26385408, 13828122, 26385408,...
## $ pci               <dbl> 452, 452, 95, 452, 95, 452, 95, 452, 95, 452, 95,...
## $ scenario          <chr> "campus", "campus", "campus", "campus", "campus",...
## $ provider          <chr> "o2", "o2", "tmobile", "o2", "tmobile", "o2", "tm...
## $ timestamp         <dttm> 2018-12-10 09:08:46, 2018-12-10 09:08:47, 2018-1...
```

```
dataset_cells = make_dataset(data_source_dir, datatype="cells")
glimpse(dataset_cells)
```

```
## Rows: 93,443
## Columns: 21
## $ time_s            <dbl> 0.07, 0.07, 0.07, 1.07, 1.07, 2.07, 2.07, 3.07, 4...
## $ timestamp_ms      <dbl> 1544432926, 1544432926, 1544432926, 1544432927, 1...
## $ distance_m        <dbl> 0.02, 0.02, 0.02, 6.80, 6.81, 14.45, 14.46, 23.01...
## $ latitude          <dbl> 51.49033, 51.49033, 51.49033, 51.49036, 51.49036,...
## $ longitude         <dbl> 7.412292, 7.412292, 7.412292, 7.412422, 7.412422,...
## $ altitude          <dbl> 161.88, 161.88, 161.88, 162.54, 162.54, 162.67, 1...
## $ velocity_mps      <dbl> 6.76, 6.76, 6.76, 7.65, 7.65, 8.57, 8.57, 10.08, ...
## $ acceleration_mpss <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ direction         <dbl> 76.84, 76.84, 76.84, 77.47, 77.47, 76.36, 76.36, ...
## $ isRegistered      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ rsrp_dbm          <dbl> -99, -103, -103, -104, -107, -104, -107, -100, -1...
## $ rsrq_db           <dbl> -12, -14, -16, -14, -17, -14, -17, -17, -17, -11,...
## $ rssnr_db          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ cqi               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ ss                <dbl> 39, 38, 39, 36, 35, 36, 35, 42, 42, 38, 38, 38, 3...
## $ ta                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ ci                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ pci               <dbl> 146, 450, 266, 450, 146, 450, 146, 450, 450, 450,...
## $ scenario          <chr> "campus", "campus", "campus", "campus", "campus",...
## $ provider          <chr> "o2", "o2", "o2", "o2", "o2", "o2", "o2", "o2", "...
## $ timestamp         <dttm> 2018-12-10 09:08:46, 2018-12-10 09:08:46, 2018-1...
```

Add the eNodeB-ID

This function converts a cell-ID to an eNodeB-ID:

```
cell_id_to_enodeb = function(cell_id) {
  result = tryCatch(
    {
      hex_string = as.character(as.hexmode(cell_id))
      enodeb_hex = str_sub(hex_string, start=1, end=-3)
      enodeb_integer = as.integer(as.hexmode(enodeb_hex))
      return(enodeb_integer)
    },
    error = function(err) {
      return(NA)
    }
  )
  return(result)
}
```

Let's test it using the example from the slides (the correct eNodeB-ID is 50464):

```
cell_id_to_enodeb(12918809)
```

```
## [1] 50464
```

A few more tests:

