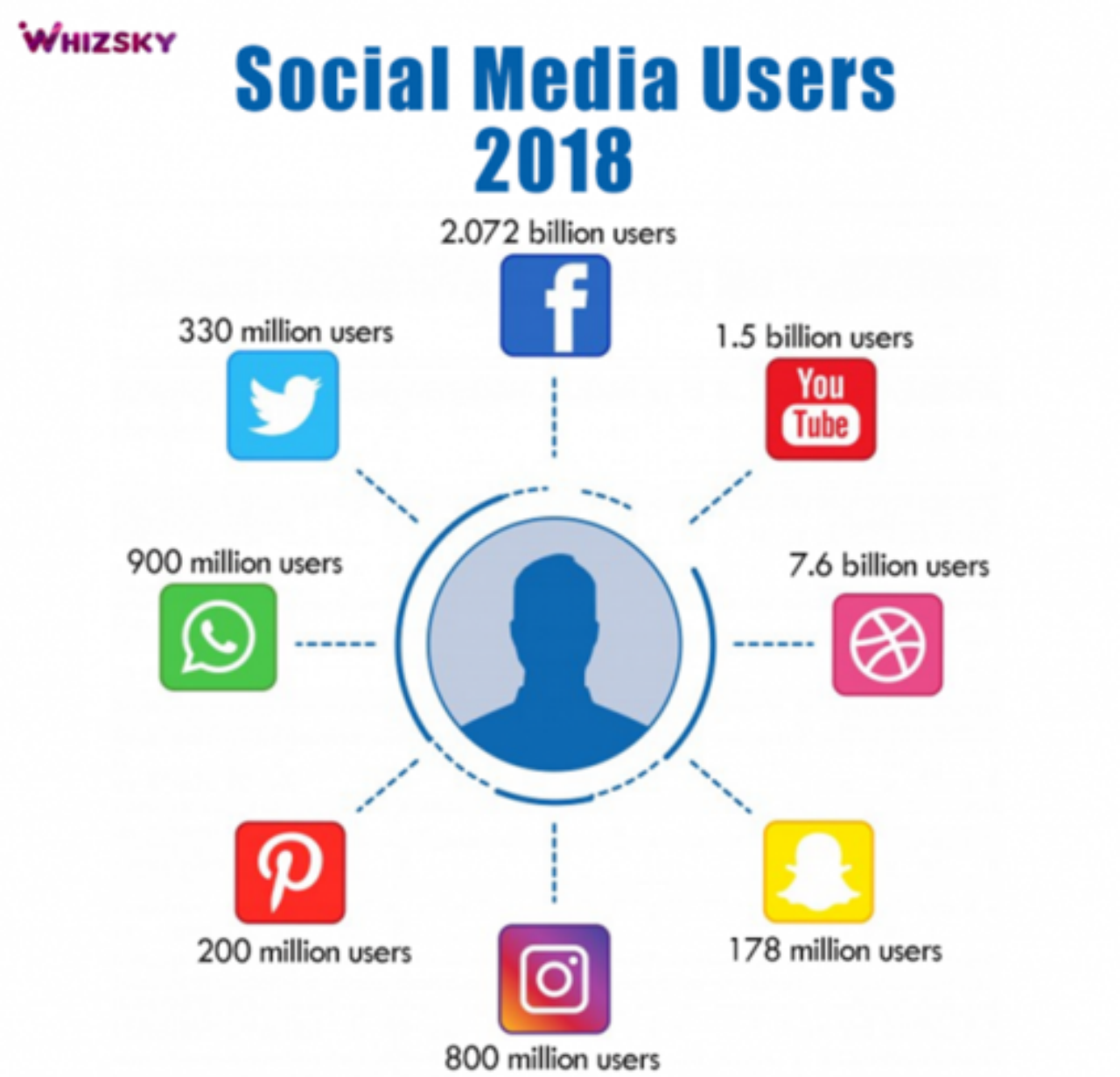
Multimedia Information Retrieval (MMIR)

Günther Specht
Michael Tschuggnall

Summer Term 2019



Source: https://www.whizsky.com/2018/05/amazing-social-media-statistics-and-facts-in-2018/

Challenges

- How to make multimedia content available to search engines and search based applications?

