CLARIAH
Common Lab Research Infrastructure
for the Arts and Humanities

CLARIN

Increasing the impact of OH with HLT

.

How can advanced HLT help to increase the findability and (re-)usability of OH-collections in scholarly work?

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

UNIVERSITY OF TWENTE.

I'm a computer scientist, Michael. That means that I can't explain about people, their motivation or their dreams. But if you want to know how to search for "words", how to combine "themes" or how to "access recordings", I'm your man.

Making sense of digital spaghetti

CLARIAH and Utrecht University are putting together a two-week immersion course on getting to grips with Big Data.

Location:	*Utrecht & Woudenberg*
Date:	*4-jan-2016 - 15-jan-2016*
Time:	*9:00 (welcome with coffee)*

Background

"Why Gramma, what big data you have!"

Increasingly, research in all disciplines, from the natural sciences to social sciences and humanities, involves big data. The availability of vast amounts of textual, audio-visual and structured data from digital sources is revolutionizing research in the humanities and social sciences. The most advanced scholarship in these areas, currently and in the foreseeable future, relies on the use of sophisticated tools for accessing, processing, analysing and presenting this data.

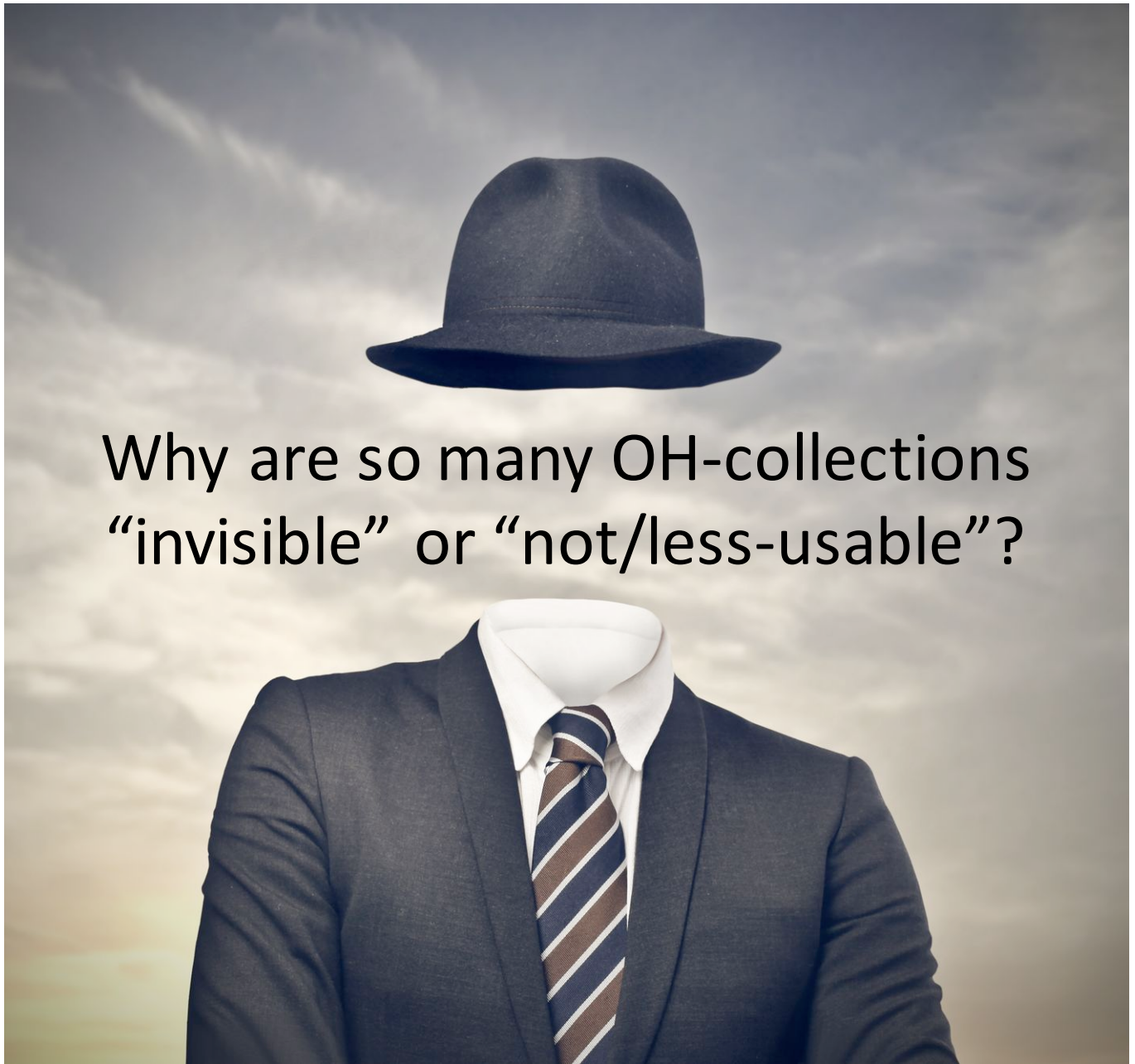
General goals

CLARIAH, together with University College Utrecht, offer a short two-week module in which we give a small group of undergraduate, graduate students, and professional researchers the opportunity to gain familiarity and experience with some common approaches to handling very large datasets. Together, we'll use a batch of current Twitter™ data as a vehicle for practicing computational thinking, and the general concepts that data analysis with computers involves.

As part of an inclusive approach to large-scale research, this module stimulates the kind of thinking that CLARIAH hopes to engender: basic programming techniques, the use of multiple paradigms to solve problems, drawing on reasoning, logic, analysis, hypothesis-testing, and formal problem-solving methods, enabling all researchers, regardless of discipline, to engage fully with their own research.

Invisibility of OH-collections

- Lack of metadata
- Non-open policy of the archives
- IPR-rules
- Not digitalized
- Only in the local language



- Lack of metadata standards
- Lack of of fine-grained descriptions
- Lack of the full recordings (*often just edited snippets*)
- Lack of context

Human Language Technology (HLT)

Automatic Speech Recognition (ASR)

Real Recognition

Alignment

Making the audio searchable

Optical Character Recognition (OCR)

"reading" the paper transcriptions

"digitizing" corresponding documents

*Landgenoten, heden morgen is **den menschheid** opgeschrikt door een brute daad*

