Principles of AI Engineering

Chapter 1: Introduction

Prof. Dr. Steffen Herbold

Credit:

Based on contents from Christian Kästner (https://github.com/ckaestne/seai)

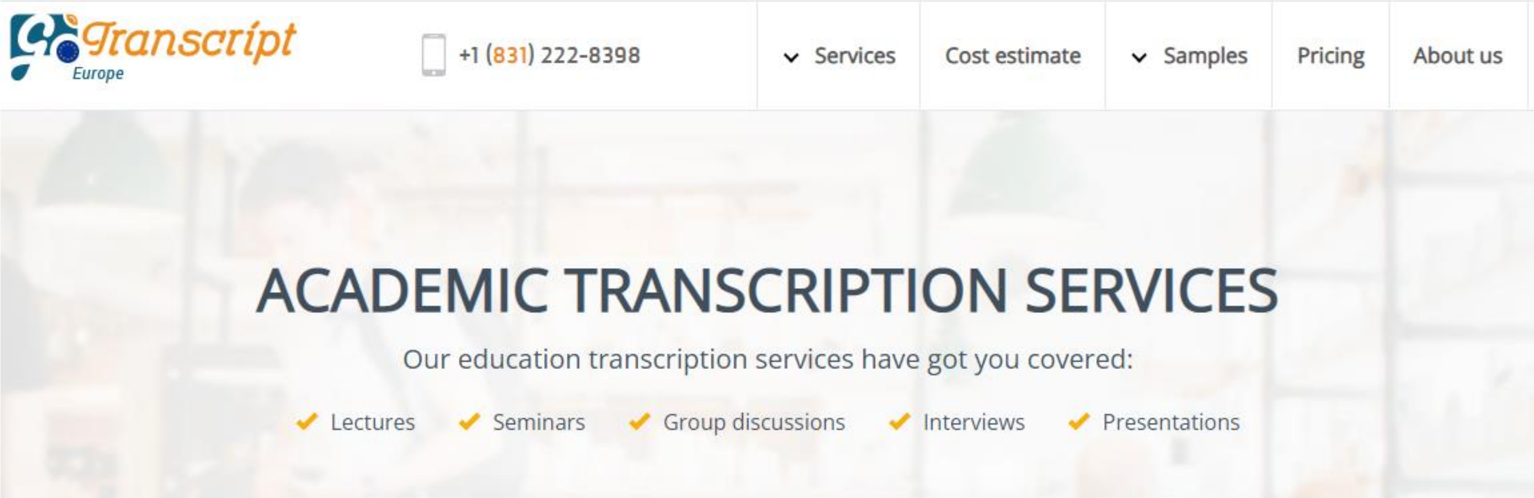
Contents

- Why is AI Engineering Important? A small case study.
- Skills required for AI Engineering
- What makes software with ML challenging?

Why is AI Engineering important?

A small case study

Case study



https://gotranscript.com/

- Take audio or video files and produce text
  - Used by academics to analyze interview text
  - For podcast show notes
  - Generation of subtitles for videos

- State of the art: Manual transcription, often mechanical turk

Idea: Let‘s use AI!

- Deep Neural Network (DNN) trained on publicly available and transcribed interviews
    - For example, from a public news channel

- Use transfer learning to support domain-specific terminology
    - Requires smaller corpus from the domain (e.g., medicine, engineering, etc.)

- Research shows that this works really well!

