Scalable Language Technologies Lab

Summer Term 2026

Course Information

Instructors

Martin Potthast

Niklas Deckers

Tim Hagen

Contact

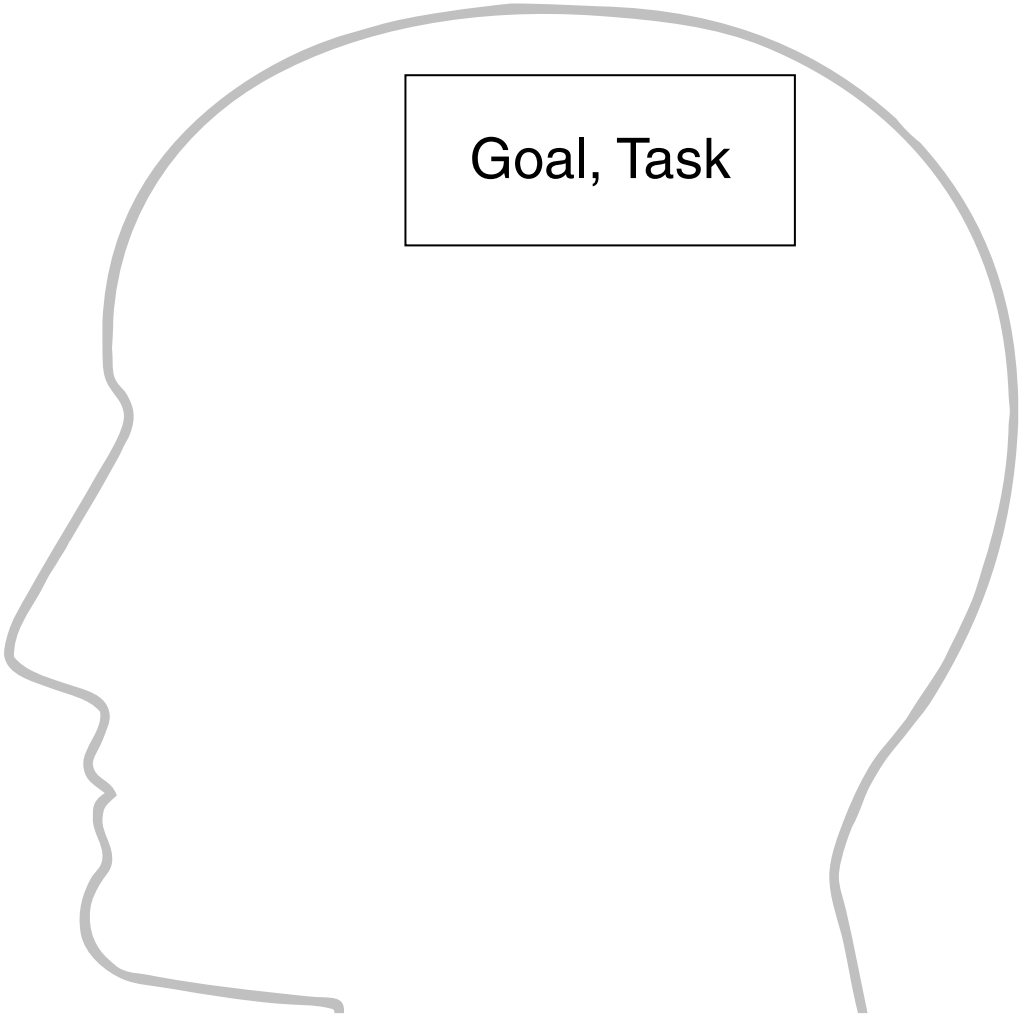
- Contact: tim.hagen@uni-kassel.de
- Office hours: by appointment
- Web: temir.org > Teaching > Scalable Language Technologies Lab

Scalable Language Technologies

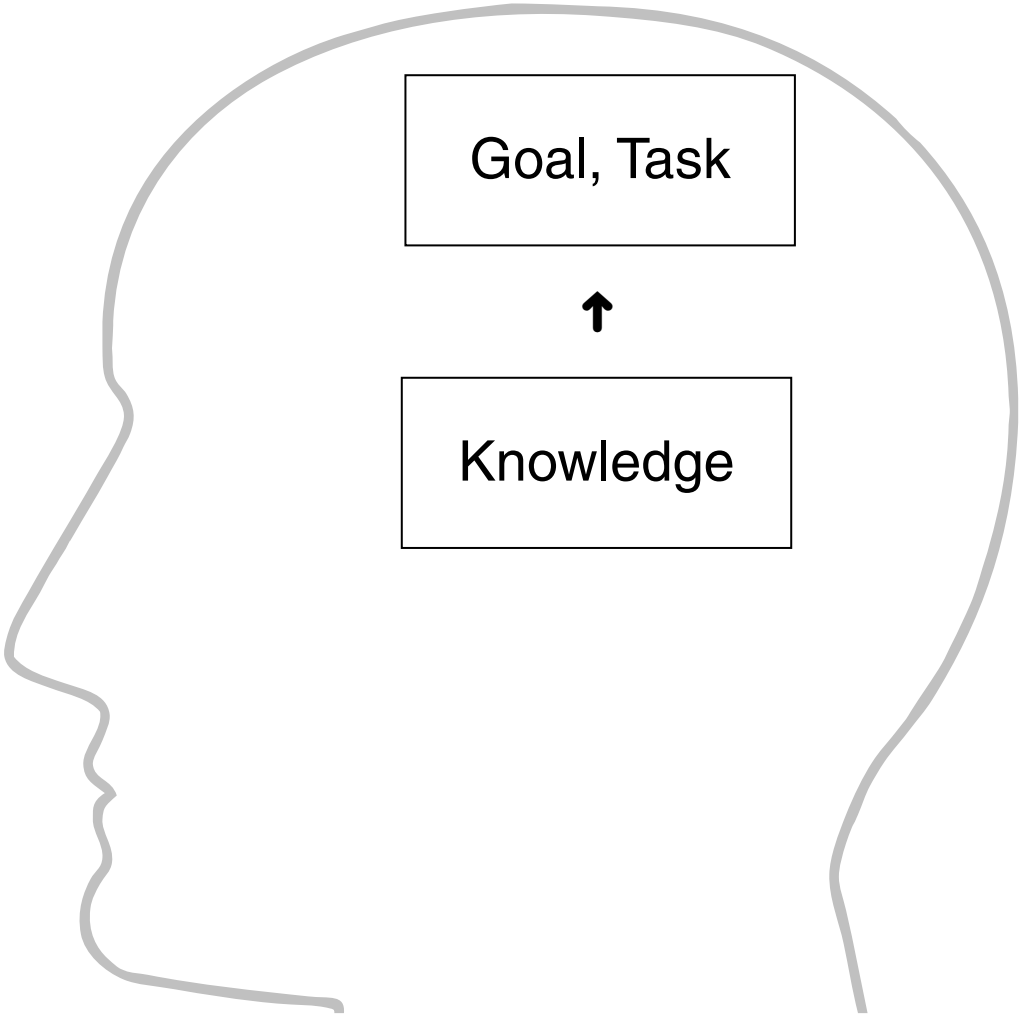
Language technologies are methods and tools for analyzing, modifying, and generating human language.

