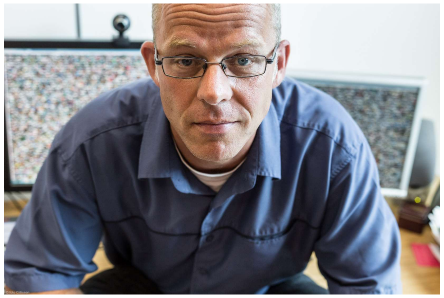
Interactive Learning

Human-Centered Machine Learning 2025/26

MARCEL WORRING

Who Am I?

- Full professor in Multimedia Analytics
- Leading the MultiX research group



We research multimedia analytics by developing AI techniques for getting the richest information possible from the data, visualizations, and interactions surpassing human and machine intelligence. We blend multi-modal data in effective interfaces for applications and social impact in public health, forensics and law enforcement, cultural heritage, and data-driven business.

Introduction

What should we aim for in AI?

- Useful AI
  - can be defined as an AI solution that can provide the functions required to satisfy target users' needs in the valid usage scenarios of their work and life.
- Usable AI
  - can be defined as an AI solution that is easy to learn and use via optimal user experience (UX) created by effective Human-Computer Interaction (HCI) design

[Xu2019]

Beyond Algorithms

- Current AI research
  - Is mostly focused on a fully automatic process
  - Is focused on improving the accuracy of models or at least well-defined quantitative metrics
- When striving for useful and usable AI
  - There is so much more than automatic methods that get a good score on one of the metrics

Useful AI is (mostly) Trustworthy AI

- Accuracy
  - How well does the AI system do on new (unseen) data compared to data on which it was trained and tested?
- Robustness
  - How sensitive is the system's outcome to a change in the input?
- Fairness
  - Are the system outcomes unbiased?
- Accountability
  - Who or what is responsible for the system's outcome?
- Transparency
  - Is it clear to an external observer how the system's outcome was produced?
- Explainability
  - Can the system's outcome be justified with an explanation that a human can understand and/or that is meaningful to the end user?
- Ethical
  - Was the data collected in an ethical manner? Will the system's outcome be used in an ethical manner?

[Mosqueira-Rey2022]

Usable AI starts with Data Usability

- Usable
  - usability in the most limited context, i.e., that we can use them because they are the right size, we have the right permissions, their license allows it, they do not contain sensitive information, etc.
- Relevant
  - cover the topic of interest at the right level of detail.
- Quality
  - completeness, provenance, accuracy, cleanliness, consistency of formatting, etc.
- Reusable
  - can be used in different studies where we would incorporate aspects such as that they are easily understandable, that there are different ways of accessing them, that there is a management of the changes produced in the data, as well as a collaborative nature in the data work processes.

[Mosqueira-Rey2022]

Usable AI needs to consider its different users

- ML experts
  - Experts with extensive knowledge in ML techniques.
    - Supervised learning: they select the data, label the data, classify them into training data and testing data, extract the features needed to feed the machine learning algorithm, create the model and refine it if the performance obtained is not optimal, etc.
    - Unsupervised learning: machine learning experts are required to interpret the clusters identified by the model so data can be converted into knowledge.
- Domain experts
  - In many domains, the designers of machine-learning-based systems do not themselves hold the expertise required to create training data. In such projects, the collaboration of domain experts is necessary.
- Data experts
  - A data expert or data scientist is a multi-disciplinary scientist that uses methods, processes and algorithms to extract knowledge from data.

[Mosqueira-Rey2022]

The algorithmic view

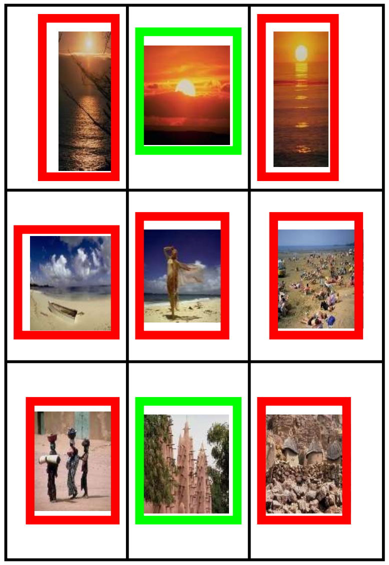
[MOSQUIRA-REY2022]

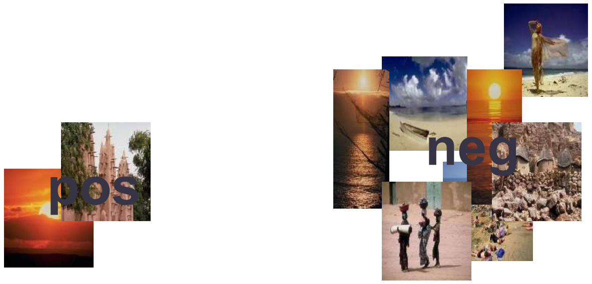
Different ways of learning

- Active learning (AL)
  - the system remains in control of the learning process and treats humans as oracles to annotate unlabeled data.
  - special case is Relevance Feedback where the user is giving feedback on the output
- Interactive machine learning (IML)
  - a closer interaction between users and learning systems, with people interactively supplying information in a more focused, frequent, and incremental way compared to traditional machine learning.
- Machine teaching
  - human domain experts have control over the learning process by delimiting the knowledge that they intend to transfer to the machine learning model.

