LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LFE Medieninformatik •

Oberseminar

What's in a history? A large-scale statistical analysis of Last.fm data

Jennifer Büttgen

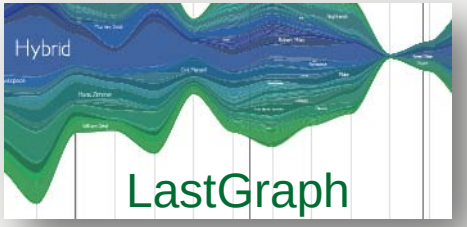
Diplomarbeit (Antrittsvortrag)

Betreuer: Dipl.-Medieninf. Dominikus Baur

Verantw. Hochschullehrer: Prof. Dr. Andreas Butz

Related Work: Previous studies

- What **communication theorists** state about media consumption:
  - Uses & Gratifications Approach, Escapism, Mood Management
- What **psychologists** found out about the uses of music:
  - Role of music in one's everyday life, preferences & personality, emotional effects, cultural differences in music perception
- How **HCI researchers** apply these findings:
  - User studies, playlist generation, shuffling and skipping, visualization of listening histories



Motivation

- A lot of studies discussed human music consumption
- But most of them…
  - Do not rely on a representative dataset, or
  - Examine human behavior over a short term only, or
  - Talk more about psychological or sociological issues
  - Describe ***why*** people listen to music, ***what music*** they listen to and ***what effects*** listening to music has.

➔ They do **not** discuss ***how*** people listen to music in real life and when looking at complete **sequences of tracks**…

Therefore, the purpose of this work is to…

- Gather a **representative dataset** of real users
- "Observe" the users' behavior over a **longer term**
- Examine complete **sequences of songs**:

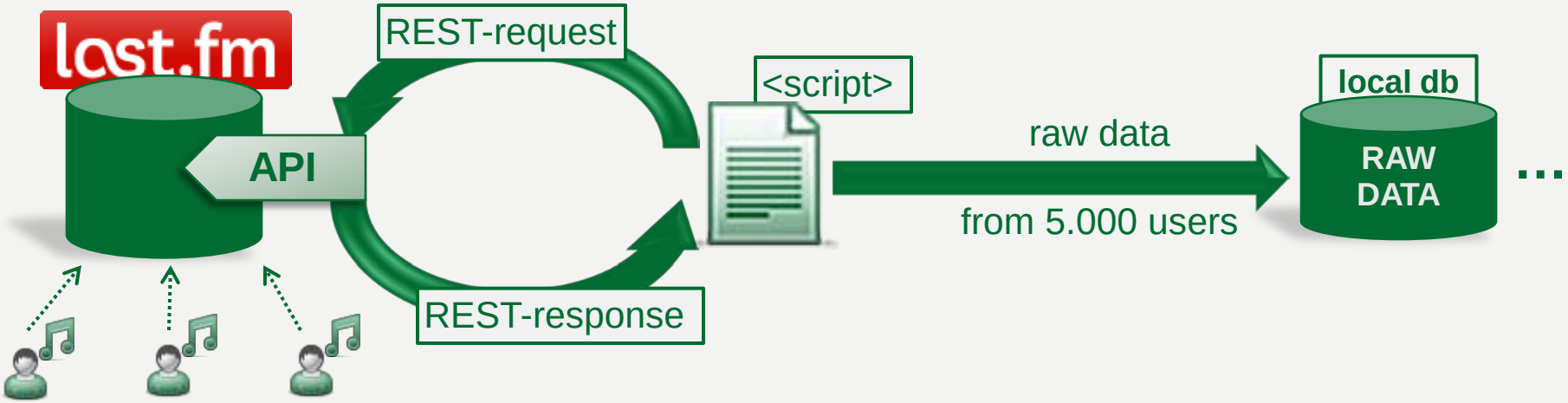
- And in case of an optimal solution:
  - Maybe even identify clusters of **user types**.

Why Last.fm?