- Support interactions between humans and machines in natural language
- Form the foundation of numerous intelligent information systems:
  - Search engines
  - Translation systems
  - Dialog and conversation systems
  - AI Agents
  - Argumentation systems
  - ...
- Research subjects of NLP and IR fields
- Rely on AI, ML, and especially Deep Learning techniques

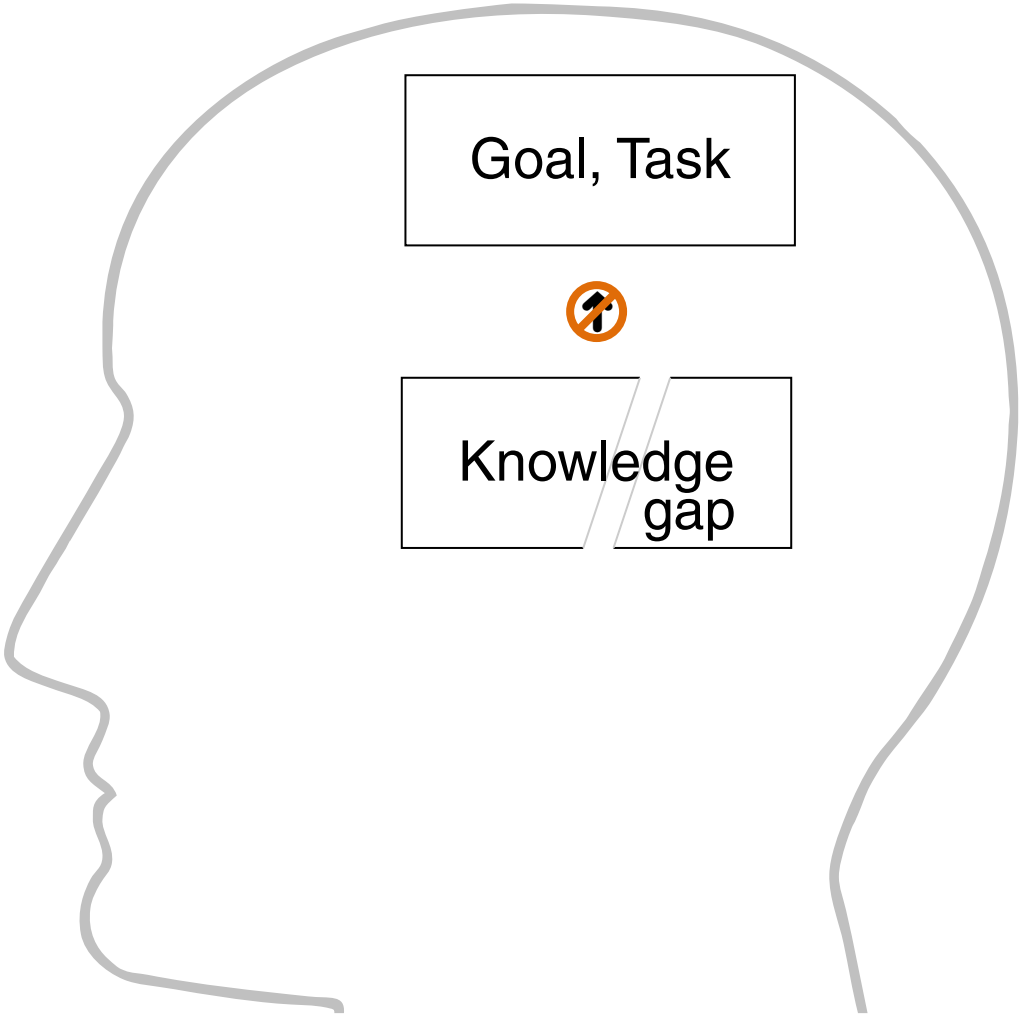
Information Retrieval



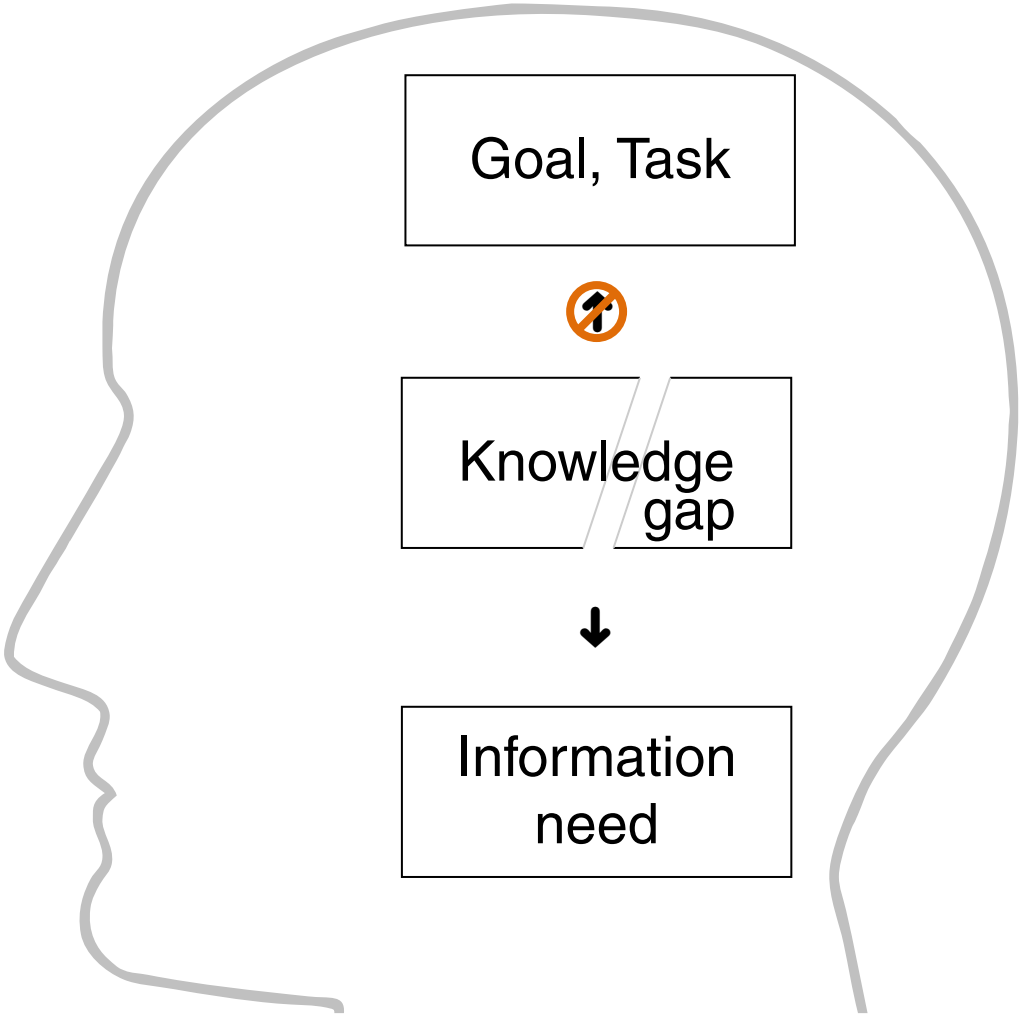
Information Retrieval



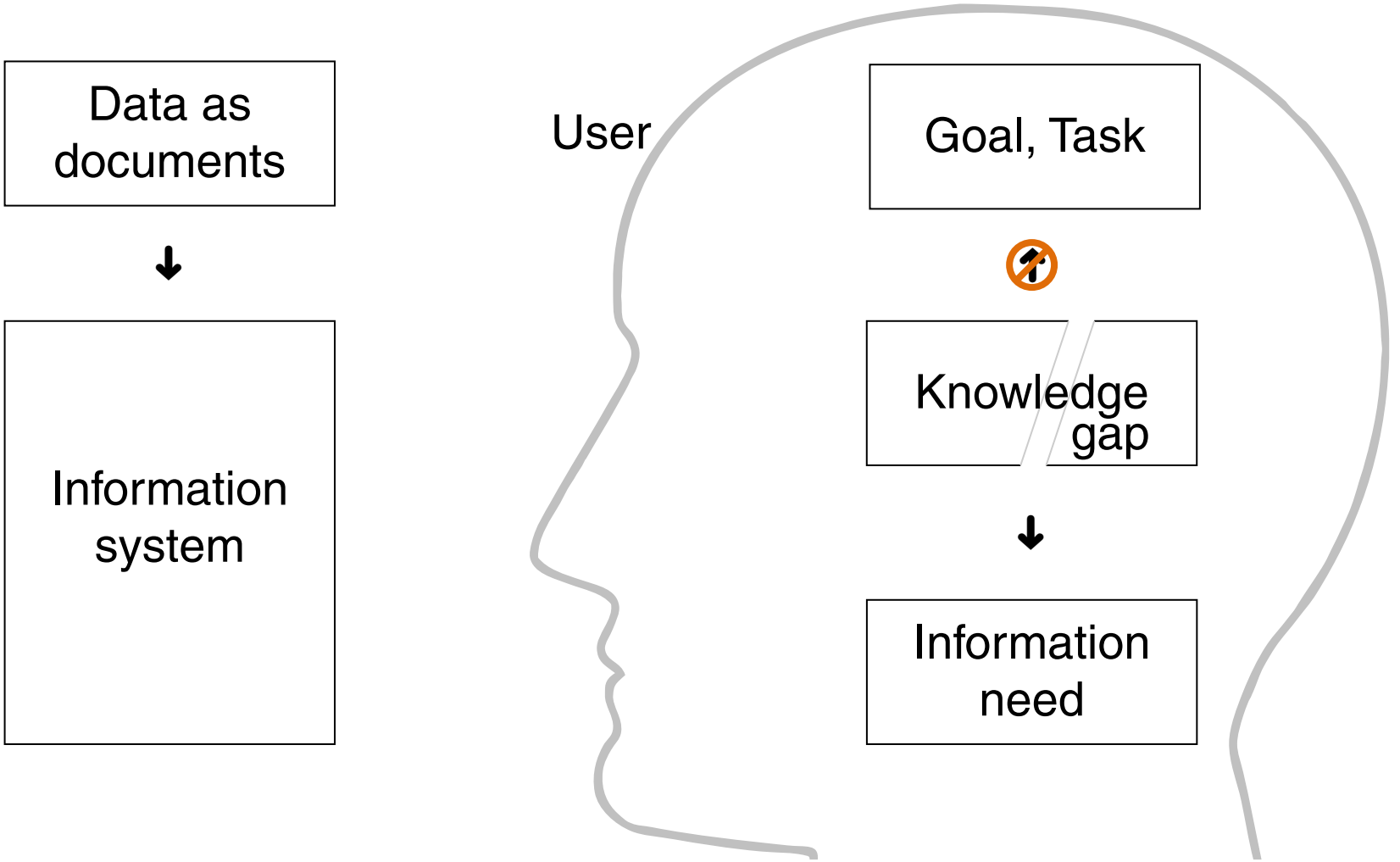
Information Retrieval



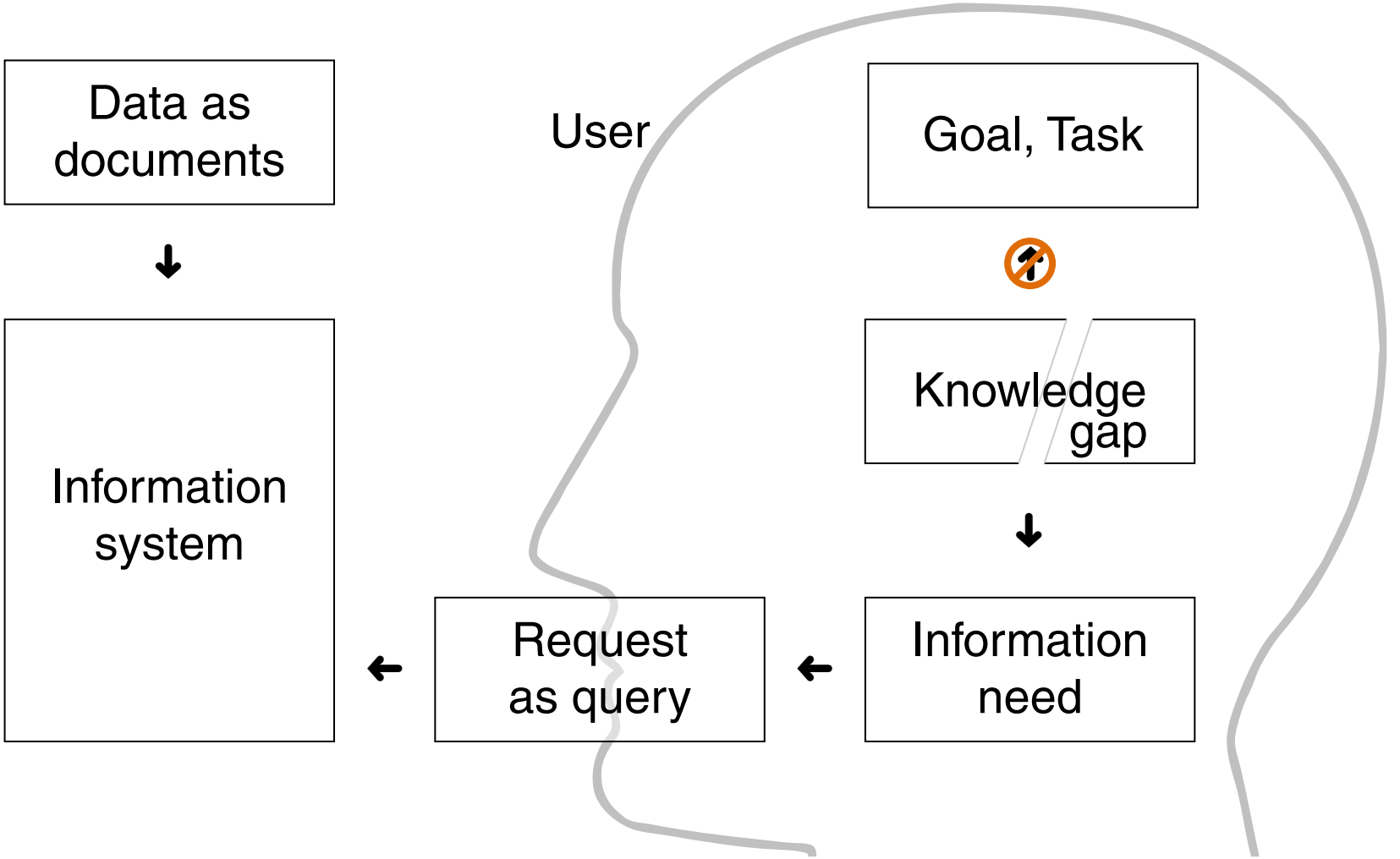