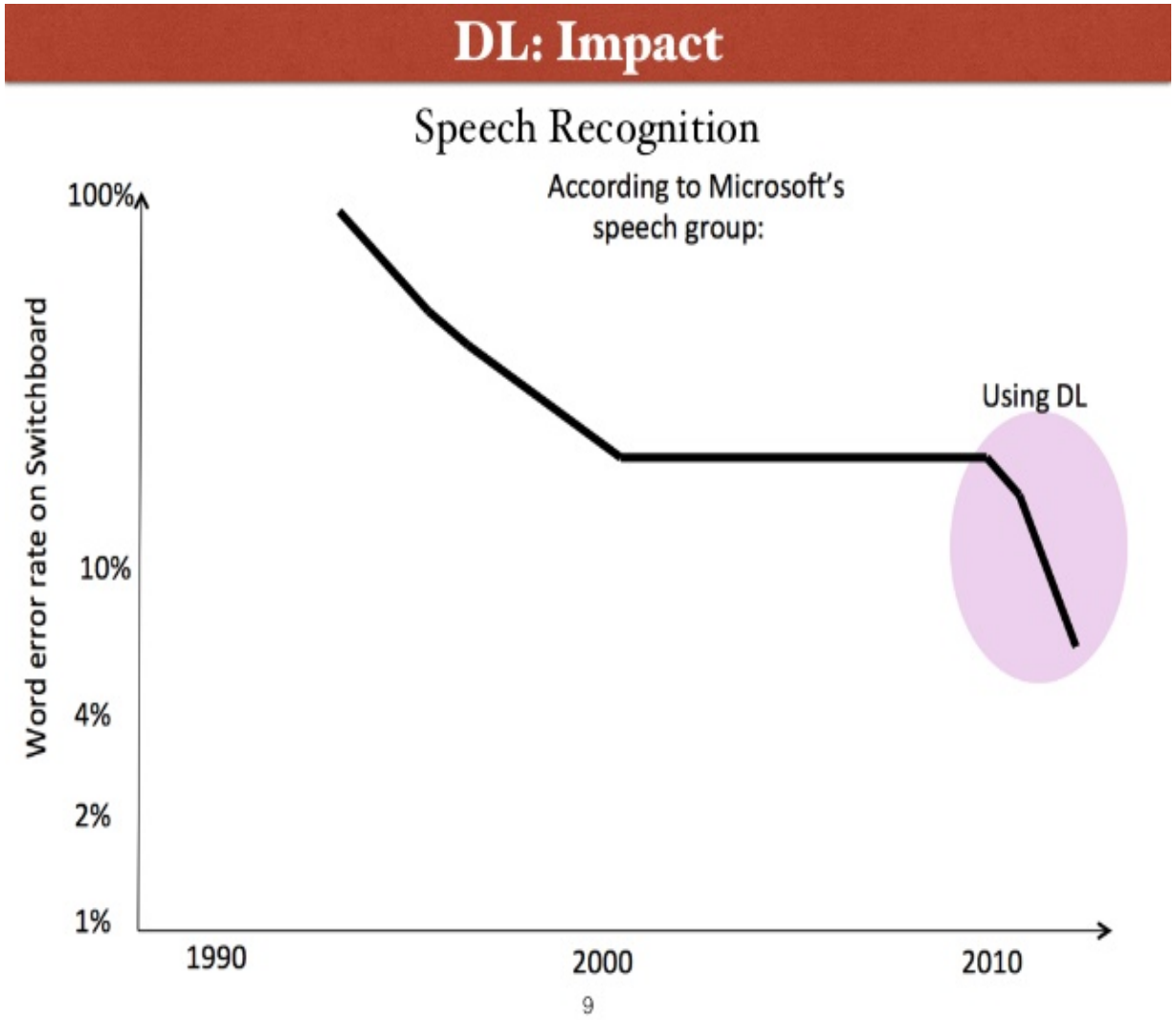
Spelling Checking

Landgenoten, vanochtend is de mensheid opgeschrikt door een brute daad

Automatic translation

Compatriots, this morning, mankind has been rocked by a brutal act

Speech Recognition



