DEIC National Supercomputing Day 07.11.2016 -
Perspectives of High Performance Computing

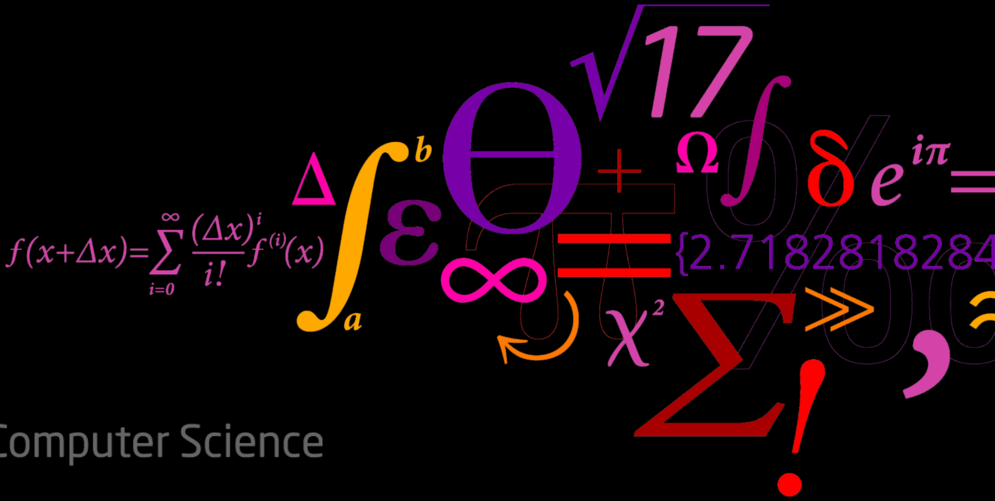
Interactive Crowdsourcing for Big Data

Jan Larsen, DTU Compute

Jens Madsen, DTU Compute

Bjørn Sand Jensen, University of Glasgow

COSOUND



DTU Compute
Department of Applied Mathematics and Computer Science

Why is audio the modality our research focus?

- Volume: Size

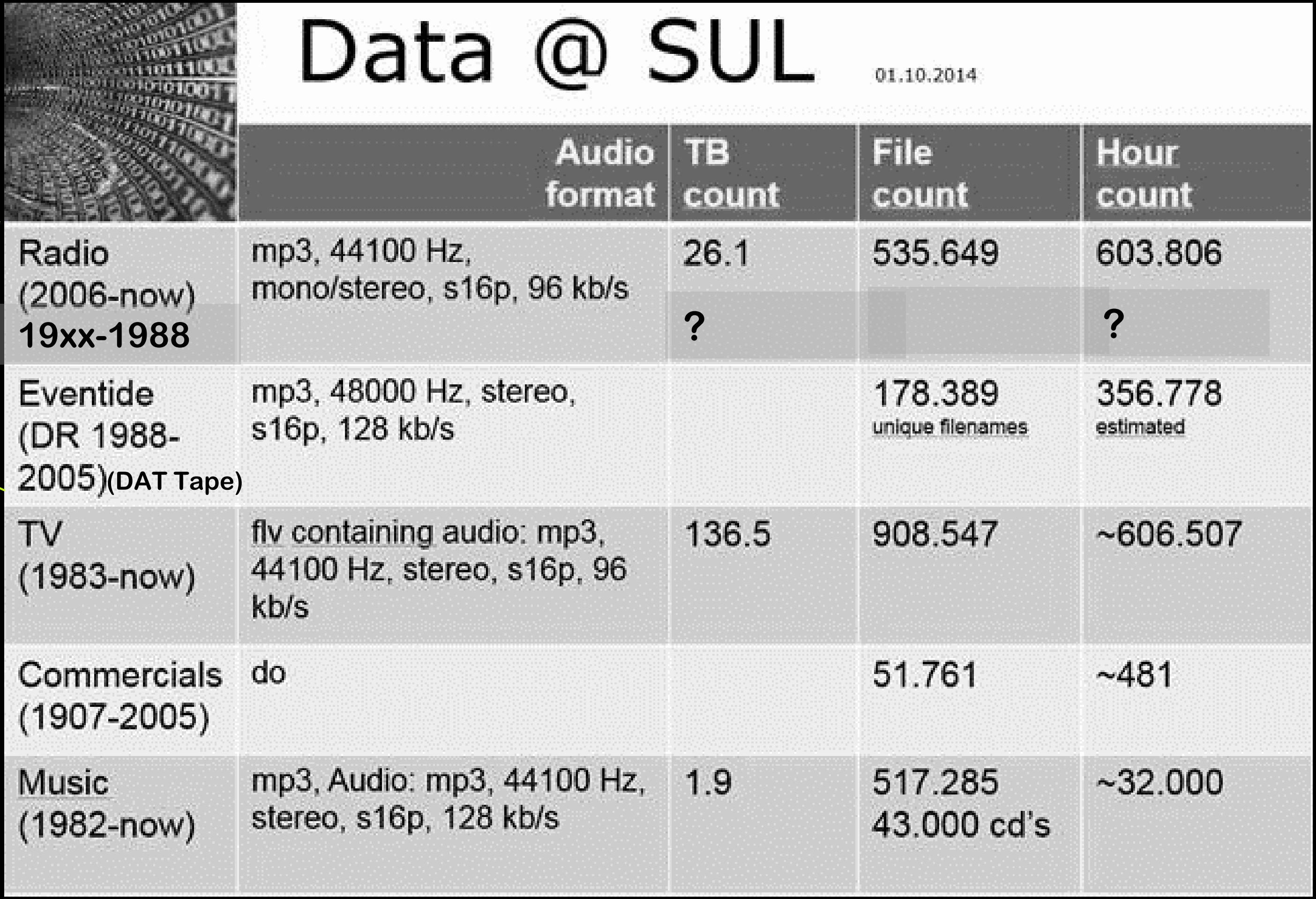
- Variety: Complexity
  - Perception
  - Affection
  - Redundancy and irrelevant information
  - Ambiguity

- Velocity: Real-time aspect – audio unfolds in time
  - Continuous speech
  - Music
  - Environmental sound

