Using Big Data for Emotionally Intelligent Mobile Services through Multi-Modal Emotion Recognition

JOHANNES KEPLER UNIVERSITY LINZ

Yerzhan Baimbetov
Ismail Khalil
Matthias Steinbauer
Gabriele Anderst-Kotsis
Institute of Telekooperation

Outline

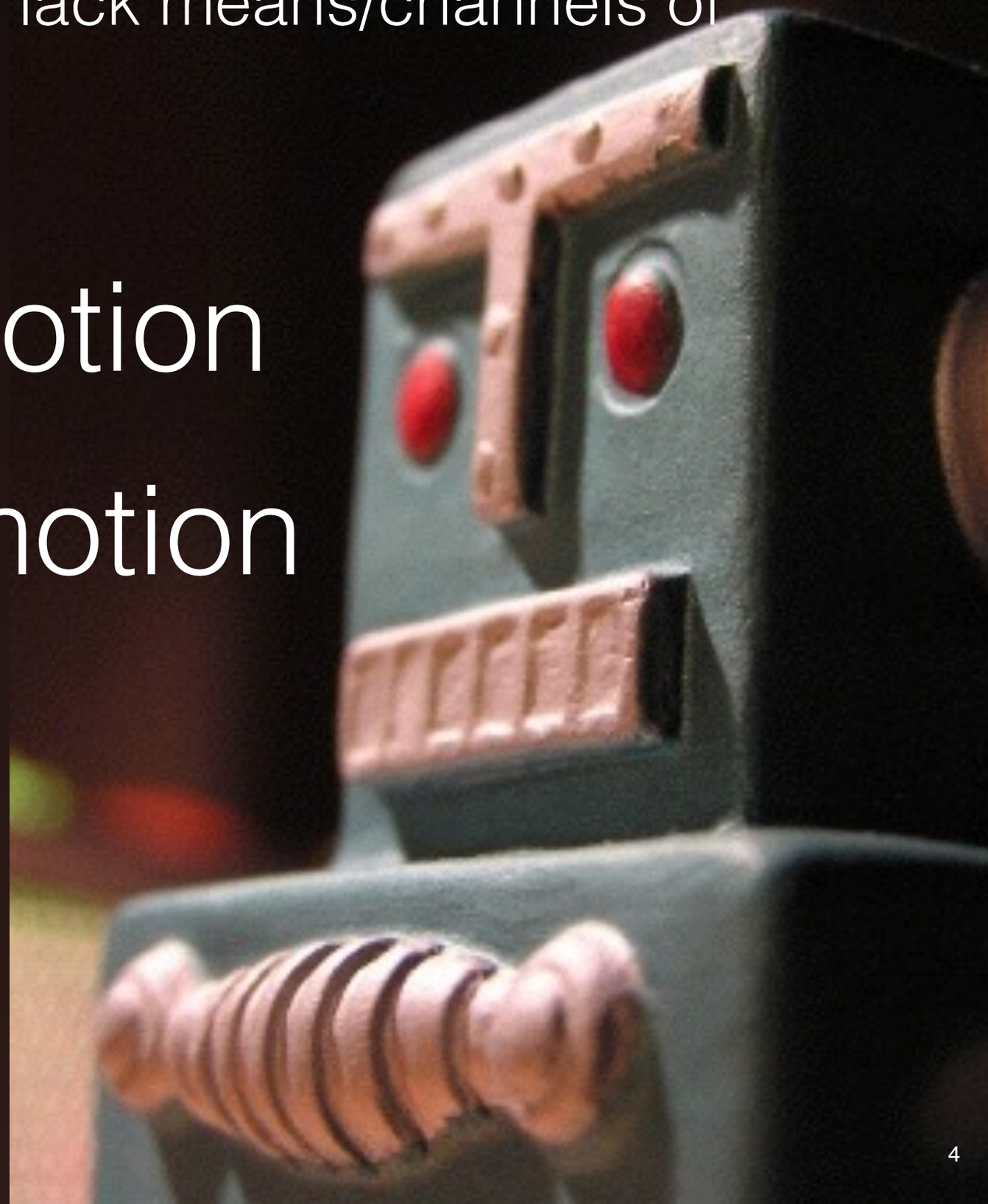
- Motivation
- Related Work
- Big Data Architecture
- Prototype Implementation
- Conclusion

human beings are highly **emotional**
emotion is a **key factor** in human-to-human interaction

Computers by default lack means/channels of

expressing emotion

interpreting emotion



Applications?