Information Retrieval



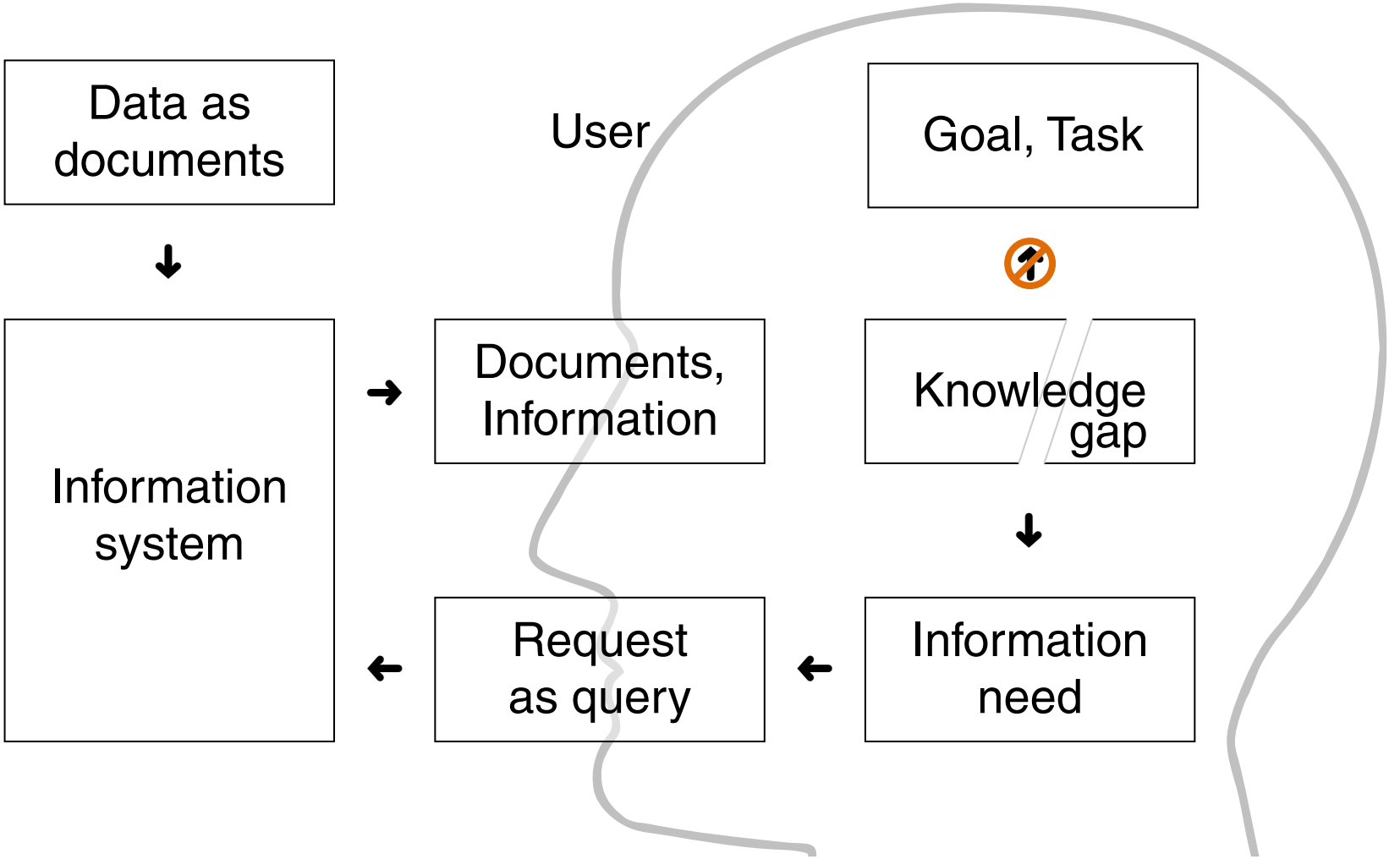
Information Retrieval



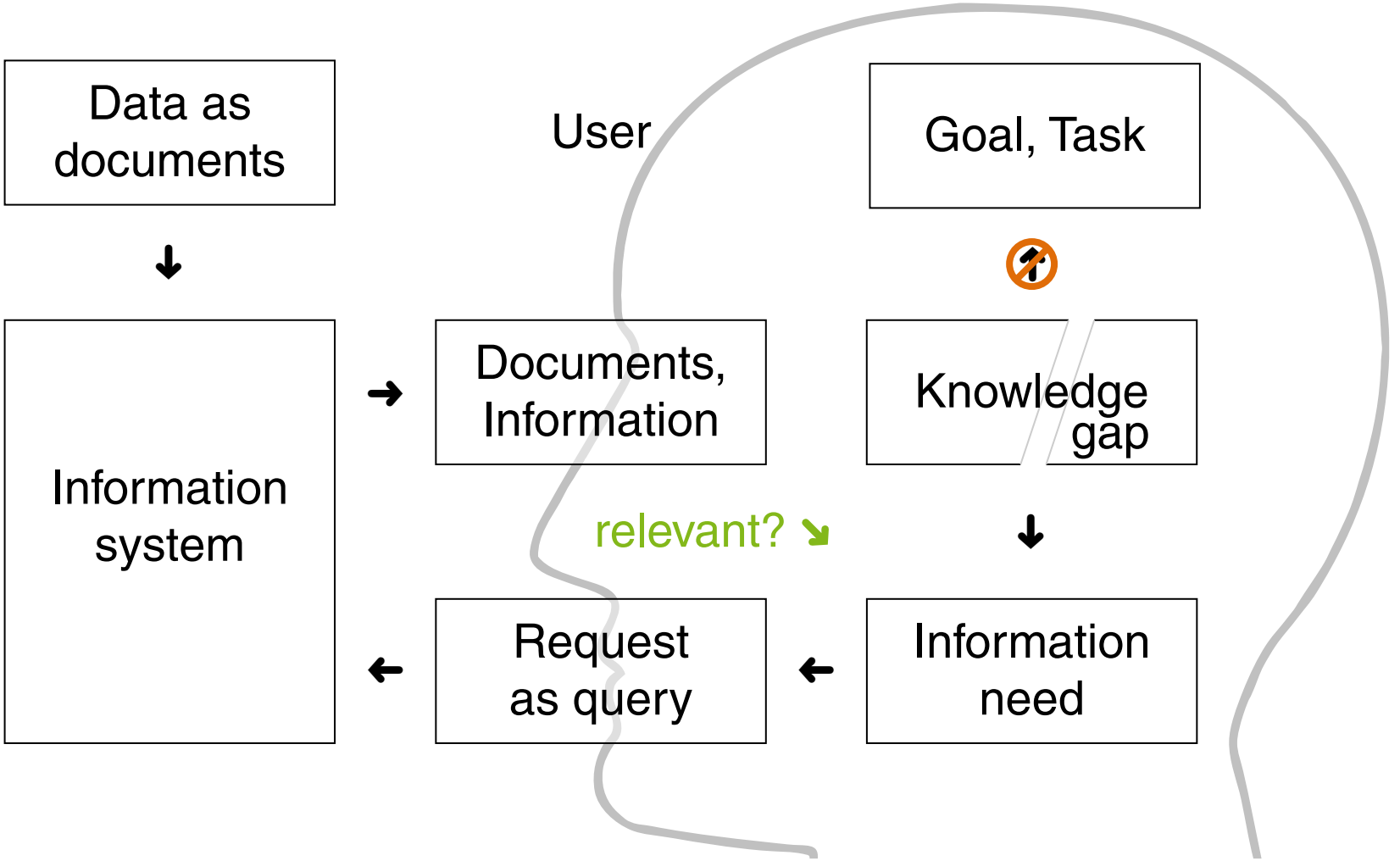
Information Retrieval



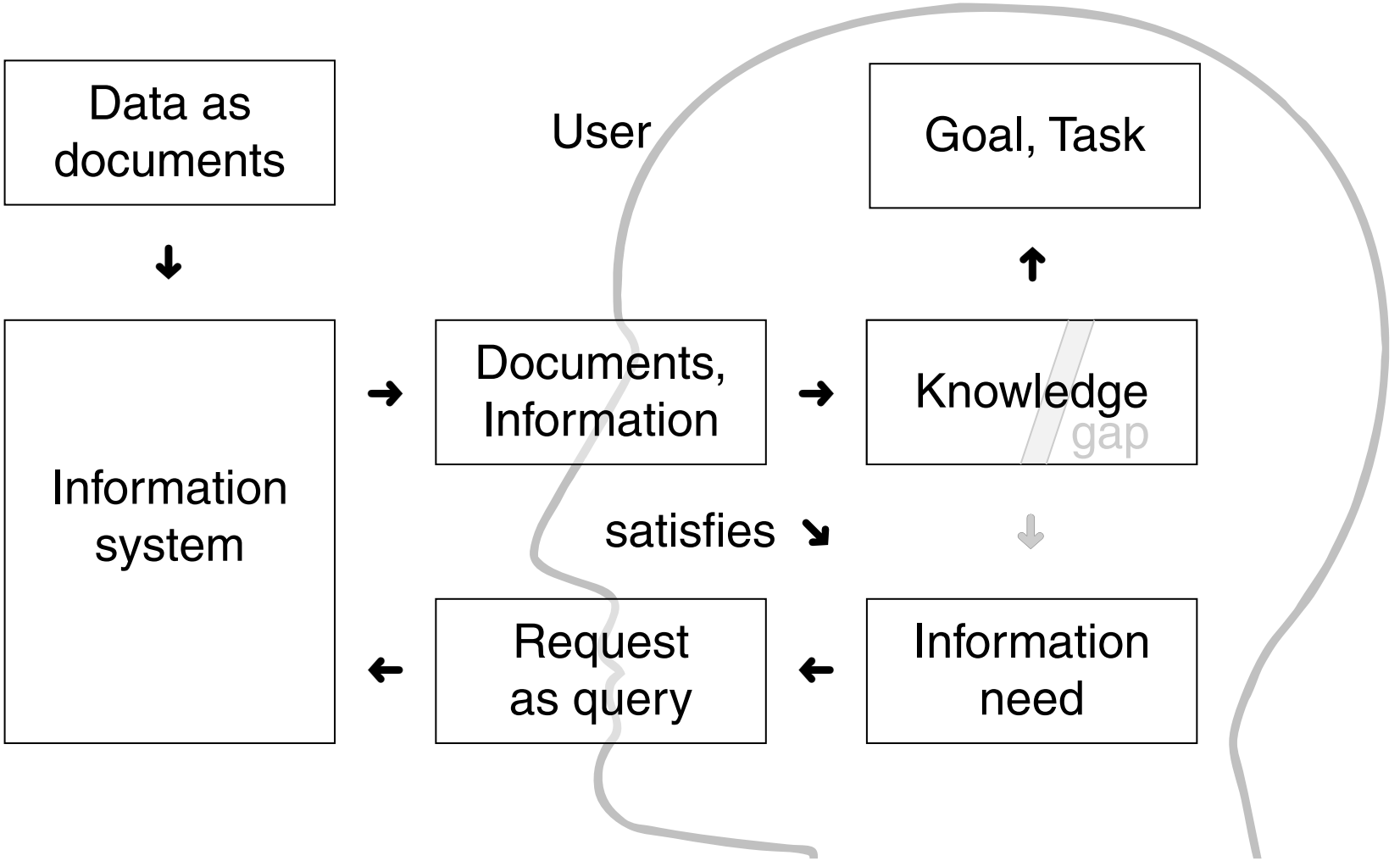
Information Retrieval



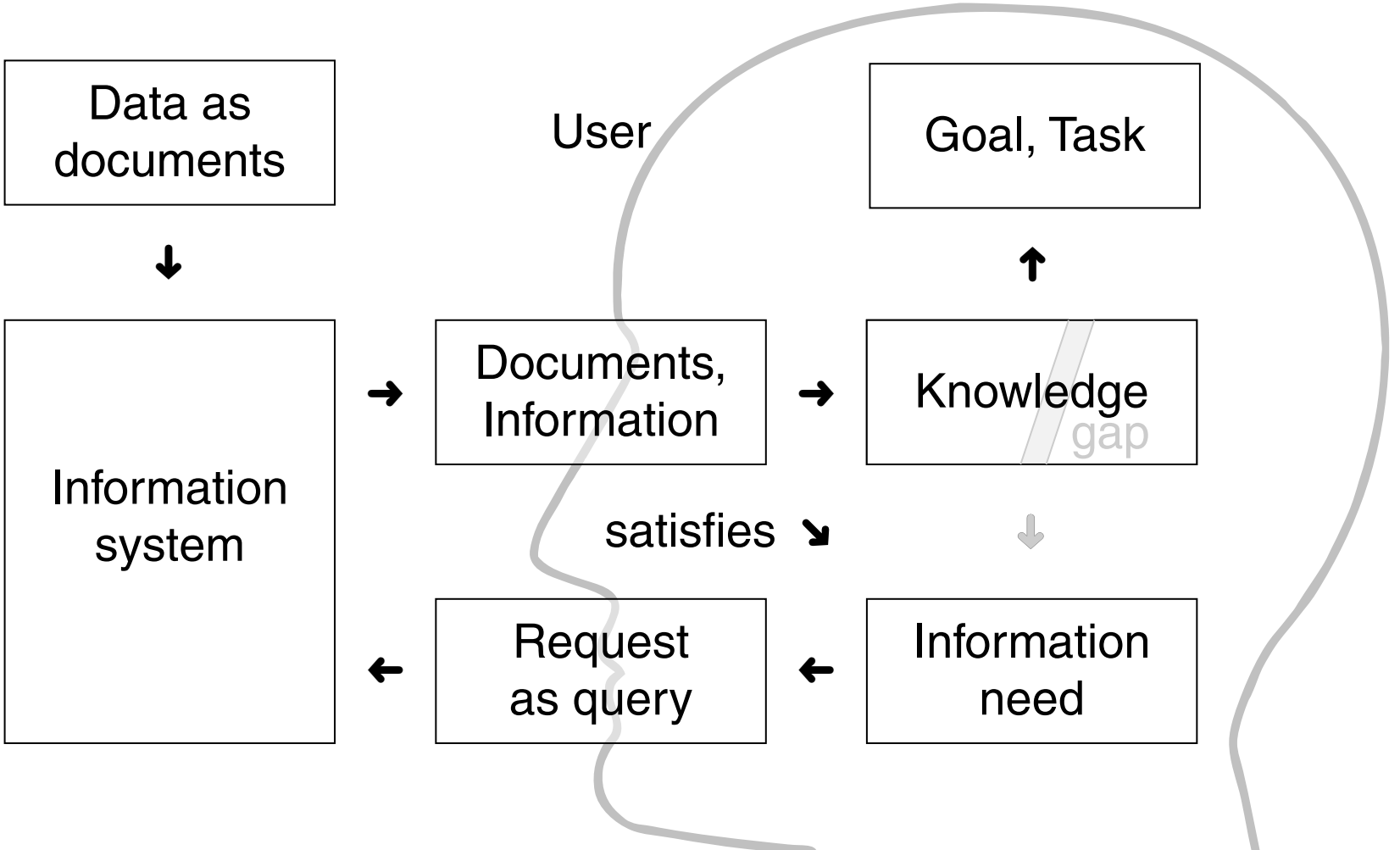
Information Retrieval



