ST. PÖLTEN UNIVERSITY
OF APPLIED SCIENCES

research

fh st. pölten

Unlocking the Power of Active Learning

A Hands-on Exploration

Fabian Kovac

fabian.kovac@fhstp.ac.at
St. Pölten University of Applied Sciences, Austria

Oliver Eigner

oliver.eigner@fhstp.ac.at
St. Pölten University of Applied Sciences, Austria

Hello There!

research

Fabian Kovac

fabian.kovac@fhstp.ac.at
St. Pölten University of Applied Sciences, Austria

Oliver Eigner

oliver.eigner@fhstp.ac.at
St. Pölten University of Applied Sciences, Austria

Timetable

General Structure

- Before break (14:00 - 15:30)
  - Introduction
  - Background to Active Learning
  - Challenges
  - From theory to practice
  - Why/when to use
  - Sampling methods/strategies
- Break (15:30 - 16:00)
- After break (16:00 - 17:30)
  - Hands-on session (demo time)
  - Road ahead and other ideas using the same concept

Workshop Material

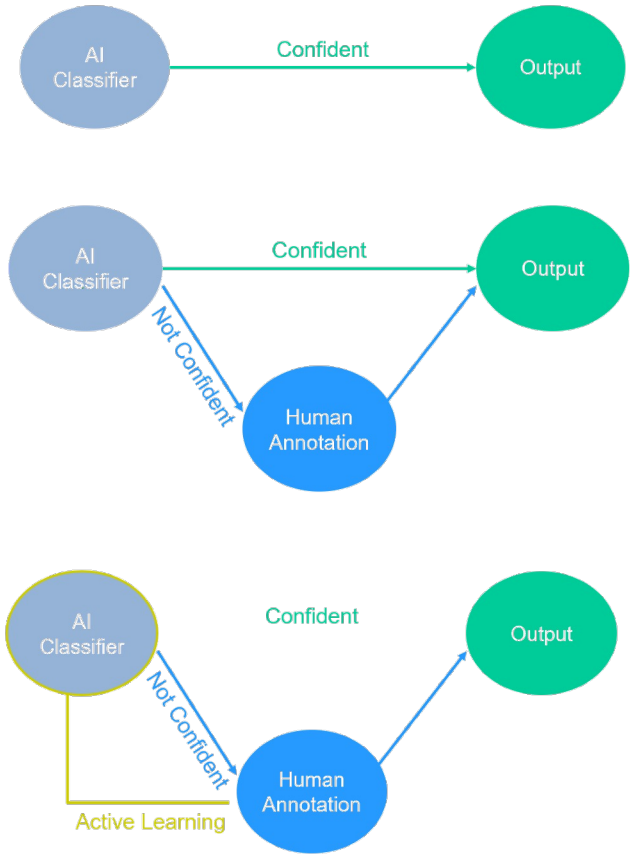
research fh st.pölten

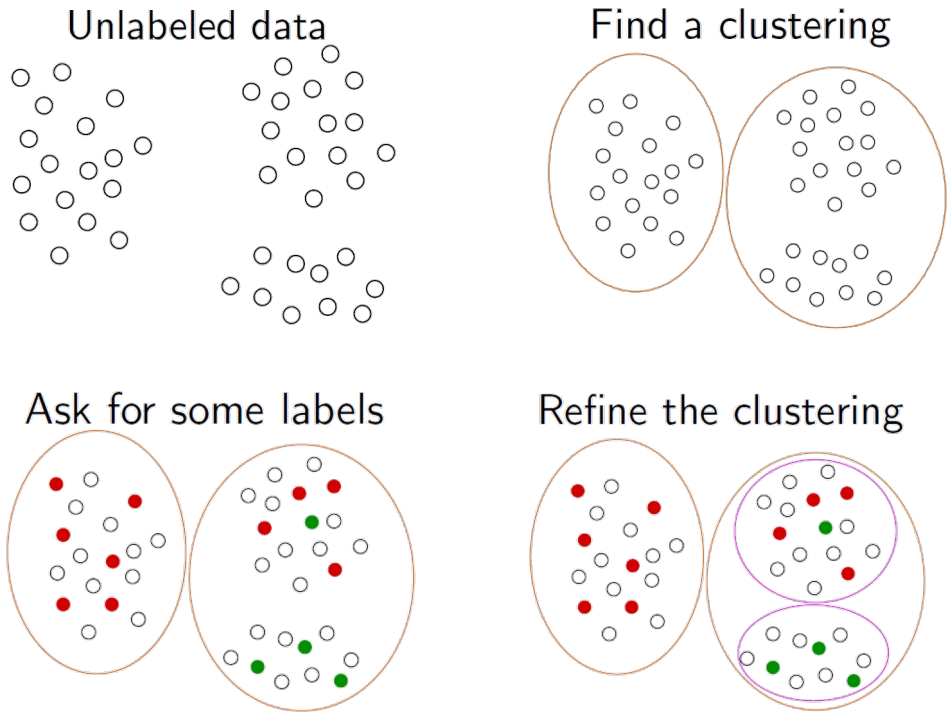
https://github.com/fkabs/mlprague_2024

„ fkabs/mlprague_2024 “

Introduction Human-in-the-loop

Transition to Active Learning





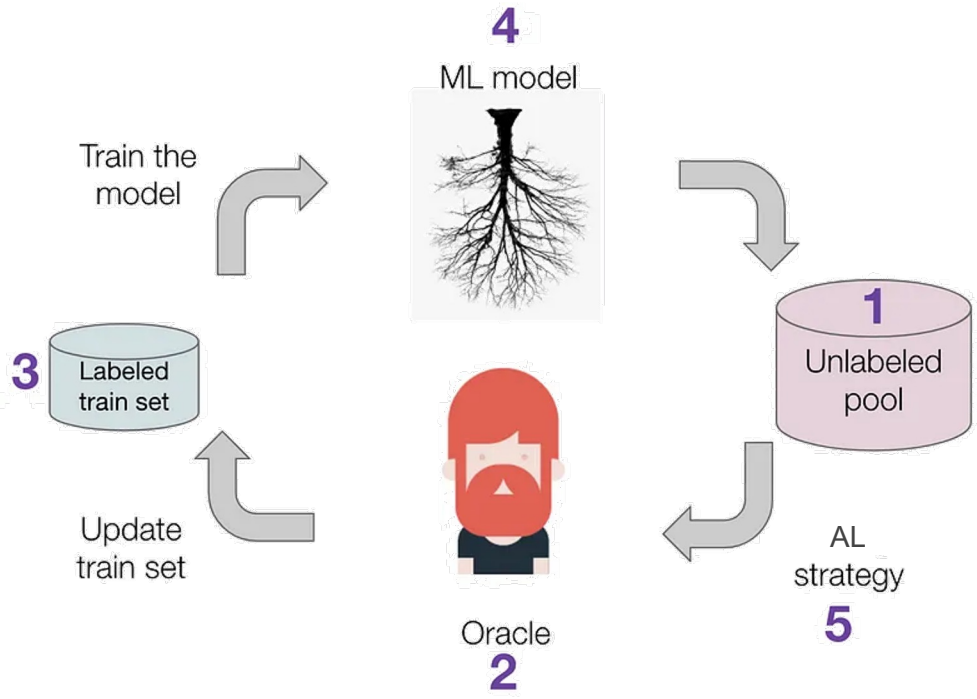
Introduction Human-in-the-loop

Classic and Active Learning ML comparison

research

fh st.pölten

	Classic ML	Active Learning ML
Data sampling for labeling	Randomly select the data from the unlabelled pool (**1**) both for train and test set	Smartly select the data for train set - with **AL** *strategy* (**5**)
Data labeling	Assign labels to data using an oracle (**2**)	
Model training	Train the ML model (**4**)	
Model evaluation	Check model performance on test/val set	



https://colab.research.google.com/drive/1G_4o-1_CkR4eNgzGVcx89IJ_IKLBQQv-?usp=sharing

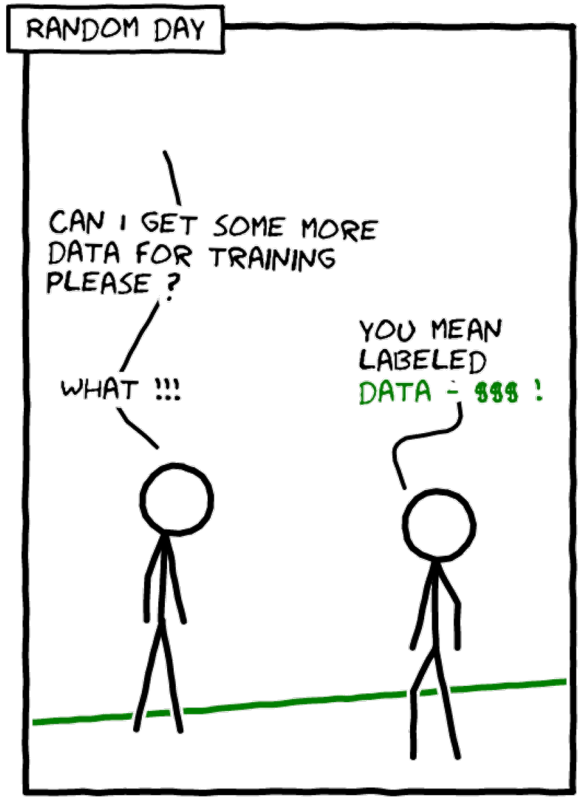
Introduction Human-in-the-loop

Active Learning - Labeling

research

- Why is labeling expensive:
  - Labeling requires human to go through your example and judge it
  - To ensure quality, you would need multiple humans to judge the same example and take a vote.
  - Preparing rating guidelines for your task and keeping them updated

- Typical Active learning setup:
  - labeled training set (seed data)
  - base estimator (active learner) trained on this set
  - instances from unlabeled pool send to human annotator