Medical

Monitoring
Treatment Decisions
Autism?

(e-)Learning

Presentation Style
Learner Attention
Better Feedback

Monitoring

Car Drivers
Angry e-Mails
ATM Disposals
Call Center

Entertainment

Virtual Reality Games
Music / Video Mood

Law

.at / Supreme Court
.de / § 244 StPO
.ch / Court Ruling

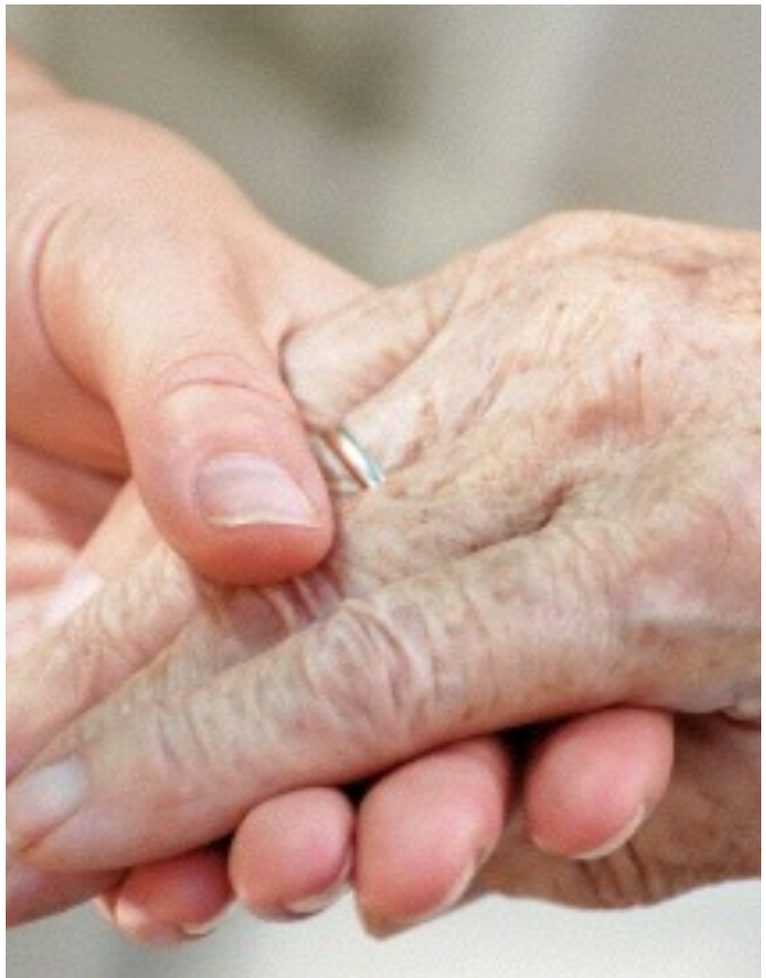
Marketing

Higher Attention >
Higher Impact
Mood Dependent

Multi-Modality of Emotions

Emotion is expressed through
a wide variety
of different
channels

facial expressions
vocal sounds