- How to gather complete metadata?
- Exploiting multimedia content requires:
  - Acquiring it
  - (Re) Formatting it
  - Indexing it
  - Querying it
  - Transmitting it
  - Browsing it

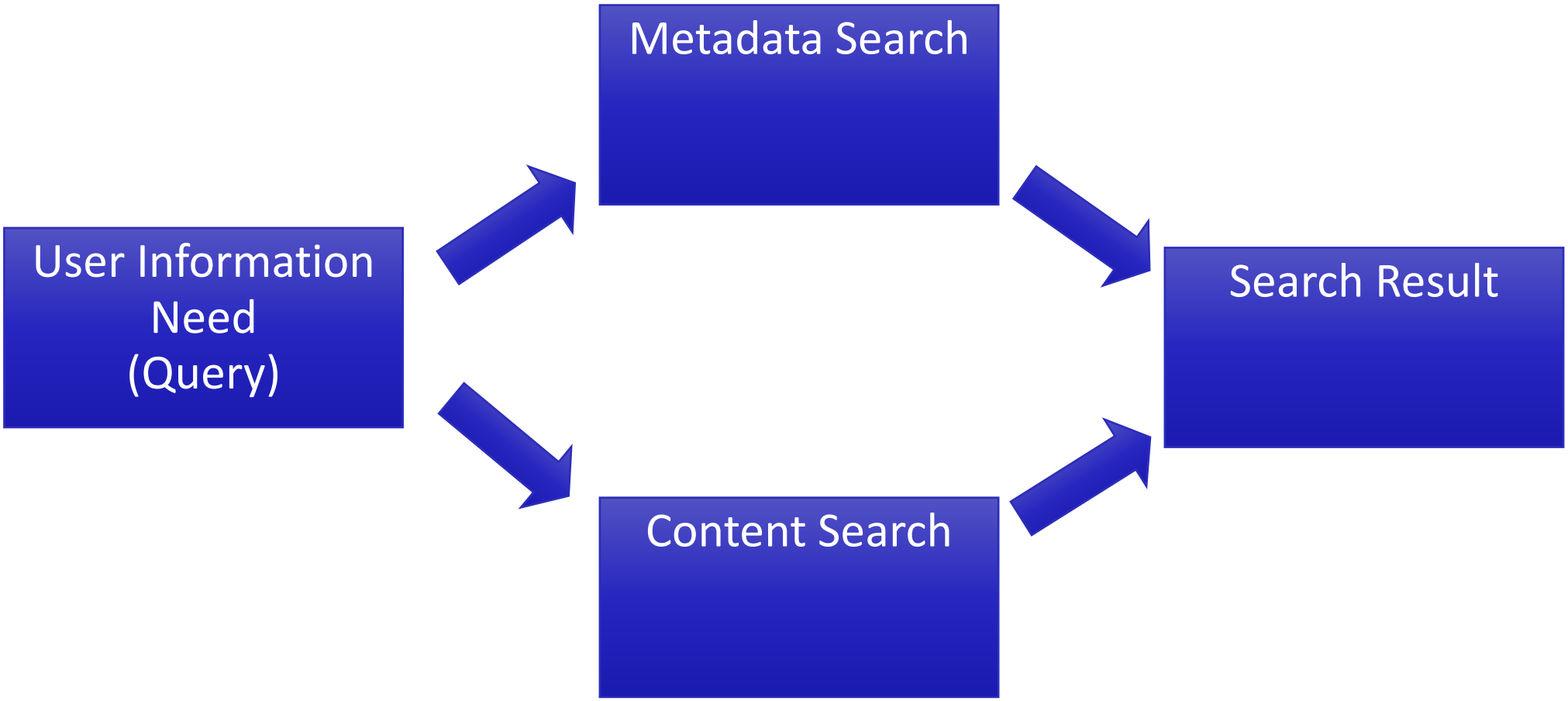
MMIR: Query Examples

- Play a few tones on a keyboard and retrieve a list of musical pieces similar to the required tune, or images matching the tones in a certain way, e.g., in terms of emotions
- Draw a few lines on a screen and find a set of images containing similar graphics, logos, ideograms,...
- Define objects, including color patches or textures and retrieve examples among which you select the interesting objects to compose your design
- On a given set of multimedia objects, describe movements and relations between objects and search for animations fulfilling the described temporal and spatial relations
- Describe actions and get a list of scenarios containing such actions
- Using an excerpt of Pavarotti's voice, obtaining a list of Pavarotti's records, video clips where Pavarotti is singing and photographic material portraying Pavarotti