Introduction Human-in-the-loop

Machine Teaching Approach

research

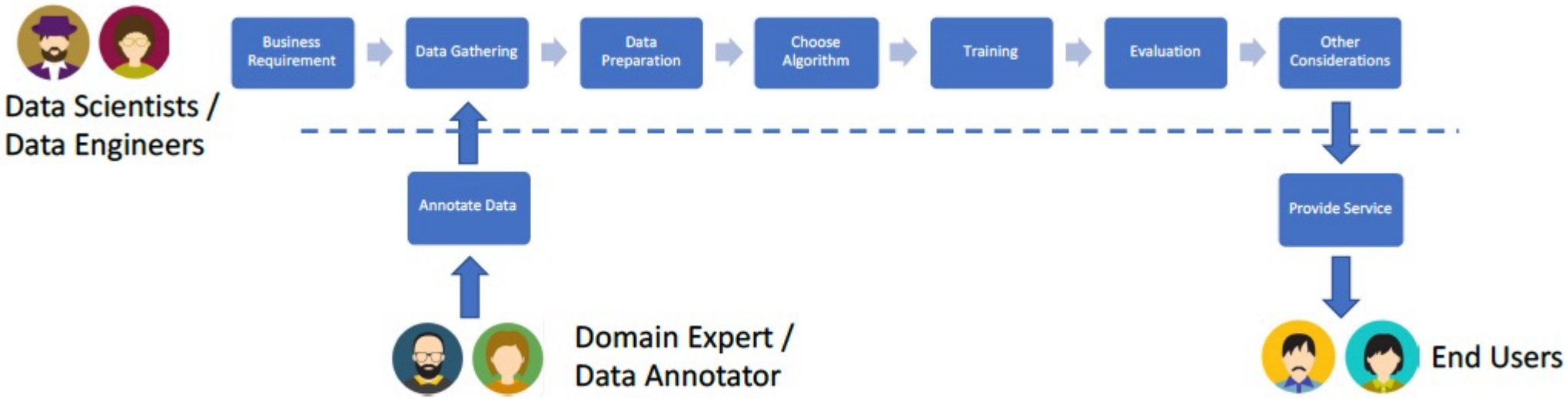


Figure 1: AI Recipe for Data Kitchen

https://lfaidata.foundation/blog/2020/12/11/human-centered-ai-for-bi-industry/

Introduction Human-in-the-loop

Machine Teaching Approach

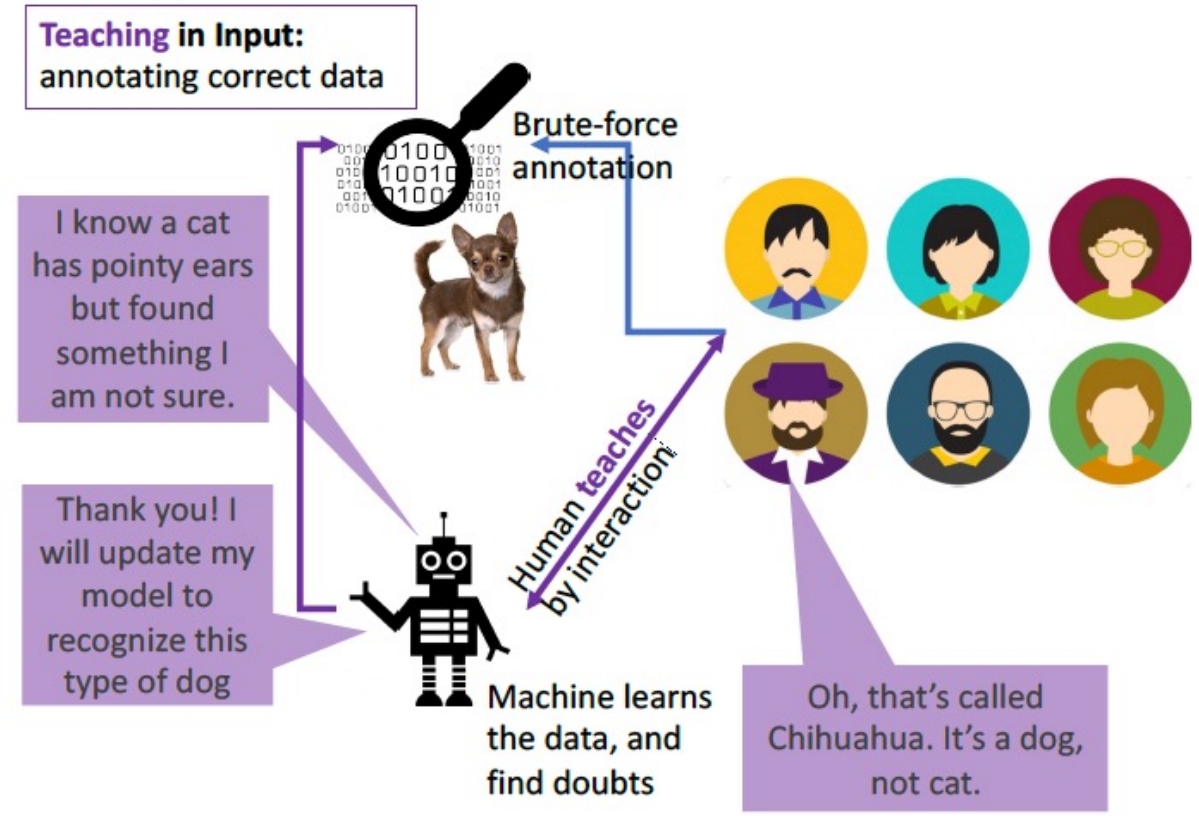


Figure 2: Machine Teaching in Data Input Process

https://lfaidata.foundation/blog/2020/12/11/human-centered-ai-for-bi-industry/

Introduction Human-in-the-loop

Machine Teaching Approach

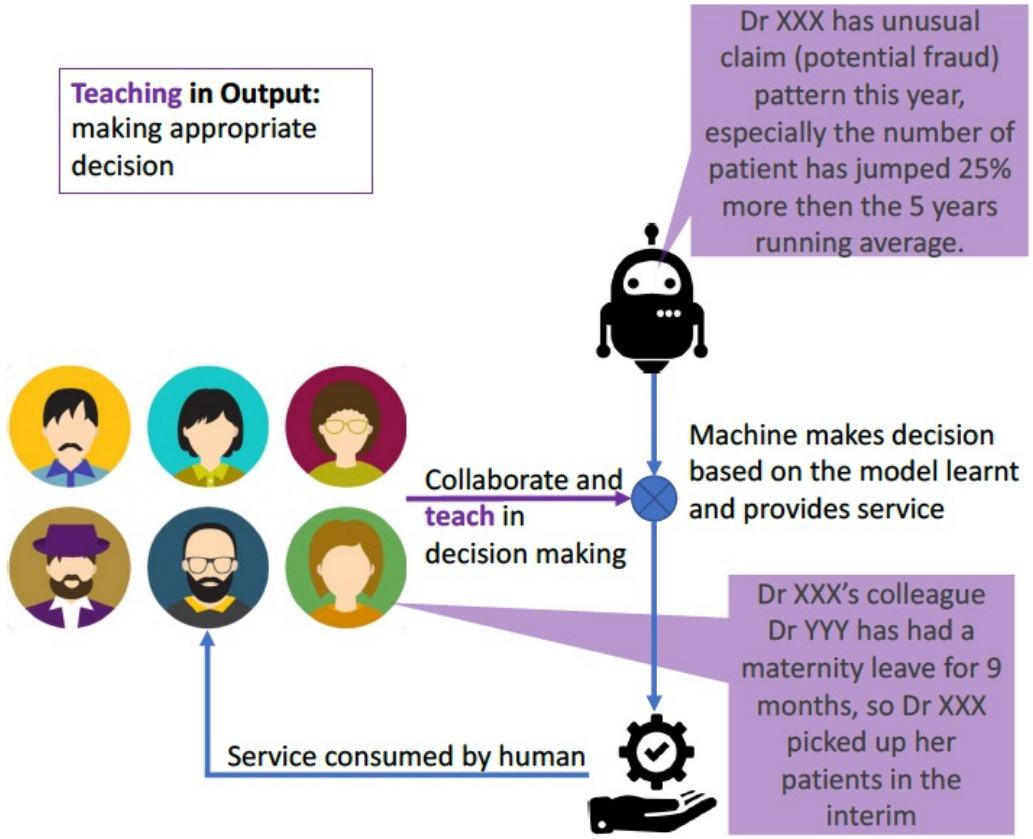


Figure 3: Machine Teaching in Model Decision Process

https://lfaidata.foundation/blog/2020/12/11/human-centered-ai-for-bi-industry/

Introduction Human-in-the-loop

Machine Teaching Approach

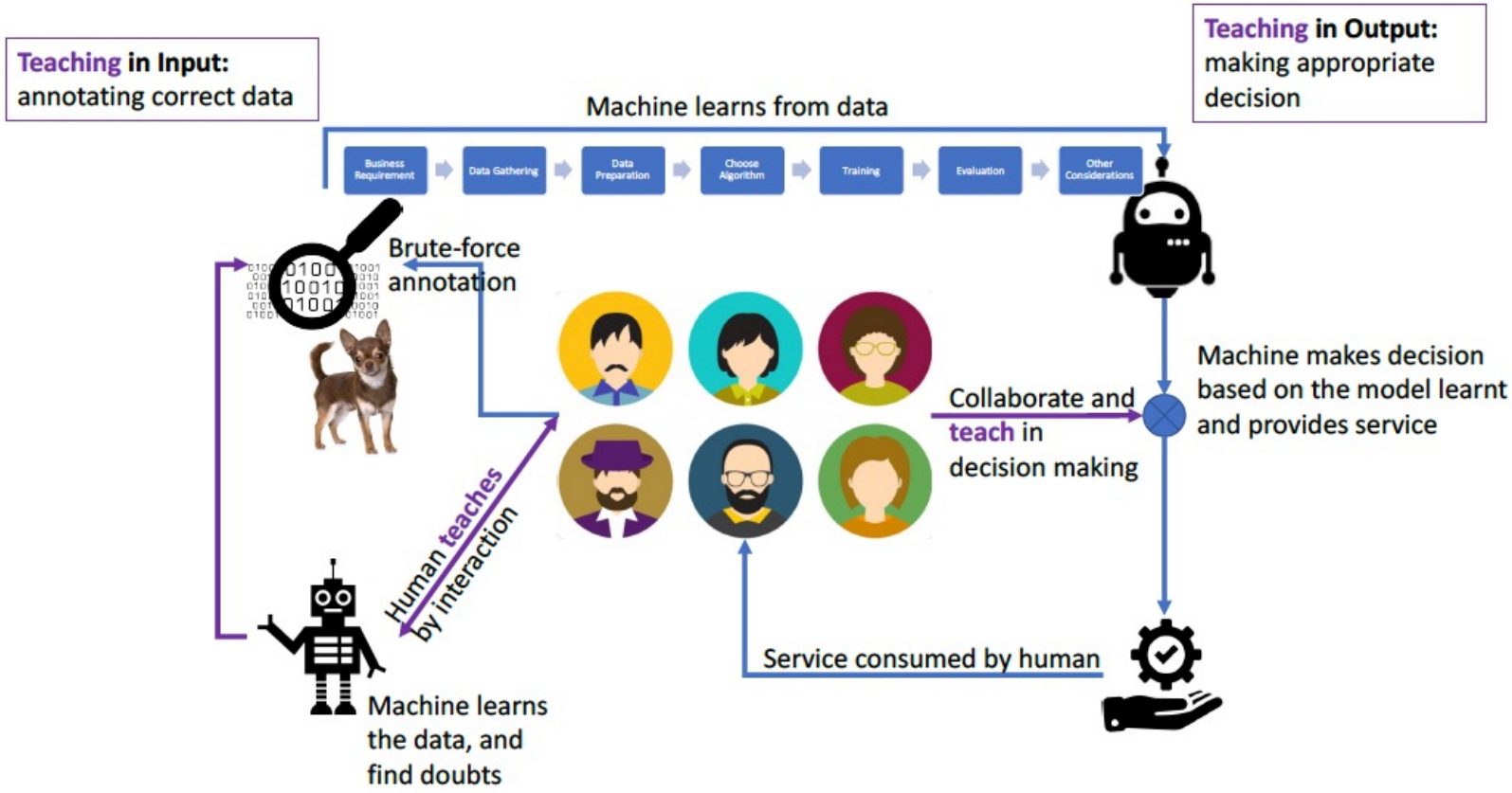


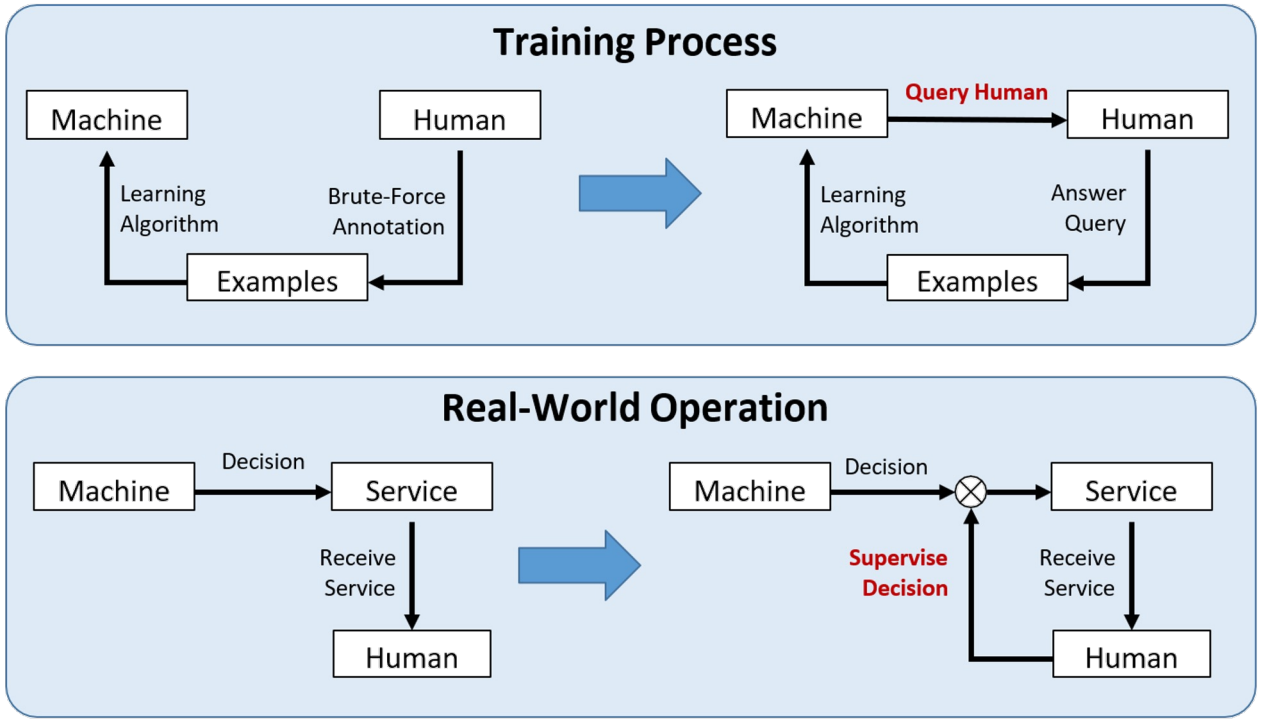
Figure 4: Putting Together of overall Machine Teaching in Human-Centered AI