gestures
postures



Bluetooth
Proximity

Location

Photos

Text Messages
e-Mail

Acceleration

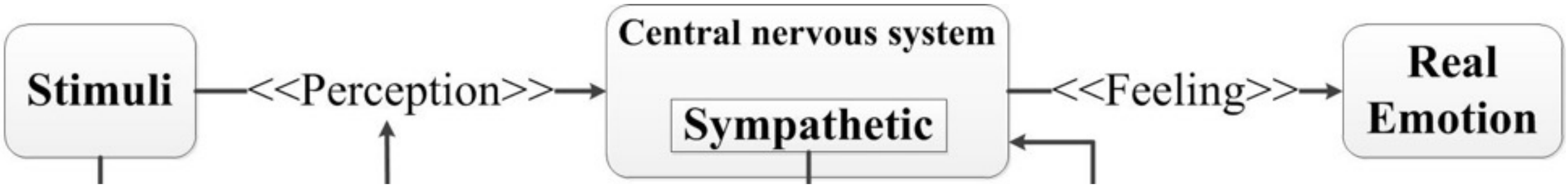
Phone Calls
Voice Messages

Videos

Sensory data can be used to

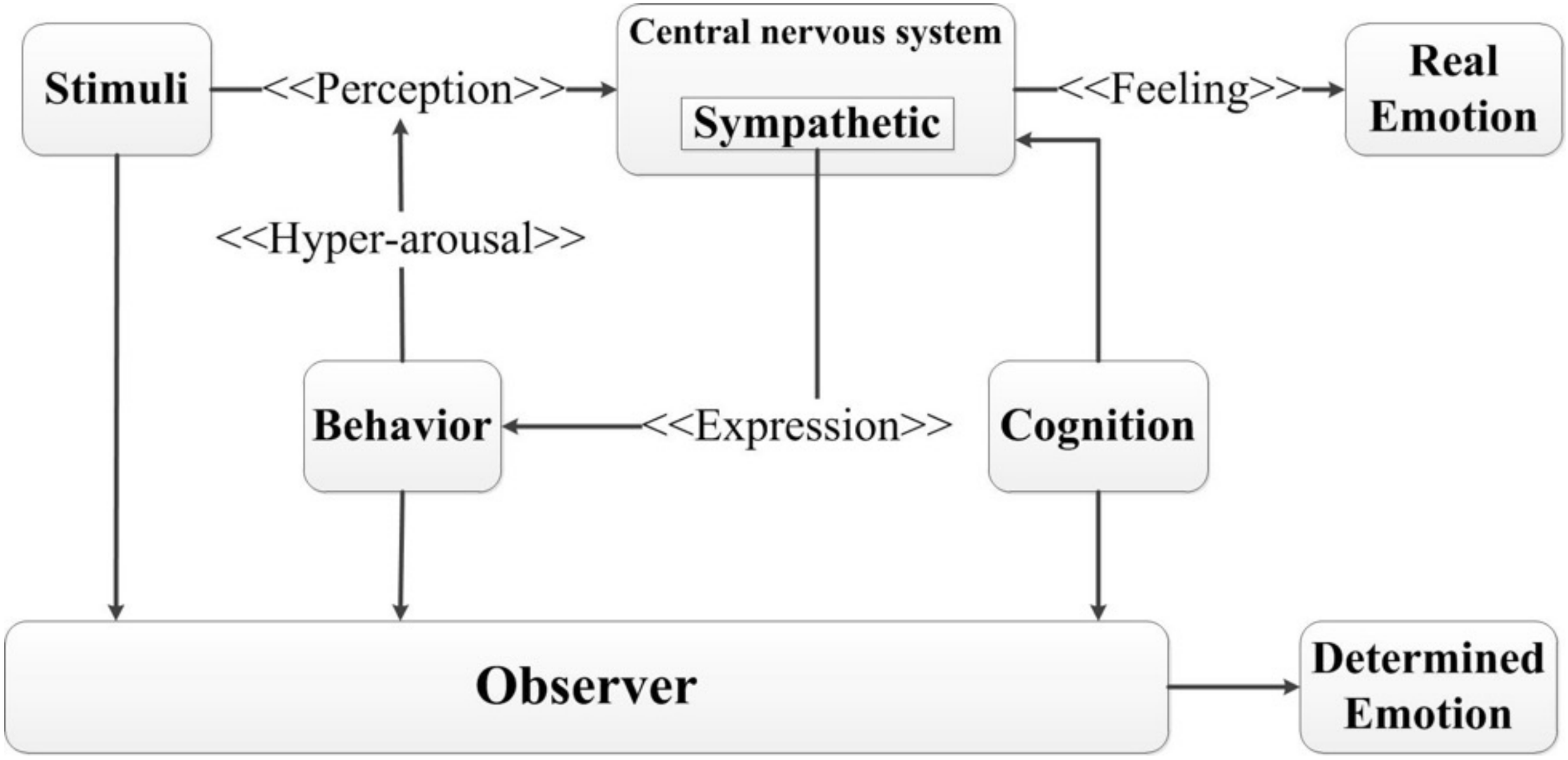
- Derive the users current **context**
  - Who is the user? With whom?
  - Where is the user? In which environment? What are the surroundings? Time of the day?
  - At home? On a bus, car, train? In the street? In the office? In a shop?
- Derive the users current **emotional state**