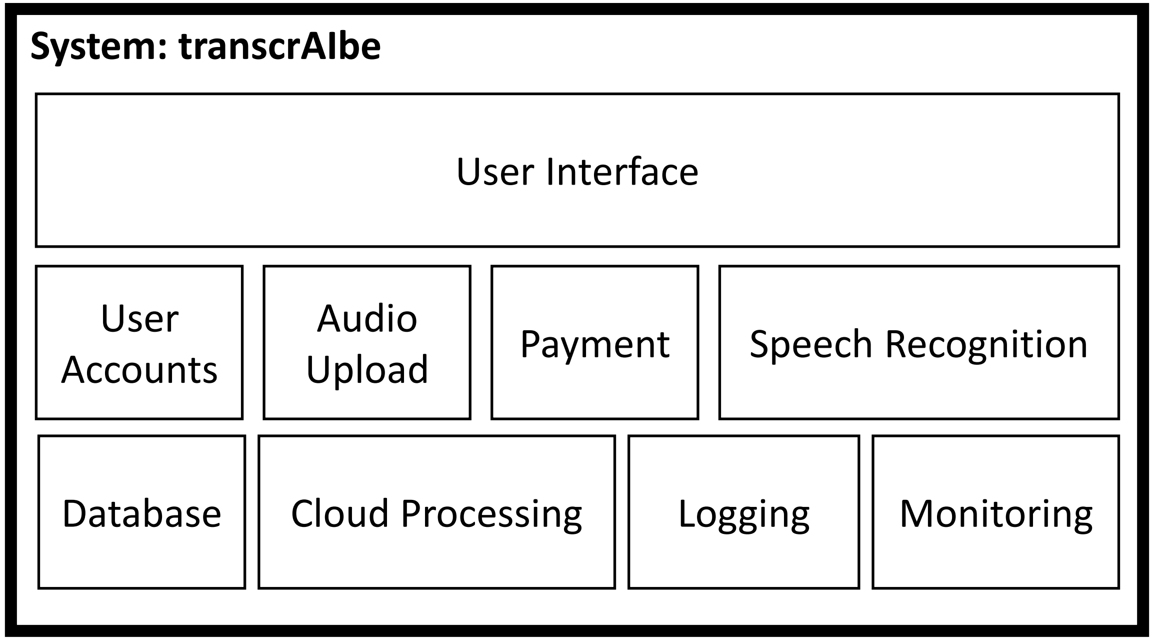
→ Let‘s commercialize this and sell this as a tool called *transcrAIbe*

Live exercise: Challenges for creating *transcrAIbe*

- One challenge for …
  - … the machine learning
  - … the engineering for building the product
  - … operating and updating the product
  - … for the team and the management
  - … the business side
  - … safety and ethics

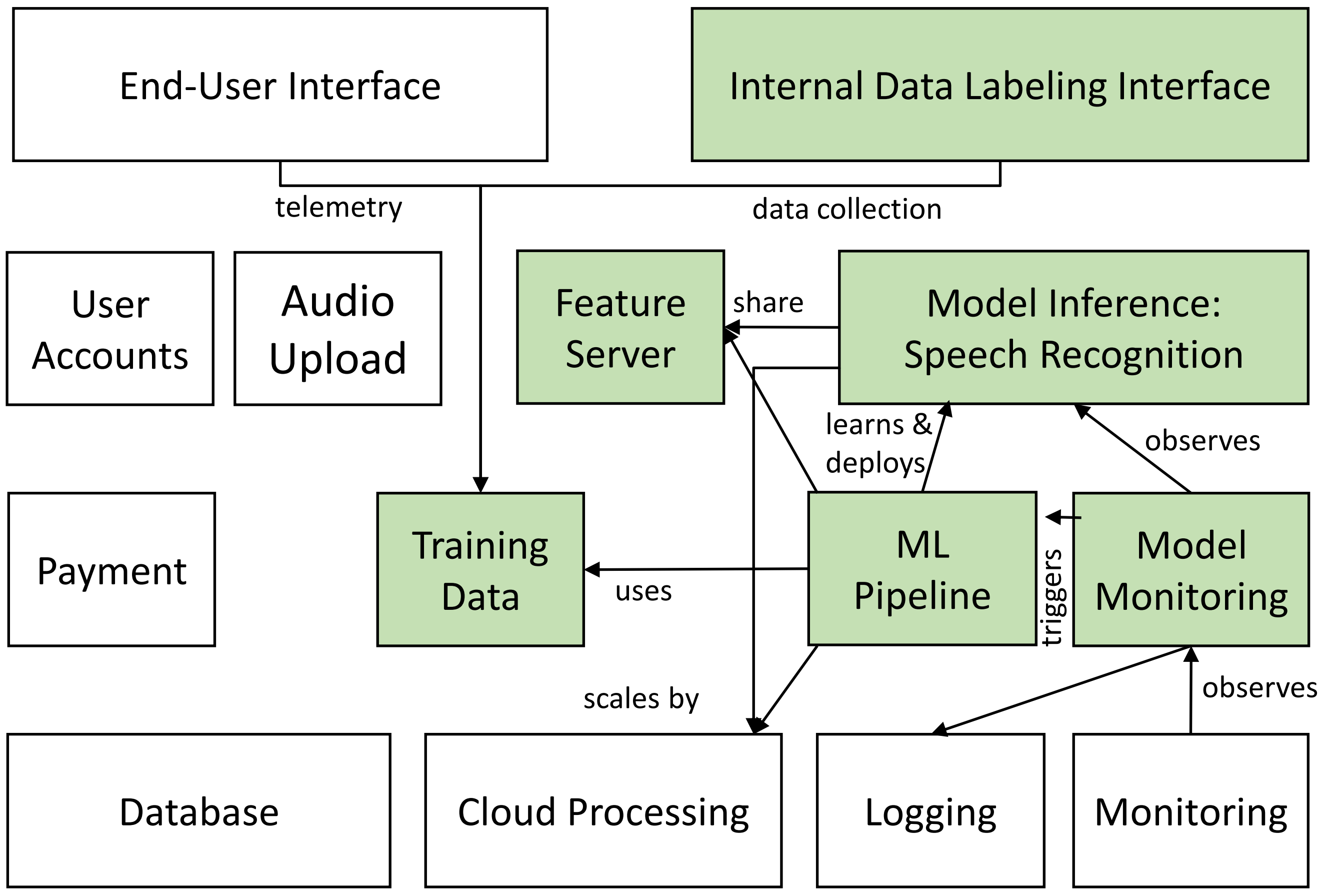
- Without this slide, how many of these aspects would you have considered?!