https://lfaidata.foundation/blog/2020/12/11/human-centered-ai-for-bi-industry/

Introduction Human-in-the-loop

Integrate Human Intelligence into the loop

- (1) training and
- (2) real-world operation of AI Systems

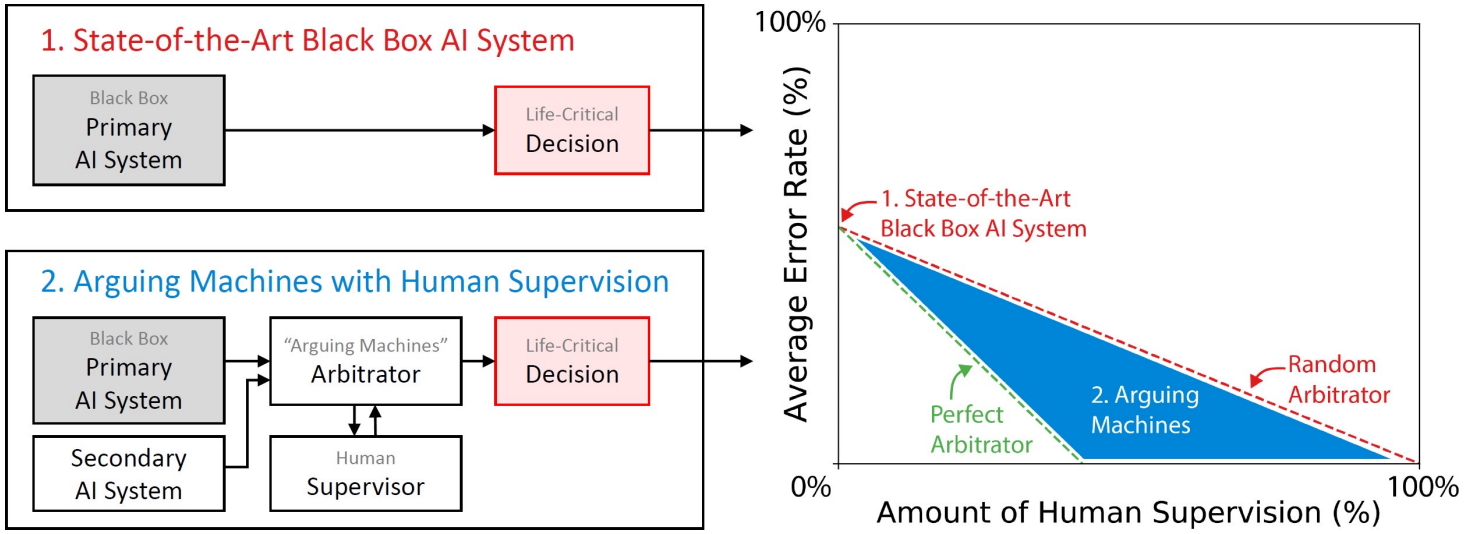


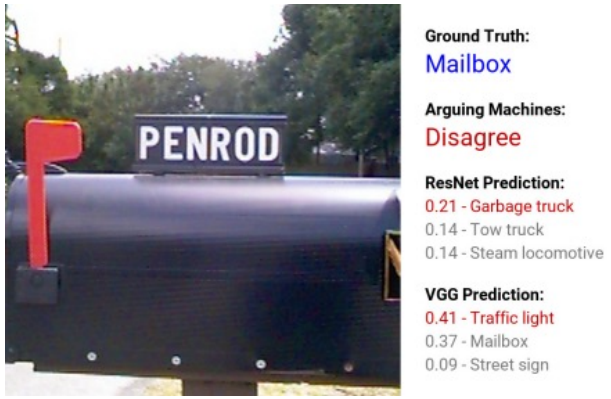
Introduction Human-in-the-loop

Challenge: AI Safety - methods for effective supervision of machines

research

- Human Supervision of AI Systems that Make Life-Critical Decision
  - Avoid catastrophic actions in exploration or unintended consequences of reward function
- Example Challenges:
  - Uncertainty estimate matches (within 5%) the error rate on large-scale image classification problem (that contains examples outside the distribution of the training set)

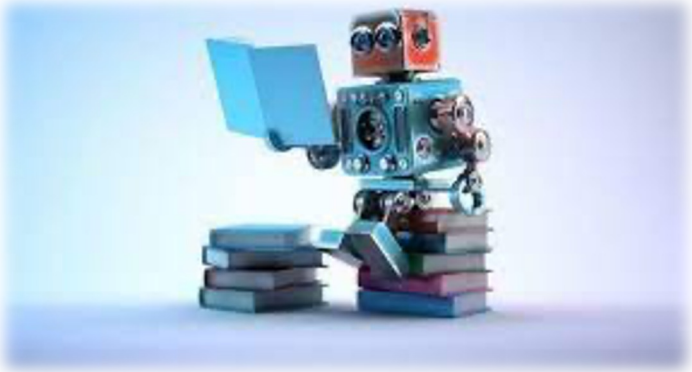




L. Fridman, L. Ding, B. Jenik, and B. Reimer, "Arguing Machines: Human Supervision of Black Box AI Systems That Make Life-Critical Decisions." arXiv, Sep. 24, 2018. doi: 10.48550/arXiv.1710.04459.

Machine Teaching

Active Learning



Active Learning

Challenges

research

theguardian

Computers now better than humans at recognising and sorting images

millions of labeled images
1000's of human hours

QUARTZ

Google says its new AI-powered translation tool scores nearly identically to human translators

trained on more texts than a human could read in a lifetime

FEB 28, 2015 @ 11:25 PM 7,767

Forbes LOG IN

Google's DeepMind Masters Atari Games

A computer that taught itself to play almost 50 video games including Space Invaders and Pong is being hailed as the pinnacle of artificial intelligence.

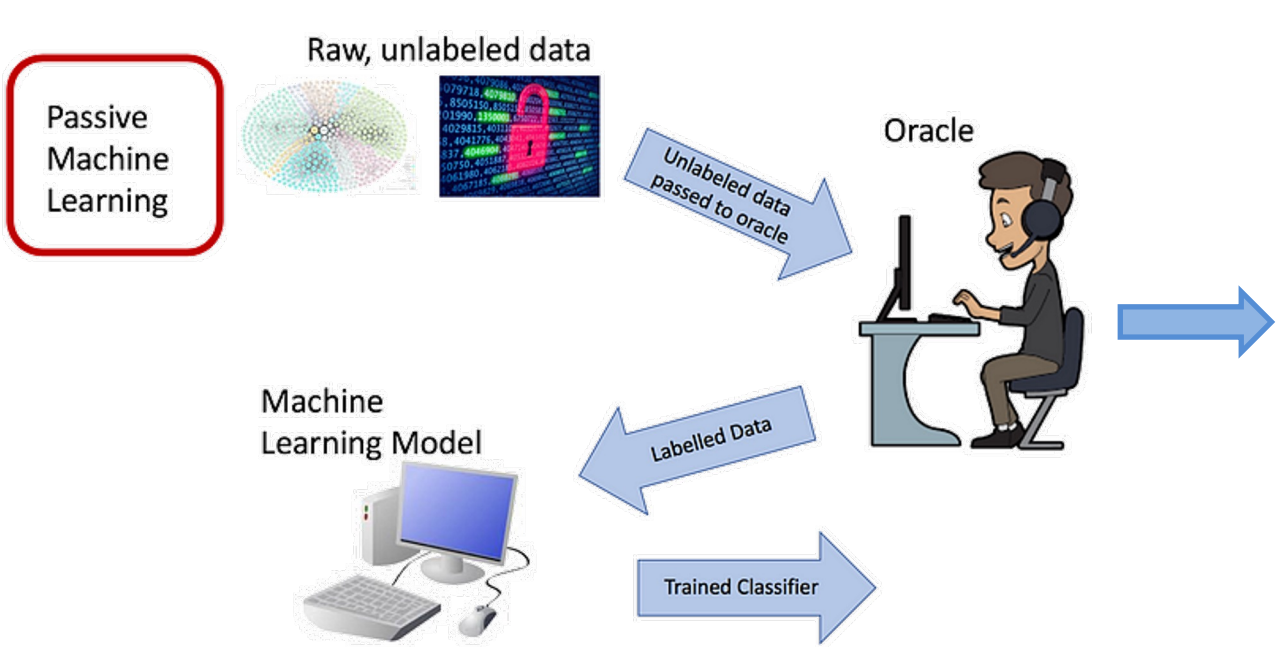
playing more games than even a teenager could stomach !

Can we train machines with less labeled data and less human supervision?