```
print(cell_id_to_enodeb(13828122)==54016)
```

```
## [1] TRUE
```

```
print(cell_id_to_enodeb(26385408)==103068)
```

```
## [1] TRUE
```

```
print(cell_id_to_enodeb(13067274)==51044)
```

```
## [1] TRUE
```

```
print(is.na(cell_id_to_enodeb(NA)))
```

```
## [1] TRUE
```

```
print(is.na(cell_id_to_enodeb(0)))
```

```
## [1] TRUE
```

Now update the datasets to also contain the eNodeB-ID:

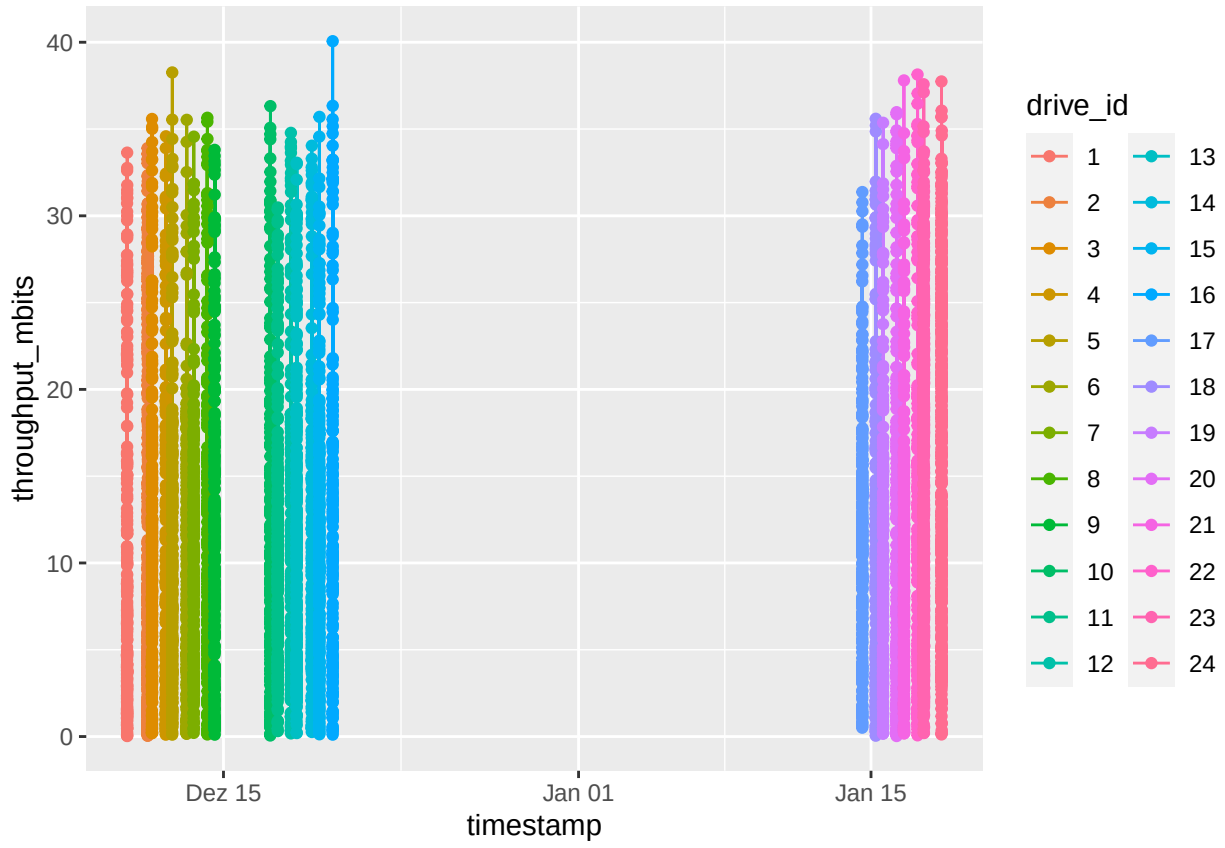
```
dataset_ul = dataset_ul %>% mutate(enodeb=map_int(ci, cell_id_to_enodeb))
dataset_dl = dataset_dl %>% mutate(enodeb=map_int(ci, cell_id_to_enodeb))
dataset_context = dataset_context %>% mutate(enodeb=map_int(ci, cell_id_to_enodeb))
dataset_cells = dataset_cells %>% mutate(enodeb=map_int(ci, cell_id_to_enodeb))
```

Add an ID for each drive

```
# this function assumes that the data is sorted by timestamp in ascending order
get_drive_ids = function(data, max_time_delta = minutes(10)) {

  num_rows = nrow(data)
  drive_ids = integer(num_rows)

  last_timestamp = as_datetime(origin) # the earliest possible date
  cur_drive_id = 0

  for(cur_row in seq_len(num_rows)) {

    cur_timestamp = data[[cur_row, "timestamp"]]

    if (cur_timestamp - last_timestamp > max_time_delta) {
      cur_drive_id = cur_drive_id + 1
    }
    last_timestamp = cur_timestamp

    drive_ids[cur_row] = cur_drive_id
  }

  return(drive_ids)
}
```

```
dataset_ul = add_column(dataset_ul, drive_id = factor(get_drive_ids(dataset_ul)))
dataset_dl = add_column(dataset_dl, drive_id = factor(get_drive_ids(dataset_dl)))
dataset_context = add_column(dataset_context, drive_id = factor(get_drive_ids(dataset_context)))
dataset_cells = add_column(dataset_cells, drive_id = factor(get_drive_ids(dataset_cells)))
```

```
ggplot(dataset_ul, aes(x=timestamp, y=throughput_mbits)) +
  geom_line(aes(color=drive_id)) +
  geom_point(aes(color=drive_id))
```



Store them:

```
write_csv(dataset_ul, str_c(data_destination_dir, "dataset_ul.csv"))
write_csv(dataset_dl, str_c(data_destination_dir, "dataset_dl.csv"))
write_csv(dataset_context, str_c(data_destination_dir, "dataset_context.csv"))
write_csv(dataset_cells, str_c(data_destination_dir, "dataset_cells.csv"))
```