Traditional ways of employing user interaction

Selection of pos/neg examples

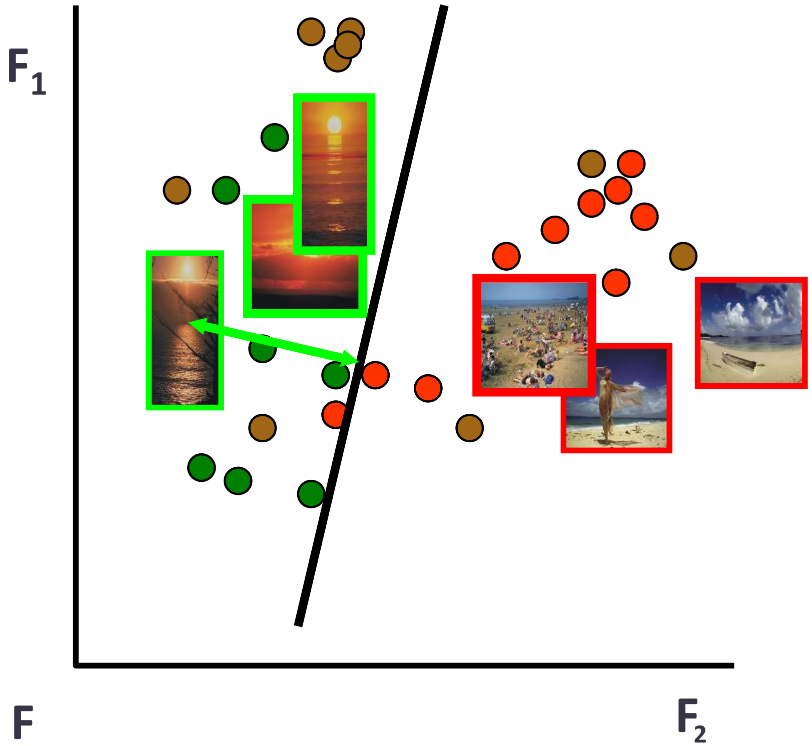




Some elements in the collection are labeled

Many are not

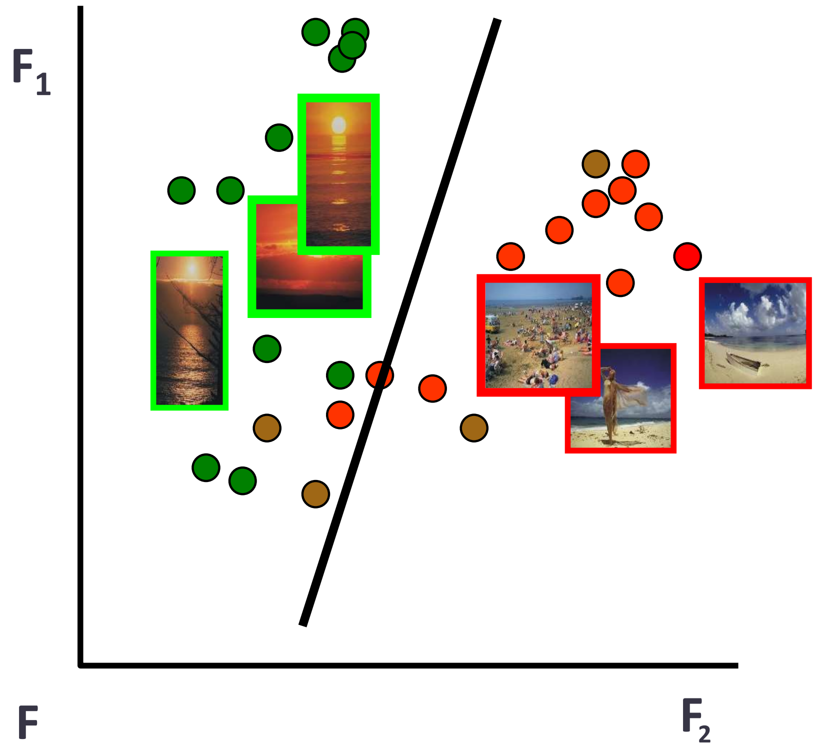
Relevance feedback



Try to find boundary in feature space best separating positive from negative examples

Measure of class membership probability

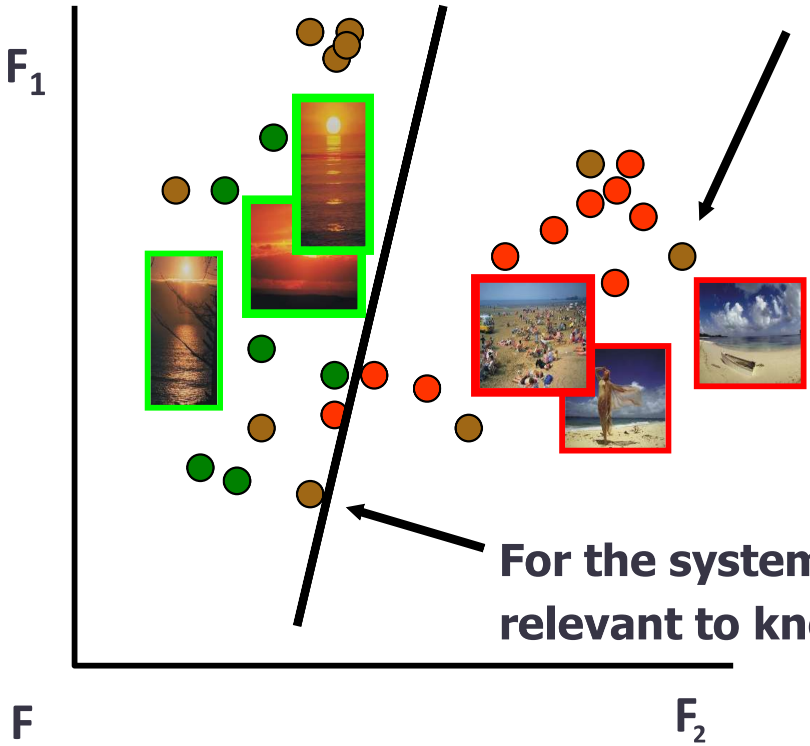
Relevance feedback



In the next iteration I will have more samples hence a better estimate of the boundary

This process is usually known as relevance feedback

Active Learning

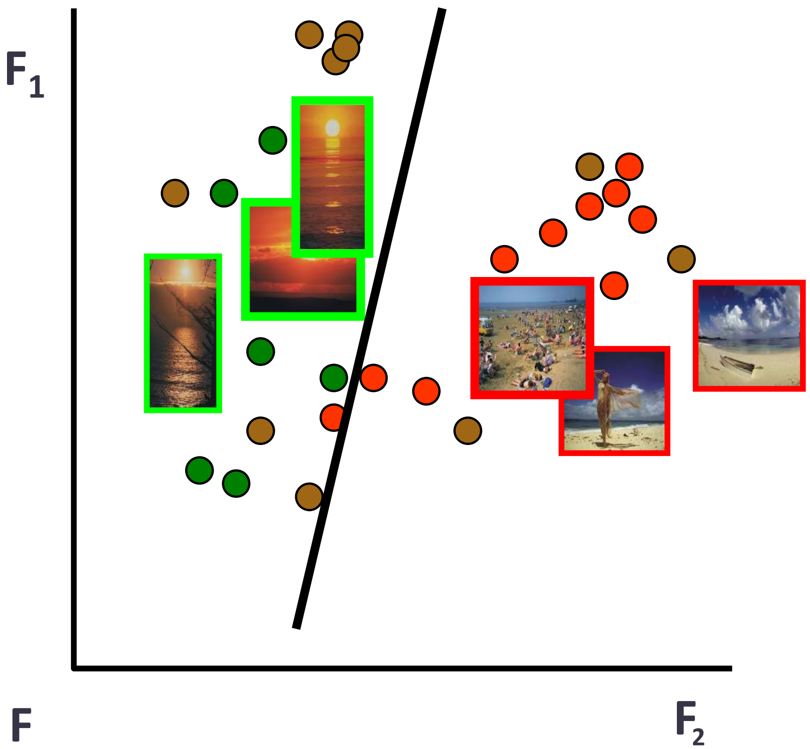


The system can safely assume this sample is also negative

In active learning the system decides which elements to show for feedback and which not.