Two-factor theory of emotion



F. Berthelon, P. Sander, “Emotion Ontology for Context Awareness”, IEEE CogInfoComm, Vol. 31, No. 1, pp39, Jan 2009

Two-factor theory of emotion



F. Berthelon, P. Sander, “Emotion Ontology for Context Awareness”, IEEE CogInfoComm, Vol. 31, No. 1, pp. 39, Jan 2009

States of Emotion / Cross Cultures? / How many different?

happy

surprised

fearful

sad

angry

disgusted

C. Potts, “Sentiment Analysis Symposium”, San Francisco, Nov. 8-9, 2011.

P. Ekman, “Universals and Cultural Differences in Facial Expressions of Emotions”, Nebraska Symposium on Motivation, Vol 19 (1972), pp. 207-283.

$$L = ER(F_1, F_2, F_3, \ldots, F_n)$$

$$L = ER(F_1, F_2, F_3, \ldots, F_n)$$

Acoustic: pitch, intensity, jitter, shimmer, speaking rate, fundamental frequency

$$L = ER(F_1, F_2, F_3, \ldots, F_n)$$

Acoustic: pitch, intensity, jitter, shimmer, speaking rate, fundamental frequency

Facial: facial action coding system, facial movements + rule system or classifier

$$L = ER(F_1, F_2, F_3, \ldots, F_n)$$

Acoustic: pitch, intensity, jitter, shimmer, speaking rate, fundamental frequency

Facial: facial action coding system, facial movements + rule system or classifier