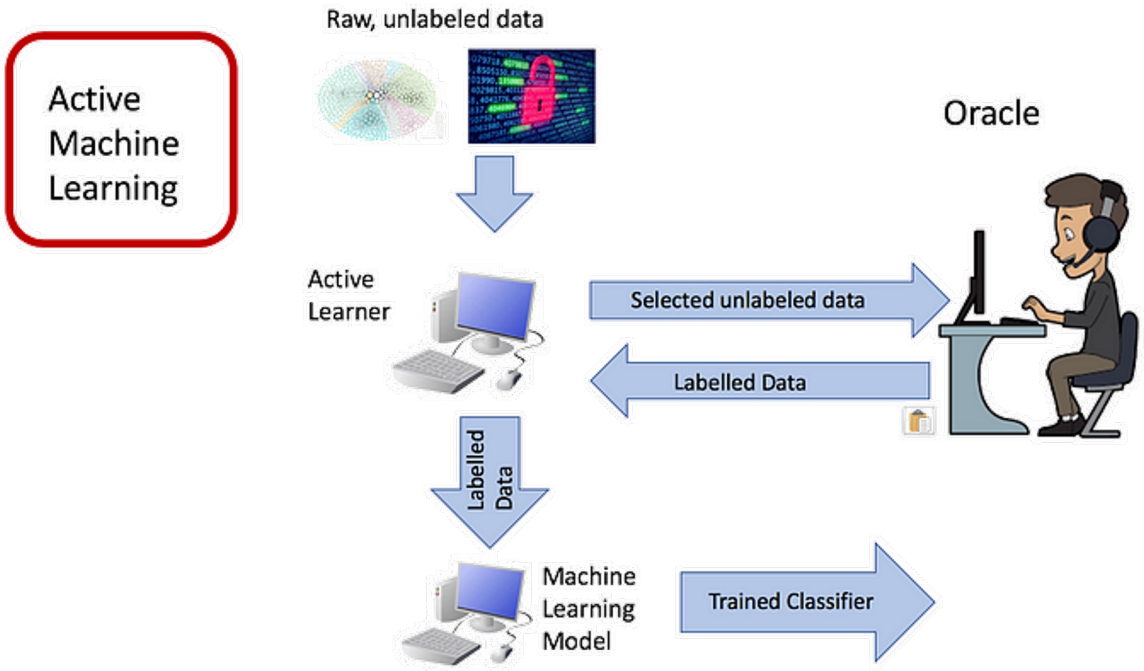
Active Learning
Background

research fh st.pölten

Passive Learning

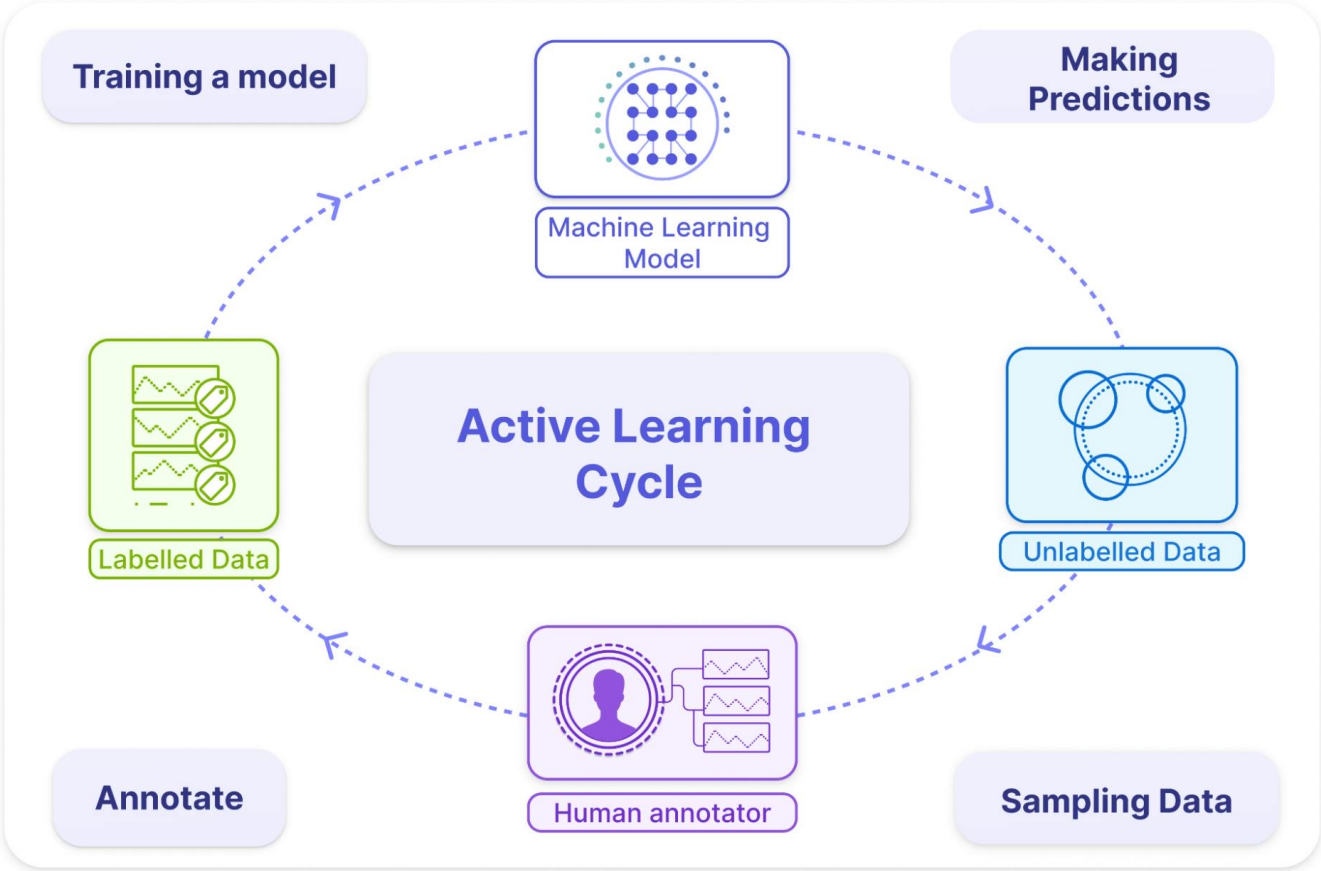


Active Learning



Active Learning

Active machine learning pipeline

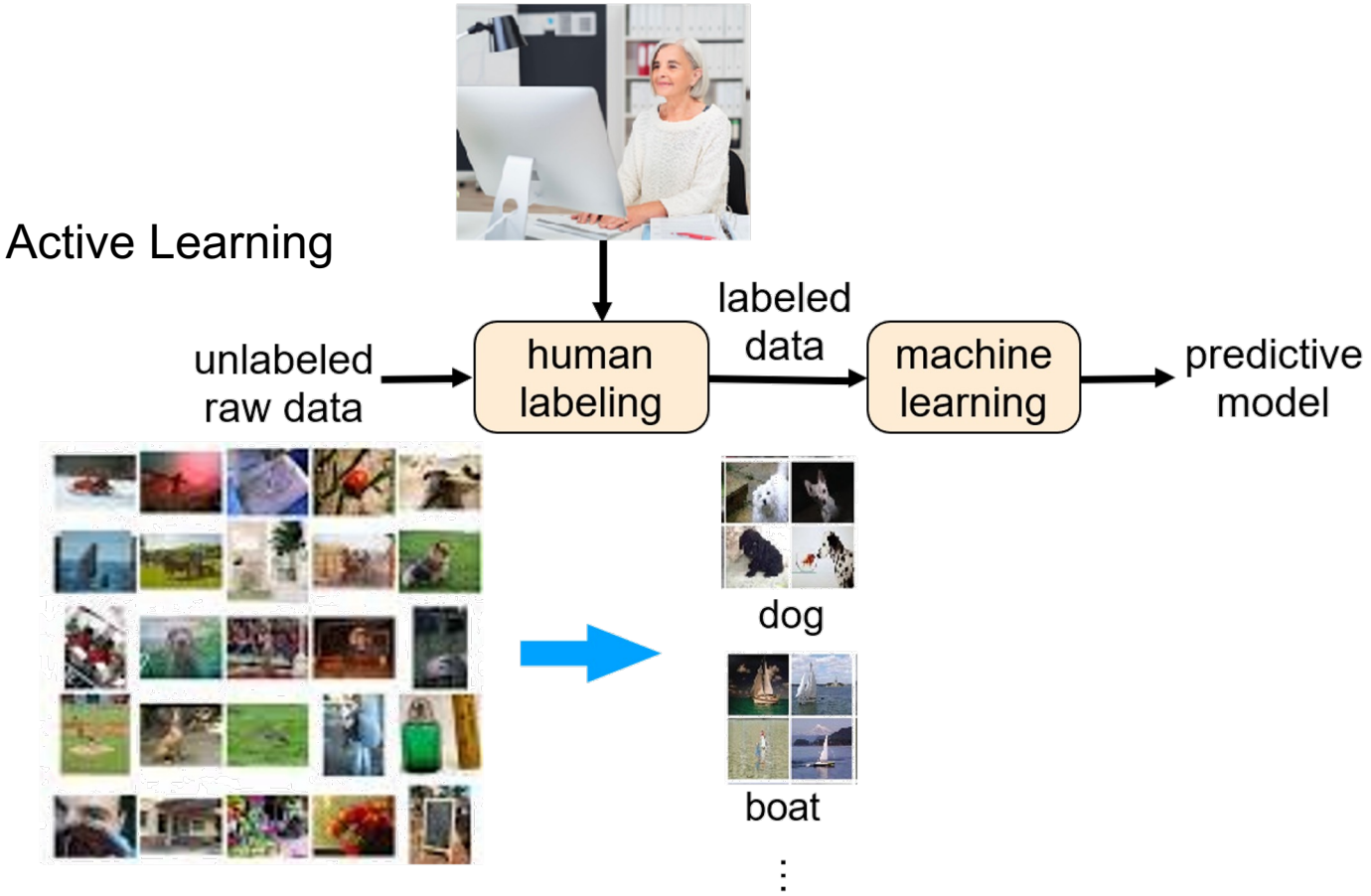


https://encord.com/blog/active-learning-machine-learning-guide/

Active Learning
from Theory to Practice

research fh st.pölten

- Conventional (Passive) Machine Learning

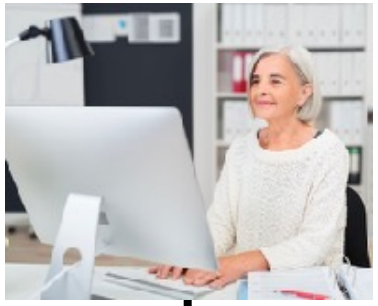


Active Learning

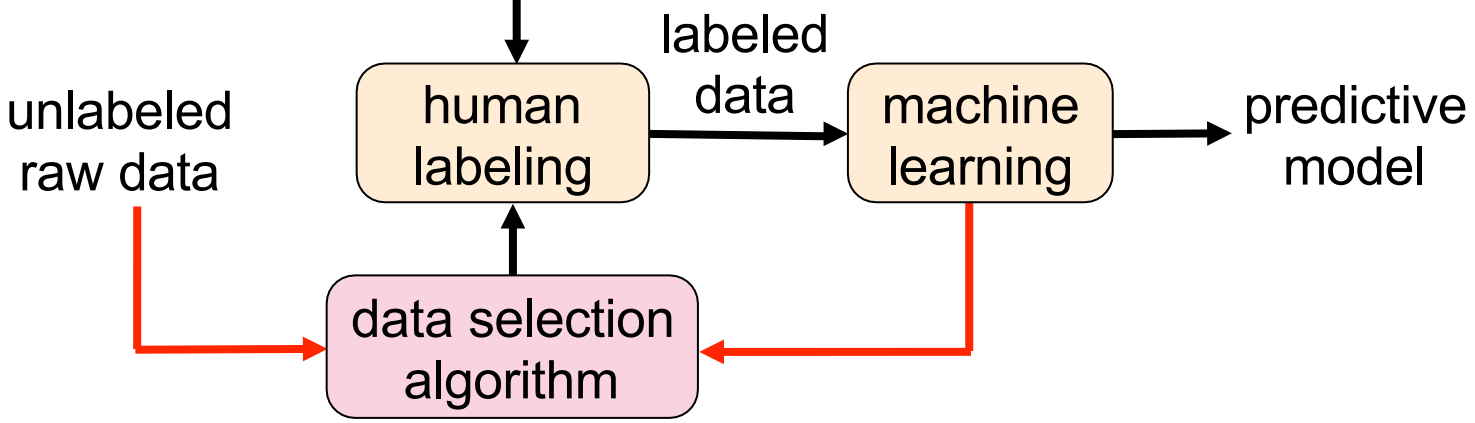
from Theory to Practice

research

fh st.pölten

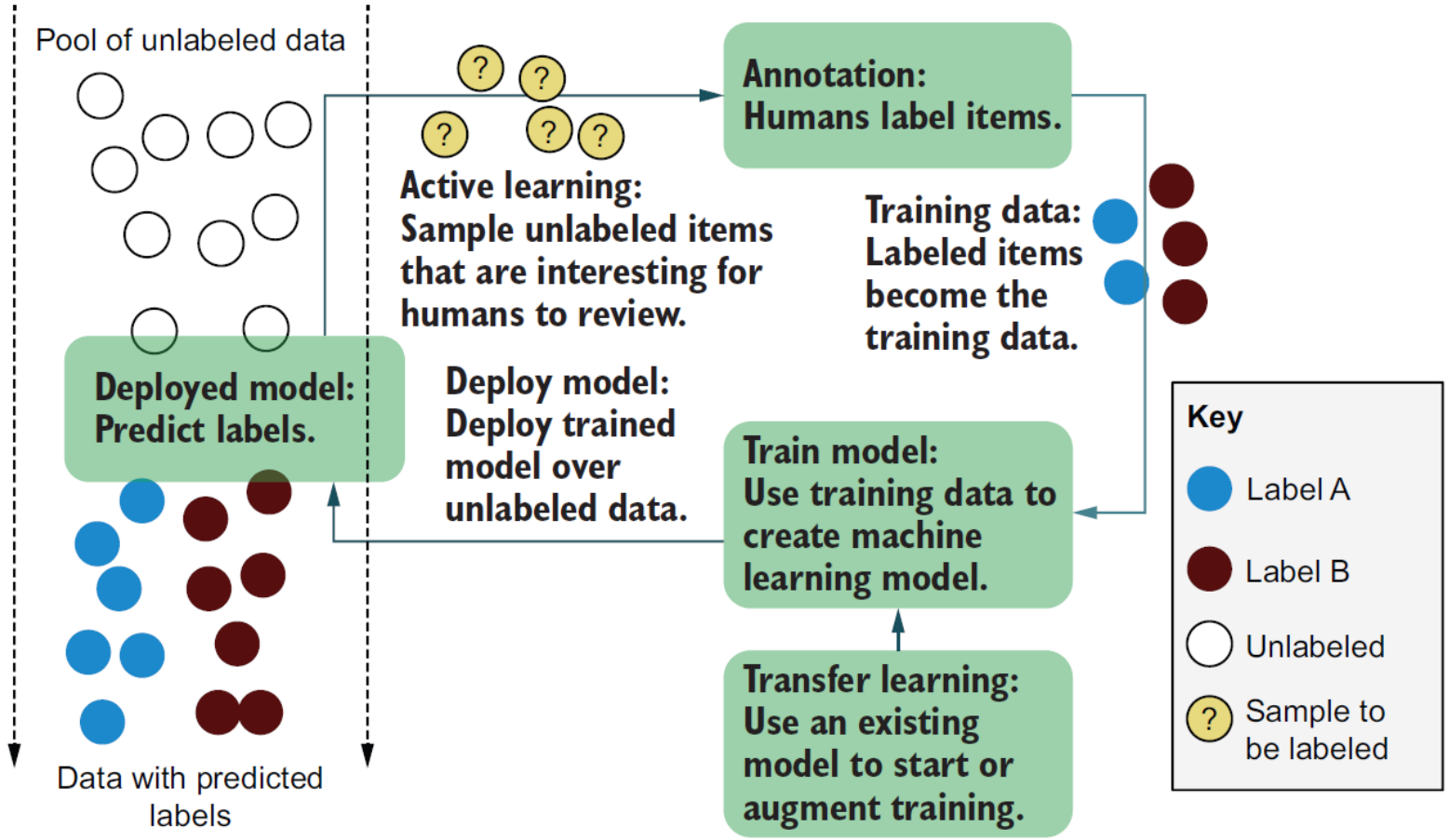


Goal: machine automatically and adaptively selects most informative data for labeling