Some Terminology

- Raw Content: the media element in native format, before any processing
- Metadata: data about data
- Global metadata: metadata that refers to an entire media element (e.g., movie title and director)
- Local metadata: metadata pertaining a specific media segment (e.g., a scene in a movie)
- Manual metadata: media descriptions edited manually
- Automatic metadata (annotations): media descriptions extracted by a software
- Derived artifacts: secondary content elements extracted (manually or automatically), e.g., video key-frames, summarizations, advertisements

Multimedia Search



Metadata

- Metadata is data about data
  - It describes properties of the data in a **structured** way
    - E.g.: creator, owner, creation date, **description**
- Some metadata is explicitly provided with the data
  - E.g.: size, file name, etc.
- Other data is **implicit**, and it must be extracted by means of algorithms and data analysis

Content Indexing

- In textual search engines, content need little (lexical) analysis before indexing
  - Index elements (words) are part of the content

- In MMIR, content cannot be indexed directly
  - Computers **cannot (yet) (fully) understand** the meaning of a multimedia content (pictures of a sunset)
  - **Pixels** and **audio samples** just bring binary information

- Indexable features must be extracted from the input data
  - **Low level features**: concisely describe physical or perceptual properties of a media element (e.g., feature vectors)
  - **High level features**: domain concepts characterizing the content (e.g., extracted objects and their properties, content categorizations, etc.)

Content Extraction:
A Quick Introduction
(or: A List of Challenges)

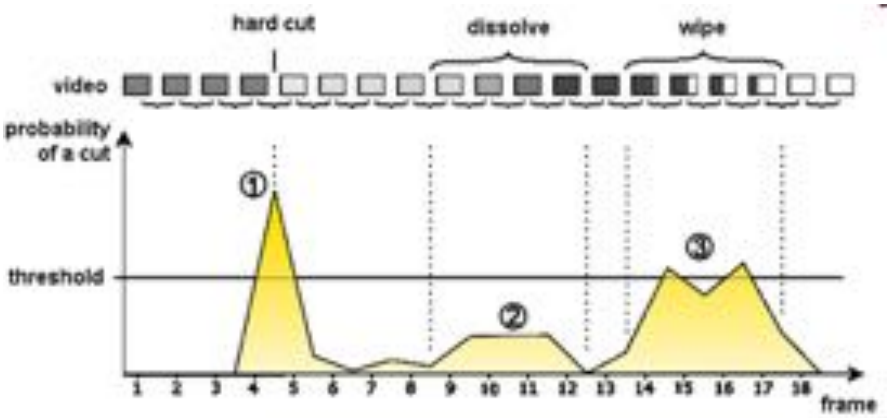
Indexing: Media Segmentation

- Media segmentation is to MMIR what tokenization is to textual IR
- Audio or video is split in units for analysis and presentation purposes
  - **ANALYSIS**: a homogeneous region can be subjected to a specific analysis pattern (e.g., a speaker's turn is processed to get speaker identification, a music or silent scene is not); music mood extraction can be performed on a subsequence of the whole track
  - **PRESENTATION**: a (possibly long) video sequence can be represented by a single thumbnail or scene

CUT

DISSOLVE

- **SHOT BOUNDARY DETERMINATION**: analyses consecutive frames for capturing typical camera motions (cut, fade in, fade out, dissolve, wipe, etc..)