For the system it is relevant to know this label

Before deep learning

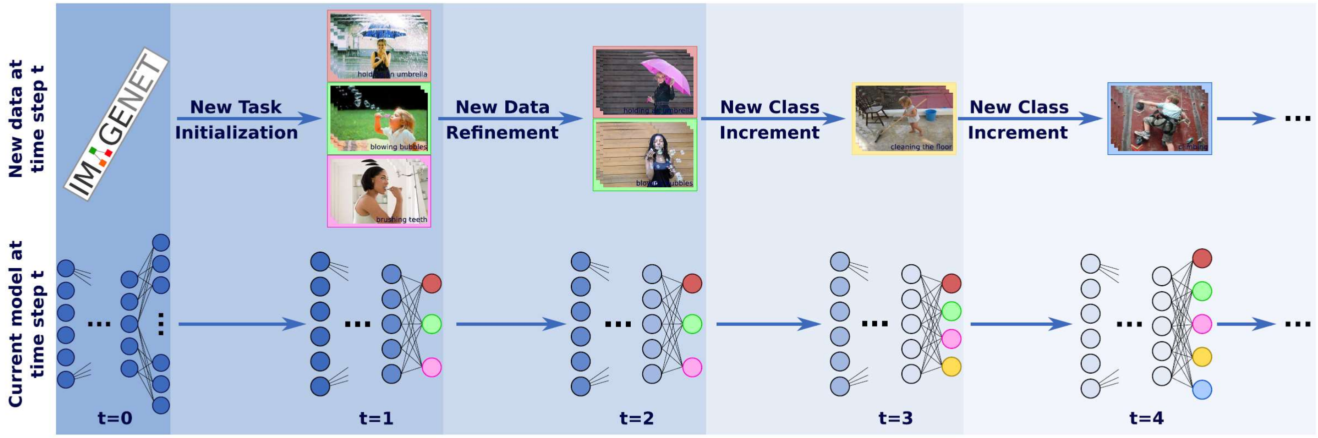


Linear or non-linear Support Vector Machine based on all current positive and negative examples

Use output of the classifier to compute probabilities and either give best output (RF) or actively select elements for labeling (AL)

With deep learning

- Two approaches
  - Apply SVM to latent spaces of the model
  - Fine-tuning deep learned models

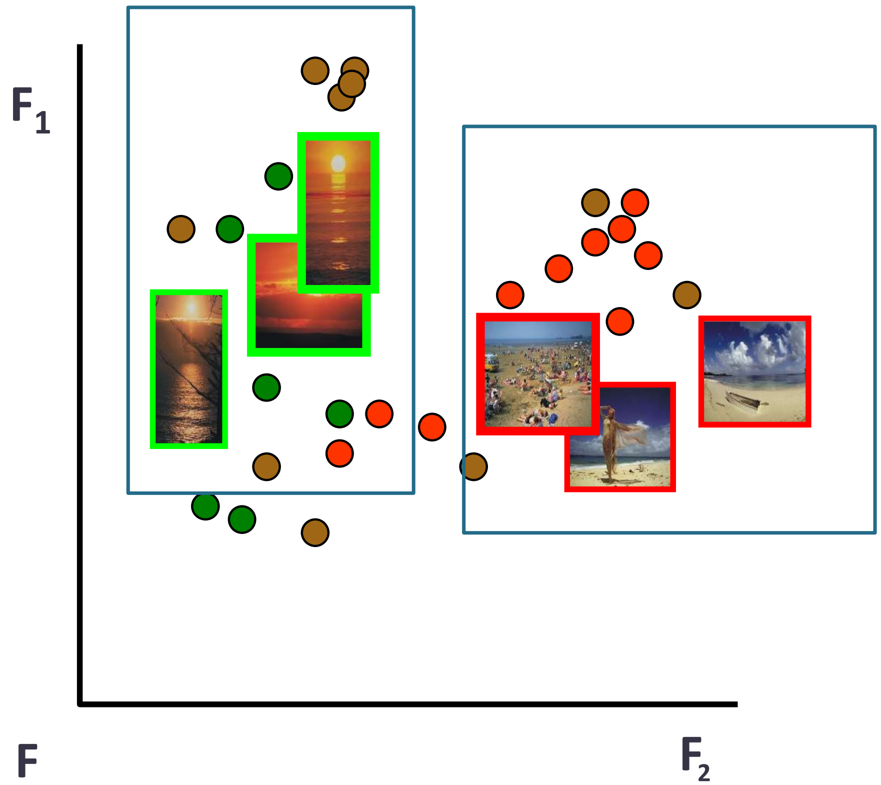


Main challenge is making this real time

[Kading2016]

Example Machine Teaching

First learn to recognize the easy sunsets



Then the more difficult beaches

Relevance feedback or Active Learning?

- Active learning – suggest items the model is *least sure* about
  - Those nearest to the decision boundary
- Adds the most useful information to the model at the cost of lowest number of labels
  - After training the model, it will be used to auto-label the rest of the data
  - Saving the user's time in the long run
- But the user doesn't see crisply-relevant items
  - So, the value of the model may be in doubt
- And for some items, it's impossible to decide even for the user
- So, relevance feedback is still what you see most

Difference Active Learning and IML

- Both focus on selecting new points for labeling by the user
- Active learning
  - Selection is driven by the model
- Interactive Machine Learning
  - Selection is driven by the user

Interactive Machine Learning: Main features

- Humans in the ML loop
  - assigned to tasks at which they are more efficient than machines.
- Humans assuming different roles
  - which can be domain experts, non-expert users, data scientists, etc.
- Incremental methodology
  - in which the model is updated iteratively and incrementally.
- The importance of the user interface
  - which influences how learning takes place and conditions learning outcomes.

Focus on much more diverse tasks than only classification

Tasks in Interactive Machine Learning

- Visual cluster analysis
  - Exploratory data analysis
  - Comparative clustering analysis
- Interactive dimensionality reduction
  - Subspace analysis
  - High-dimensional data exploration
  - Progressive dimensionality reduction
- Interactive model analysis
  - Model understanding
  - Model diagnosis

Tasks in Interactive Machine Learning

- Interactive classification
  - Interactive labeling
  - Interactive feature engineering
  - Parameter space analysis
- Interactive regression
  - Interactive correlation analysis
  - Interactive numerical prediction
- Interactive information retrieval
  - Structured information retrieval
  - Unstructured information retrieval