- Last.fm „scrobbles“ a lot of data about its various users‘ music preferences and listening behavior.
- Most of this data can be accessed with the **Last.fm API**
  - E.g. via REST-style requests and responses

See also: http://last.fm/api ; Icons from http://iconza.com/

The working process part 1: Gathering data



- **Basic algorithm**: *next user = current user's last neighbour*
- **Problems and weaknesses of the data**:
  - Songs are only scrobbled if played at least 30 seconds
  - Wrong ID3-tags from users (typing errors, „The"-Band or not?, …)
  - Users can (and do!) turn off scrobbling sometimes. But when?

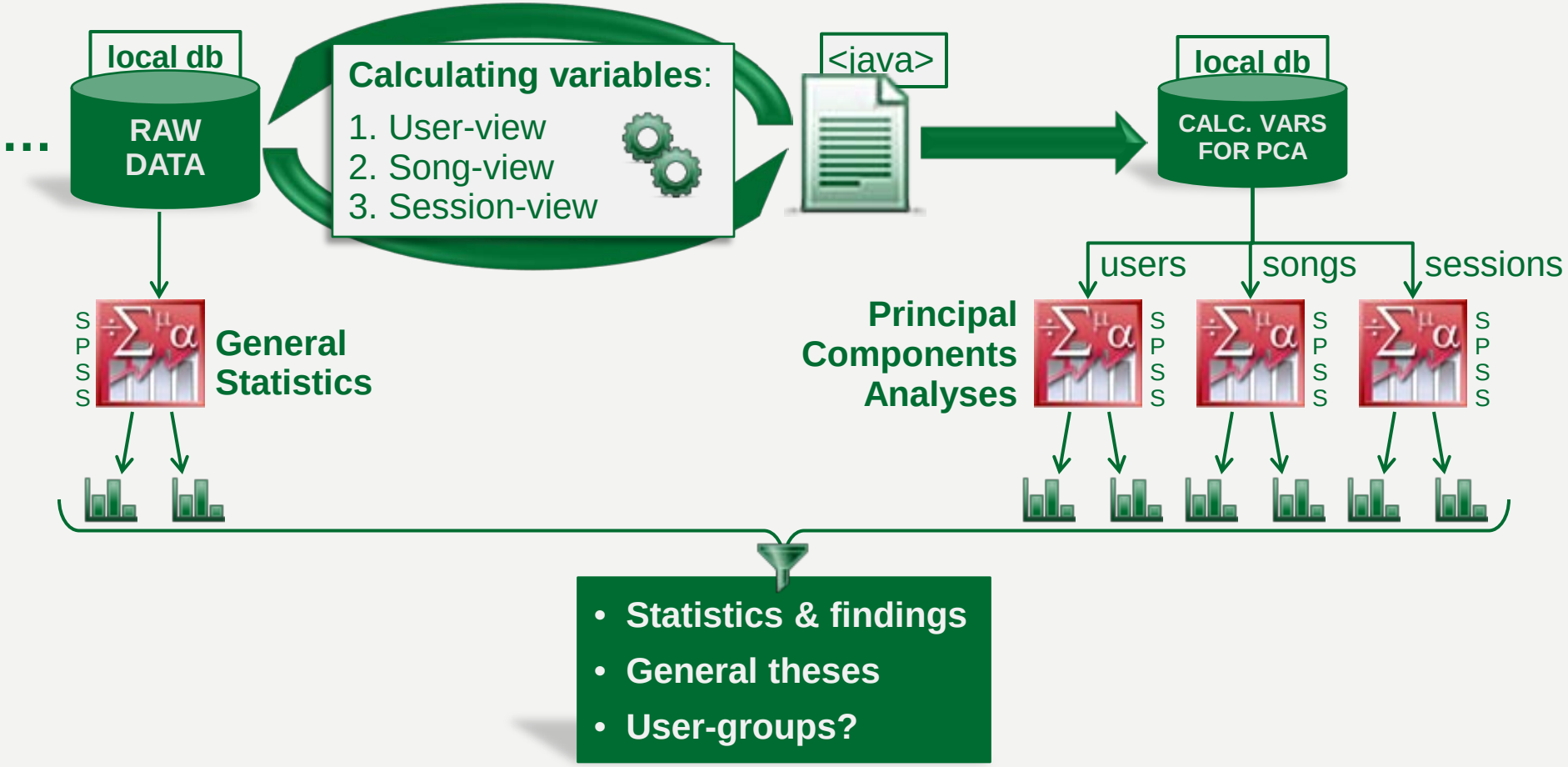
Determining what to analyze

- Challenge:
  - Find reasonable **variables** that describe the specific issues,
  - and suitable **algorithms** to calculate them,
  - so that the calculated variables can be analyzed and evaluated statistically with a **Principal Components Analysis**.
- Assistance in finding variables:
  - Brainstorming group sessions and results from previous studies