- Veracity: Uncertainty
  - Elicitation of human knowledge

IBM, www.ibmbigdatahub.com, The Four Big V's of Big data

Big Audio Data – the Danish media archive



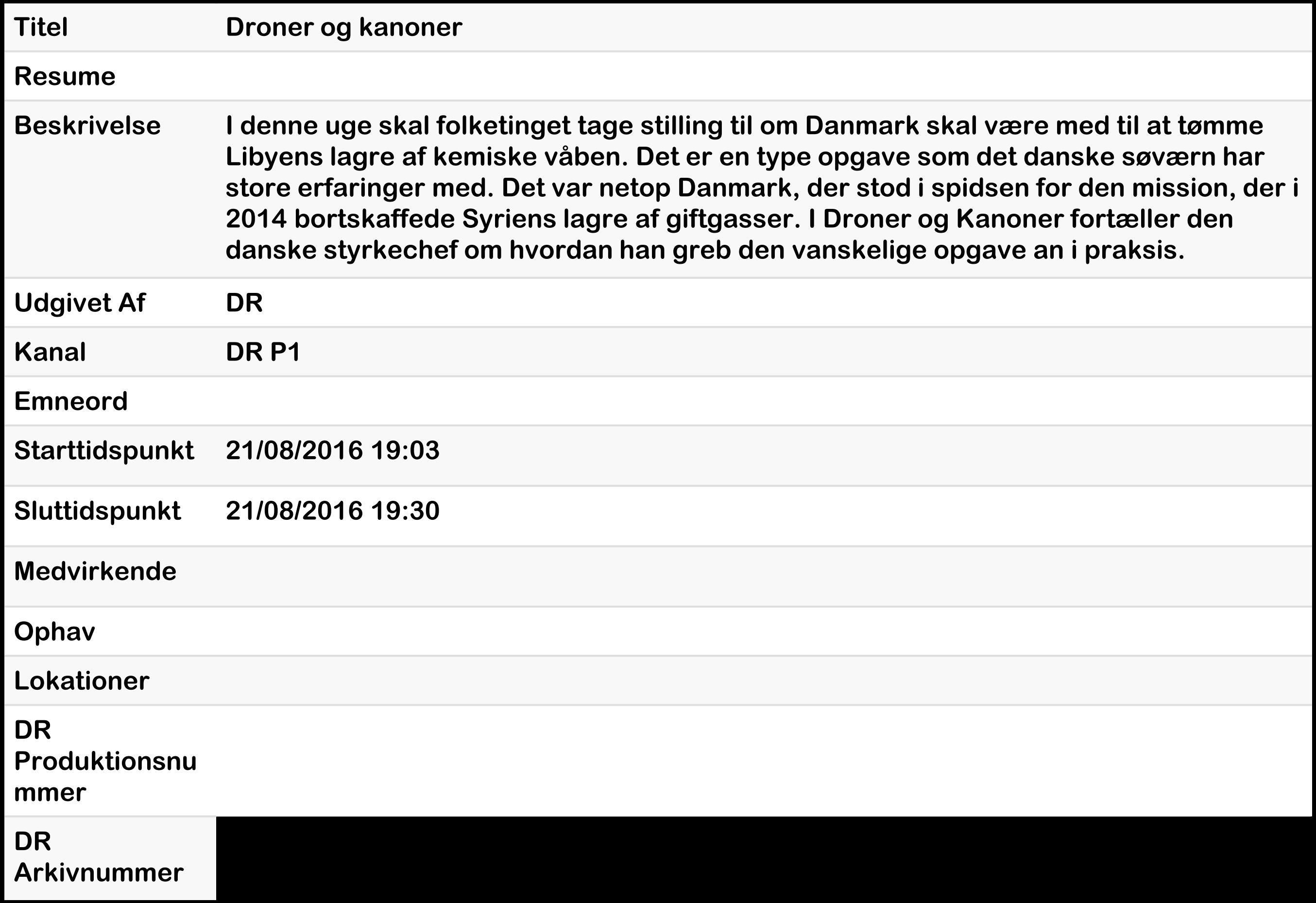
Data @ SUL 01.10.2014

	Audio format	TB count	File count	Hour count
Radio (2006-now) 19xx-1988	mp3, 44100 Hz, mono/stereo, s16p, 96 kb/s	26.1 ?	535.649	603.806 ?
Eventide (DR 1988-2005)(DAT Tape)	mp3, 48000 Hz, stereo, s16p, 128 kb/s		178.389 unique filenames	356.778 estimated
TV (1983-now)	flv containing audio: mp3, 44100 Hz, stereo, s16p, 96 kb/s	136.5	908.547	~606.507
Commercials (1907-2005)	do		51.761	~481
Music (1982-now)	mp3, Audio: mp3, 44100 Hz, stereo, s16p, 128 kb/s	1.9	517.285 43.000 cd's	~32.000

Radio

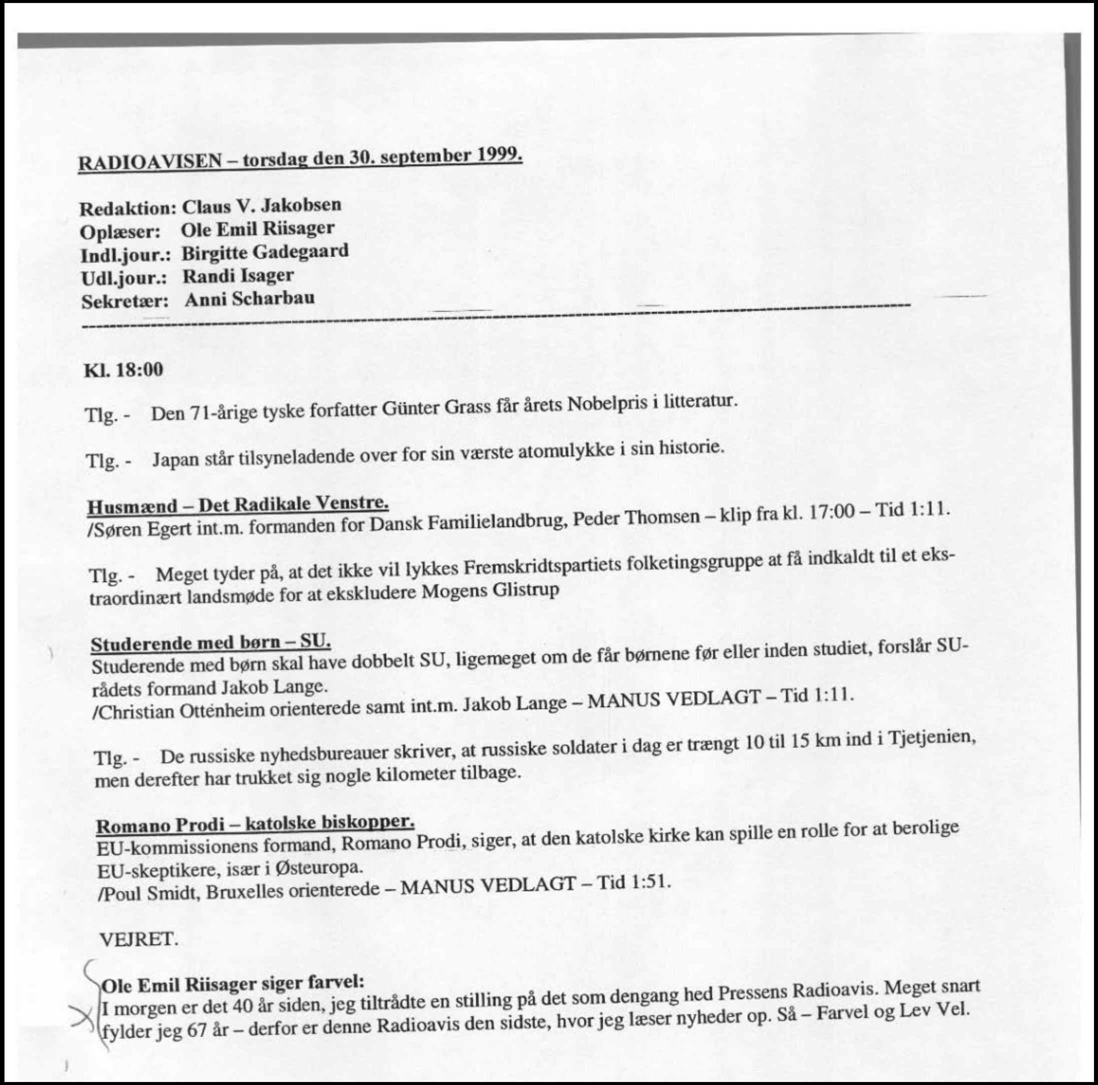
Almost 1 mio. hours

Existing unstructured, unsegmented metadata in radio archive



Titel	Droner og kanoner
Resume	
Beskrivelse	I denne uge skal folketinget tage stilling til om Danmark skal være med til at tømme Libyens lagre af kemiske våben. Det er en type opgave som det danske søværn har store erfaringer med. Det var netop Danmark, der stod i spidsen for den mission, der i 2014 bortskaffede Syriens lagre af giftgasser. I Droner og Kanoner fortæller den danske styrkechef om hvordan han greb den vanskelige opgave an i praksis.
Udgivet Af	DR
Kanal	DR P1
Emneord	
Starttidspunkt	21/08/2016 19:03
Sluttidspunkt	21/08/2016 19:30
Medvirkende	
Ophav	
Lokationer	
DR Produktionsnummer	
DR Arkivnummer	

Existing unstructured metadata in radio archive



RADIOAVISEN – torsdag den 30. september 1999.

Redaktion: Claus V. Jakobsen
Oplæser: Ole Emil Riisager
Indl.jour.: Birgitte Gadegaard
Udl.jour.: Randi Isager
Sekretær: Anni Scharbau

Kl. 18:00

Tlg. - Den 71-årige tyske forfatter Günter Grass får årets Nobelpris i litteratur.

Tlg. - Japan står tilsyneladende over for sin værste atomulykke i sin historie.

Husmænd – Det Radikale Venstre.
/Søren Egert int.m. formanden for Dansk Familielandbrug, Peder Thomsen – klip fra kl. 17:00 – Tid 1:11.

Tlg. - Meget tyder på, at det ikke vil lykkes Fremskridtspartiets folketingsgruppe at få indkaldt til et ekstraordinært landsmøde for at ekskludere Mogens Glistrup

Studerende med børn – SU.
Studerende med børn skal have dobbelt SU, ligemeget om de får børnene før eller inden studiet, forslår SU-rådets formand Jakob Lange.
/Christian Otténheim orienterede samt int.m. Jakob Lange – MANUS VEDLAGT – Tid 1:11.