Information Retrieval

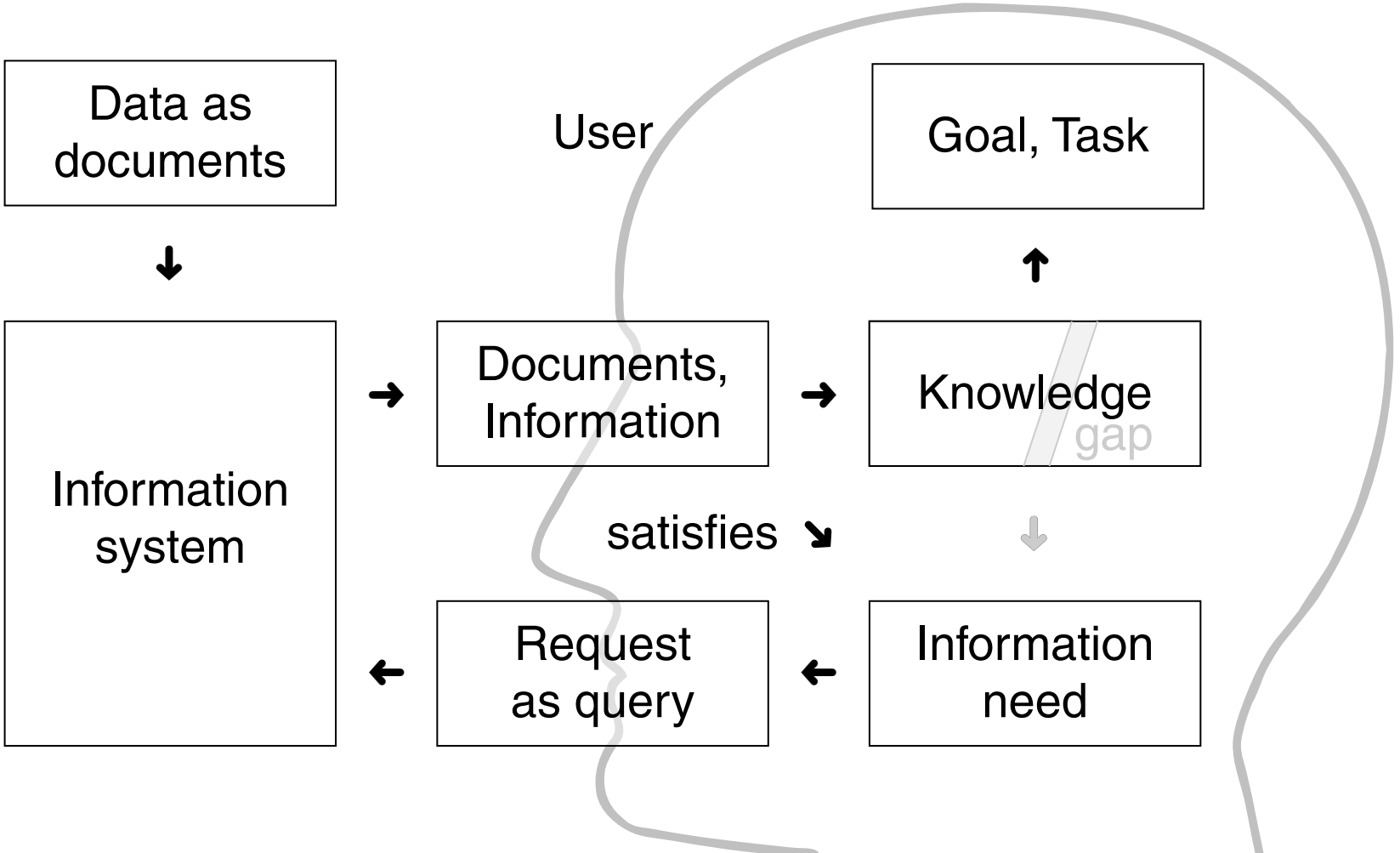


Information Retrieval



What Search Engines do you know?

Information Retrieval



What Search Engines do you know?

Google/Bing/... (Web Search), Library Index, Generative Models (e.g., GPT), Netflix

Topic

- Generative search is information retrieval on an “infinite index”
  - Web search query $\rightarrow$ search engine result page (SERP) $\rightarrow$ query refinement
  - Prompt to a generative model $\rightarrow$ generated candidate images $\rightarrow$ prompt refinement
- Generative models continuously map from an embedding of the prompt and a random noise input to the image space
- Examples: Stable Diffusion, Imagen, Midjourney, Banana Pro, DALL-E
- Prompting is difficult: the map from the prompt tokens to the image space is not intuitive
- Sense of direction; lack of control

Why is User Behavior Interesting?