Strong improvement due to:

- Massive availability of data (speech and text)
- Unlimited computing power (cloud computing)
- Deep Learning Algorithms

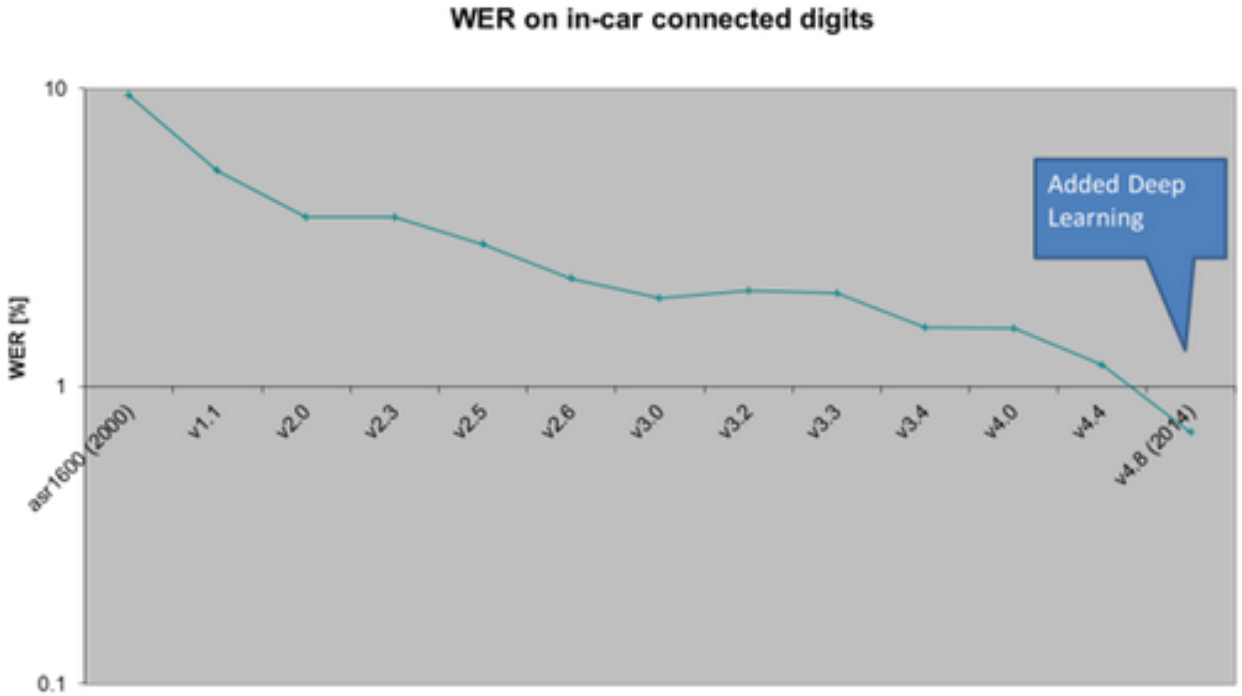
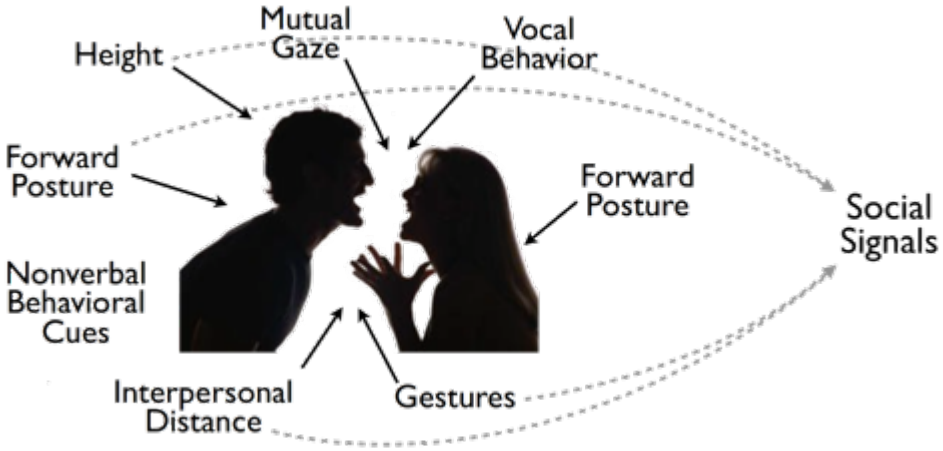