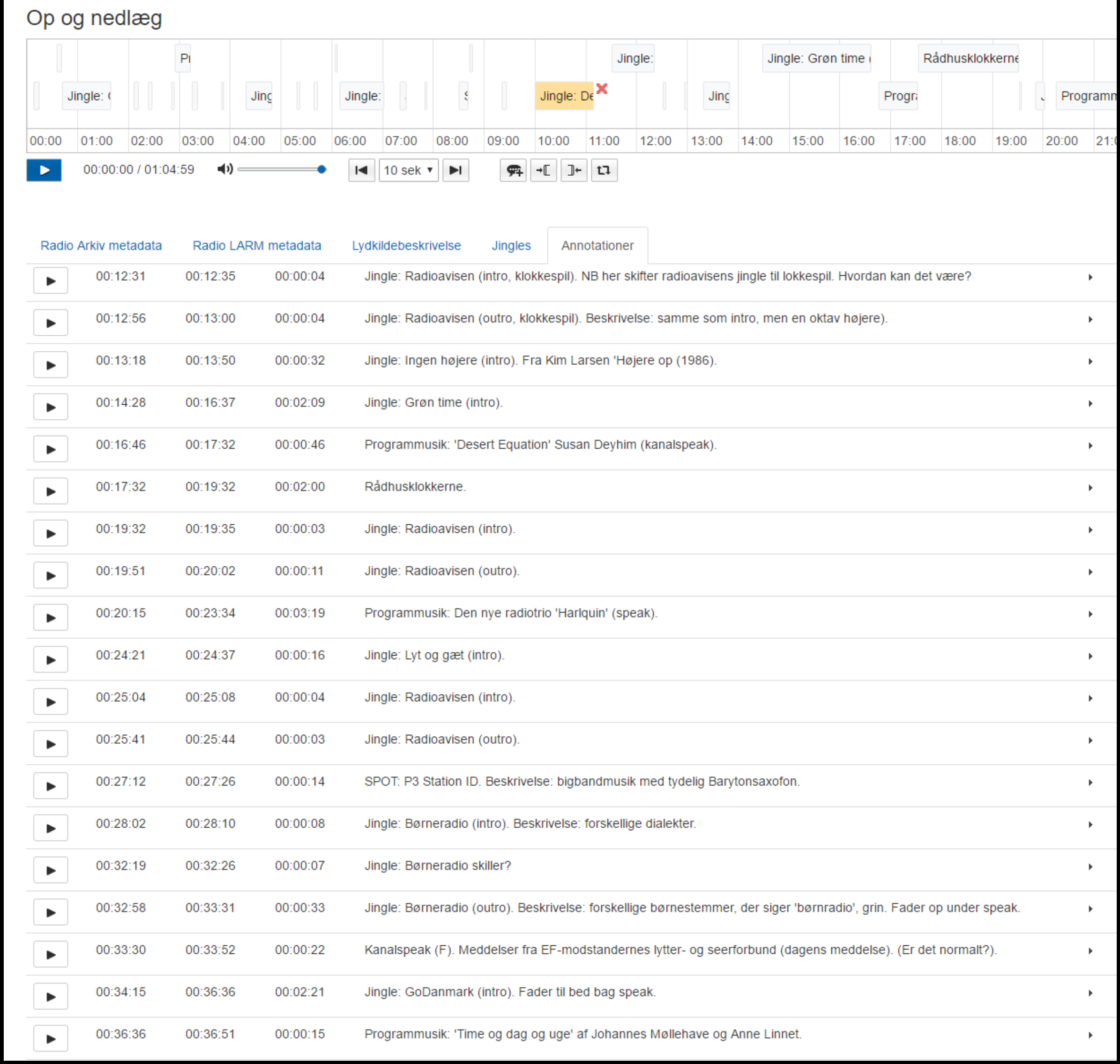
Tlg. - De russiske nyhedsbureauer skriver, at russiske soldater i dag er trængt 10 til 15 km ind i Tjetjenien, men derefter har trukket sig nogle kilometer tilbage.

Romano Prodi – katolske biskopper.
EU-kommissionens formand, Romano Prodi, siger, at den katolske kirke kan spille en rolle for at berolige EU-skeptikere, især i Østeuropa.
/Poul Smidt, Bruxelles orienterede – MANUS VEDLAGT – Tid 1:51.

VEJRET.

Ole Emil Riisager siger farvel:
I morgen er det 40 år siden, jeg tiltrådte en stilling på det som dengang hed Pressens Radioavis. Meget snart fylder jeg 67 år – derfor er denne Radioavis den sidste, hvor jeg læser nyheder op. Så – Farvel og Lev Vel.

Custom cultural research metadata schema



Op og nedlæg

▶	00:12:31	00:12:35	00:00:04	Jingle: Radioavisen (intro, klokkespil). NB her skifter radioavisens jingle til lokkespil. Hvordan kan det være?	▸
▶	00:12:56	00:13:00	00:00:04	Jingle: Radioavisen (outro, klokkespil). Beskrivelse: samme som intro, men en oktav højere).	▸
▶	00:13:18	00:13:50	00:00:32	Jingle: Ingen højere (intro). Fra Kim Larsen 'Højere op (1986).	▸
▶	00:14:28	00:16:37	00:02:09	Jingle: Grøn time (intro).	▸
▶	00:16:46	00:17:32	00:00:46	Programmusik: 'Desert Equation' Susan Deyhim (kanalspeak).	▸
▶	00:17:32	00:19:32	00:02:00	Rådhusklokkerne.	▸
▶	00:19:32	00:19:35	00:00:03	Jingle: Radioavisen (intro).	▸
▶	00:19:51	00:20:02	00:00:11	Jingle: Radioavisen (outro).	▸
▶	00:20:15	00:23:34	00:03:19	Programmusik: Den nye radiotrio 'Harlquin' (speak).	▸
▶	00:24:21	00:24:37	00:00:16	Jingle: Lyt og gæt (intro).	▸
▶	00:25:04	00:25:08	00:00:04	Jingle: Radioavisen (intro).	▸
▶	00:25:41	00:25:44	00:00:03	Jingle: Radioavisen (outro).	▸
▶	00:27:12	00:27:26	00:00:14	SPOT: P3 Station ID. Beskrivelse: bigbandmusik med tydelig Barytonsaxofon.	▸
▶	00:28:02	00:28:10	00:00:08	Jingle: Børneradio (intro). Beskrivelse: forskellige dialekter.	▸
▶	00:32:19	00:32:26	00:00:07	Jingle: Børneradio skiller?	▸
▶	00:32:58	00:33:31	00:00:33	Jingle: Børneradio (outro). Beskrivelse: forskellige børnestemmer, der siger 'børnradio', grin. Fader op under speak.	▸
▶	00:33:30	00:33:52	00:00:22	Kanalspeak (F). Meddelser fra EF-modstandernes lytter- og seerforbund (dagens meddelse). (Er det normalt?).	▸
▶	00:34:15	00:36:36	00:02:21	Jingle: GoDanmark (intro). Fader til bed bag speak.	▸
▶	00:36:36	00:36:51	00:00:15	Programmusik: 'Time og dag og uge' af Johannes Møllehave og Anne Linnet.	▸

Limitations of existing tools in exploiting the radio archive

- Unstructured, unsegmented meta data

- LARM.fm is a custom built search and visualization tools not intended for automated big-data analytics

- Kulturarvscluster?

Challenges

- End-users
  - Danish cultural heritage researchers
  - Danish broadcasting corporation
  - Hindenburg systems

- Needs
  - We have all this data, we want to do something with it!
  - Dialog between end-users and engineers
    - End-user: What is possible?
    - Engineer: What do you want?

- Overall need
  - Making the archive searchable
  - What to search for is unlimited

VISION

Smart crowd sourcing can effectively enrich media achieves with high quality metadata by using machine learning, gamification and interaction with users

Implicit crowdsourcing for Distributed Human Intelligence Tasking

COSOUND

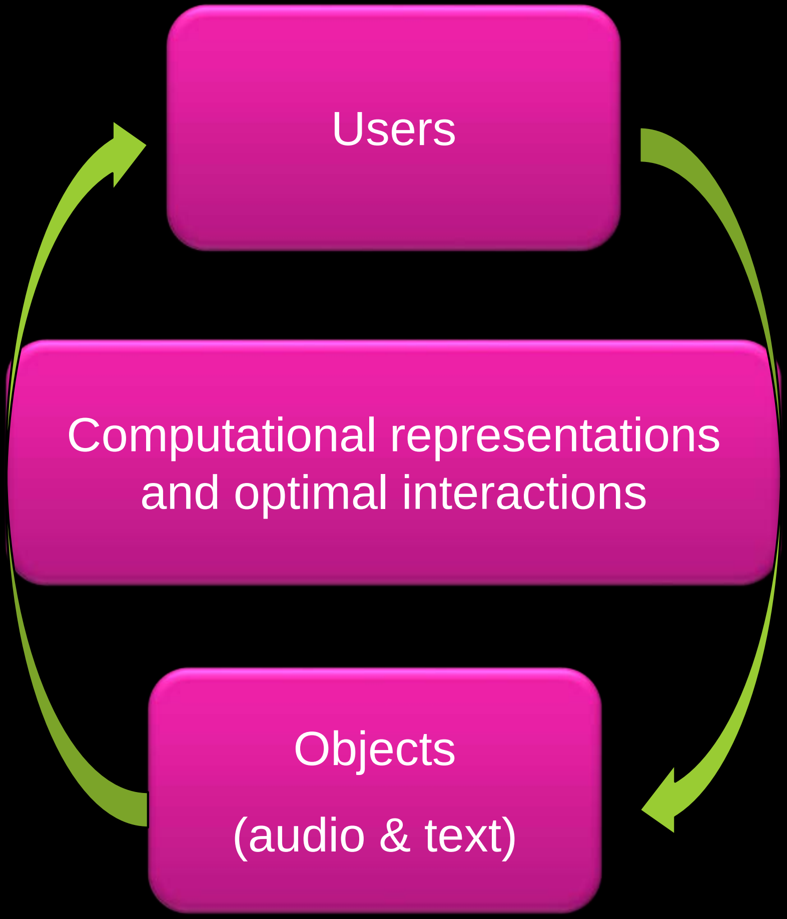
- Strategic research council (Innovation Fund Denmark) project 2012-2016

- Academic partners
  - Technical University of Denmark
  - University of Glasgow
  - University of Copenhagen - School of Library and Information Science and Humanities
  - University of Aalborg
  - Queen Mary University of London

- Industrial partners/end-users
  - Danish broadcast corporation (DR)
  - Bang & Olufsen
  - Hindenburg Systems