Active Learning

from Theory to Practice - idea



Munro, Robert, and Robert Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

Active Learning

Sampling Strategies

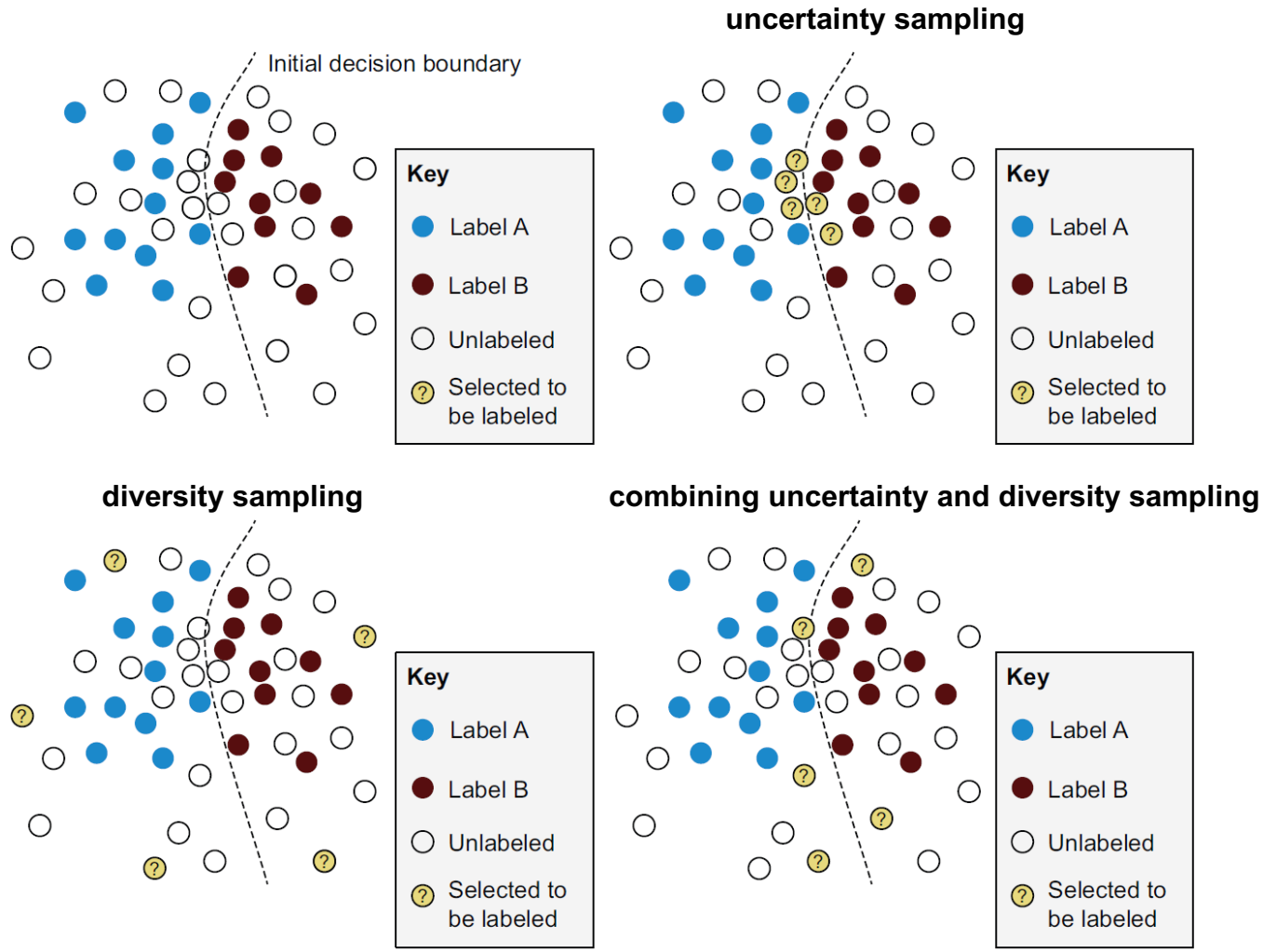
research

- Uncertainty sampling (Exploitation)
  - set of strategies for identifying unlabeled items that are near a decision boundary
- Diversity sampling (Exploration)
  - set of strategies for identifying unlabeled items that are underrepresented or unknown to the model
  - also called outlier detection or anomaly detection
- Random sampling
  - simplest but can be the trickiest

Active Learning

Pros and cons of different strategies

research



Monarch, Robert Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

Active Learning

Pros and cons of different strategies

research

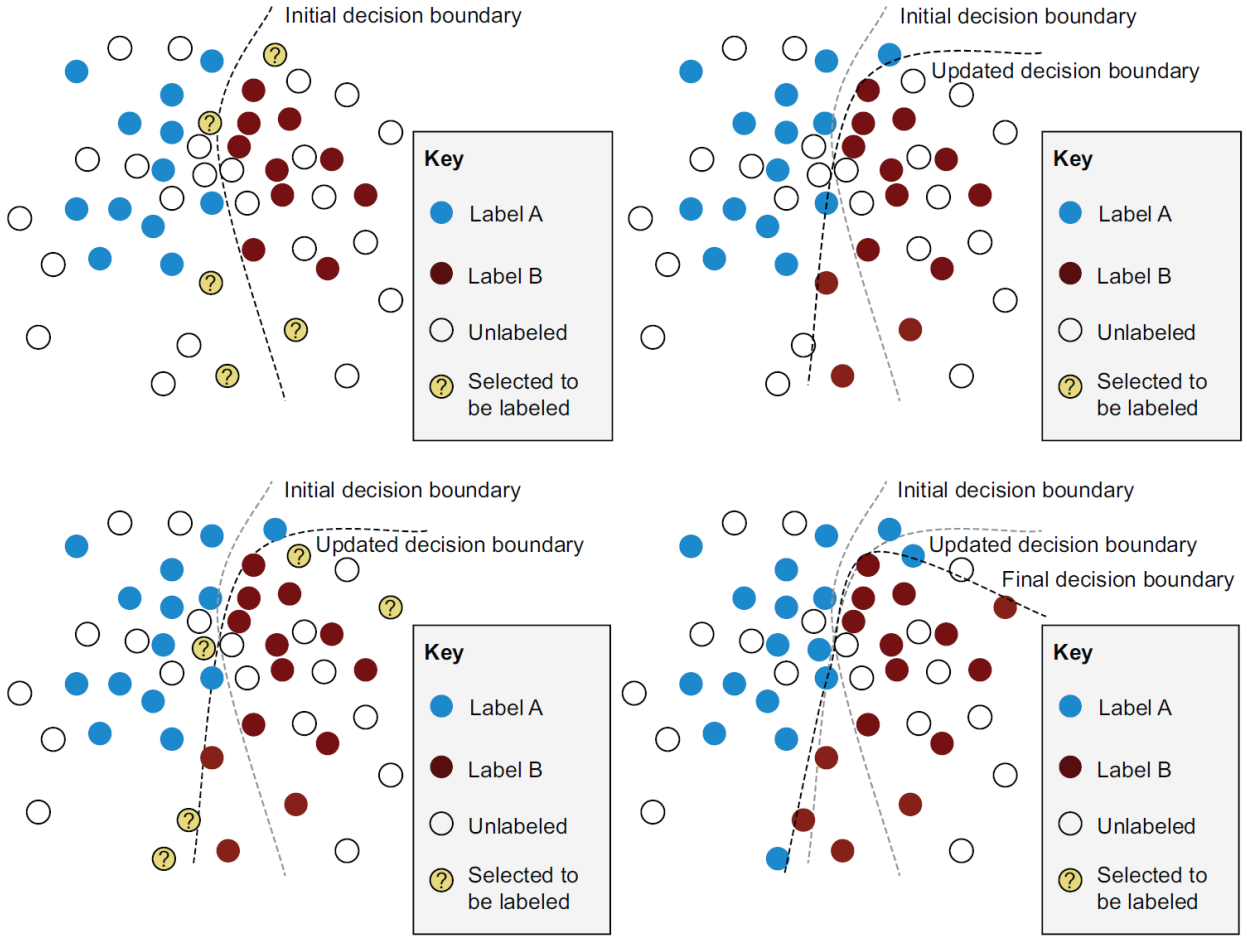
- Important: active learning process is iterative
- In each iteration of active learning, a selection of items is identified and receives a new human-generated label.
- Then the model is retrained with the new items, and the process is repeated
  - e.g., two iterations for selecting and annotating new items, resulting in a changing boundary
- The number of iterations and the number of items that need to be labeled within each iteration depend on the task.

Active Learning

Iterative active learning process

research fh st.pölten

combining uncertainty and diversity sampling



Monarch, Robert Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

Active Learning

When to use?

research

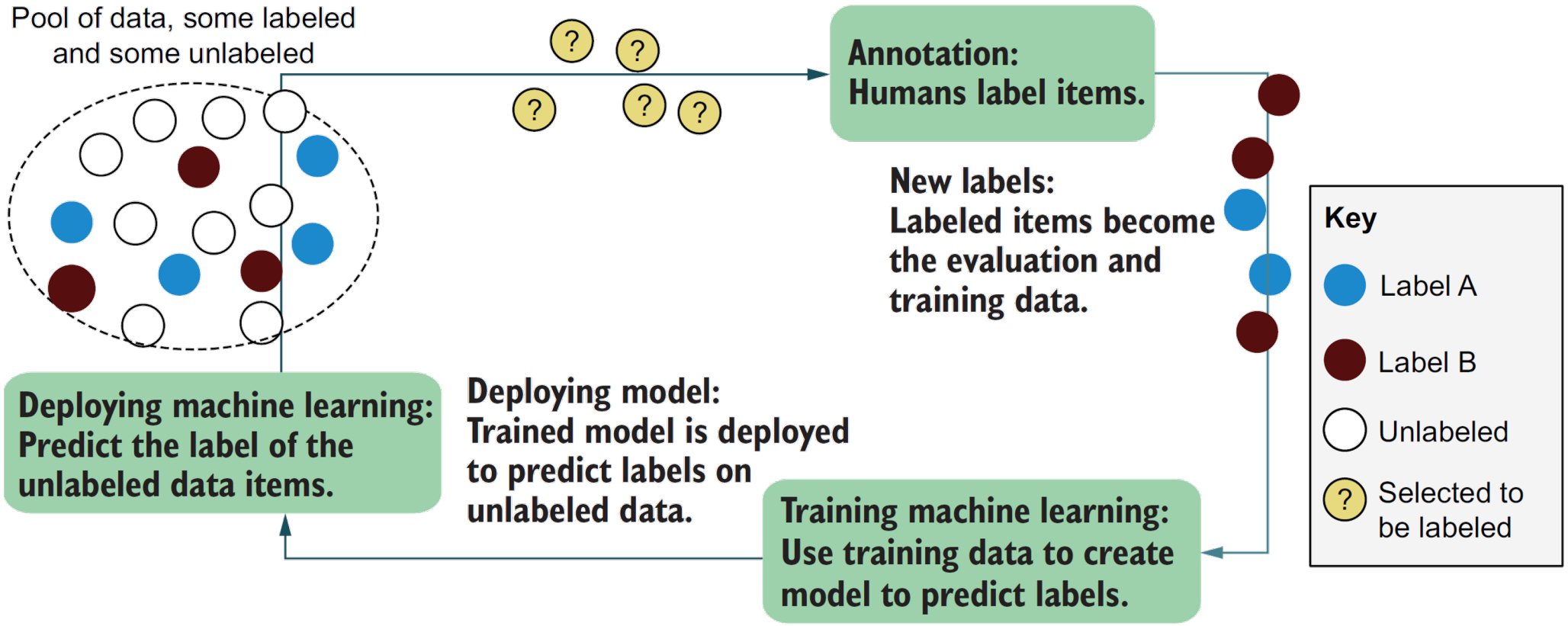
- when you can annotate only a small fraction of your data and when random sampling will not cover the diversity of data.
- this recommendation covers most real-world scenarios, as the scale of the data becomes an important factor in many use cases.