Tasks in Interactive Machine Learning

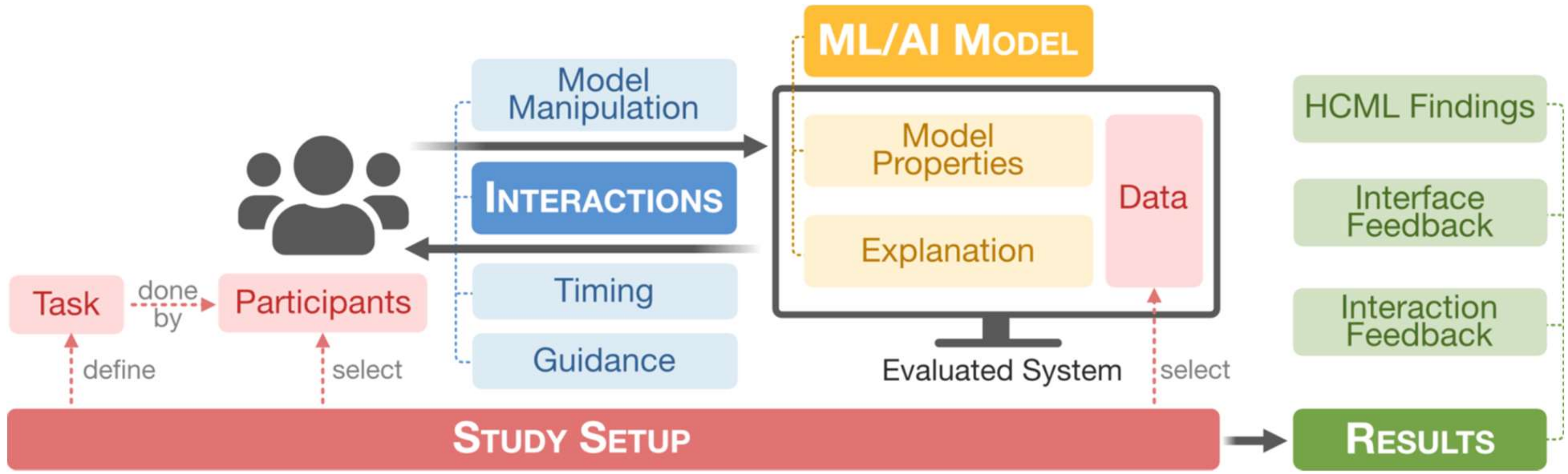
- Visual pattern mining
  - Exploratory event analysis
  - Mobility pattern analysis
- Visual topic analysis
  - Flat topic analysis
  - Hierarchical topic analysis
  - Topic evolution analysis
- Interactive anomaly detection
  - Anomalous points detection
  - Anomalous sequences detection

Evaluation of Interactive Learning

Evaluation

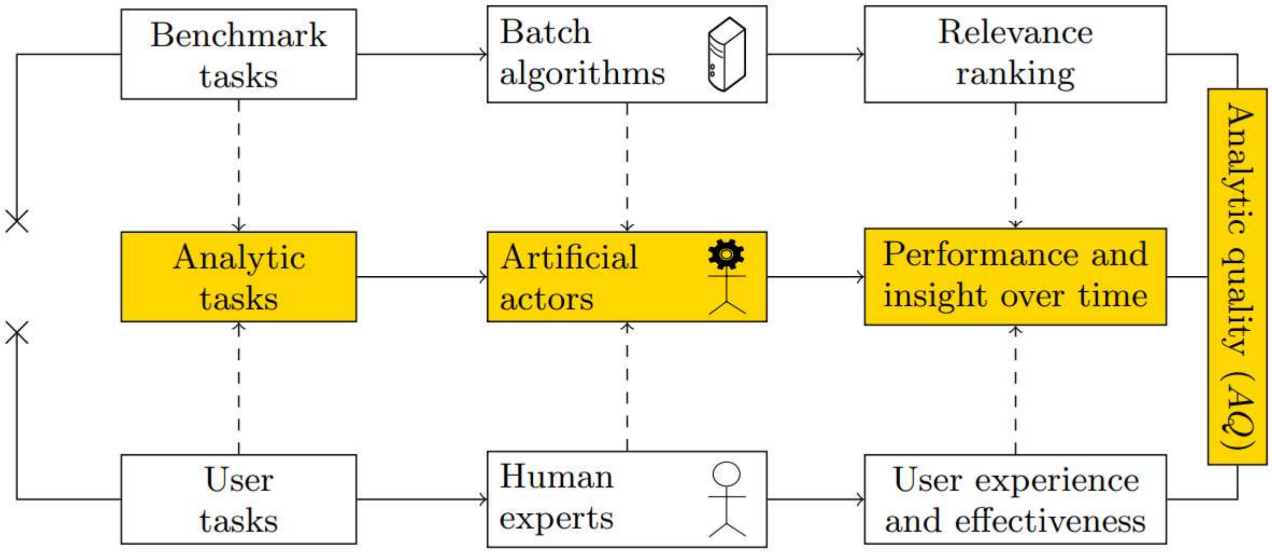
- What we need to do
  - Evaluate the quality of the model
  - Evaluate the user experience
  - Evaluate effort from both user and system
  - Take into account time

Many methods for evaluation



[Sperlle2021]

Analytic Quality



[Zahalka2015]

Two Example Systems

The MH17 Investigation

MH17 Investigation Photo Database
(Dutch Safety Board)

! Large and complex image collection

! Contains image categories never seen before

? How to explore and search within such an image collection?

Explore
What kind of images, which geolocations, how many, what categories?

Search
Where was the left wing tip located? Are there more images of the avionics? Is there an image with a better angle?

[Gisolf2021]