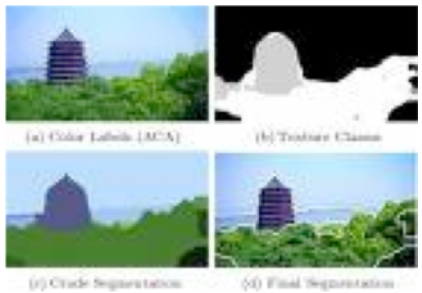


Indexing: Feature Extraction

- In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction.
- It works by transforming input data into smaller output data, so that some “features” are retained in the output that serve for a given task (e.g., OCR, face detection, …)
- Employed techniques are either general purpose (e.g., Principal Component Analysis [PCA]) or task-specific (e.g., edge detection filters).

Indexing: Image

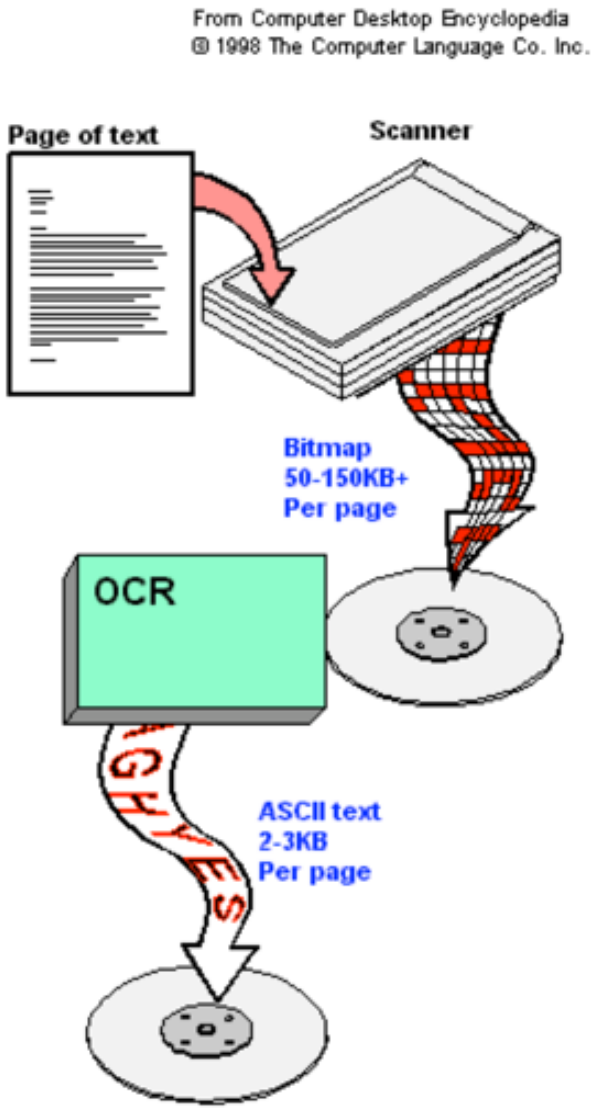
- Text/Image segmentation



- Low-level feature extraction (color (e.g., histogram), texture)

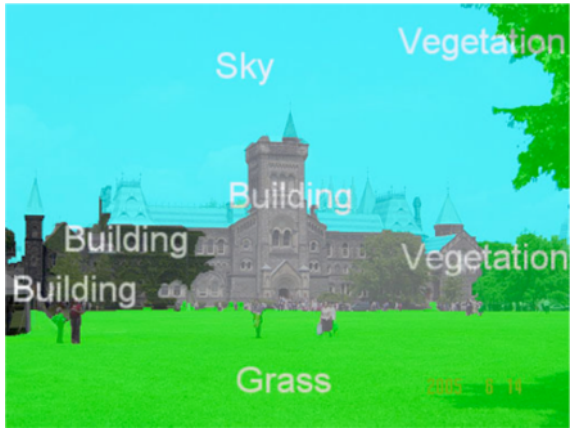
Indexing: Optical Character Recognition (OCR)

