# Smart Water

## Large Scale Data Labeling with Active Learning and $H_2O$.ai

FSDL Final Project

Yufeng Wang

May 15th 2021

# Menu

- Intuition
- Active Learning
- Solution Architecture
- Demo
- Next Steps

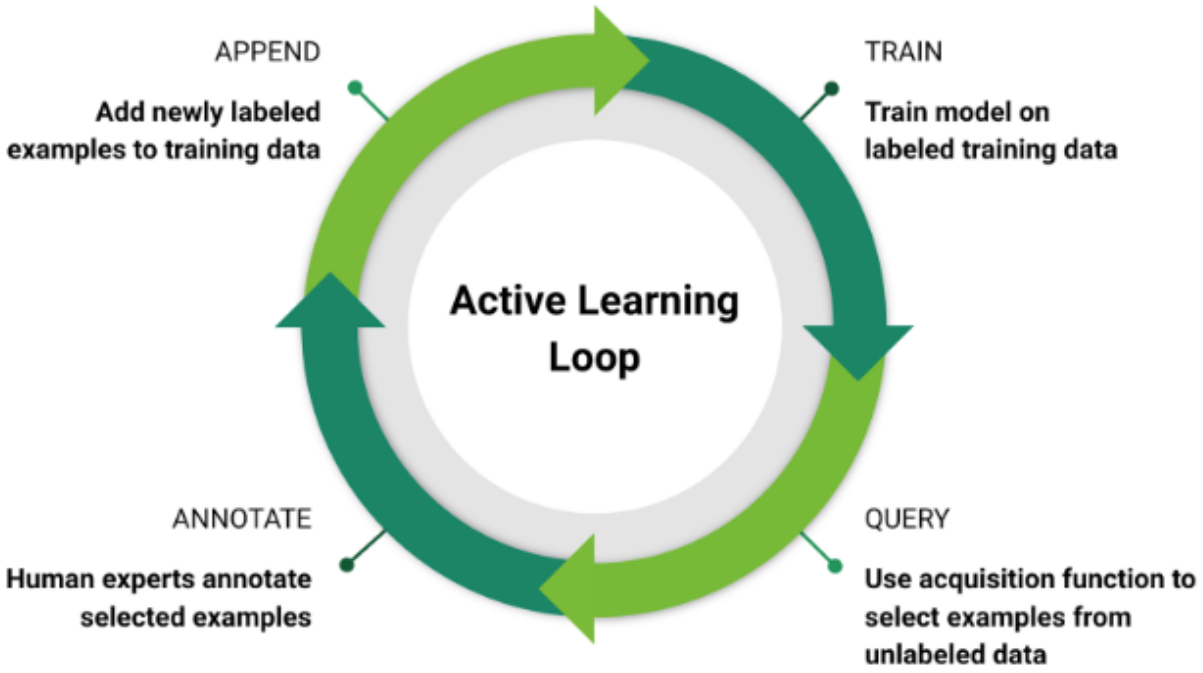
smart water

# Intuition

- Data is Food for AI.
  - Source and Prepare high quality data.
- From Model-centric to Data-centric AI
  - From Big Data to Good Data
  - Labeling data better can improve model performance
- Data Labeling is expensive
  - Human expert: domain knowledge, time and money
  - Should label the data smartly