The method

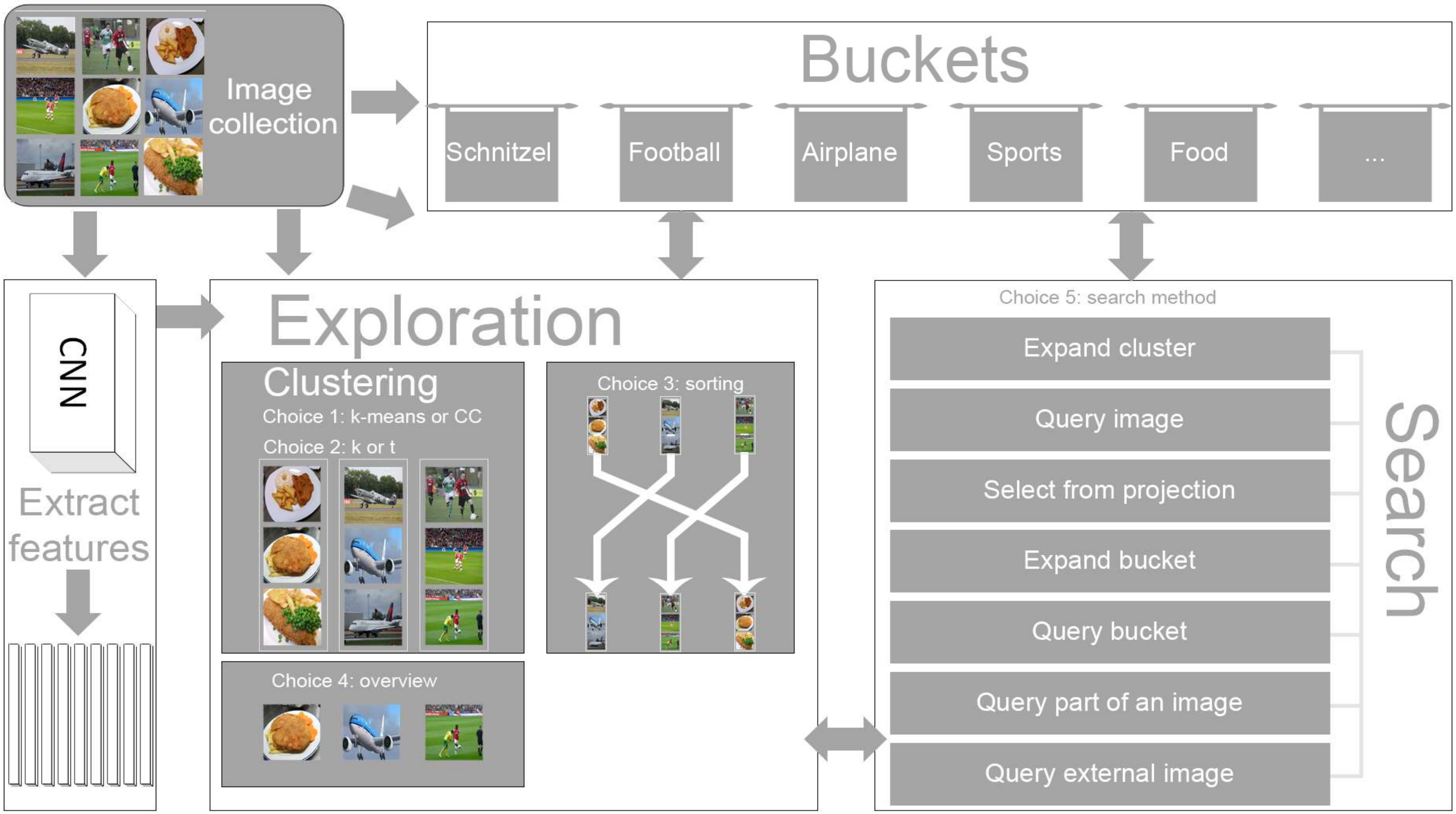


Image collection
Buckets
Schnitzel
Football
Airplane
Sports
Food
...
CNN
Extract features
Exploration
Clustering
Choice 1: k-means or CC
Choice 2: k or t
Choice 3: sorting
Choice 4: overview
Choice 5: search method
Expand cluster
Query image
Select from projection
Expand bucket
Query bucket
Query part of an image
Query external image
Search

The datasets

MH17 image collection	Real world dataset for actual investigation
	15,000 images
	28 categories determined by experts
	Classified
Google Open Image Dataset	10,000 images
	37 categories not present in pretrained dataset

Malaysia Airlines Flight 17 (MH17) was a scheduled passenger flight from Amsterdam to Kuala Lumpur that was shot down on 17 July 2014 while flying over eastern Ukraine.

(Wikipedia)

The artificial analyst

Has access to ground truth

Uses a combination of the 5 choices as a **strategy**

5 choices
1. KM or CC
2. *k* or threshold
3. Sorting or not
4. Overview or not
5. Search method

Strategy example:
Use CC with a threshold of 0.5. Don't use sorting, but do use the overview. Use *query bucket* as the search method

Categorizes all images in the image collection into buckets

Each simulation is *one vs the rest*

One predetermined bucket for relevant images

OID has 37 categories, thus per strategy 37 simulations

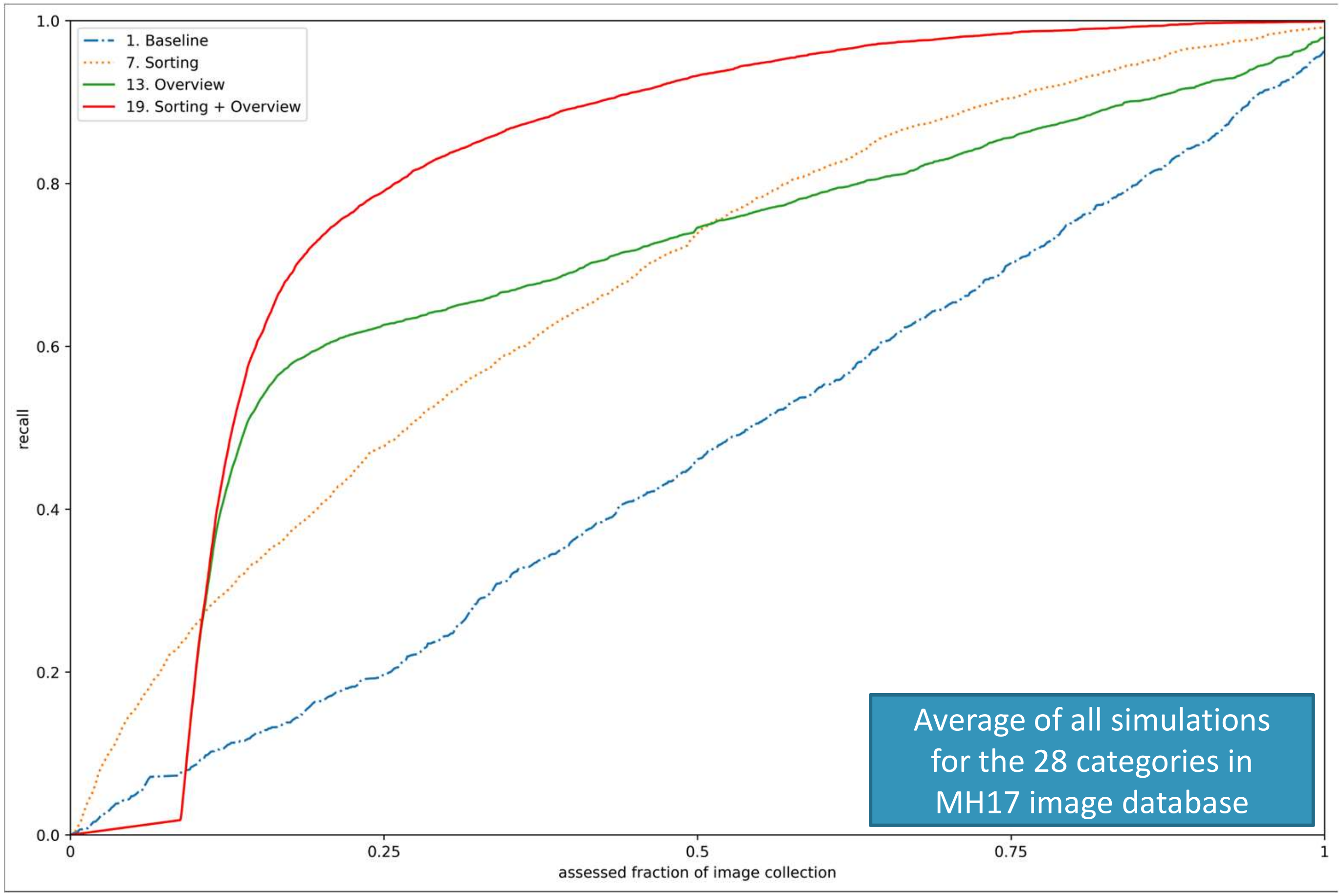
The AA sees each image 1 by 1 and only once

Order determined by strategy

Results: Exploration

5 choices for Baseline
1. KM
2. *k* = 1221
3. No sorting
4. No overview
5. No search method

5 choices for Sorting + Overview
1. KM
2. *k* = 1221
3. Used sorting
4. Used overview
5. No search method