- Other partners

State and University Library, Chaos Insight, LARM.fm, Syntonetic



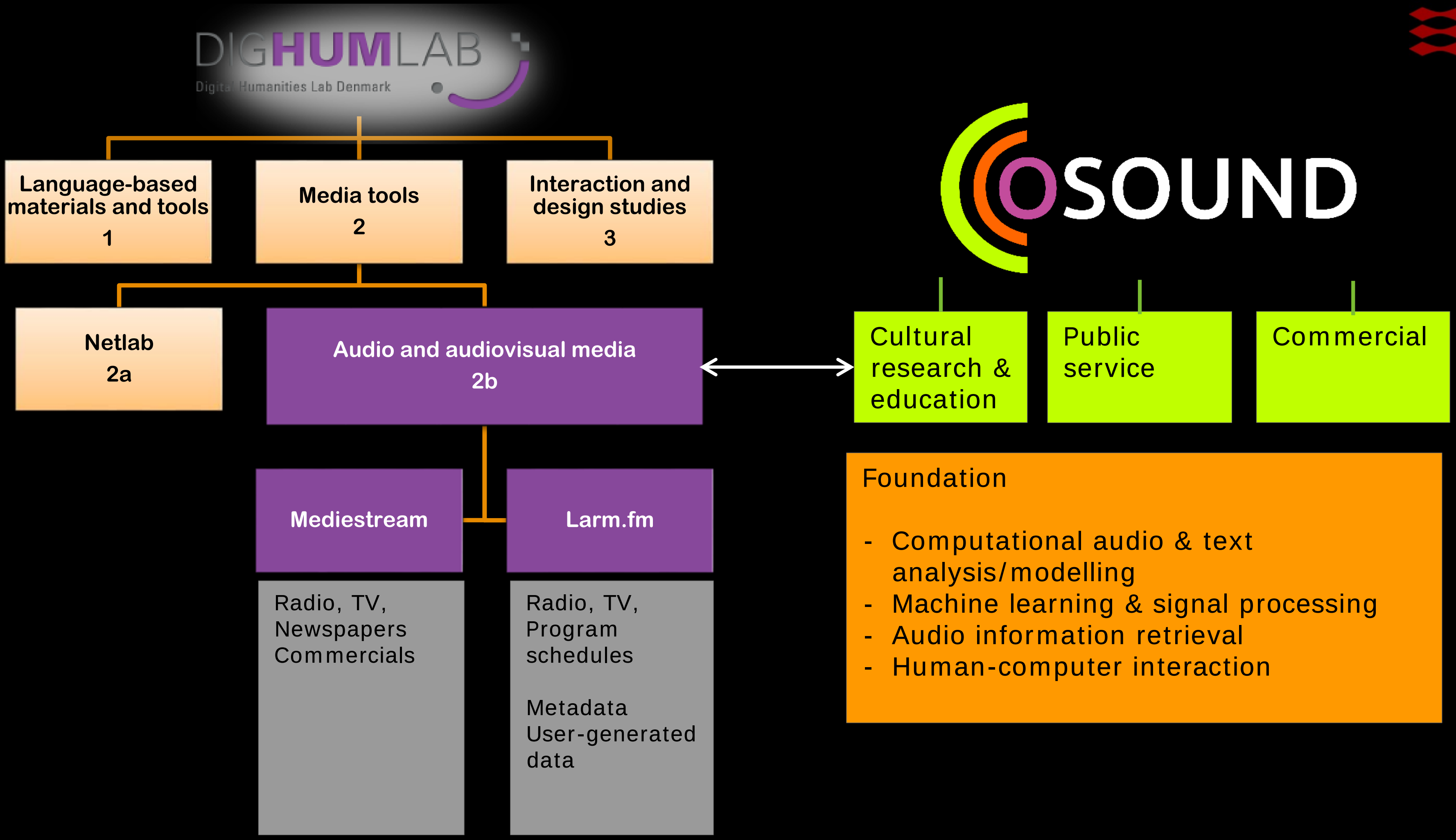
Users

Computational representations and optimal interactions

Objects

(audio & text)

The main hypothesis is that the integrating of bottom-up data, derived from audio streams, and top-down data streams, provided by users, will enable leaned and actionable semantic representations, which will positively impact and enrich user interaction with massive audio archives, as well as facilitating new commercial success in the Danish sound technology sector.