smart water

# Active Learning Strategy

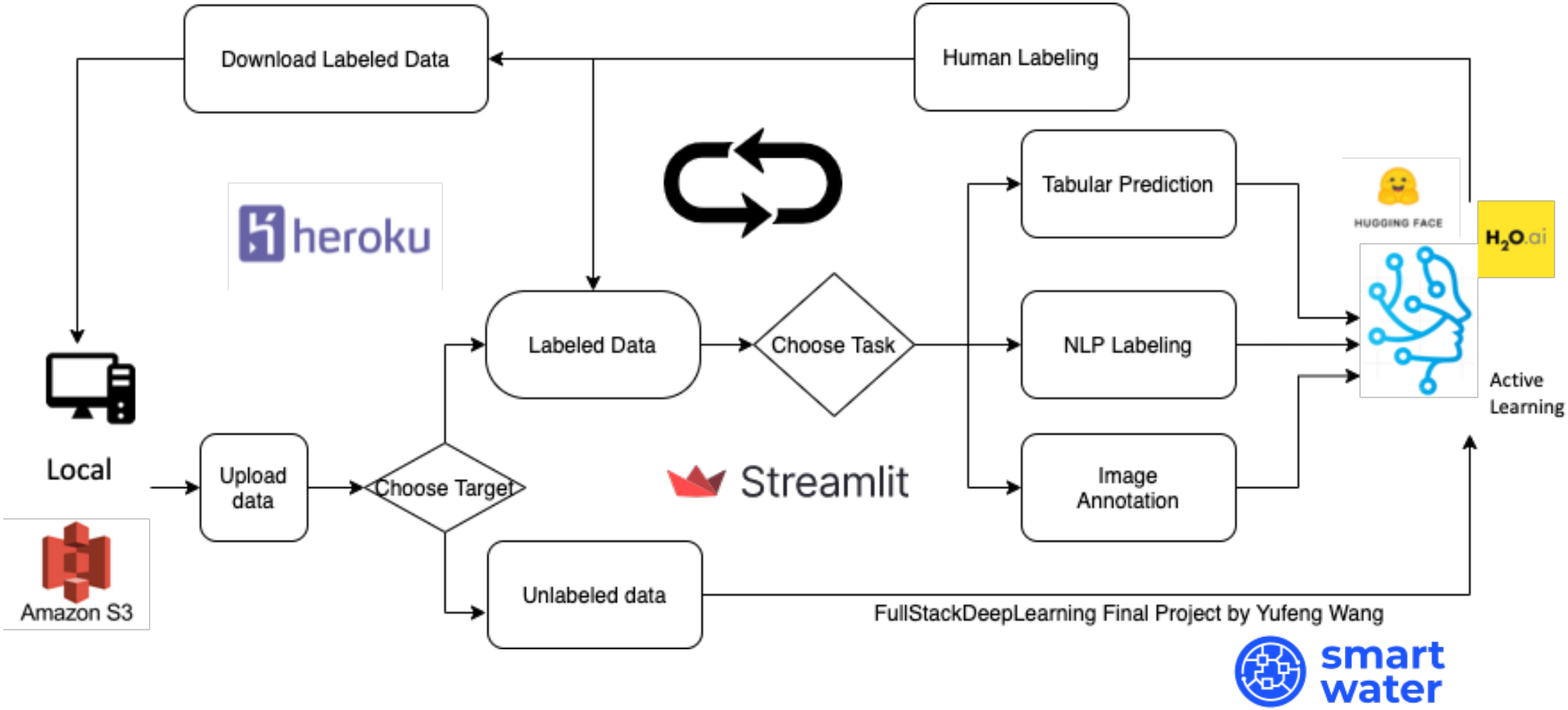



- Margin Sampling
  - select k samples with the lowest difference between two highest class probabilities.
- Entropy Selection
  - select ‘k’ samples with the highest entropy, i.e., with high uncertainty.
- Least Confidence
  - select ‘k’ samples with the least confidence (max probability class)

smart water

# Solution Architecture: Smart Water

## Large Scale Data Labeling with Active Learning and $H_2O.ai$



# Demo

- Local: http://localhost:8501/
- Heroku: https://smartwater001.herokuapp.com/

# Next Steps

- Optimization for computing resources to get labeling faster
  - GPU Parallel computing (now CPU)
  - Code efficiency (now no code review, just by myself)
  - Cloud Computing ( now with local machine)
- Build For more and more tasks
  - NLP based tasks: Sentiment analysis, Named Entity Recognition, Topic Modeling, etc.
  - Image based tasks: image recognition, multi-object classification, etc
- More human-friendly Layout and user interaction