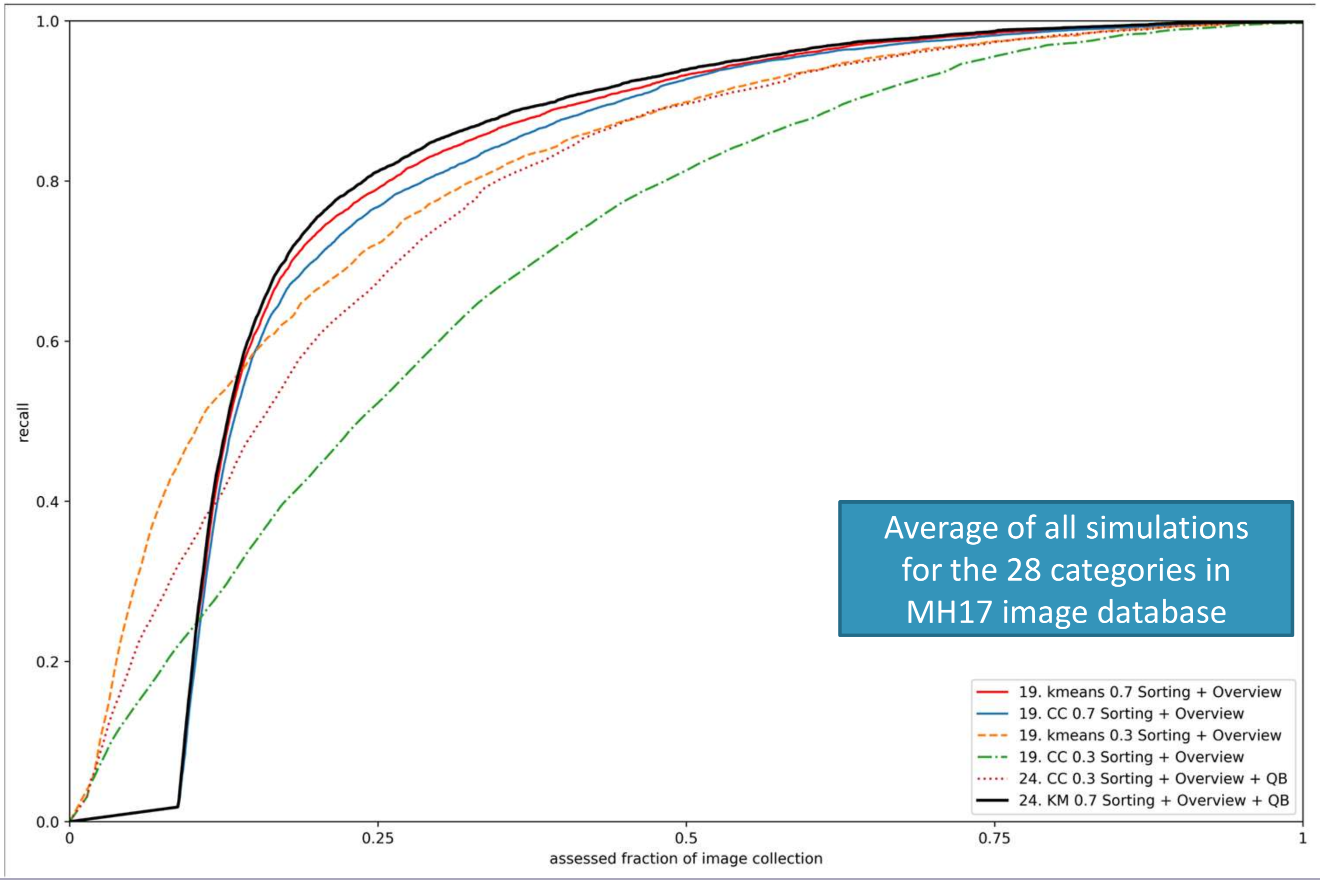


Average of all simulations for the 28 categories in MH17 image database

Results: Search

5 choices for KM 0.7 Sorting + Overview
1. KM
2. *k* = 1221
3. Used sorting
4. Used overview
5. No search method

5 choices for KM 0.7 Sorting + Overview + QB
1. KM
2. *k* = 1221
3. Used sorting
4. Used overview
5. Query Bucket as search method



Average of all simulations for the 28 categories in MH17 image database

Results: User experiments

User group	Image collection	Task	Result
4 aviation accident investigators	MH17 15,000 images	Use application to organize MH17 image collection and think aloud 2 hours	Clusters usable and of high quality. Would increase their efficiency. Confidence in method
6 groups of Forensic Science students	OID 10,000 images	1. Find image of old blue car 2. Find 40 images that best summarize the image collection 3. Find the food item most common in the image collection and find the most images	1. All groups found the correct image. *External query* was the fastest method. 2. Best performance by setting threshold such that there were 40 clusters. Then pick one image per cluster. 3. 4 out of 6 groups identified the correct food. Best performance with 86% recall and 10% false positives.

The Video Browser Showdown

Dataset:

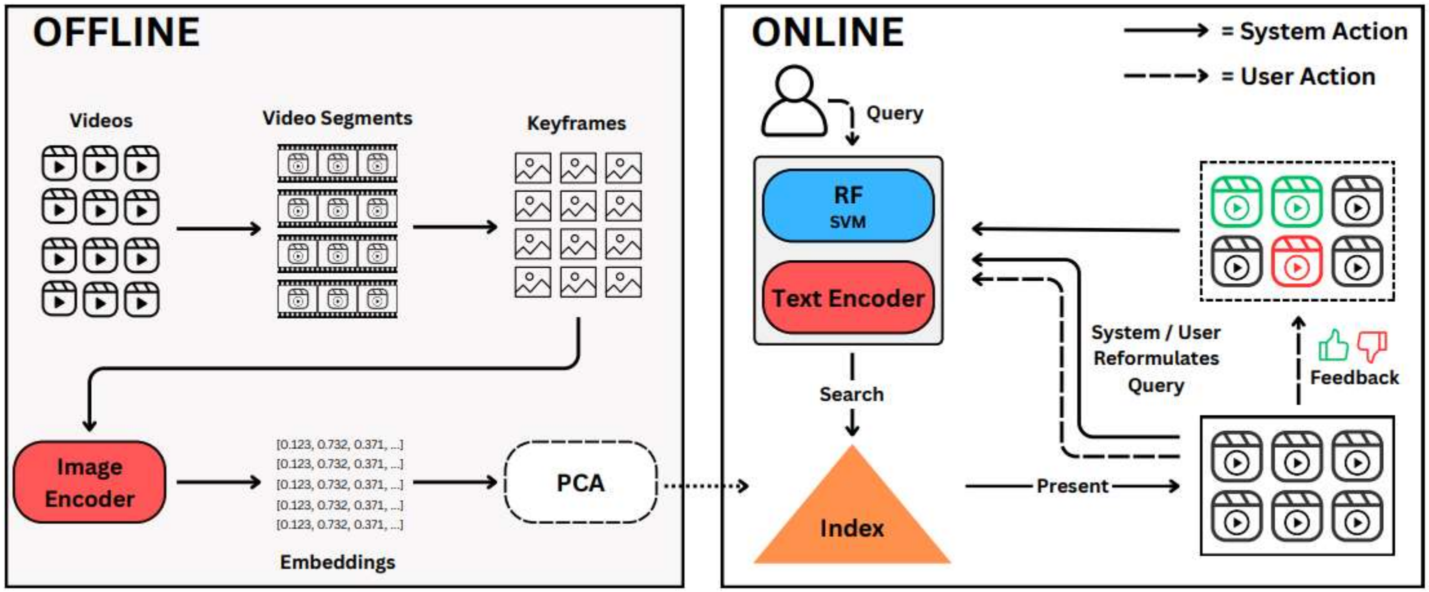
4800 h of the Vimeo Creative Commons Collection

Known-Item Search (KIS): a single video clip (a few seconds long) is randomly selected from the dataset and visually presented – this is known as **visual KIS (KIS-V)**. The participants need to find exactly the single instance presented. A variation of this task is **textual KIS (KIS-T)**, where instead of a visual presentation, the searched segment is described only by text given by the moderator (and presented as text via the projector). Another variant of this is **KIS-C**, where the target scene is also textually described, but only with minimal details in the beginning. Further details are revealed after 60 seconds based on questions/chats from participants.