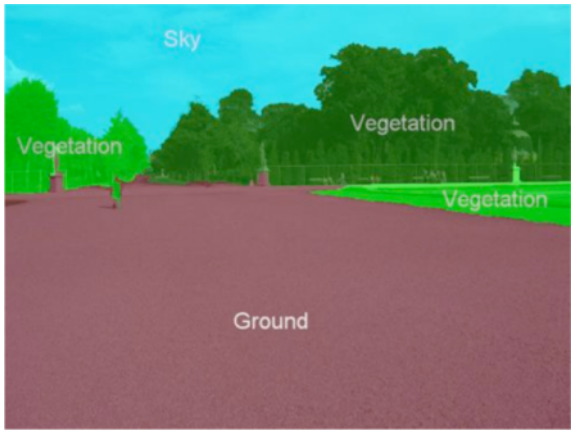
- OCR is a technique for translating images of typed or handwritten text into symbols
- Solved problem for typewritten text (99% accuracy)
- Commercial solutions for handwritten text (e.g., MS Tablet PC)
- Video OCR has specific problems, due to low resolution, small text size, and interference with background
- Detection is normally done on the most representative image of an entire shot, rather than frame by frame
- Approach: filter for enhancing resolution + pattern matching for character identification



Indexing: Concept Detection

- Image analysis extract low level features from raw data (e.g., color histograms, color correlograms, color moments, co-occurrence texture matrices, edge direction histograms, etc..)
- Features can be used to build discrete classifiers, which may associate semantic concepts to images or regions thereof
- Concepts can be detected also from text (e.g., from manual or automatic metadata) using NLP techniques like simple POS-tagging or NER



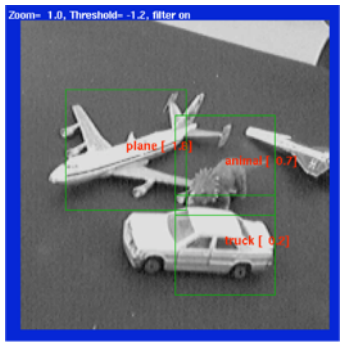


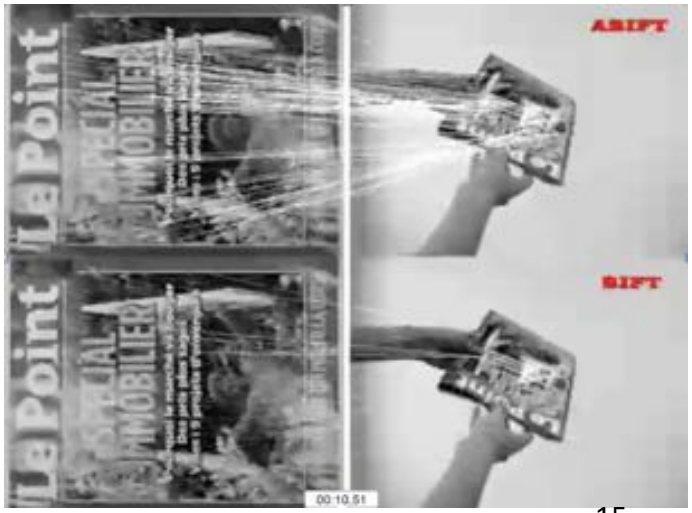
Indexing: Image

- Face recognition and identification

CREDITS: Thorsten Hermes@SSMT2006

- Object Recognition





Indexing: Video

- Motion detection and identification

- Object Tracking

CREDITS: Thorsten Hermes@SSMT2006

Indexing: Video

- Shot/Scene detection

- Video OCR

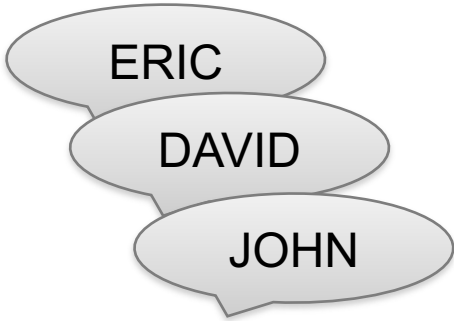
Indexing: Multimodal Annotation Fusion

- Media segmentation and concept extraction are probabilistic processes
- The result is characterized by a confidence value
- Significance can be enhanced by comparing the output of distinct techniques applied to the same or similar problems
- Examples:
  - Media segmentation: shot detection + speaker's turn identification
  - Person recognition: voice identification + face detection
  - Concept detection: image based classification (e.g., "outdoor" & "water" + object extraction: "bird", "boat")