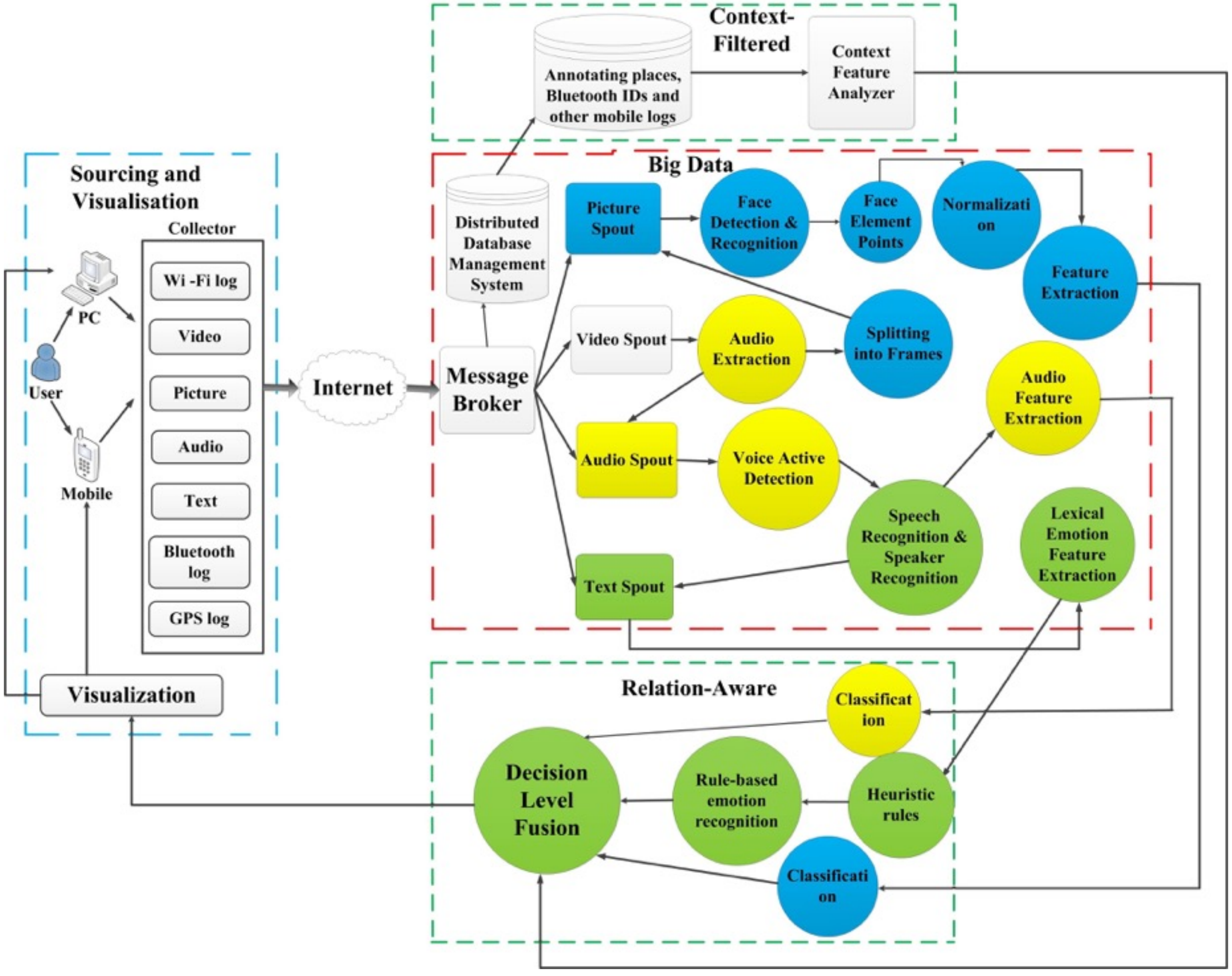
lorem ipsum

Textual: language dependent dictionary, emoticons, imprecisions / sarcasm

m ipsum
lorem ipsum
psum
EM IPSUM

Label Fusion

Global
Emotion
Label

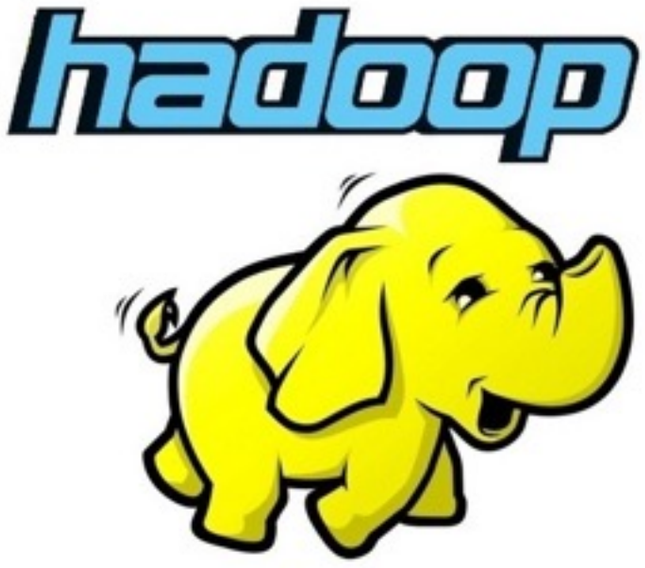

Context-Filtered
Annotating places, Bluetooth IDs and other mobile logs
Context Feature Analyzer
Sourcing and Visualisation
Collector
Wi -Fi log
Video
Picture
Audio
Text
Bluetooth log
GPS log
PC
User
Mobile
Visualization
Internet
Big Data
Distributed Database Management System
Message Broker
Picture Spout
Face Detection & Recognition
Face Element Points
Normalizati on
Feature Extraction
Video Spout
Audio Extraction
Splitting into Frames
Audio Feature Extraction
Audio Spout
Voice Active Detection
Speech Recognition & Speaker Recognition
Lexical Emotion Feature Extraction
Text Spout
Relation-Aware
Classificat ion
Decision Level Fusion
Rule-based emotion recognition
Heuristic rules
Classificati on

location
text messages, e-mail
photos, videos
voice recordings

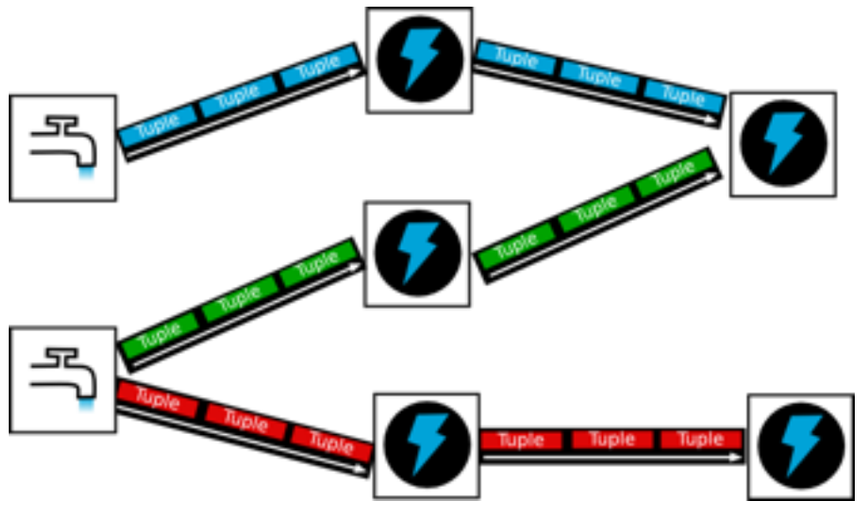
Sensor Data Retrieval

New data received from funf is

(1) stored on Hadoop DFS for distributed processing



(2) generating a processing command in Apache Storm



Pre-Processing

Video data is split into their respective video and audio streams, the video frames are split into individual frames. Those frames are fed back into the picture stream (Xuggler and FFMPEG).