Possible components of *transcrAIbe*



Only one AI component!

A closer look at ML part of the components



Terminology

- No standard term for referring to building systems with AI components
  - ML-enabled systems
  - **Production systems with ML components**
  - AI-enabled systems
  - ML-infused systems
  - Software engineering for AI (SE4AI)
  - Software engineering for ML (SE4ML)
  - **AI Engineering**

My personel favorites

- Related terms
  - MLOps: technical infrastructure for automating ML pipelines
  - ML systems engineering: building distributed, scalable ML and data storage platforms

Skills required for AI Engineering

Data Scientists vs. Software Engineers

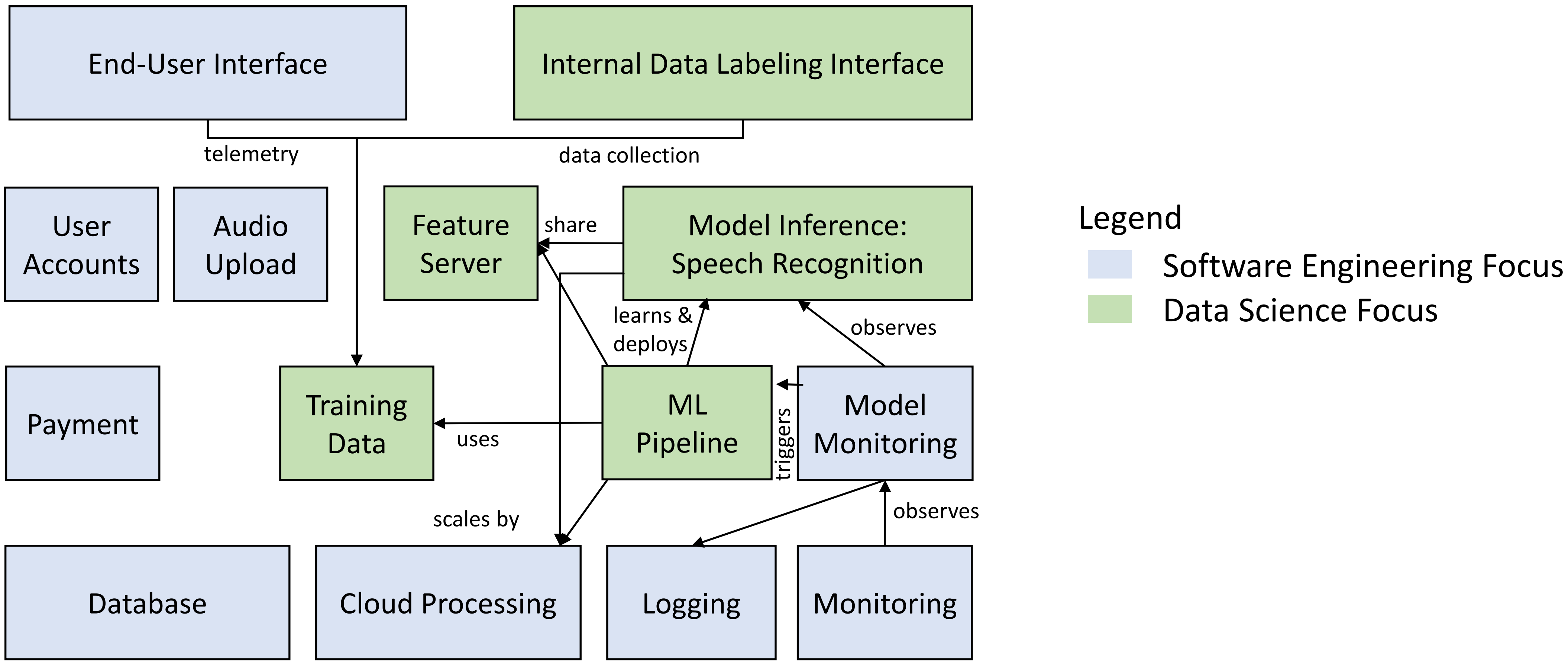
Data Scientist

- Often fixed datasets for training and testing
- Focused on accuracy
- Expert in data modelling and feature engineering
- Code often prototypical and hacky (e.g., Jupyter Notebook)
- Model size, updateability, implementation stability usually ignored

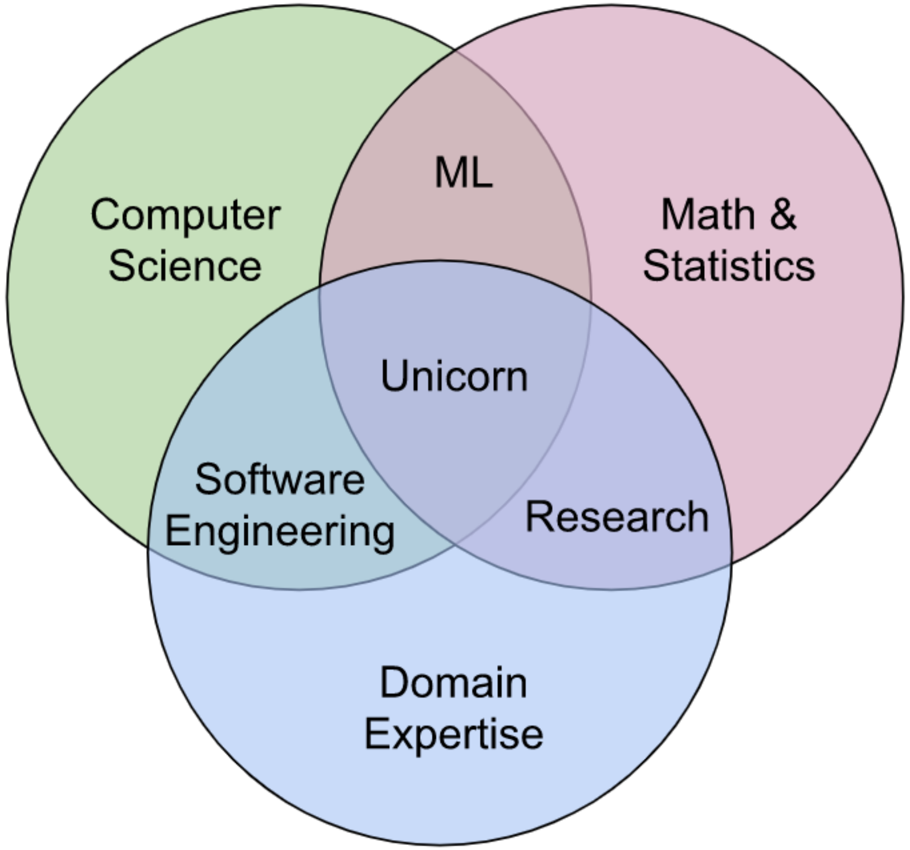
Software Engineer

- Builds a product
- Concerned about cost, performance, stability, release time
- Measures quality through user satisfaction
- Detects and handles mistakes
- Maintains, evolves, and extends product
- Considers non-functional requirements such as security and fairness

Different focus



Extremely rare skillset required!