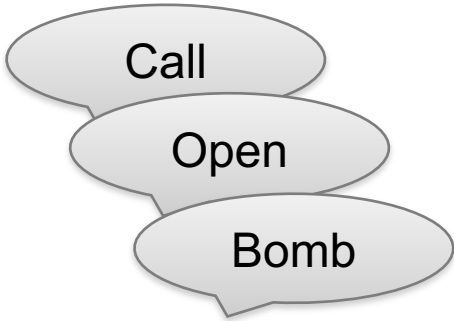
Music Information Retrieval (MIR)

Examples

- Speaker Identification



- Word Spotting



- Speech to text

Identification and Classification

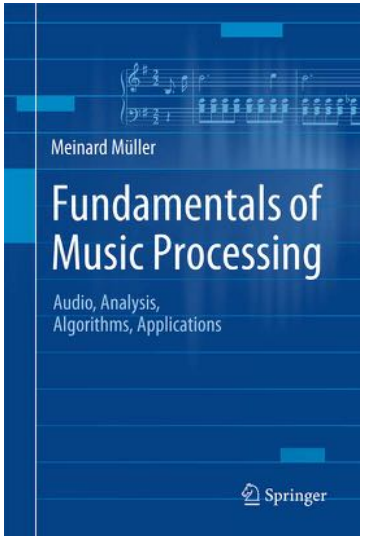
- Audio event identification

- Classification of music genres, mood, instruments, rhythmic features, structure, …

MIR

- In the following a short overview about the following subtopics is given:
  - Music Representation
  - Audio Features
  - Problem types overview

- We refer to the slides provided by **Meinard Müller** and **Christof Weiß** as part of their tutorial at ISMIR 2017, which accompany the excellent book „**Fundamentals of Music Processing**“



https://www.audiolabs-erlangen.de/resources/MIR/2017_TutorialAudioMIR_ISMIR/

Examples: Synchronization

- **A Web-Based Interface for Score Following and Track Switching in Choral Music**
  Frank Zalkow, Sebastian Rosenzweig, Johannes Graulich, Lukas Dietz, El Mehdi Lemnaouar, and Meinard Müller

  Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR), 2018

https://www.audiolabs-erlangen.de/resources/MIR/2018-ISMIR-LBD-Carus/

Examples: Categorization

- **Finding drum breaks in digital music recordings.**
  Patricio López-Serrano, Christian Dittmar, and Meinard Müller.

  International Symposium on Computer Music Multidisciplinary Research. Springer, Cham, 2017.

https://www.audiolabs-erlangen.de/resources/MIR/2017-CMMR-Breaks

Examples: Automatic Transcription

- **A Review of Automatic Drum Transcription.**
  Wu, Chih-Wei et al.

  IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 26.9 (2018): 1457-1483.

https://www.audiolabs-erlangen.de/resources/MIR/2017-DrumTranscription-Survey

Examples: Source Separation

- **Unifying Local and Global Methods for Harmonic-Percussive Source Separation.**
  Dittmar, C., López-Serrano, P., & Müller, M.

  2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

https://www.audiolabs-erlangen.de/resources/MIR/2018-ICASSP-HPSS_KAM_NMF

Examples: Structure Detection

- **Music Structure Analysis for the Song Cycle "Die Winterreise" (D911) by Franz Schubert.**
  Harald G. Grohganz, Polina Gubaidullina, Michael Clausen, Meinard Müller

https://www.audiolabs-erlangen.de/resources/MIR/METRUM-winterreise/

Examples: Modification/Creativity

- **Let it Bee - Towards NMF-inspired Audio Mosaicing.**
  Driedger, Jonathan, Thomas Prätzlich, and Meinard Müller

  Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 2015

https://www.audiolabs-erlangen.de/resources/MIR/2015-ISMIR-LetItBee

Genre Detection @ DBIS (1)