On each picture frame face detection is executed. Each individual frame is fed back into the system for face feature detection (Luxand FaceSDK).

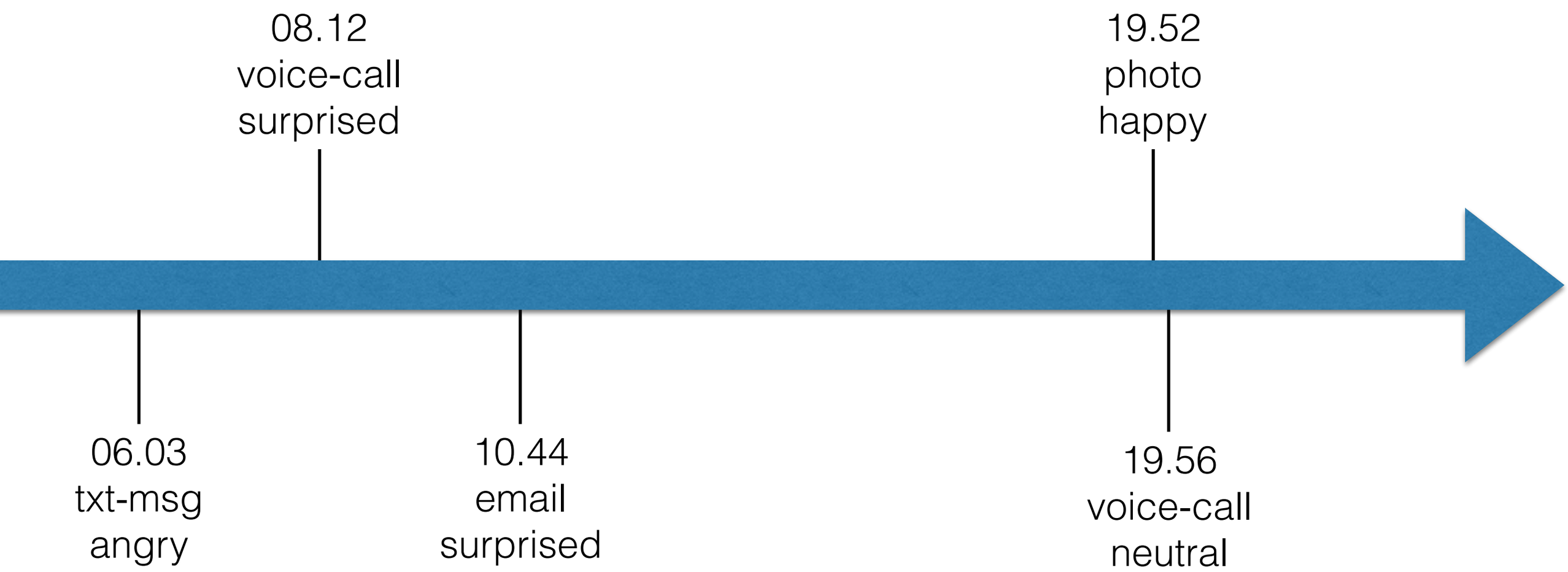
Acoustic: yet to be implemented

Facial: Using Fraunhofer SHORE Demo and eyeris Emovu API for facial emotion recognition