By Steven Geringer, via Ryan Orban. Bridging the Gap Between Data Science & Engineer: Building High-Performance Teams. 2016

T-shaped people

I-shaped

Expert at one thing

Generalist

Capable of a lot of things
but not expert in any

T-shaped

Capable of a lot of things
and expert in one of them

Example of T-shaped skill set

T-shaped

Basic skills in software engineering, distributed computing, and communication

Expert in deep neural networks for computer vision (technique expertise) and their use in the automotive domain (domain expertise)

What makes software with ML challenging?

ML models make mistakes

NeuralTalk2: A flock of birds flying in the air
Microsoft Azure: A group of giraffe standing next to a tree
Image: Fred Dunn, https://www.flickr.com/photos/gratapictures - CC-BY-NC

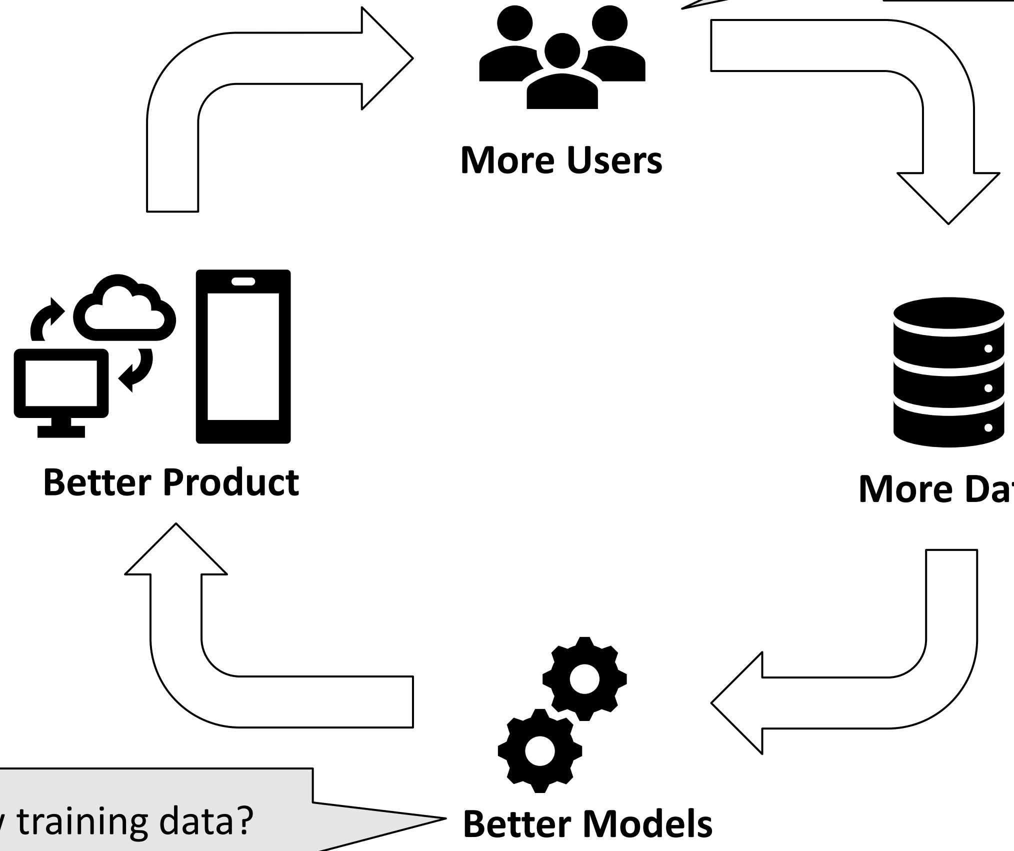
Usually no clear specification that can be tested

```java
/**
 * Returns the text spoken within the audio file
 * @param audioFile handle of a file with a supported audio format (mp3, wav, ogg)
 * @return String with the extracted text
 */
public String transcript(File audioFile) {
    ...
}
```

How would you write a software test for the correctness this function!?