- Implementation steps - Example:
  - Ranking predictions by model confidence to identify confusing items
  - Finding unlabeled items with novel information
  - Building a simple interface to annotate training data
  - Evaluating changes in model accuracy as you add more training data

Active Learning
Example

research

- Assumption: architecture of your first human-in-the-loop machine learning system

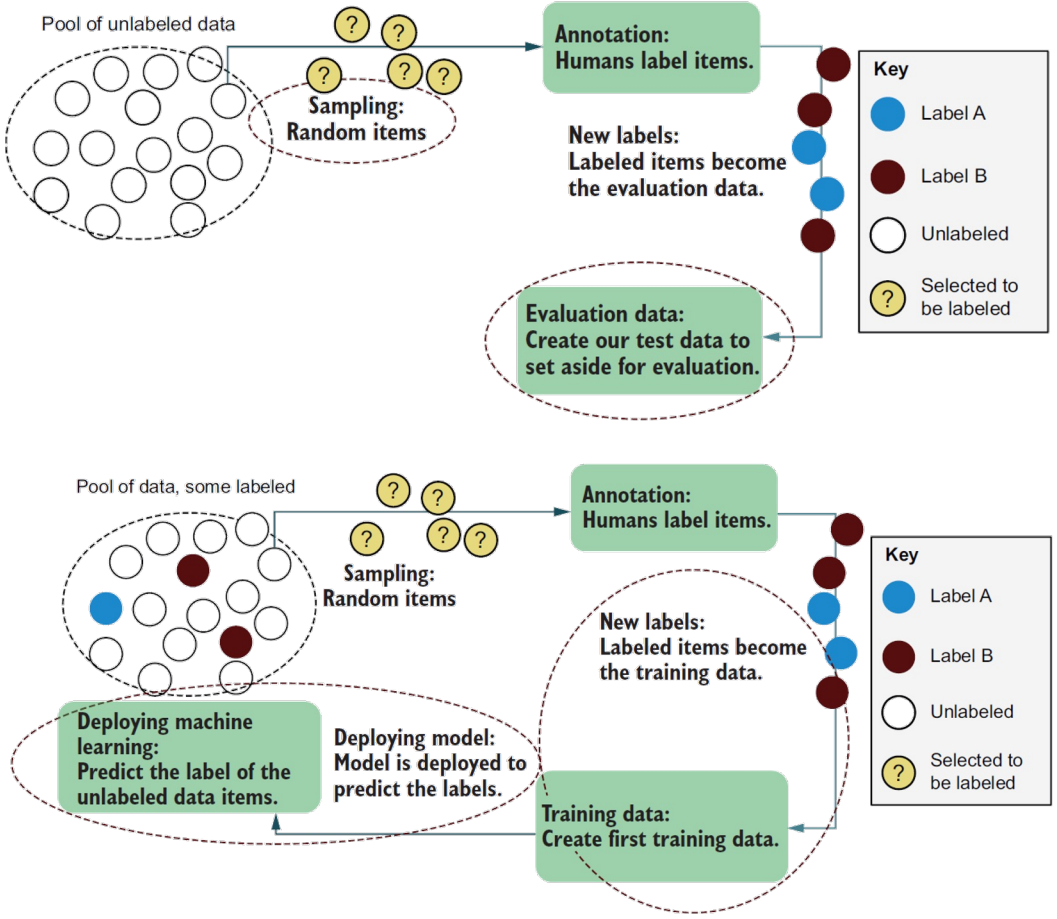


Monarch, Robert Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

Active Learning
Example

research | fh st.pölten

- Process for building your first human-in-the-loop machine learning model



Monarch, Robert Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

Active Learning

Example

research fh st.pölten

- Our system will label a set of news headlines as
  - “disaster-related” or
  - “not disaster-related.”
  - → messages from several past disasters
- this real-world task could have many application areas:
  - help identify disaster-related news articles in real time to help with the response
  - Adding a new “disaster-related” tag to news articles to improve the searchability and indexability of a database
  - Supporting a social study about how disasters are reported in the media by headline analyzing

https://github.com/rmunro/pytorch_active_learning

Active Learning

Example

research fh st.pölten

- active learning process notes:
  - First iteration
    - We are annotating mostly “not disaster-related” headlines, which can feel tedious
  - Second iteration
    - We have created your first model! Your F-score is probably terrible, maybe only 0.20
  - Third and fourth iterations
    - We should start to see model accuracy improve, as we are now labeling many more “disaster-related” headlines, bringing the proposed annotation data closer to 50:50 for each label
  - Fifth to tenth iterations
    - Our models start to reach reasonable levels of accuracy, and we should see more diversity in the headlines; F-Score goes up by a few percentage points for every ~100 annotation

Active Learning
Example

research fh st.pölten

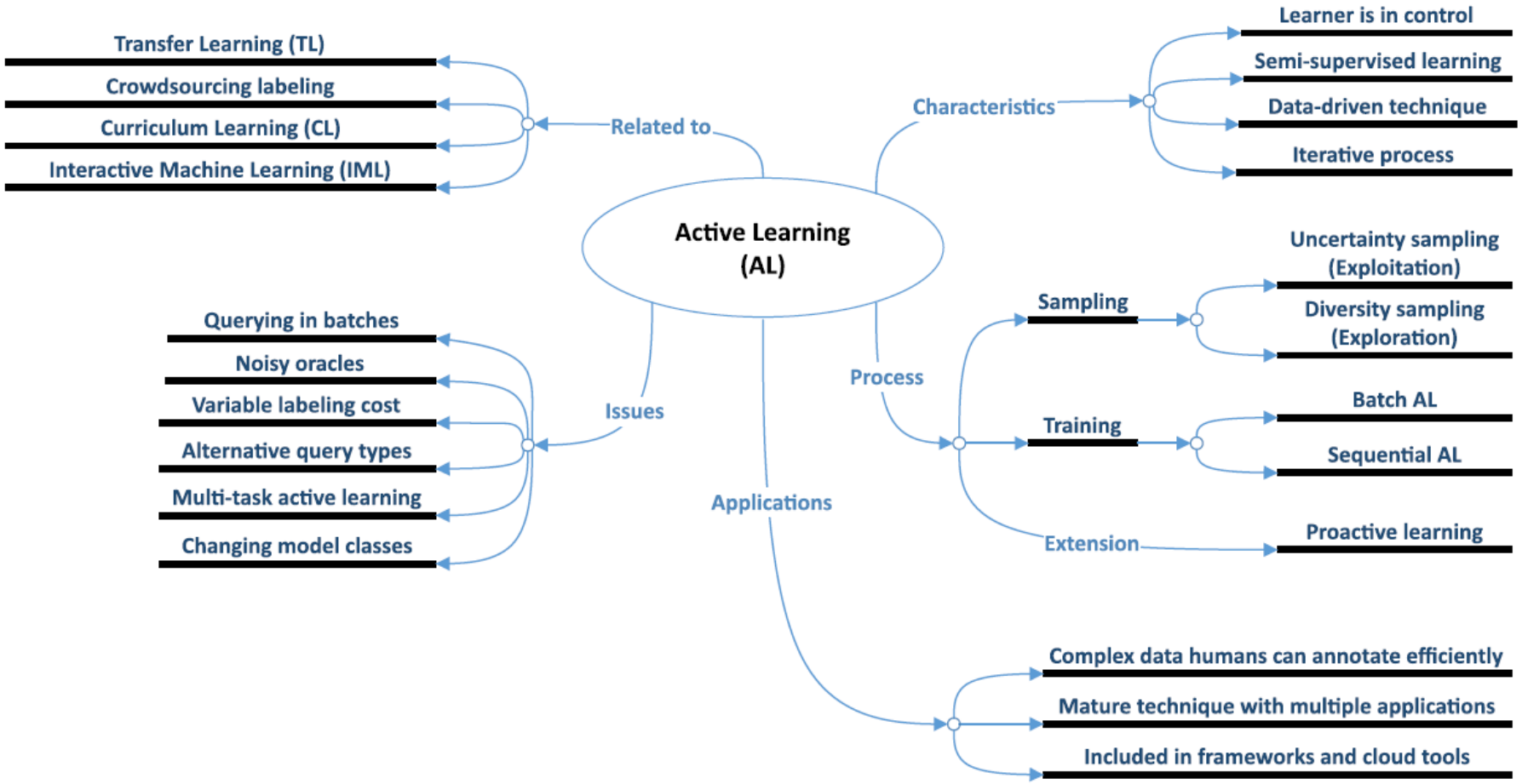
- Building an interface to get human labels
  - The right interface for human labeling is as important as the right sampling strategy.
  - e.g., a simple interface for labeling text

```
Please type 1 if this message is disaster-related, or hit Enter if not.
Type 2 to go back to the last message, type d to see detailed definitions, or type s to save your annotations.

Concerns risk of terror attacks in Europe is high

> 1
```

https://github.com/rmunro/pytorch_active_learning/blob/master/active_learning_basics.py

Active Learning

Mind Map



Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D. *et al.* Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev* 56, 3005–3054 (2023).

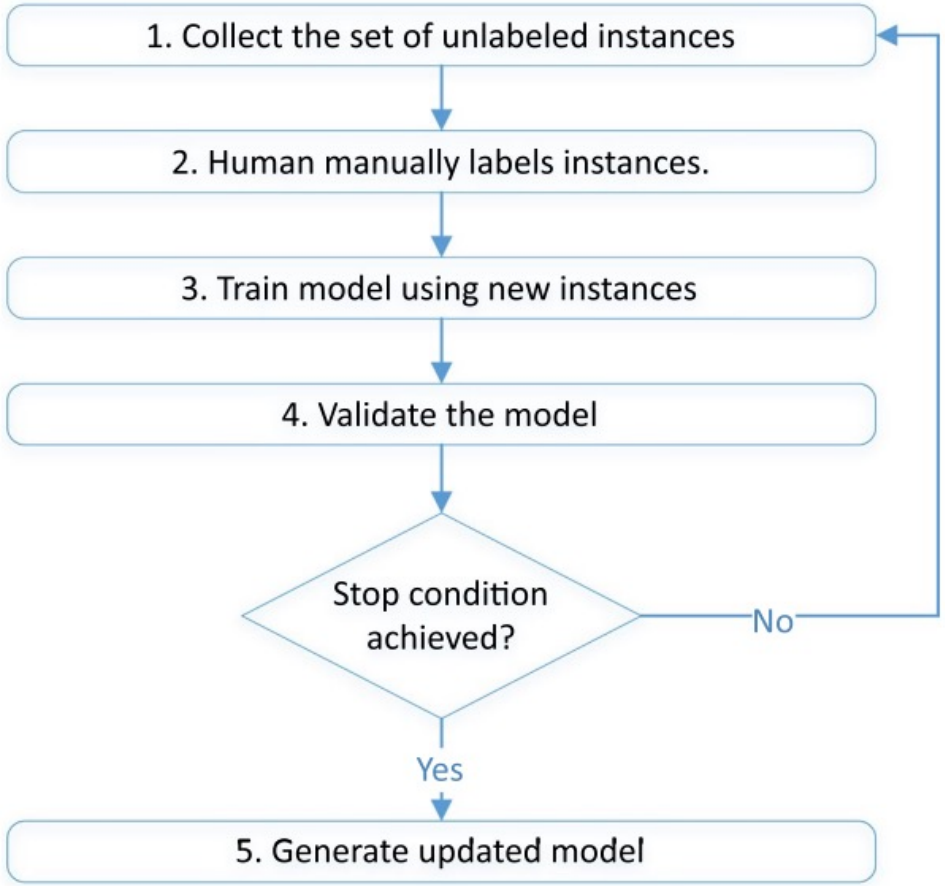
Why/When to use

research /fh/// st. pölten

- Reduced Annotation Time
- Cost Savings
- Efficient Resource Allocation
- Improved Model Performance
- Scalability
- No labeled data …

Active Learning

Steps taken in order to update the model in AL



Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D. *et al.* Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev* 56, 3005–3054 (2023).