Figure 3: Exponential error rate reduction since 14 years

Human Language Technology (HLT)



Social Signal Processing (SSP)

Face

Body

Prosody

Social Cues & Meaning in the Signal

What need to be done?

CLARIN

Facilitating the useful HLT-technology

Webservice to make/ edit the metadata

Broadcasting the metadata of the OH-collections

DynamicMetadata

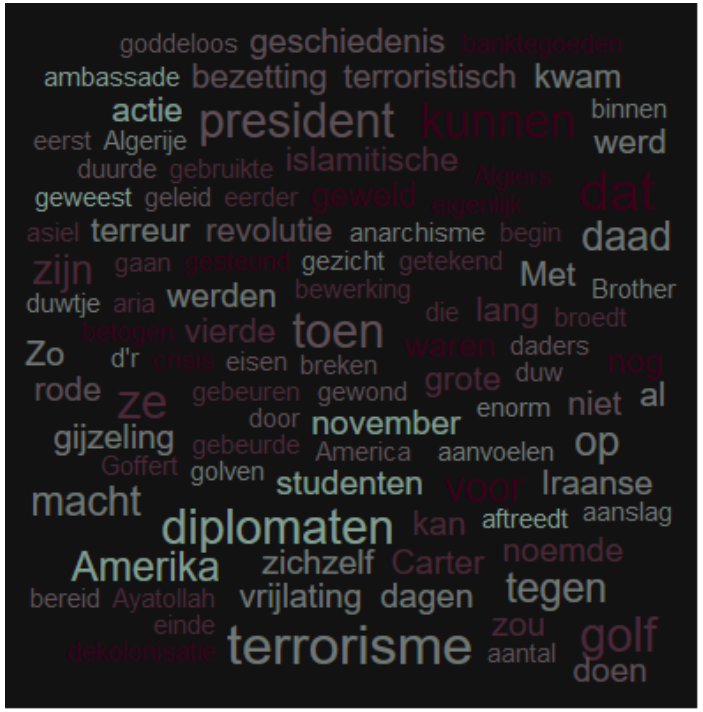
Static Metadata

Use WebVTT as the default subtitling format.

Creation of a flexible, comprehensive metadata scheme (CMDI?) with obligatory and potential fields.

WebVTT files provide captions or subtitles for video content, and also text video descriptions, chapters for content navigation, and more generally any form of metadata that is time-aligned with audio or video content.

Dynamic metadata



First 3 min

Middle 3 min

Last 3 min

TIME

Metadata

Static Metadata

Collection level

- File level
- File level
- File level
- File level
- File level

What are they talking about?
Who are talking?
Which language(s)?
How can I access it?