Ad-hoc Video Search (AVS): here, a rather general description of many shots is presented by the moderator (e.g., „Find all shots showing cars in front of trees“) and the participants need to find as many correct examples (instances) according to the description.

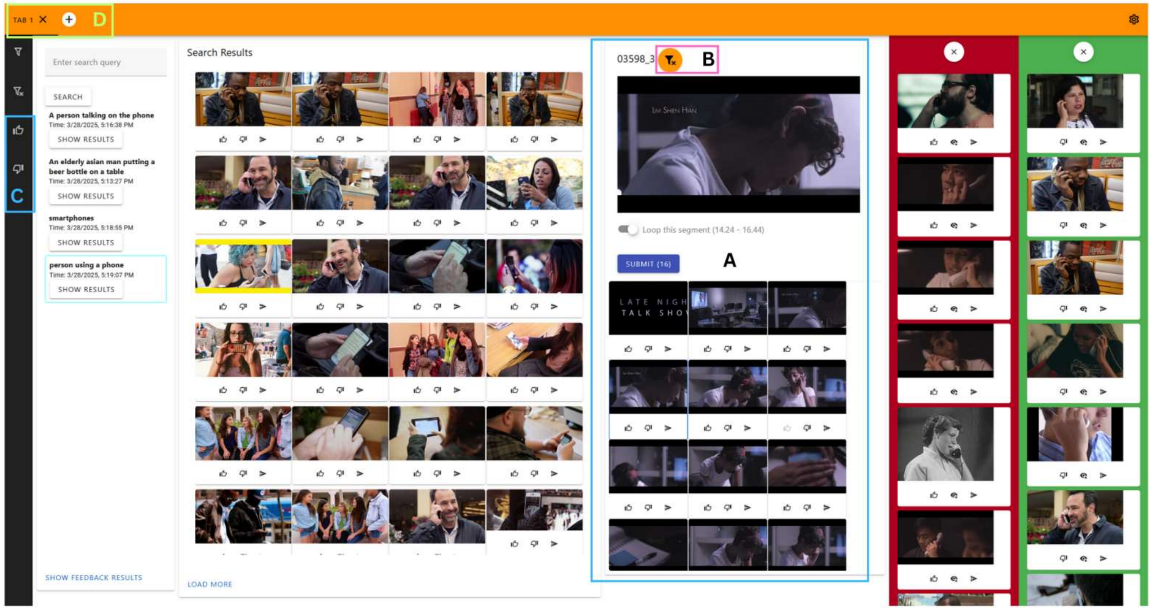
Visual Question Answering (VQA): this task type asks specific questions about a particular video or video collection, which are intended to be submitted as a manually entered text (by a human). For example, it could show a video clip and ask „*How many nights do we see passing in the video until this segment?*„. It requires manual inspection and interactive exploration.

Exquisitor



[Sharma2025]

