Task 1
Dataset:
Different CSV files containing historical financial data for cryptocurrencies.

Goal:
The major goal is to find as much historical data on cryptocurrencies as possible and bring them together in a suited database in suited representations.

Problems (just a selection):

- Different time intervals (recorded every minute or daily)
- Different stock exchanges may value the price differently for the same currency
- Gaps in the recording, inconsistency between sources, slightly different timestamps
- Value could potentially refer to different currencies

Lasse, Adrian

Task 2
Zurich Insurance, gave me a very large (>40 GB) dataset essentially containing huge amounts of vehicle GPS positions across different trips and drivers (~125 million). The whole data is chunked into 500 CSVs, and to deduce something like whether trips ended in an accident would require matching all the data by their trip ids first, joining them etc.

We were thinking of getting the data into one piece, moving it into a relational database to be able to query trips as a whole, and then visualize them on a map. From that point we're sure we could find plenty of things to add on such as merging segmented trips (due to gas stop etc.), create driver profiles, or even perform deeper analytics as all the GPS pings include driving data such as RPM, speed etc. as well.

Daniel, Alexander

Task 3
Dataset: "Good Morning Tweets"
Tweets captured over ~24 hours with the text 'good morning' in them
https://www.kaggle.com/tentotheminus9/good-morning-tweets

Task: Extract keywords (hashtags), convert CSV into relational format (normalisation):
- For each tweet, extract the hashtags that are contained in it.
- Convert the CSV into these four tables:
  1. users: this table should contain all information that belongs to the twitter user.
  2. tweets: this table should contain all information that belongs to the tweet. It should also contain a foreign key to the author of the tweet.
  3. hashtags: this table contains the name and id of a hashtag, for each (distinct) hashtag previously extracted.
  4. tweets_hashtags: this table stores a many-to-many relation between tweets and hashtags.

Hao, Jakob

Task 5
My proposed task is to scrape the HPI websites for course-related data like:
schedule
course-type
# of LP
Link to all slides / materials
language
deadline for enrolment
room (maybe even room changes)
etc.
Then combine, visualise, integrate these into a Database.
This would enable quick search for specific courses to get LPs in a specific field.
Maybe combine it with data (slides) from Moodle.
Also batch-downloading the slides of a course at the end of semester before learning for exams.

Felix, Margaux

Task 6
In the Enron Corpus(and basically every other email corpus) has a lot of partially duplicates with every answer and forwarding of massages. Furthermore, it contains a lot of personal data and it is not obvious which person received which message in which order. This means I have 3 possible goals in these kind of data sets:

- Remove partially duplicates
- Replace personal data with pseudonyms (not only the header but also in the body of the messages)
- Structure (and visualize?) the information in a way that the information flow is more visible.

Lukas, Oliver, Lisa

Task 7
Integrating several movie rating datasets like IMDB, Rotten Tomatoes or MovieLens into a single (relational) dataset.
Given different sources you could aggregate the ratings, in order to retrieve a less biased rating for a movie of choice.
As well you could apply simple heuristics to the aggregation, in the case you trust one of the sources more than another source.

Lando, David

Task 8
Integrate DBpedia and Wikidata, and match entities between the two.
Datasources: DBpedia (https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10) and Wikidata (https://www.wikidata.org/wiki/Wikidata:Database_download)

Axel, Jan

Task 10
I would like to integrate EU data from Eurostat with US data from data.gov.
Very interesting would be the different water quality level from US and EU compared or combined in one data set.
One would have to find what similar attributes two such data sets have and combine them using the same data representation.
Another interesting topic would be Biodiversity, but Eurostat datasets are not very extensive.

Justus, Theresia

Task 13
Merge the Datasets and load it into a Spark Dataframe / relational Database.
Use case: This allow to analyse the distribution of death causes over different states and races and to put the leading causes of deaths per state into relation with the total population of the state.

1. Leading Causes of Death per State 1999-2016
2. Population per State 2010-2017
3. Population per State 2000-2010

Torben, Nico

Task 14
Join multiple datasets and save them into a database in order to conduct a detailed analysis of a major US city, e.g. Chicago.
For this purpose many nation-wide, zip-code-based datasets (e.g Department of Treasury: income tax returns) could be used as well as datasets which focus only on Chicago itself (e.g. Crimes from 2001 present).
The resulting dataset then could be used to make zip-code-based visualizations of different features all around the city.
The necessary datasets can be retrieved form data.gov.

Hendrik, Nils