Topic 2 (T1 → T2)

Topic 2 (T3 → T4)

Topic 2 (T5 → T6)

Chap. 1 (T1* → T2*)

Chap. 2 (T3* → T4*)

Speaker 1 (T1” → T2”)

Speaker 2 (T3” → T4”)

Speaker 1 (T5” → T6”)

Speaker 1&2 (T7” → T8”)

?

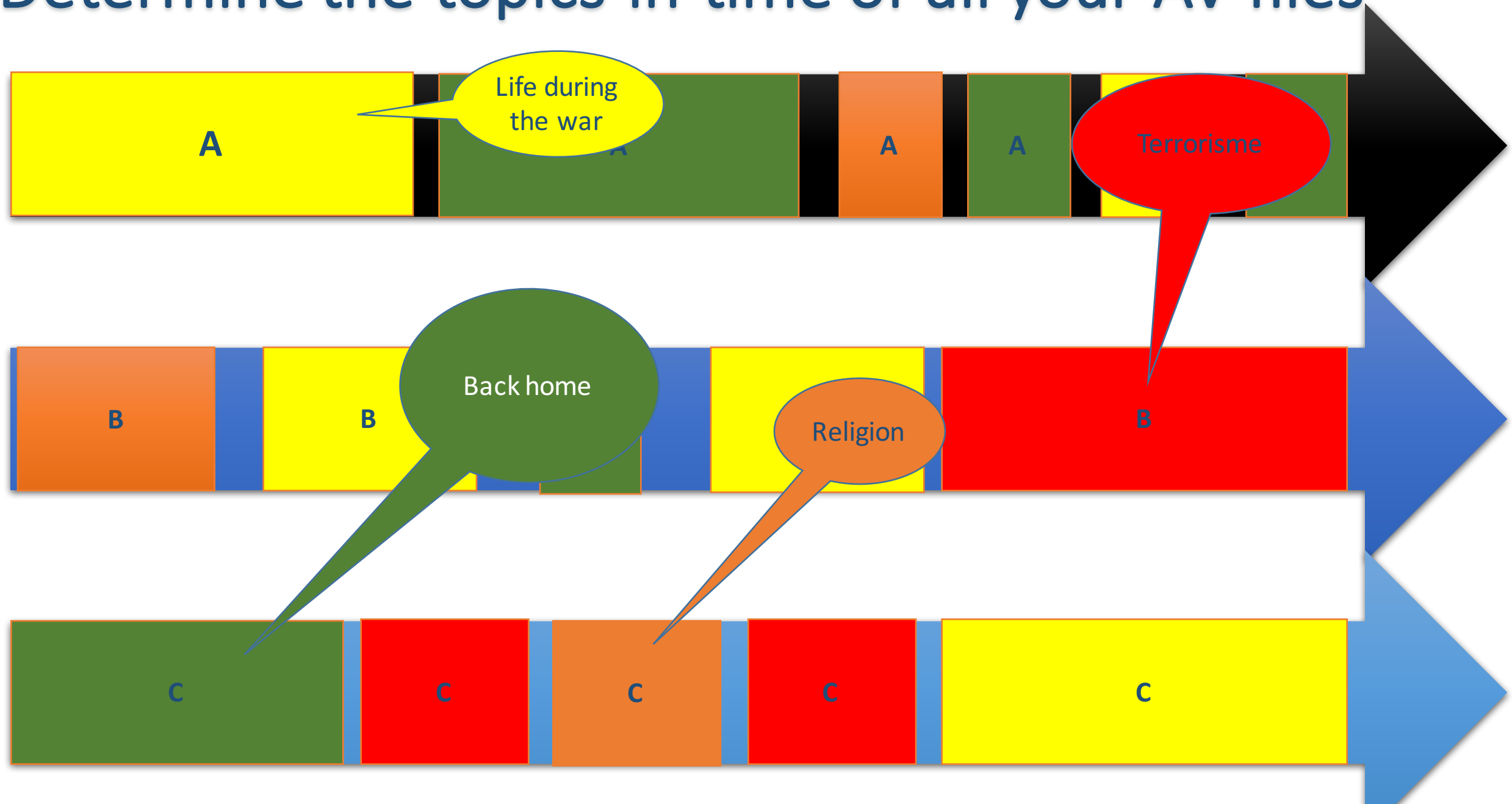
?