DIGHUMLAB

Digital Humanities Lab Denmark

Language-based materials and tools 1

Media tools 2

Interaction and design studies 3

Netlab 2a

Audio and audiovisual media 2b

Mediestream

Larm.fm

Radio, TV,
Newspapers
Commercials

Radio, TV,
Program schedules

Metadata
User-generated data

COSOUND

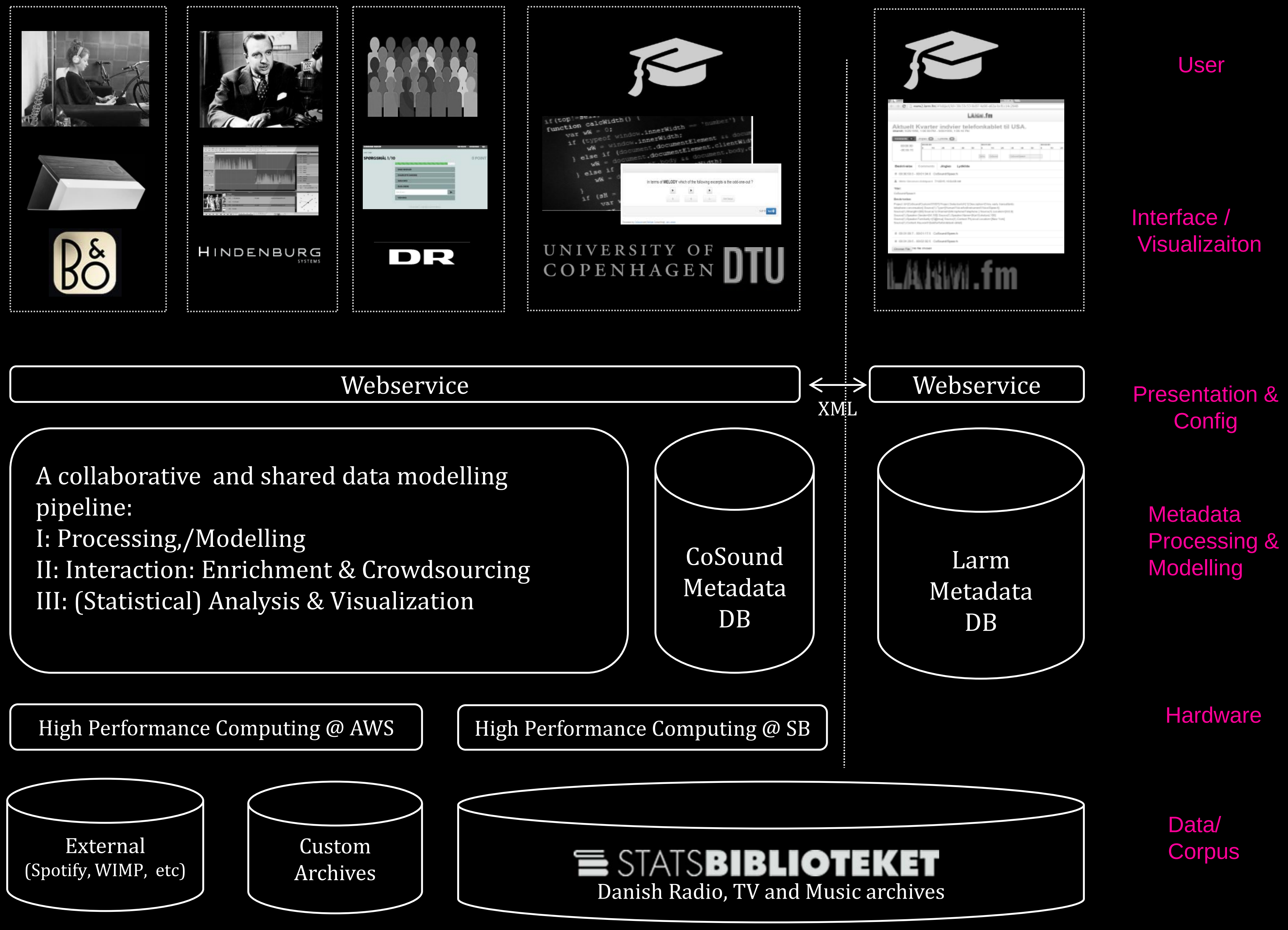
Cultural research & education

Public service

Commercial

Foundation

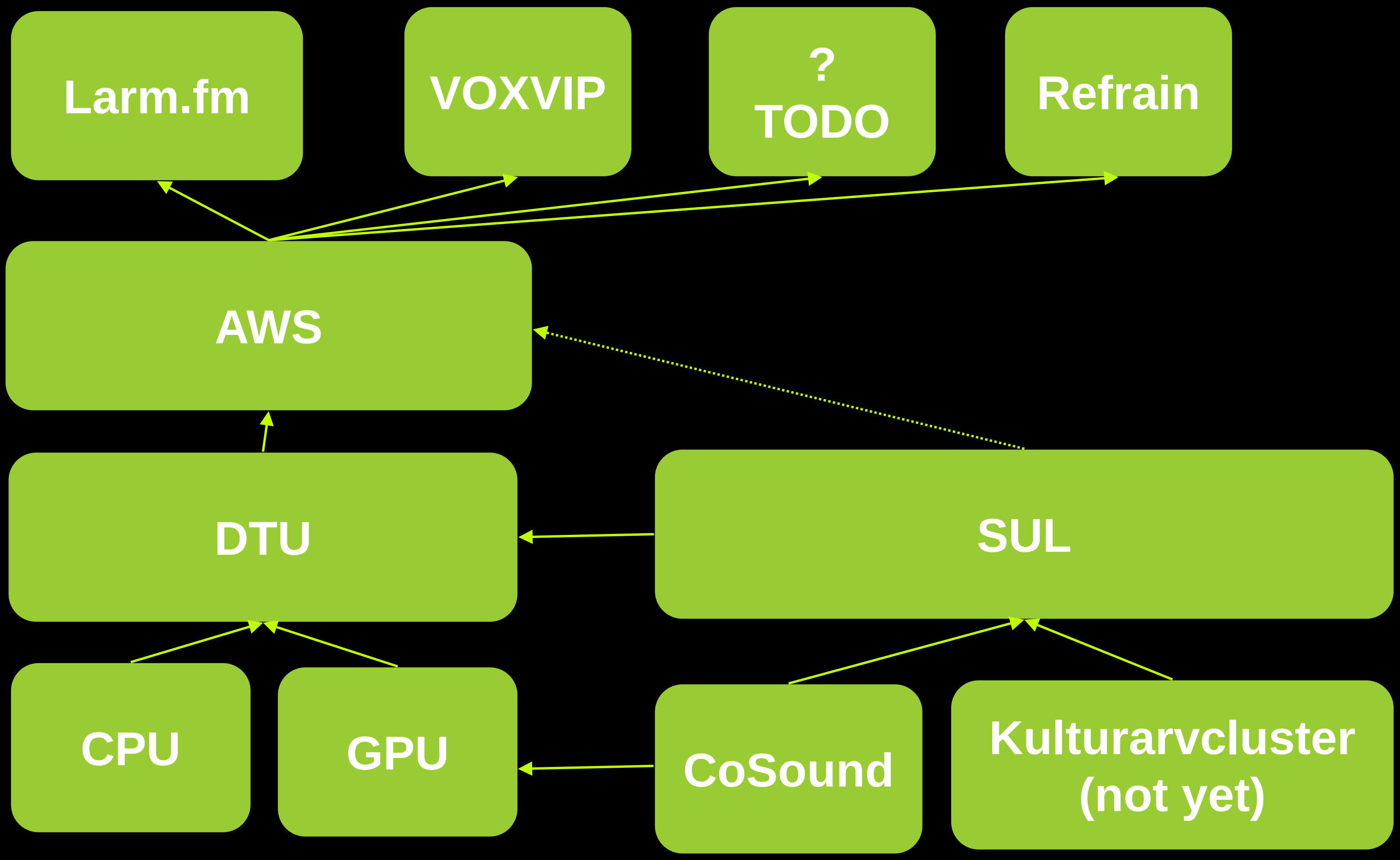
- Computational audio & text analysis/modelling
- Machine learning & signal processing
- Audio information retrieval
- Human-computer interaction



A collaborative and shared data modelling pipeline:
I: Processing,/Modelling
II: Interaction: Enrichment & Crowdsourcing
III: (Statistical) Analysis & Visualization

CoSound Metadata DB
Larm Metadata DB
Metadata Processing & Modelling
High Performance Computing @ AWS
High Performance Computing @ SB
Hardware
External (Spotify, WIMP, etc)
Custom Archives
STATSBIBLIOTEKET
Danish Radio, TV and Music archives
Data/ Corpus

CoSound Computing Infrastructure



The CoSound hardware @ SUL

Established in 2012/13

Purpose: Archive analysis at SUL

8 X Blade servers

- Centos 6.4
- 96GB ram per server
- 2 cpu w/6 cores pr. cpu
- 1Gbit network access to archive
- Que system: Octopus
  - Custom, polling based (due to DRM and SUL policies)
- Execution:
  - Plugin based, pre-approval

The CoSound hardware @DTU

- Algorithm Development
- Split processing of archive material on GPU cluster

- 1400 +972 Std Cores with a total of 200TB ram
- 8 + 24 GPUs
- Que system: Torque
- Scientific Linux 6.4 / Ubuntu

GBAR (general purpuse):
45 x Huawei XH620 V3
2x Intel Xeon Processor 2660v3 (10 core)
128 GB memory
FDR-Infiniband
1 TB-SATA disk

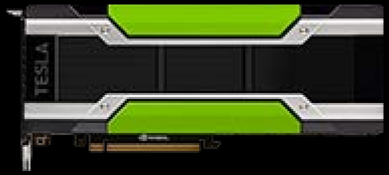
42 x IBM NeXtScale nx360 M4 nodes
2x Intel Xeon Processor E5-2680 v2 (ten-core, 2.80GHz, 25MB L3 Cache)
128 GB memory
QDR Infiniband interconnect
500 GB internal SATA (7200 rpm) disk for OS and applications

64 x HP ProLiant SL2x170z G6 nodes
2x Intel Xeon Processor X5550 (quad-core, 2.66 GHz, 8MB L3 Cache)
24 GB memory
QDR Infiniband interconnect
500 GB internal SATA (7200 rpm) disk for OS and applications

4 x HP ProLiant SL390s G7 nodes – GPGPU
2x Intel Xeon Processor X5650 (six-core, 2.66GHz, 12MB L3 Cache)
2x Tesla S2050 GPUs
48 GB memory
QDR Infiniband interconnect

DTU Compute nodes
27 x Huawei XH620 V3
•2x Intel Xeon Processor 2660v3 (10 core)
•128 GB memory
•FDR-Infiniband
•1 TB-SATA disk

21 nodes each equipped with:
•2 Sockets – 8 Core Intel Xeon E5-2665 2.4GHz – HP ProLiant SL230s G8
•64GB RAM
•500 GB internal SATA (7200 rpm) disk for OS and applications
•QDR-Infiniband
6 nodes each equipped with:
•2 Sockets – 8 Core Intel Xeon E5-2665 2.4GHz – HP ProLiant SL230s G8
•256GB RAM
•500 GB internal SATA (7200 rpm) disk for OS and applications
•QDR-Infiniband



DTU Compute (CogSys) for machine learning

6 x nodes:
- 64 GB memory
- Linux
- 4 Tesla or K40 GPUs (total 24 GPUs)

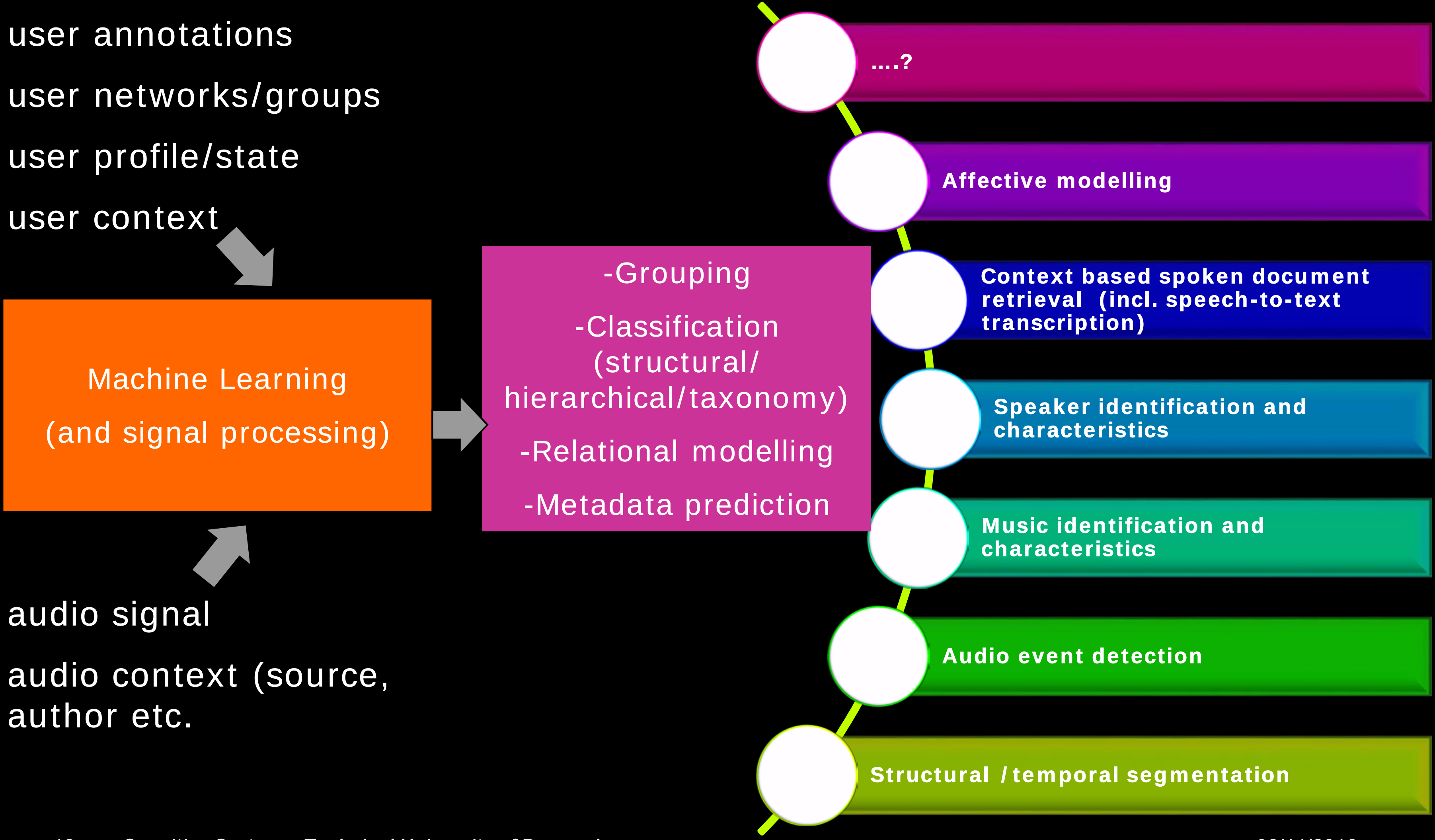
The CoSound hardware @AWS

Front-end

- Webservers
- Databases
- HPC nodes for low latency model-based interaction
- Ad-hoc, elastic for specific applicaitons (e.g. Refrain)