m ipsum
lorem ipsum
psum
EM IPSUM

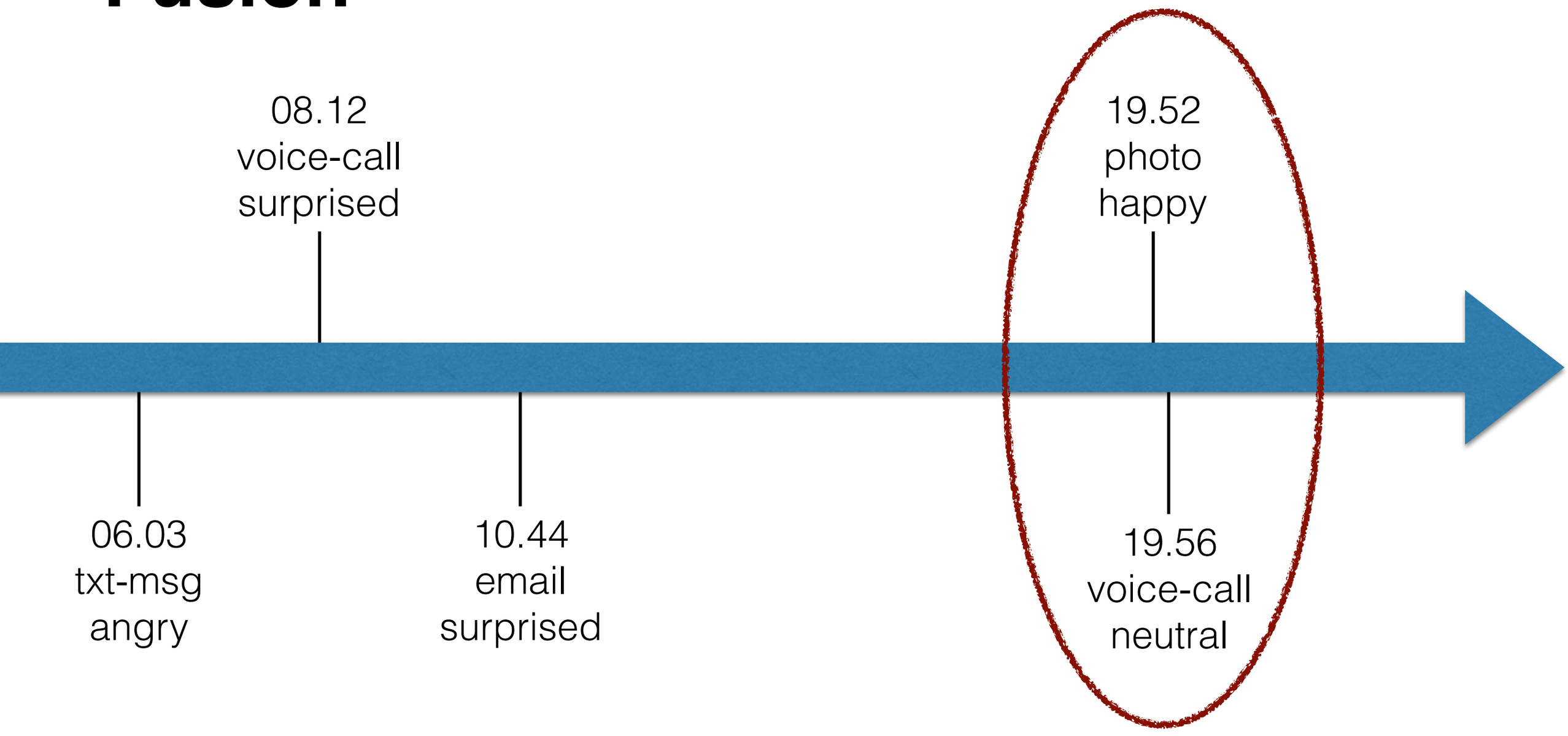
Textual: Emotion labels for English language are computed by the Synesketch library and a simple table for emoticons

Fusion



From the collected data-points it is clear that at most times of the day only a single modality generates an emotion label

Fusion



For emotion labels in clear proximity simple rules are used for fusion (facial overrules text, state overrules neutral, etc.)

Conclusions

- Big Data architecture has proven to be scalable for time consuming tasks (pre-processing, assignment of labels) that can be run in parallel
- Results are promising enough to make us believe that one will be able to build emotionally intelligent services in the near future
- Data retrieval is still very clumsy
- Privacy concerns can hinder the success of such systems

Privacy Concerns

- Very many personal data is collected in a centralised service
  - User needs to have trust in the service
  - Service is a promising target for attackers
- Systems is built in a way such that users easily opt-out (delete all their data)
- System is currently run in a private cloud setup which is completely under our control

JKU
JOHANNES KEPLER
UNIVERSITY LINZ

Slides at
https://steinbauer.org/

Matthias Steinbauer
matthias.steinbauer@jku.at
@climbinggeek