?

?

This was just one AV-file

Determine the topics-in-time of all your AV-files



Search

Social Justice

Topic 7

Interview X (T1-T2)
Interview X (T3-T4)
Interview Q (T5-T6)
Documentary W (T7-T8)
Discussion 12 (T10-T12)
Discussion 19 (T1-T7)
Keynote 99 (T9-T11)

Some Examples

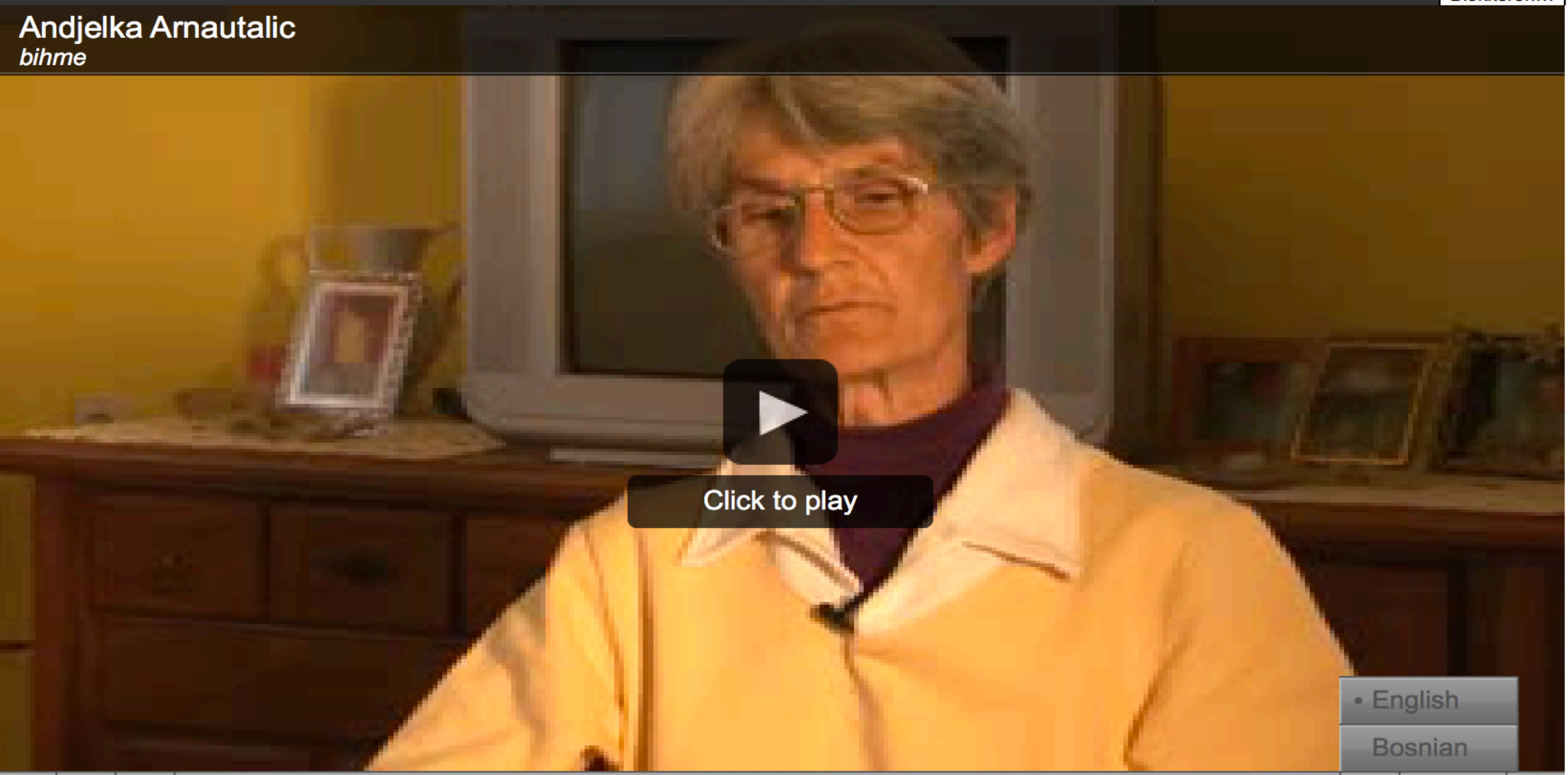
Bosnian Memories

University of the Netherlands

Blokkeren…

Andjelka Arnautalic

bihme



University of the Netherlands

Increasing the impact

ASR

Imperfect result

Crowd sourcing

Perfect result

Google Translate

Break through language barriers.