CoSound level 1: Processing, Modelling & Prediction

What, when, where, who, to whom… and how?



user annotations

user networks/groups

user profile/state

user context

Machine Learning

(and signal processing)

audio signal

audio context (source, author etc.

-Grouping

-Classification (structural/ hierarchical/taxonomy)

-Relational modelling

-Metadata prediction

….?

Affective modelling

Context based spoken document retrieval (incl. speech-to-text transcription)

Speaker identification and characteristics

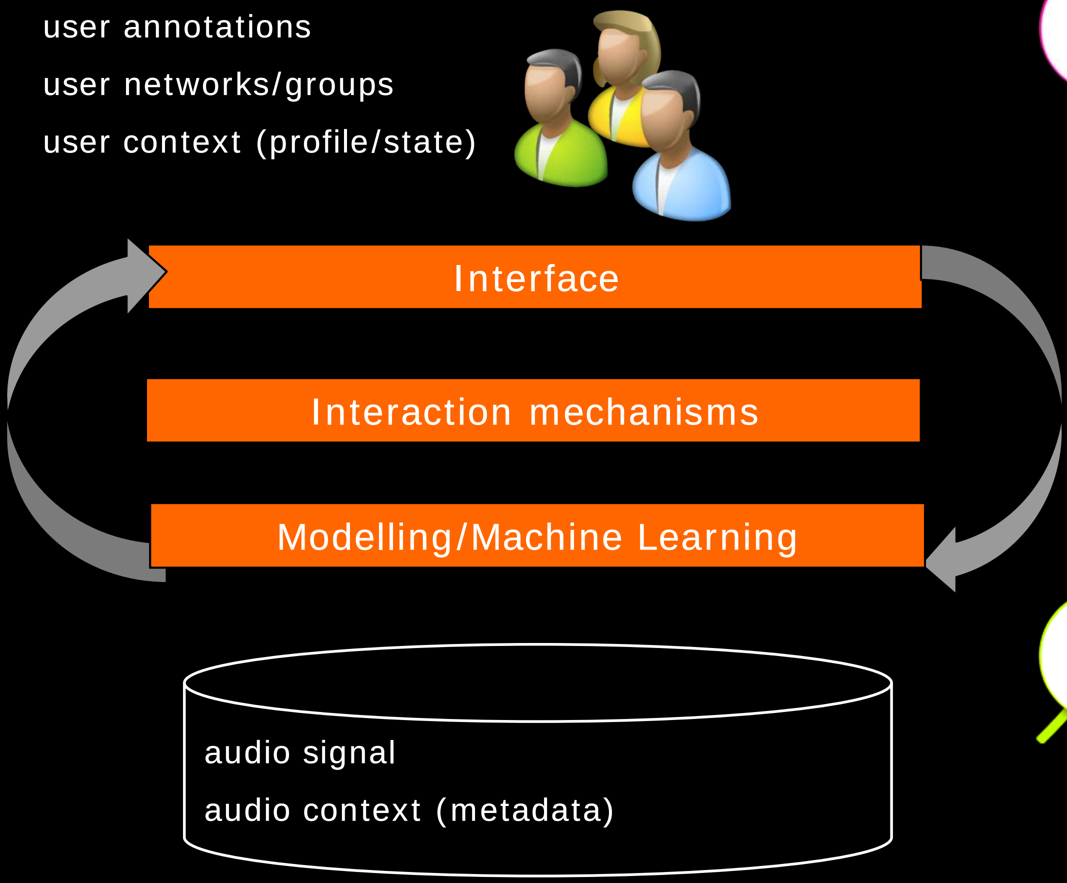
Music identification and characteristics

Audio event detection

Structural / temporal segmentation

CoSound Level 2: Model-based interaction - users in the loop

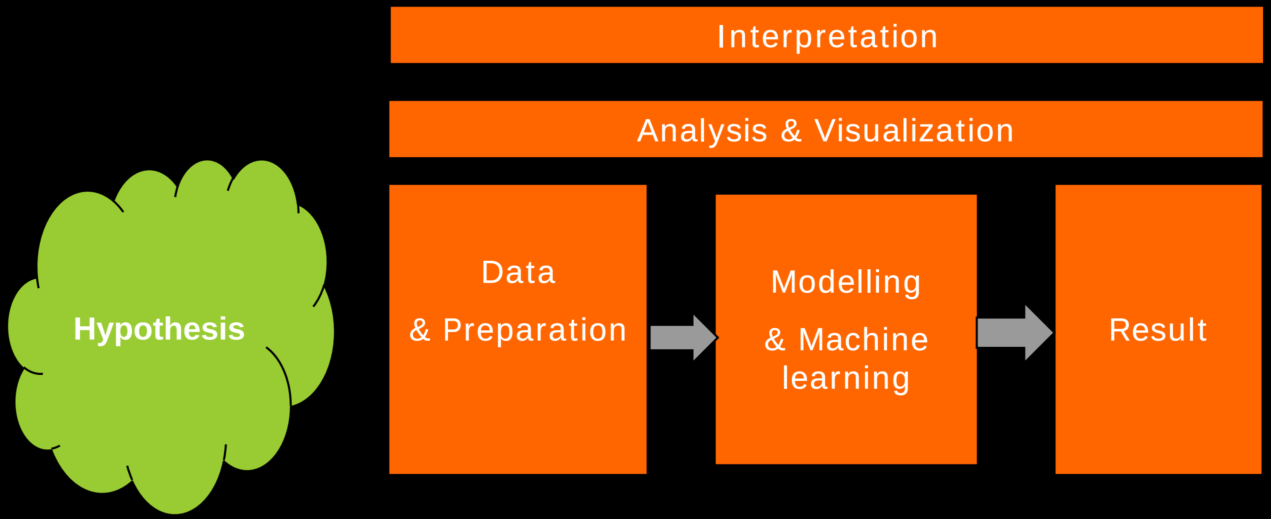
...for dissimination, enrichment, discovery



- **Modular interaction and experimentation (generic UI components, easy configuration via webservice)**
- **Crowdsouring (public/community / experts)**
- **Controlled experiments (public/community / experts)**
- **Optimal experimental design**
- **Sequential experimental design - active learning**

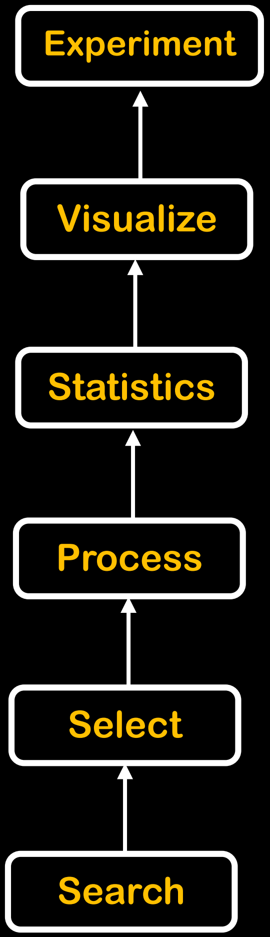
CoSound level 3: Analysis, visualization & interpretation

The missing link/tools…?



- **Visualization**
- **Statistical hypothesis testing**
- **Performance evaluation (generalization)**
- **Robustness / complexity**
- **Exploration and hypothesis refinement/formulation**

Big data tools for research



Experiment

Visualize

Statistics

Process

Select

Search

Create experiments to acquire data

Visualize results for researcher

Summarize data using statistics (summary, longitudinal, etc.)

Use big audio data tools (segmentation, features, speaker ID, ASR, …)

Selection and curation of material (multiple searches, …)

Find material (longitudinal, specific, fuzzy, etc.)

CoSound Research Projects

- Structual segmentation and grouping [technical/humanities]
- Music analysis using computational methods [digital humanities]
- Music affect/emotion prediction [technical, music perception]
- Multi-modal music similarity [technical, music perception]
- Radio genre modelling and prediction [technical/humanities]
- Phone voice detection [technical/humanities]

- Speaker identificaiton and modelling [technical]

- Transcription & topic modelling [technical]

What is metadata?