- We want to build information systems that assist users in their information need formulation process
- Usually, search is not a single-step process
- Analysis of user behavior can help us infer intentions

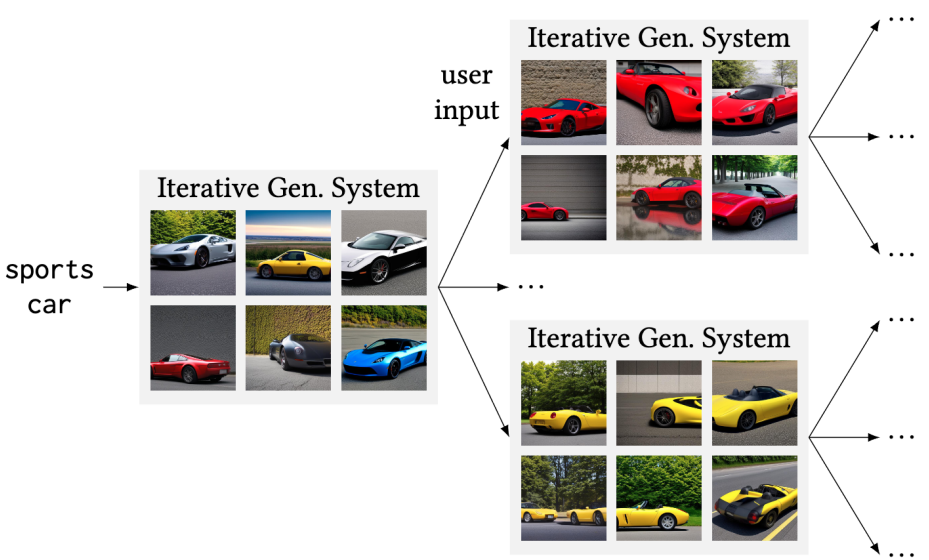
➜ Improved (more direct) maps from query to desired output

- Holds for (classic) web search and search on the infinite index

Walking Around Inside the Index

- Discretely: Prompt Engineering / Query Refinement
- Continuously: Inside the embedding space

- Steps can be controlled iteratively through user input



Query Logs und Prompt Logs

- Provide a history of user prompts
- Types of behaviors:
  - Query reformulation/refinement
    narrowing / broadening / paraphrasing
  - Topic Switching
- AOL Query Log
- User behavior can give us interesting insights about the search engine's quality
  - Is the returned information relevant but not specific enough?
  - Is the returned information non-relevant and the user tries again with a paraphrased query?
  - Is the user now interested in something else entirely?

Hierarchical Embeddings

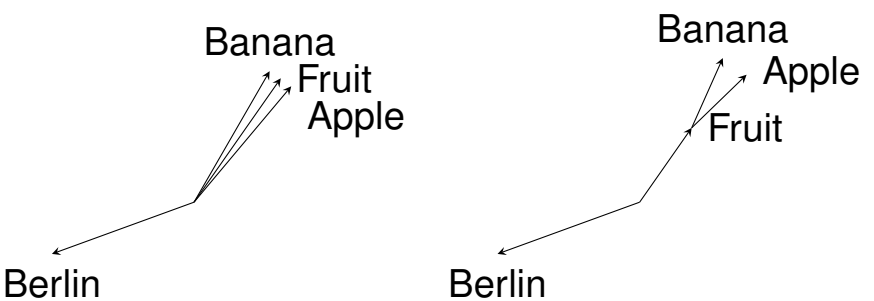
- We often need to represent text as numbers to apply deep learning methods

➜ Map words into a vector space (“embeddings”)

- Similar words lie in similar directions from the origin
- But: No hierarchical information

➜ Order Embeddings: Distance from the origin signals specificity

Embed into a Poincaré space (distances increase further away from the origin)



How does tracking query refinement inside hierarchical spaces help to discover user intentions?

Tasks

Develop an application that solves the following tasks:

- Predicting next refinement (/search trajectory)
- Classifying different refinement types
- Identifying satisfaction (Stopped prompting because satisfied or gave up?)
- Deviations from strictly hierarchical methods—creative inventions?
- How do I move in a hierarchical embedding space when I refine a query? Move into a similar direction further away from the origin?
- How can we measure creativity?

Questions

- How can we quantify how much more specific a user's reformulated query is?
- When the reformulation does not go deeper into the hierarchy, did the user have a "creative" new idea? What may have triggered the latteral movement?

Hypothesis: *Humans search on the infinite index as they would on a text corpus.*

Learning Objectives

- Work in a structured and self-supervised manner
- Apply current research in language technologies
- Develop and carry out experiments at scale
- Work with large corpora via state-of-the-art infrastructure
- Collaborate effectively in a group
- Scientific writing and presentation
- Create demonstrable software solutions

Labor: Scalable Language Technologies

Organization

- **Workload:** 4 SWS
- **Schedule:** Monday, 10:00–12:00 + consultations by individual appointment
- **Location:** Room -1605; individual appointments online
- **Communication:** Email, Discord

- **Materials:** Slides, papers, and resources on the course website

Deliverables

- Code/Demo
- Report (template will be provided)
- Presentation

Details on the deliverables will be provided later.

Getting to Know Each Other

What experience do you have in the following subjects?

- Information Retrieval
- Text-to-image models like Stable Diffusion
- Machine learning, deep learning
- CLIP
- Programming in Python
- PyTorch
- Git, SSH, Slurm

Introductory Exercise

- We provide a Jupyter Notebook that iteratively accepts a prompt as input and generates and displays an image [notebook]
- Prompt the following image as closely as possible:

- Prompt for an image to visualize: The feeling of being watched by something you cannot see.
- Write down your prompts in order
- Next Session: We will compute the prompt embeddings and visualize them.

Submit Jupyter Notebooks with your prompt logs until Sunday to tim.hagen@uni-kassel.de.