- Result:
  - Three approaches: Analysis from the view of the individual…
  -  **User**,  **Song**, and  **Session.**

Examples of developed variables

- **User-view**: 
  - Does the user play tracks from the same album in order?
  - Does the user repeatedly listen to the same song?
  - When does the user play music? At weekends? In the morning?
- **Song-view**: 
  - How many unique users did play this song? Total plays?
  - What follows after this song? A song from the same album? A break? A streamable song (on Last.fm's radio stations)?
- **Session-view**: 
  - How long is the session?
  - How many repeated songs?
  - …

The working process part 2: Analyzing data



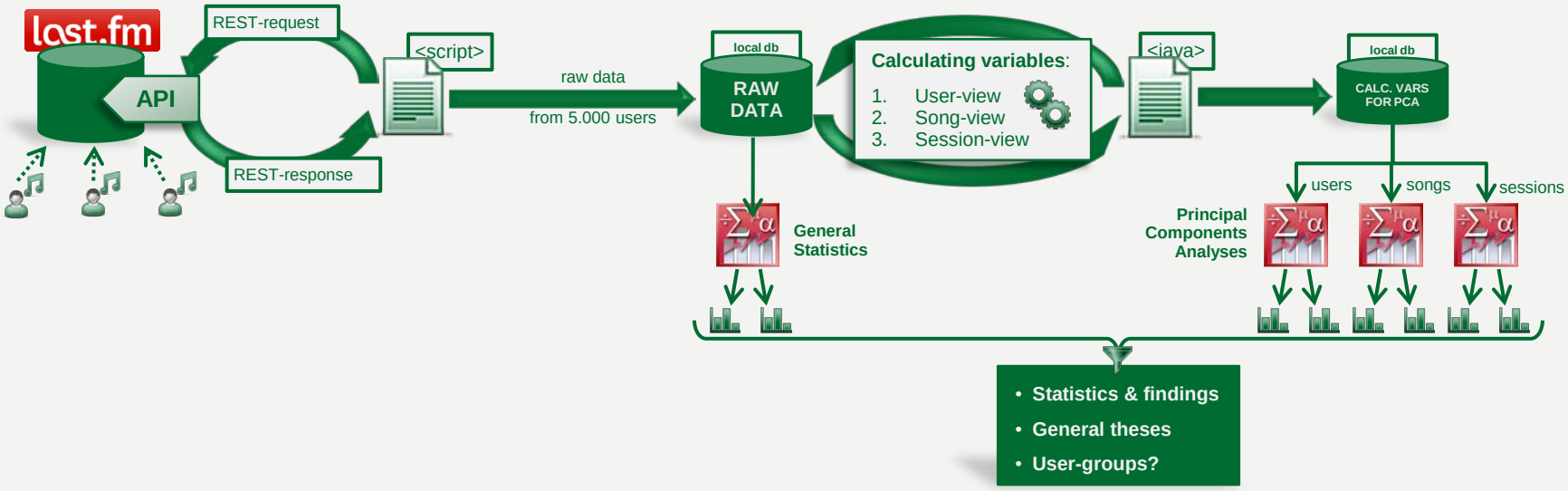
Next steps and Todos

- Finishing the Principal Components Analyses
  - Analysis from the view of individual sessions
- Evaluating & clustering the results to general theses
  - Maybe even identifying specific user types?
- Analyzing some findings in more detail
  - Depending on the time left…

Thank you for your attention!