Unlimited information to be extracted about each audio stream and across the archive

Objective

Audio type? (Segmentation)
Who is talking? (Speaker ID)
What is being said?
What are they talking about?

Subjective

Does it sound happy?
Do you like what they are saying?
Does it sound good?
Which clip do you prefer?

How can meta information be created?

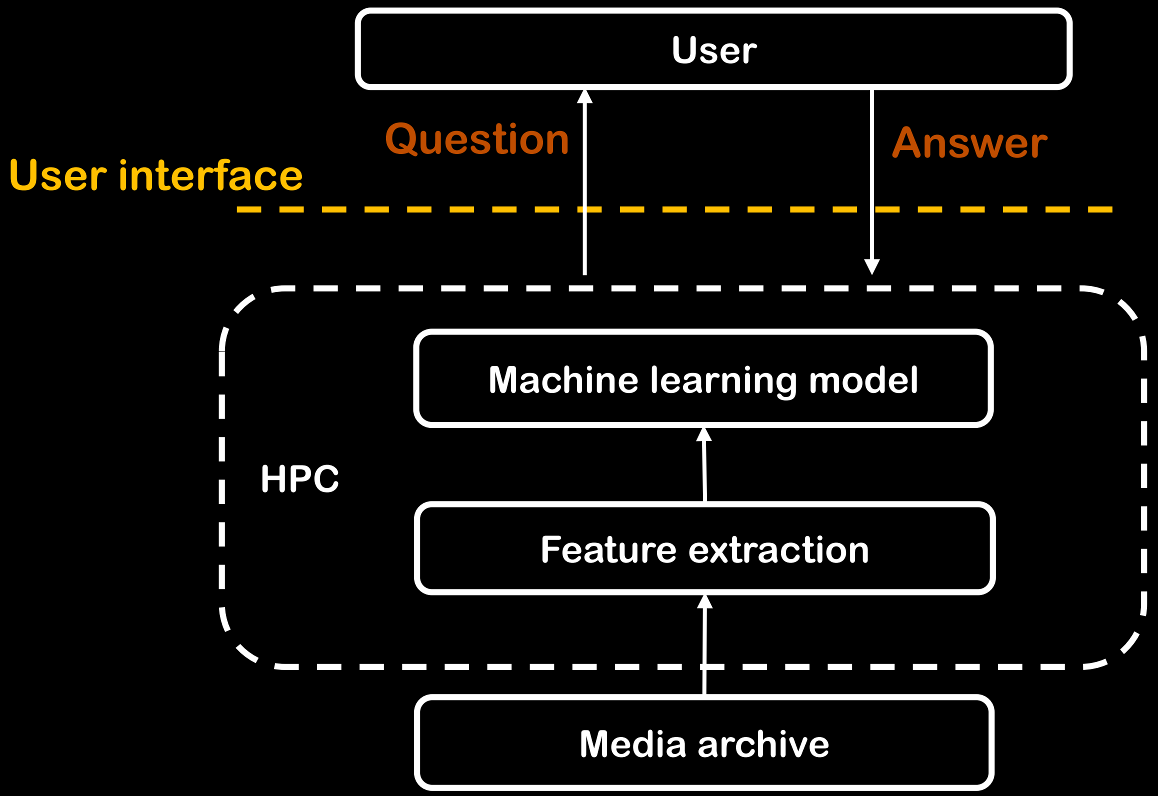
Lack of specific annotations requires prior knowledge

Manual annotation is limited or impossible due to the size of the archive, human resources, or annotators qualifications.

Semi-automatic machine learning can be used to predict information in the ensure archive based on limited number of annotations.

Smart crowdsourcing exploits machine learning to predict information in the entire archive based on ‘crowd annotators’ annotations. The individual clip is selected based on uncertain information about the label, the annotators’ qualifications and engagement based on active learning mechanisms.

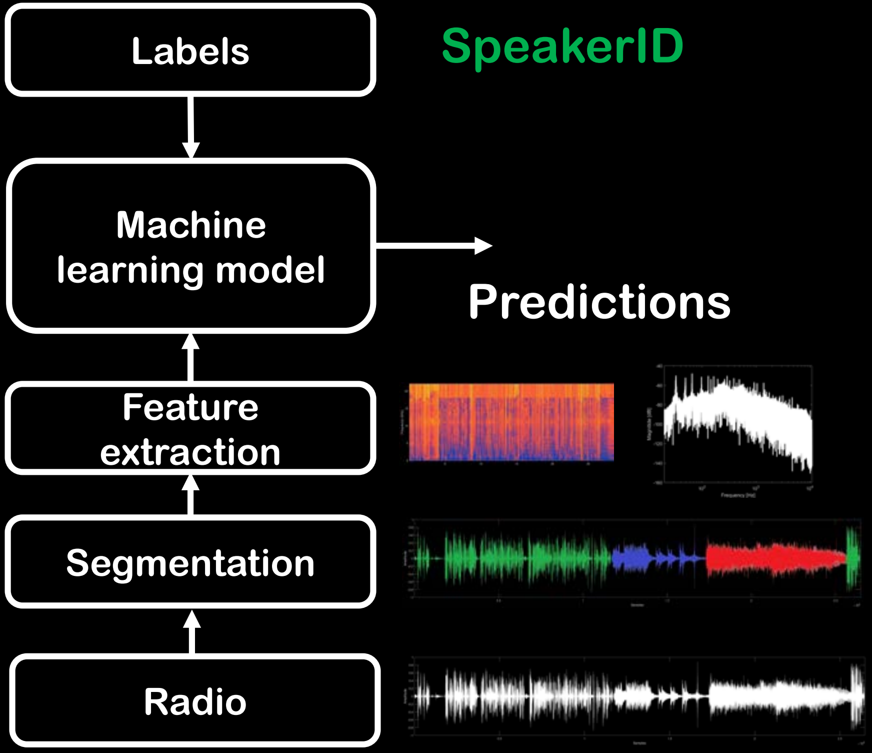
~~Traditional~~ modelling
Interactive



HPC is required to do real-time interaction with complex audio objects...

WHO'S TALKING?

Traditional speaker identification model



Labels

SpeakerID

Machine learning model

Predictions

Feature extraction

Segmentation

Radio

Crowdsourcing

- Crowdsourcing is a type of participative online activity in which one proposes to a crowd the voluntary undertaking of a task.
- The crowd has varying knowledge, heterogeneity, and number.
- The task has variable complexity and modularity in which the crowd should engage
- The crowd brings their work, money, knowledge and/or experience and always entails mutual benefit.

Estellés-Arolas, Enrique; González-Ladrón-de-Guevara, Fernando (2012), "Towards an Integrated Crowdsourcing Definition“, Journal of Information Science 38 (2): 189–200.

Crowdsourcing challenges

- Varying quality of annotations (variance)
- Varying quality of annotators (bias)
- What should be rated?
- How can we make crowdsourcing fulfil the needs of the crowd and still get information?

Smart crowdsourcing

- Smart crowdsourcing – combining machine learning and gamification

Gamification

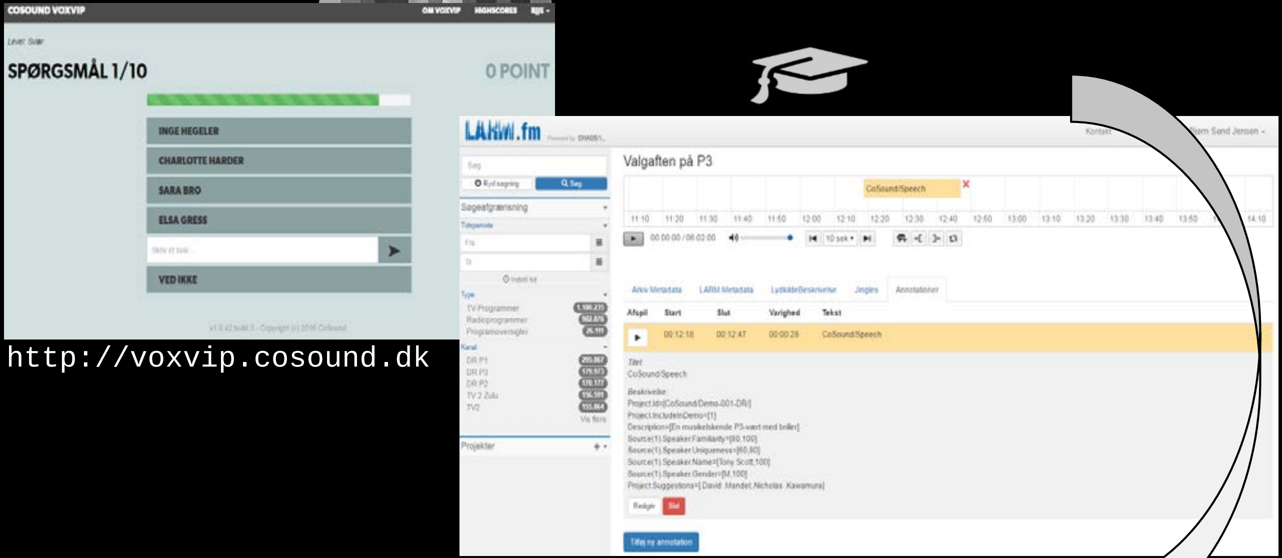
The application of game-design elements and game principles in non-game contexts.