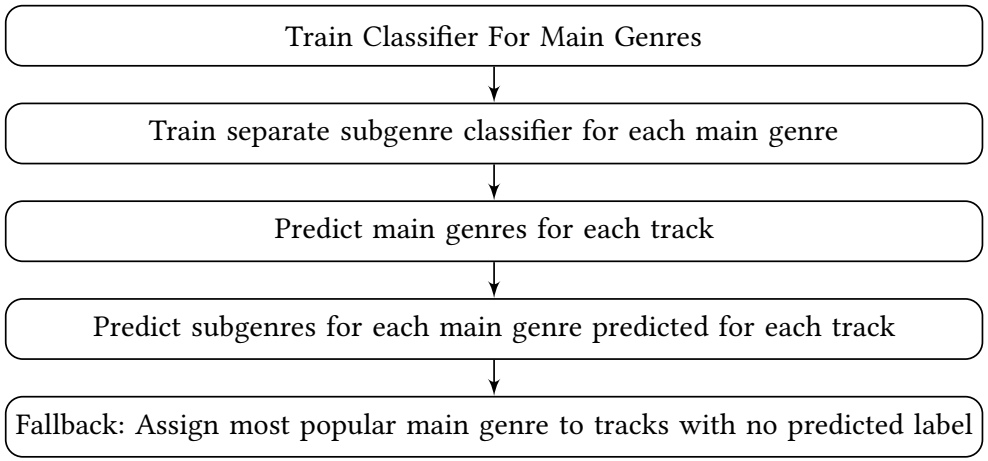
- **Hierarchical Multilabel Classification and Voting for Genre Classification**

  Benjamin Murauer, Maximilian Mayerl, Michael Tschuggnall, Eva Zangerle, Martin Pichl, Günther Specht

  CEURS *Working Notes of MediaEval 2017 Workshop,* 2017

  Use a reduced set of low level features and classify (with SVM‘s)