Questions?

Suggestions?

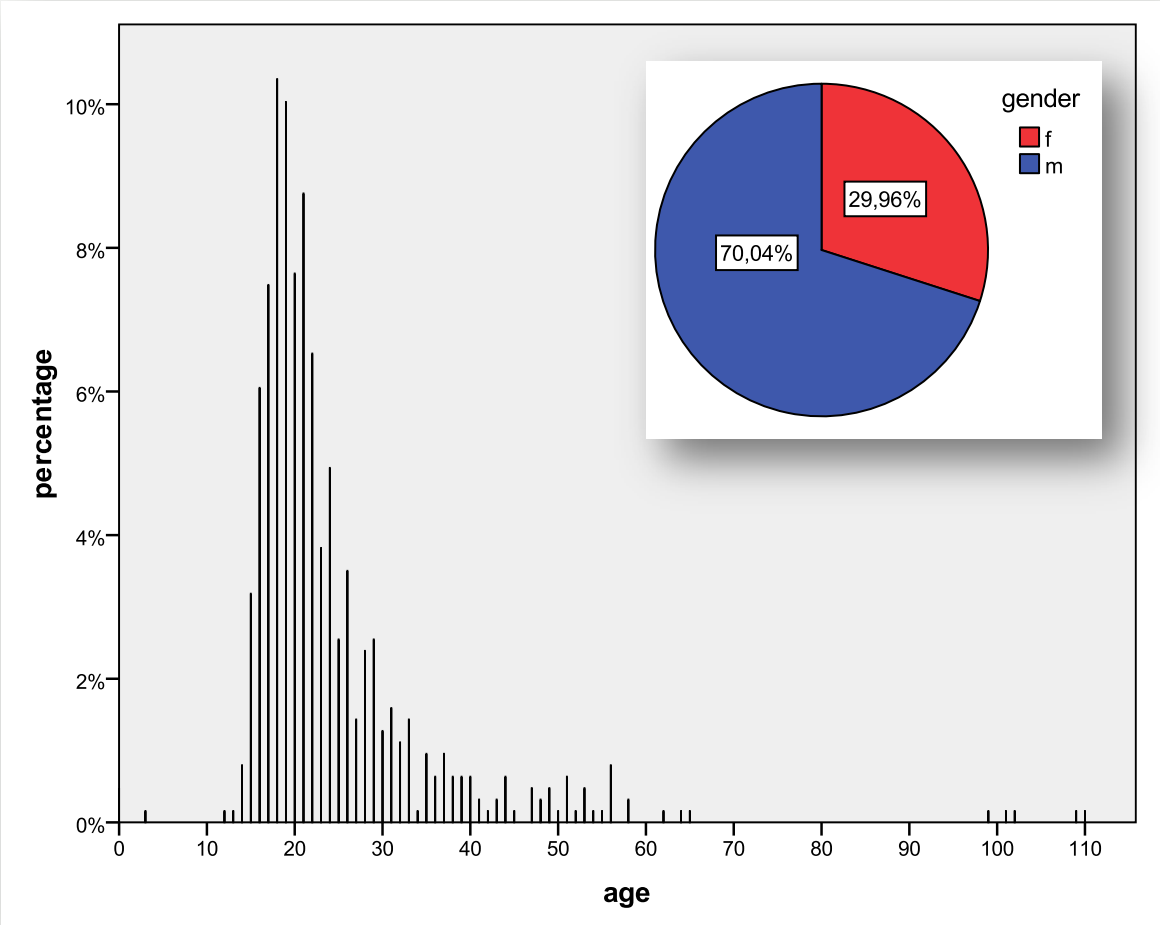


Appendix

First impressions of the raw dataset

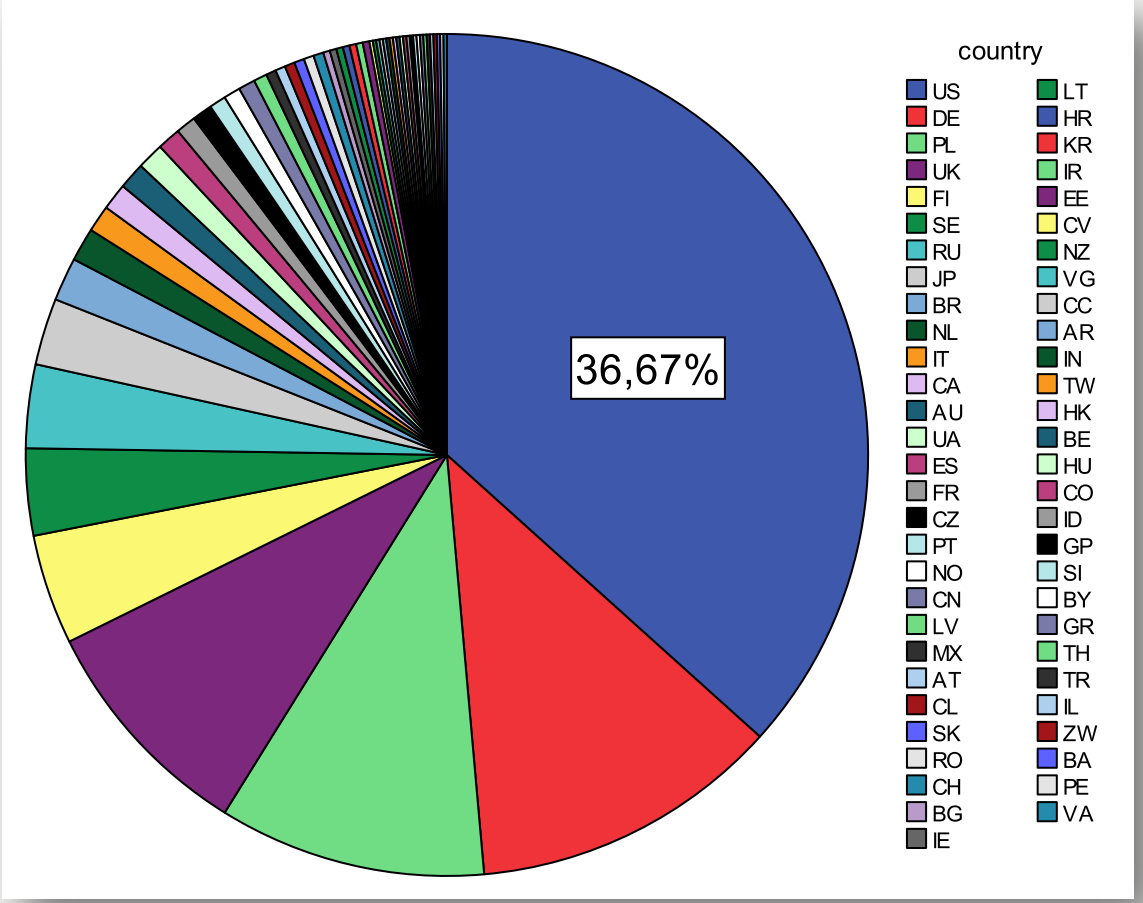
First impressions of the raw dataset: Demographics

- The users are very young
  - Ø 24.3 years,
  - Std. Dv. 11.6
- with a few cheaters (0 years or over 100 years)
- The users are predominantly male (about two thirds)



First impressions of the raw dataset: Home country

- The users come from various different countries
- The majority come from the United States (36.67%)
- Followed by
  - Germany (12%)
  - Poland (10%)
  - United Kingdom (9%)



First impressions of the raw dataset:
User activity

- There is a clear variation noticeable in general daily listening activity.

- Here: The total amount of tracks played by users from the United States at certain hours of day (**UTC**).