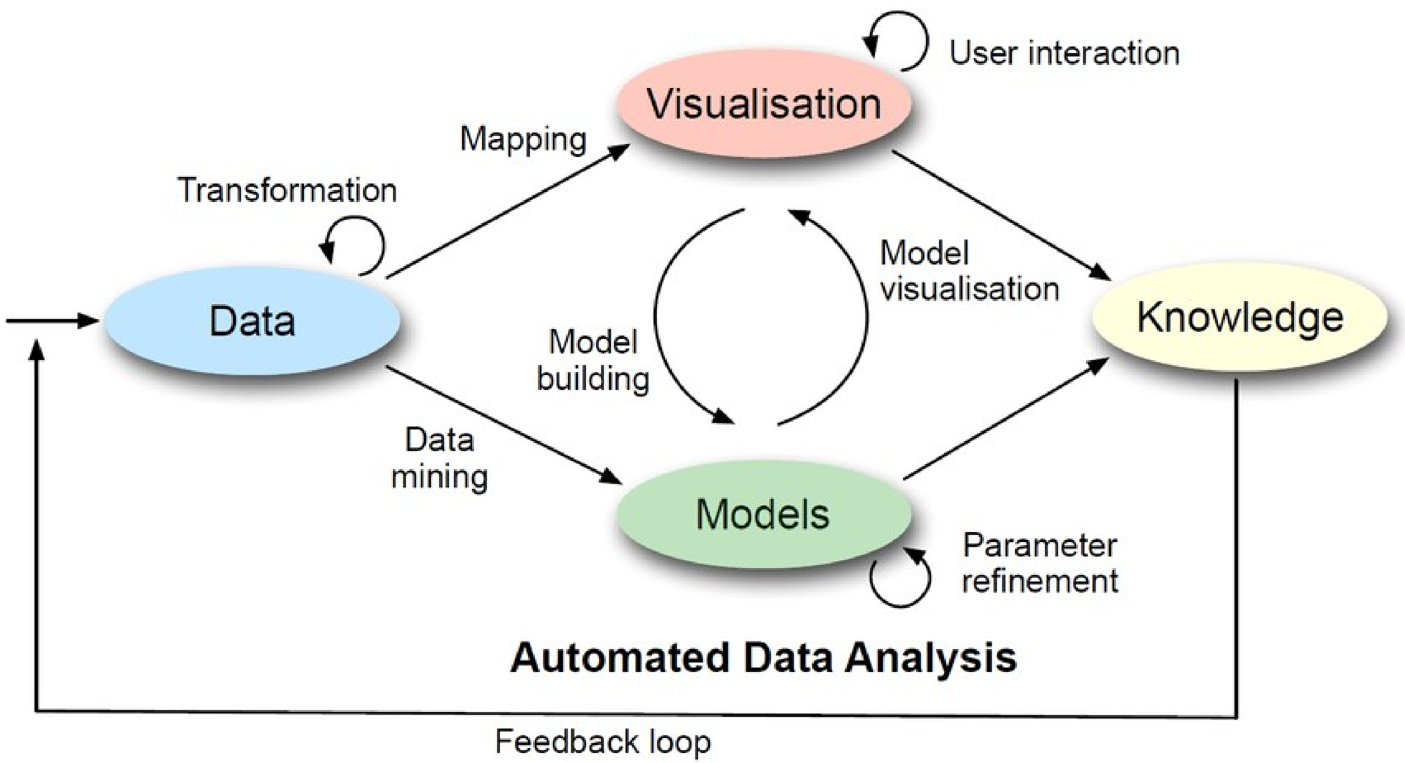
Exquisitor



Conversational Search + Relevance feedback to refine results

Visual Analytics

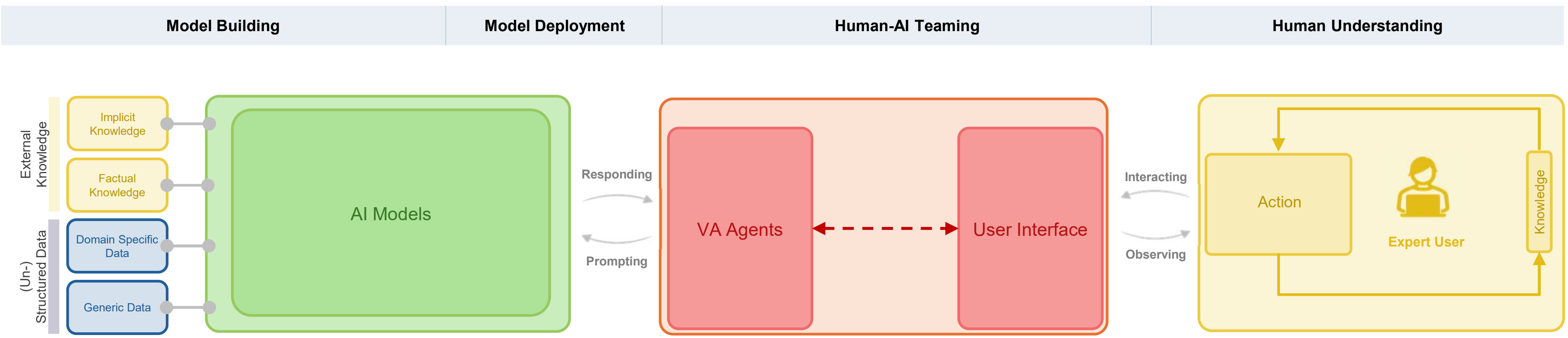
The basic visual analytics model



Bringing together information visualization and data mining

[Keim2008]

And in the foundation model Era



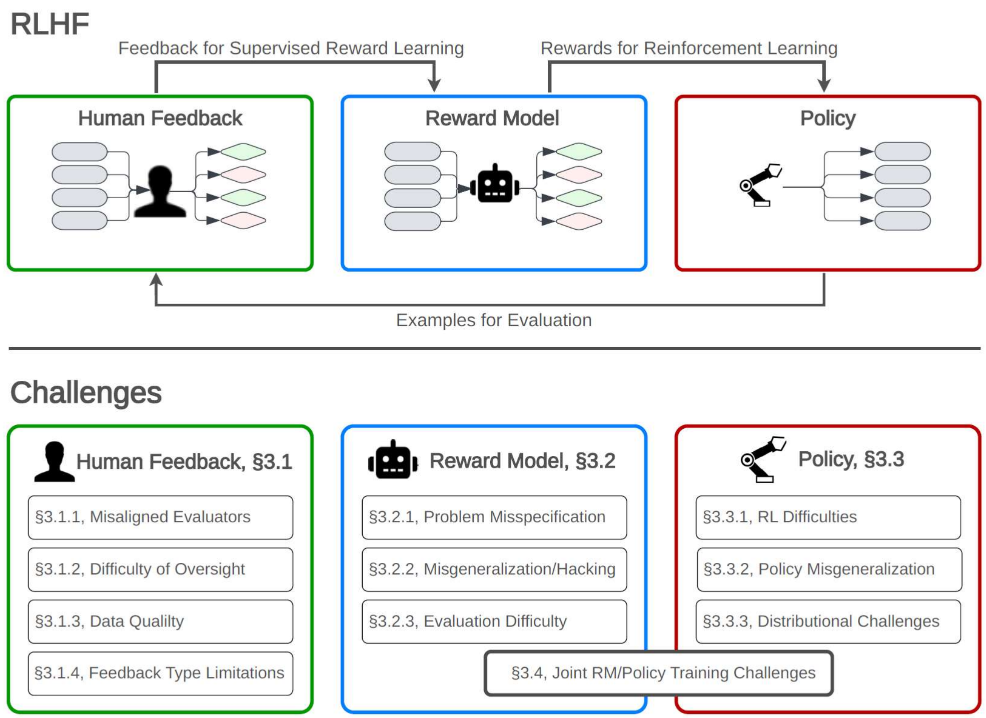
[Worring2025]

VA Model: Modelling technique categories

- Modify parameters (MP)
  - The user directly manipulates the model parameters through the visualization
  - The more populous category across all techniques
  - Pros: easier to implement, exact meaning
  - Cons: requires stats/machine learning knowledge from the user, non-intuitive
- Define analytical expectations (DAE)
  - The user interacts within the domain of expertise (using domain knowledge), the model behaves semantically: translating between the user's language and the ML/stats language
  - Fewer approaches exist
  - Pros: meaningful and intuitive to the user, no or little knowledge of stats/ML required
  - Cons: difficult to implement, knowledge gap between the developer and the user

[Endert17]

Reinforcement learning seems good model for DAE



But has its challenges, more on RLHF in subsequent lecture

[Casper2023]

Conclusions

Conclusions

- Usable and Useful AI
- Hybrid Intelligence needed
  - Combination of ML, VIS and HCI needed
- Reinforcement learning a promising avenue
  - Explicit setting of goals
  - Multiple agents (including humans) aiming to reach the goals
- Many interesting research challenges

References

- [Casper2023] S. Casper et.al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. Transactions on Machine Learning Research (12/2023)
- [Endert2017] Endert, Alex, et al. "The state of the art in integrating machine learning into visual analytics." Computer Graphics Forum. Vol. 36. No. 8. 2017.
- [Gisolf2021] F Gisolf, Z Geradts, M Worring. Search and explore strategies for interactive analysis of real-life image collections with unknown and unique categories. International Conference on Multimedia Modeling, 244-255
- [Keim08] D. Keim et al.: Visual Analytics: Definition, Process, and Challenges. In Information Visualization, Lecture Notes in Computer Science, vol. 4950, Springer, Berlin.
- [Mosqueria2022] E. Mosqueira-Rey et.al Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review (2023) 56:3005–3054
- [Sharma2025] U Sharma, OS Khan, S Rudinac, BÞ Jónsson. Can Relevance Feedback, Conversational Search and Foundation Models Work Together for Interactive Video Search and Exploration?. Proceedings of the Computer Vision and Pattern Recognition Conference, 3740-3749
- [Sperlle2021] F. Sperrle, M. El-Assady, A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. Guo, R. Borgo, D. Horng Chau, A. Endert, and D. Keim. EUROVIS 2021. Volume 40 (2021), Number 3. STAR – State of The Art Report
- [Xu2019] Xu W Toward human-centered AI: a perspective from human–computer interaction. Interactions 26(4):42–46.
- [Worring2025] M. Worring, J. Zahálka, S. van den Elzen, M.T. Fischer, D.A. Keim. A Multimedia Analytics Model for the Foundation Model Era. ARXIV.2504.08138, 2025
- [Zahalka2015] J Zahálka, S Rudinac, M Worring. Analytic quality: evaluation of performance and insight in multimedia collection analysis. Proceedings of the 23rd ACM international conference on Multimedia, 231-240