Active Learning

Approaches

research

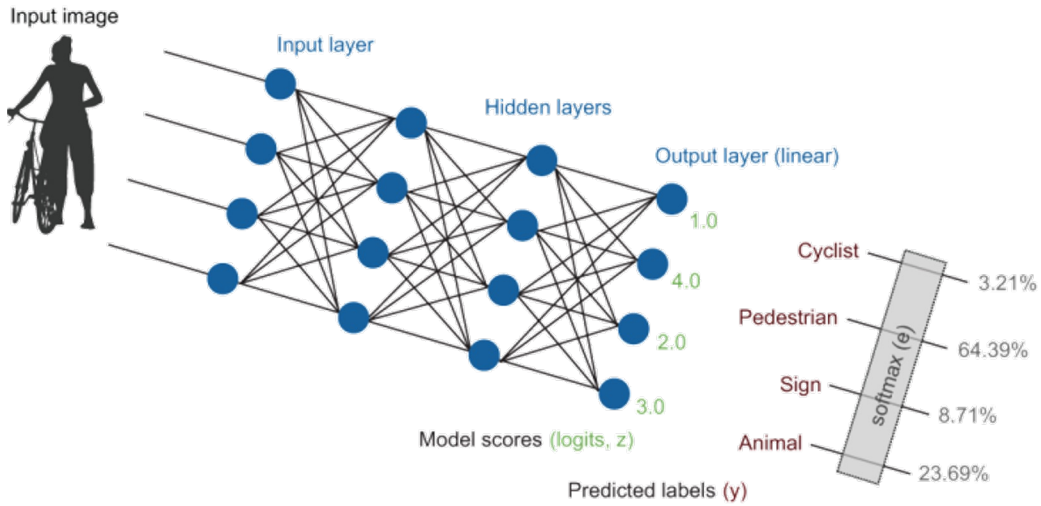
- Uncertainty sampling (Exploitation)
  - selects instances which have the least label certainty under the current trained model
    - Least confidence, which takes the example with the lowest confidence in their most likely class label
    - Margin of confidence, that uses the smallest difference between the top two highest probabilities for each possible label.
    - Ratio of confidence, which uses the ratio between the top two most confident predictions.
    - Entropy, that uses the difference between all predictions.
- Diversity sampling (Exploration):
  - selects unlabeled items that are rare or unseen in the training data to increase the picture of the problem space
    - Model-based outliers, that samples for low activation (e.g. hidden layers).
    - Cluster-based sampling, which uses unsupervised learning to cluster the data to find outliers that are not part of any trend.
    - Representative sampling, that finds items most representative of the target domain.
    - Real-world diversity, which increases fairness with data supporting real-world diversity.

Active Learning

Uncertainty Sampling - Example

research fh st.pölten

- Uncertainty Sampling
- For example, imagine that you are building a self-driving car
  - You want to help the car understand the new types of objects (pedestrians, cyclists, street signs, animals, and so on) that it is encountering as it drives along
  - To do that, however, you need to understand when your car is uncertain about what object it is seeing and how to best interpret and address that uncertainty
  - Car spends most of its time on highways
    - limited number of objects (cyclists or pedestrians)



Active Learning

Uncertainty Sampling - approaches

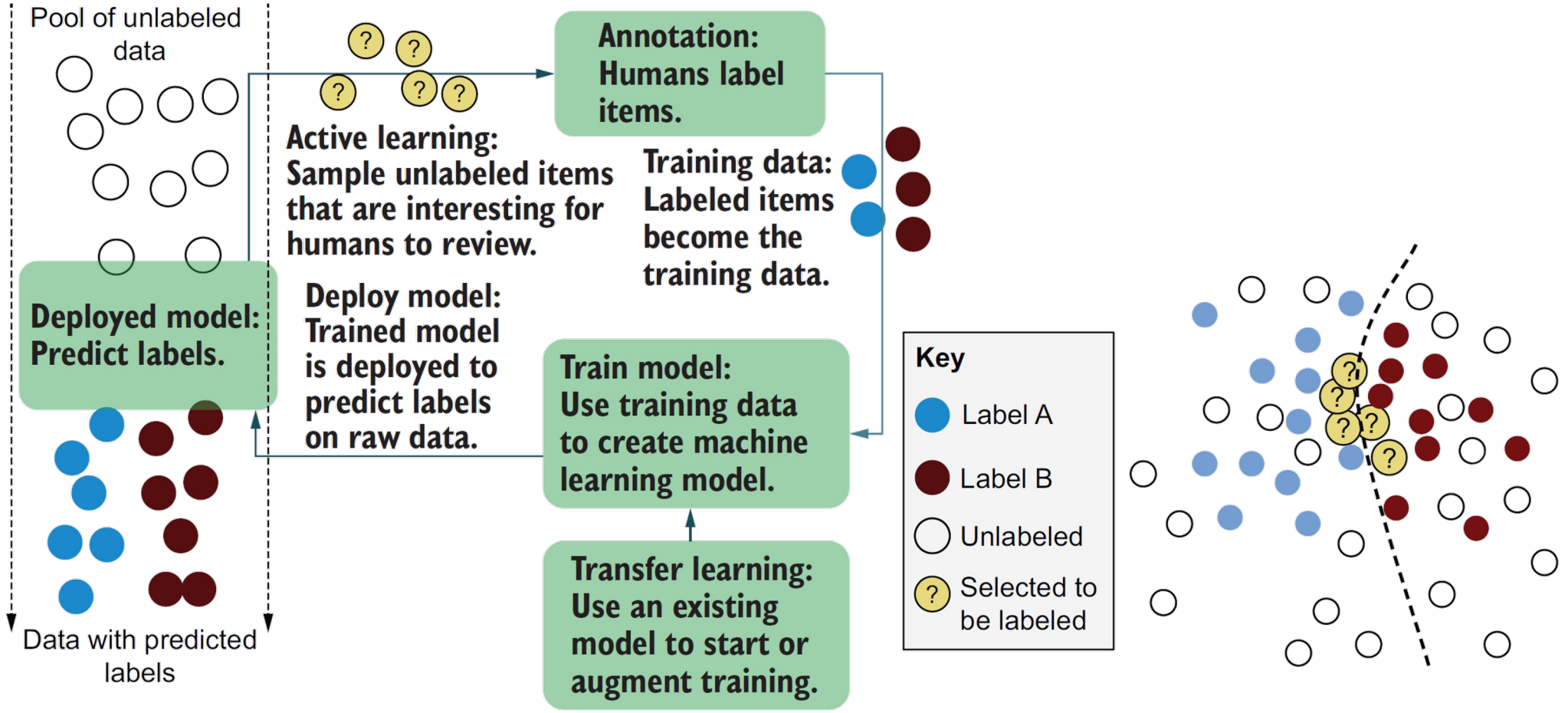
research

- selects instances which have the least label certainty under the current trained model
- four approaches to uncertainty sampling:
  - Least confidence, which takes the example with the lowest confidence in their most likely class label; Difference between the most confident prediction and 100% confidence
  - Margin of confidence, that uses the smallest difference between the top two highest probabilities for each possible label; Difference between the two most confident predictions
  - Ratio of confidence, which uses the ratio between the top two most confident predictions
  - Entropy, that uses the difference between all predictions.

Active Learning

Uncertainty Sampling

research



Monarch, Robert Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

Active Learning

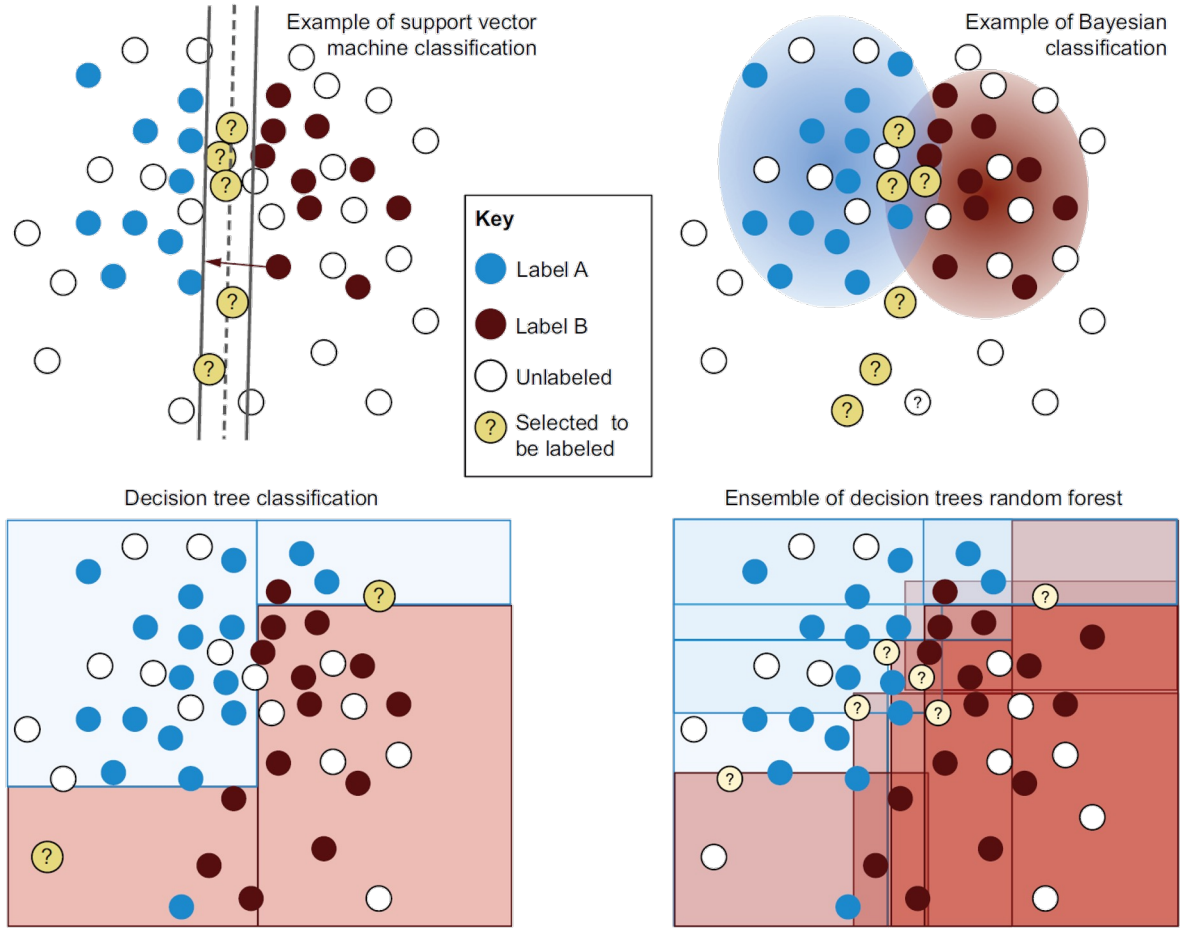
Uncertainty Sampling

research fh st.pölten

- many algorithms for calculating uncertainty
- all follow the same principles
  1. Apply the uncertainty sampling algorithm to a large pool of predictions to generate a single uncertainty score for each item.
  2. Rank the predictions by the uncertainty score.
  3. Select the top N most uncertain items for human review.
  4. Obtain human labels for the top N items, retrain the model with those items, and iterate on the processes.

Active Learning

Uncertainty Sampling from different supervised machine learning algorithms



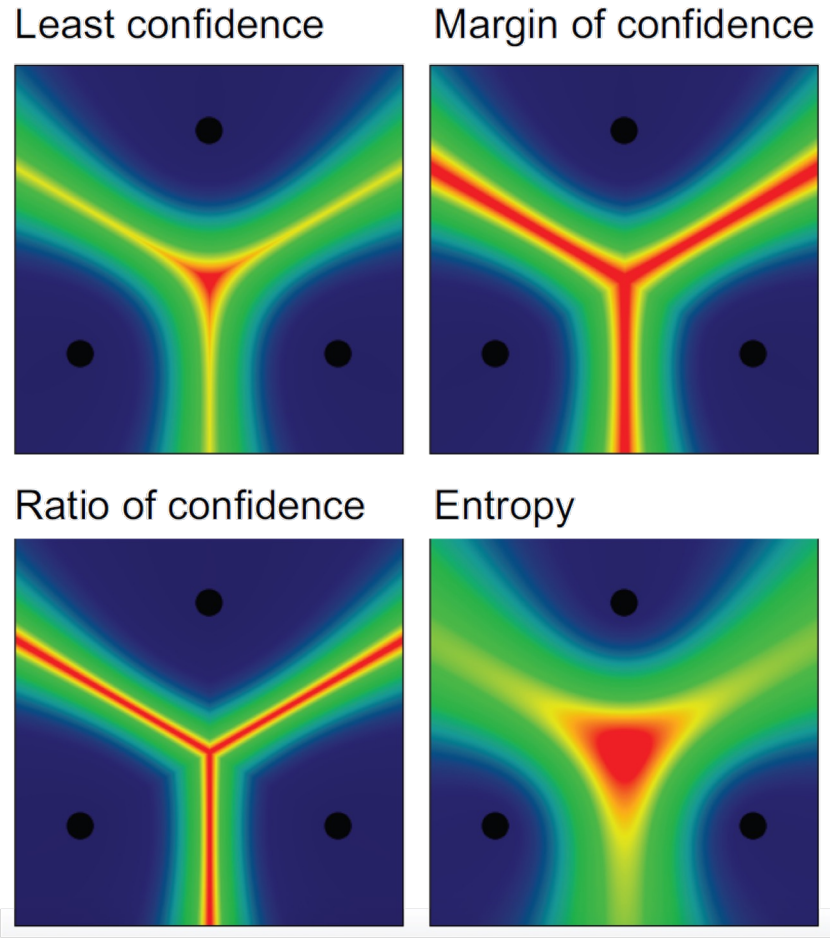
Monarch, Robert Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