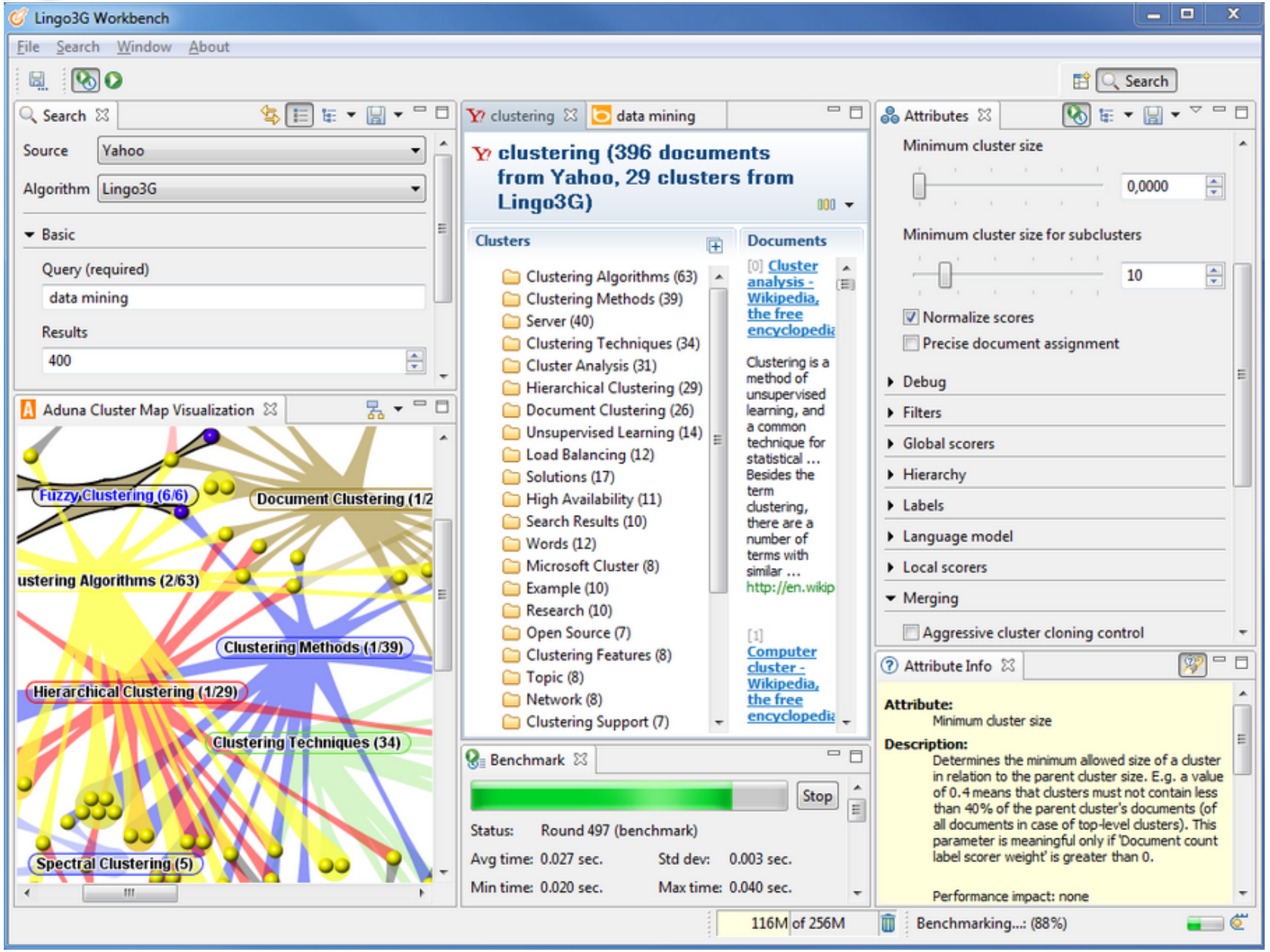
Imperfect result

Crowd sourcing

Perfect result

Clustering software

MALLET
machine learning for language toolkit



Questions?