Approach

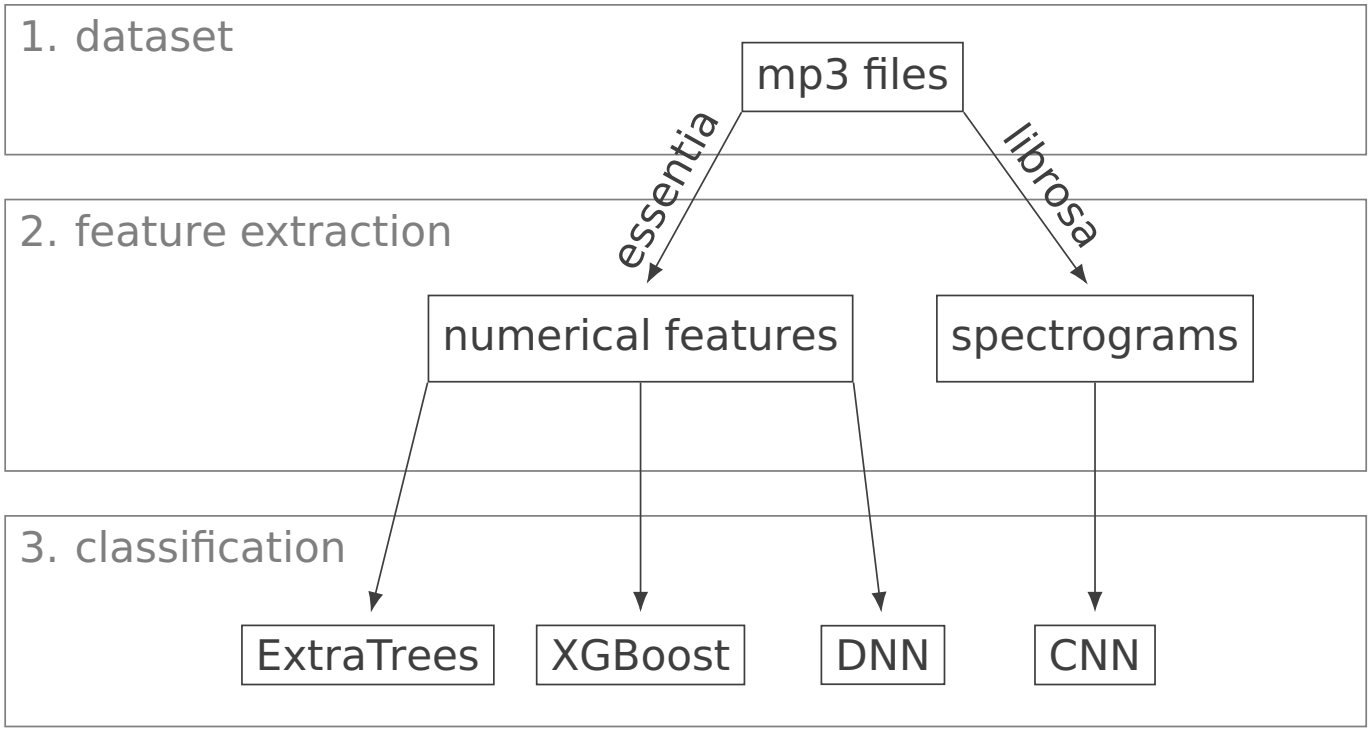


Results

Goal	AllMusic	Discogs	Lastfm	Tagtraum
Per Track (all)	0.249	0.374	0.340	0.363
Per Track (genre)	0.587	0.680	0.512	0.478
Per Track (subgenre)	0.193	0.219	0.251	0.303
Per Label (all)	0.070	0.144	0.155	0.153
Per Label (genre)	0.266	0.441	0.313	0.345
Per Label (subgenre)	0.065	0.129	0.139	0.131

Genre Detection @ DBIS (2)

- **Detecting Music Genre Using Extreme Gradient Boosting**
  Benjamin Murauer and Günther Specht
  Companion of the The Web Conference 2018 (WWW 2018), pages 1923-1927, 2018

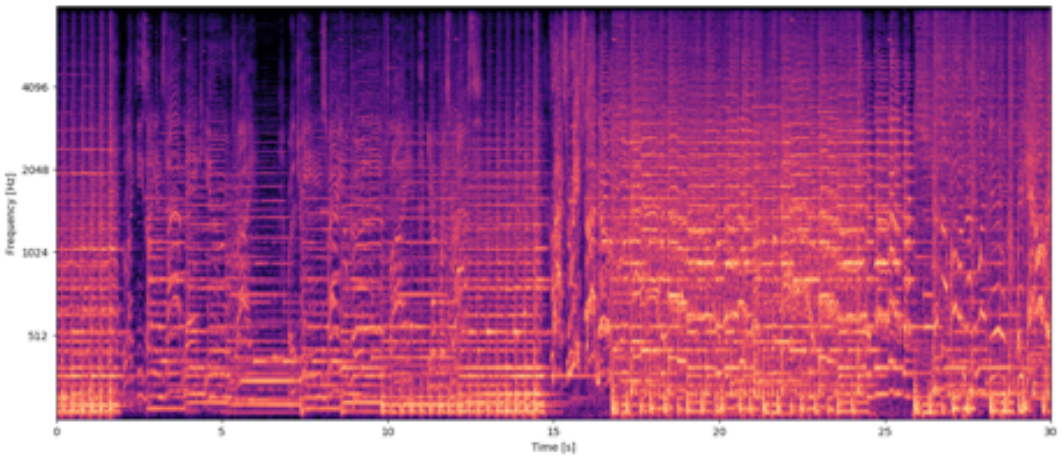


Genre Detection @ DBIS (2)

- **Detecting Music Genre Using Extreme Gradient Boosting**
  Benjamin Murauer and Günther Specht
  Companion of the The Web Conference 2018 (WWW 2018), pages 1923-1927, 2018

  Two approaches
  - Low-level features, classified with ExtraTrees, XGBoost, DNN
  - Spectograms and perform image classification with a CNN (how do genres *look*?)

Example Spectogram



Results

classifier	L	F_1^m
XGBoost	**0.82**	0.74
XGBoost	0.85	**0.78**
ExtraTrees	0.92	0.74
DNN	1.44	0.77
CNN	1.65	0.48
iyjeong	0.33	0.91
hglim	0.33	0.92