Gamification employs game design elements to improve user engagement, productivity, flow, learning, ease of use, and usefulness.

Active machine learning

Create a probabilistic machine learning model that can predict e.g. who is talking in a clip

Use the models uncertainty about who is speaking in other clips to select candidates for annotation



COSOUND VOXVIP · OM VOXVIP · HIGHSCORES

Level: Svær

SPØRGSMÅL 1/10 · 0 POINT

- INGE HEGELER
- CHARLOTTE HARDER
- SARA BRO
- ELSA GRESS
- Skriv et svar...
- VED IKKE

http://voxvip.cosound.dk

LARM.fm

Søg · Ryd søgning · Søg

Søgeafgrænsning

Tidsperiode: Fra · Til

Type: TV Programmer 1.100.235 · Radioprogrammer 902.876 · Programoversigter 26.111

Kanal: DR P1 295.867 · DR P3 179.973 · DR P2 170.177 · TV 2 Zulu 156.591 · TV2 155.864 · Vis flere

Projekter

Valgaften på P3

CoSound/Speech

11:10 · 11:20 · 11:30 · 11:40 · 11:50 · 12:00 · 12:10 · 12:20 · 12:30 · 12:40 · 12:50 · 13:00 · 13:10 · 13:20 · 13:30 · 13:40 · 13:50 · 14:10

00:00:00 / 06:02:00 · 10 sek

Arkiv Metadata · LARM Metadata · LydkildeBeskrivelse · Jingles · Annotationer

Afspil	Start	Slut	Varighed	Tekst
▶	00:12:18	00:12:47	00:00:28	CoSound/Speech

Titel

CoSound/Speech

Beskrivelse

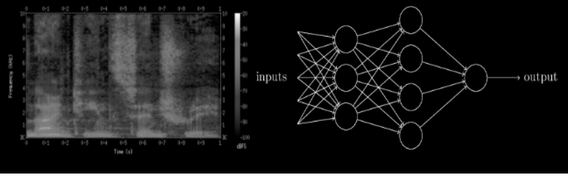
Project.Id=[CoSound/Demo-001-DR/]
Project.IncludeInDemo=[1]
Description=[En musikelskende P3-vært med briller]
Source(1).SpeakerFamiliarity=[80,100]
Source(1).SpeakerUniqueness=[60,80]
Source(1).SpeakerName=[Tony Scott,100]
Source(1).SpeakerGender=[M,100]
Project.Suggestions=[,David ,Mandel,Nicholas ,Kawamura]

Rediger · Slet

Tilføj ny annotation

Webservice(s)

- Speaker modelling / analysis
- User modelling / analysis
- Interaction/ gamification
- Visualization



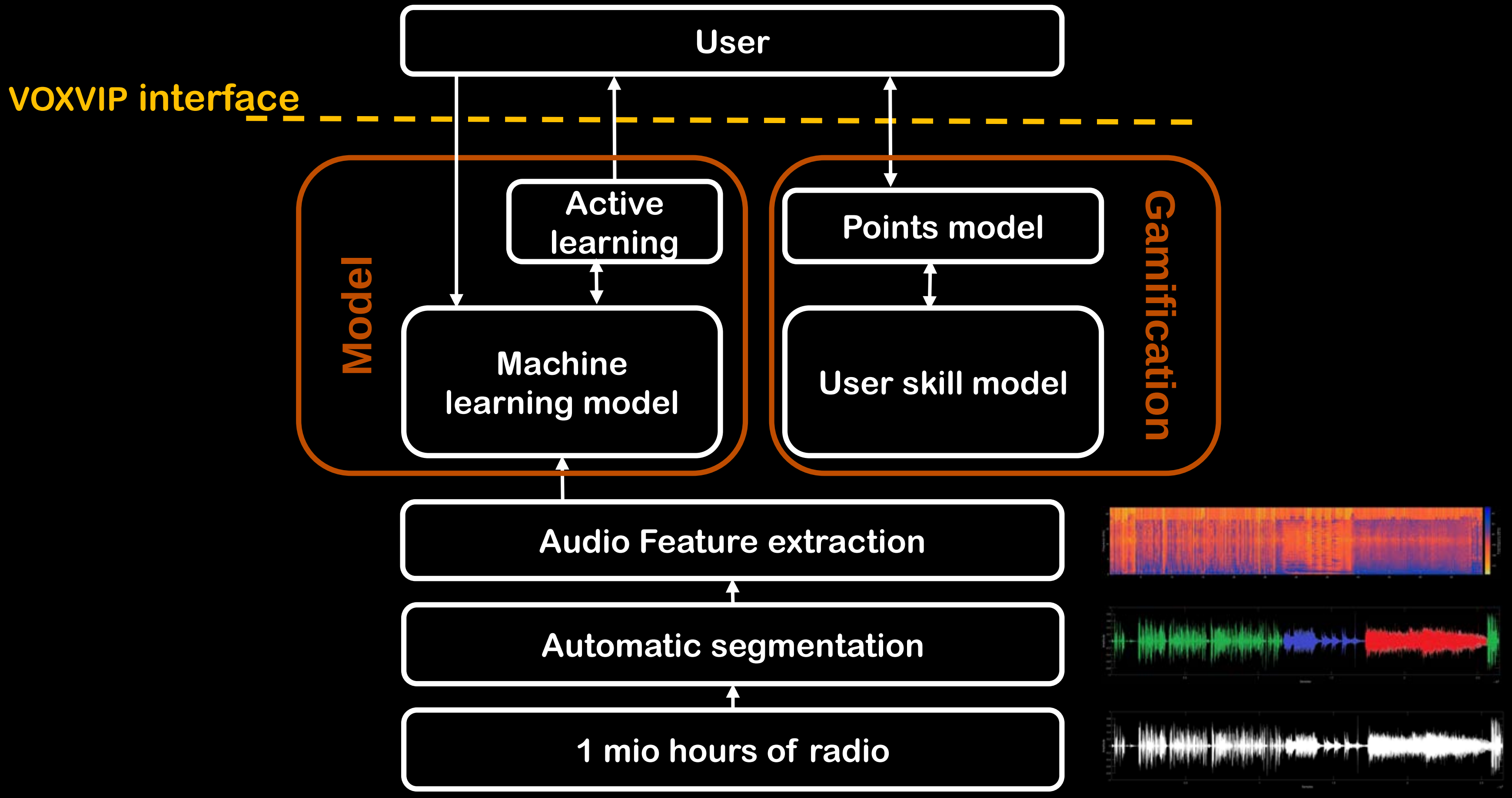
High Performance Computing @ SB + AWS

STATSBIBLIOTEKET

Danish radio archives

Currently more than 200 known and unknown speakers in 1000+ segments from 1963 to 2012

VOXVIP model



User

VOXVIP interface

Model

Active learning

Machine learning model

Points model

User skill model

Gamification

Audio Feature extraction

Automatic segmentation

1 mio hours of radio

Technical research questions

- Are model-based active learning mechanisms suitable for smart crowdsourcing?
- Is optimal performance wrt. time used achieved?
- Is age, sex or position relevant for recognition of specific voices?
- Gamification: How does levels, difficulty and point assignment influence the quality and quantity of annotations?

Conclusion VOXVIP

VOXVIP - Version 1

- 500 people have played VOXVIP
- We have identified 200 VIP people

VOXVIP - Version 2

- Speakers > 3000
- Sound clips > 10.000

- We are currently segmenting ~1 mio. hours of audio (takes a lot of CPU/GPU time)

- Building custom visualization front-ends to end-users.

Transcription: What are people talking about?

< 200ms
(for research < 1s)

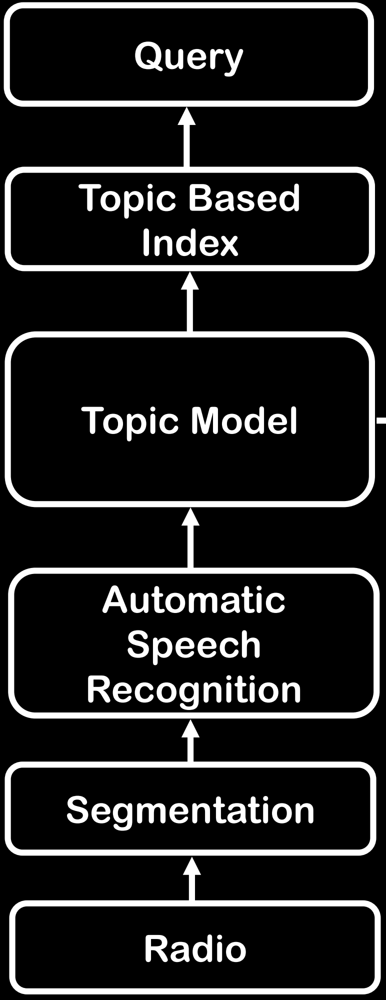
20h (building/updating – selection)

1/100 x real-time
(training, model-selection)

2-4 x real-time
(large vocab, adaption, Danish)

½-1 x real-time

Real-time=30.000h



Question: can higher-level content meta data improve searchability and information retrieval

Topics

This is the story of a little man that couldn't walk but really wanted to. Although he had thought about it, it never occurred he should just try.