Scaling and Concept Drift

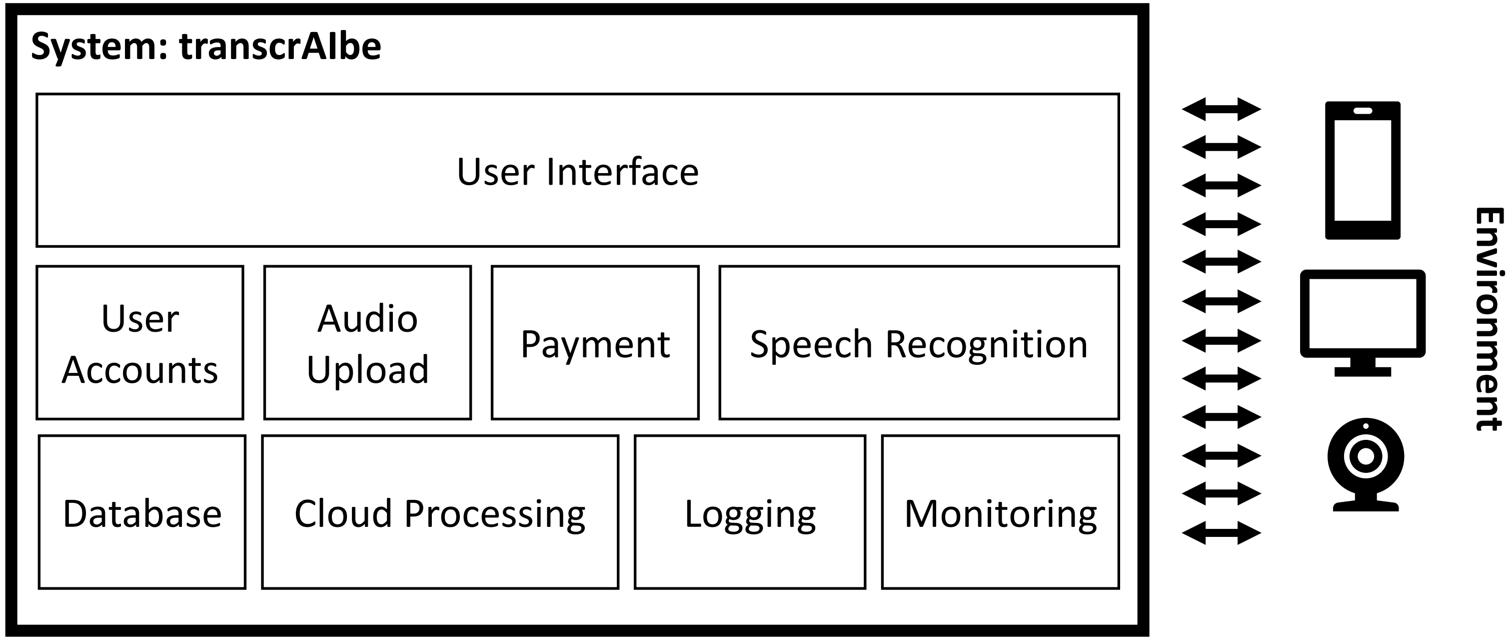
"Normal" up-scaling of applications, e.g., through load balancing
→ Common SE problem



How can this data be used to detect when the model must/should be updated?
→ Requires SE and ML considerations

Same model with new training data? New Model? How to ensure quality?
→ Requires SE and ML considerations

Interactions with the Environment



Not everything needs to be re-invented!

- Software can be safe, with unreliable components
  - E.g., through redundancy
- Cyberphysical systems often have similar properties
- Many data and scaling issues also present without ML
  - Big data, cloud computing
- ML only needs to be “good enough” and “fit for the purpose”, not “correct”

→ ML is just one more challenge for software engineering!

Questions?