Active Learning

Uncertainty Sampling – Algorithms Comparison

research fh st.pölten

- heat map of the four main uncertainty sampling algorithms
- example of target areas for the different algorithms when there are three labels
- each dot is an item with a different label, and the heat of each pixel is the uncertainty
- margin of confidence and ratio sample some items that:
  - have only pairwise confusion
  - which reflects the fact that the algorithms target only the two most likely labels
  - entropy maximizes for confusion among all labels, which is why the highest concentration is between all three labels



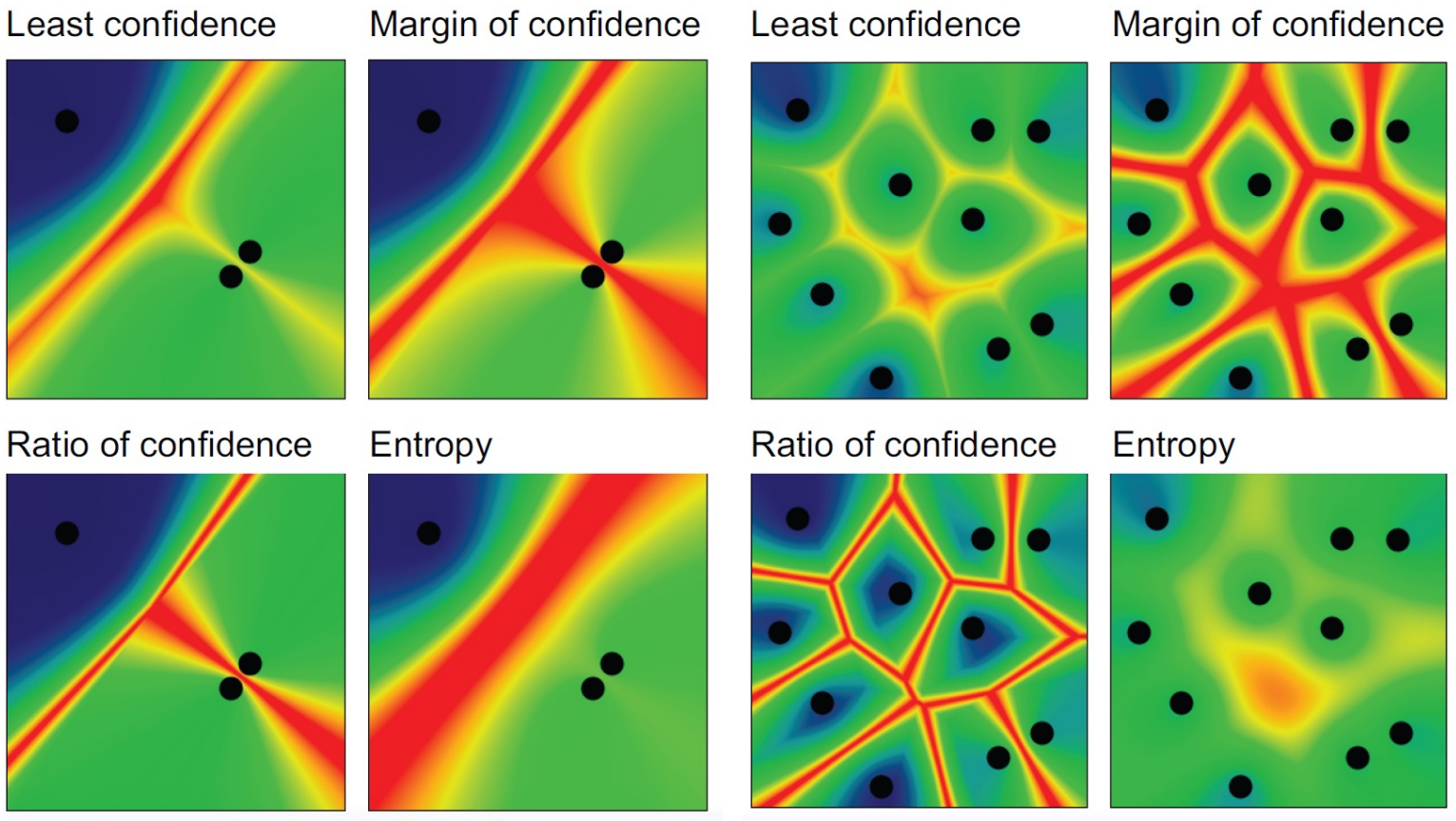
https://robertmunro.com/uncertainty_sampling_example.html

Active Learning

Uncertainty Sampling – Algorithms Comparison

research

- difference between the methods becomes even more extreme with more labels

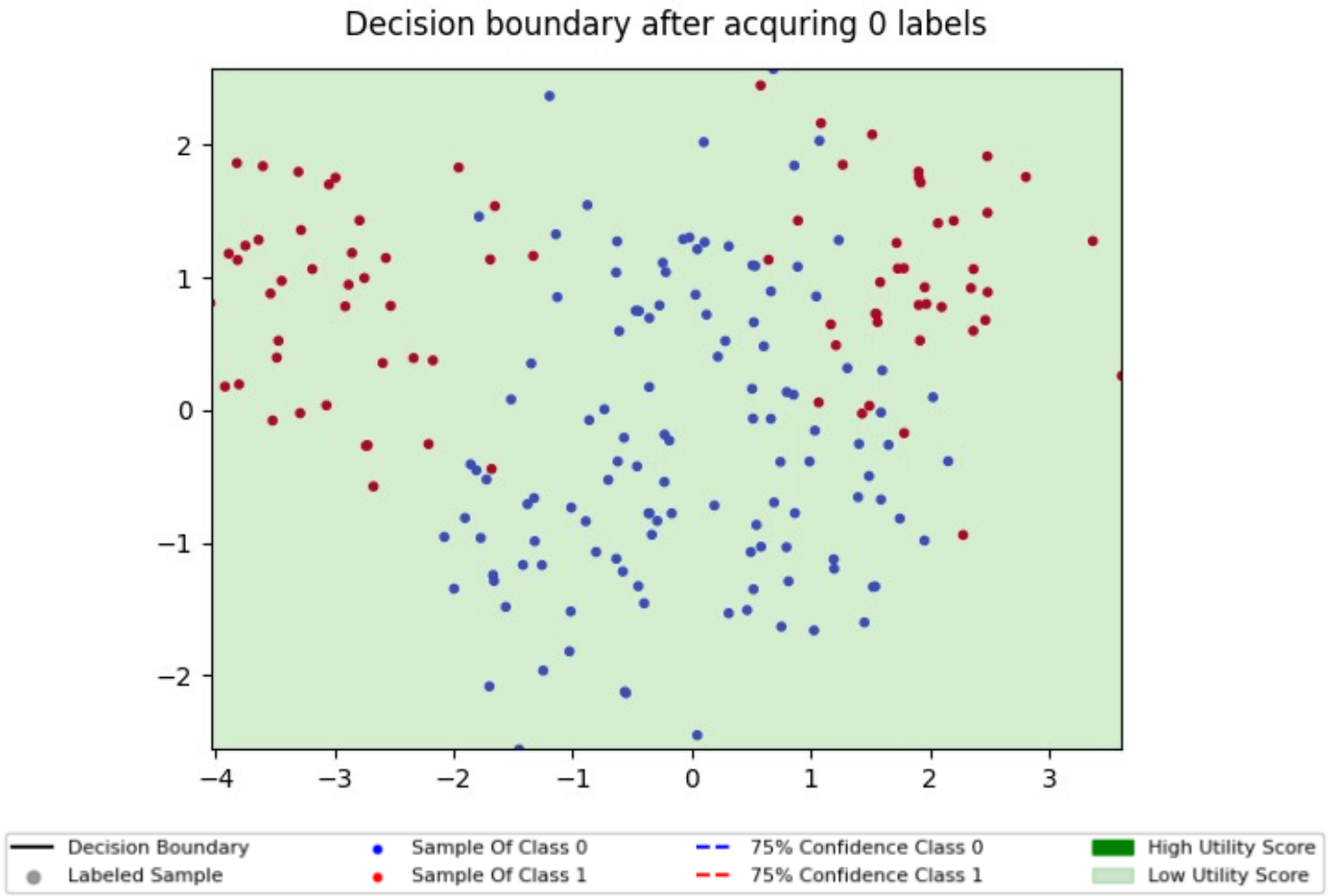


https://robertmunro.com/uncertainty_sampling_example.html

Active Learning

Uncertainty Sampling with Margin - Example

research fh st.pölten

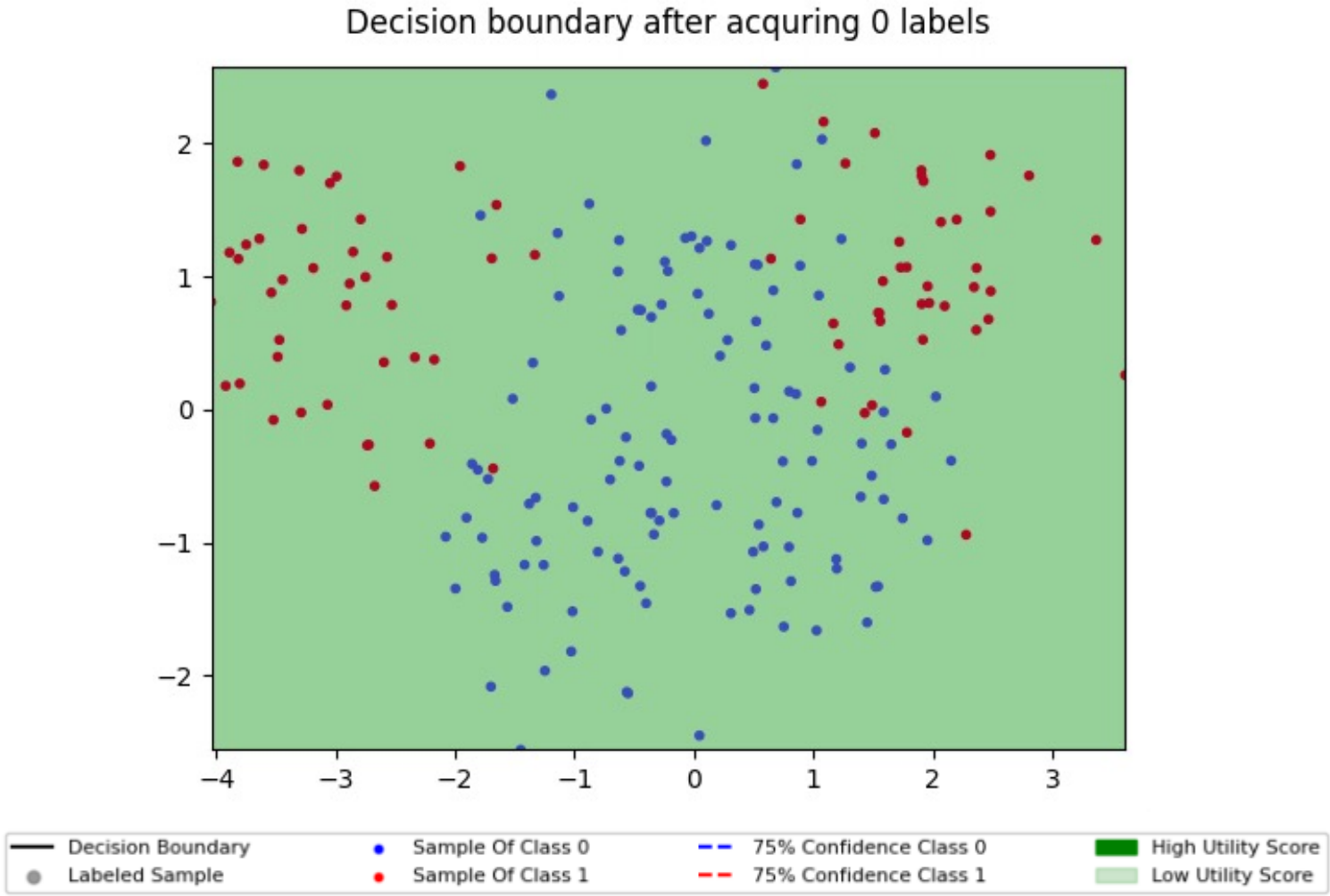


https://scikit-activeml.github.io/scikit-activeml-docs/index.html

Active Learning

Uncertainty Sampling with Least Confidence - Example

research fh st.pölten

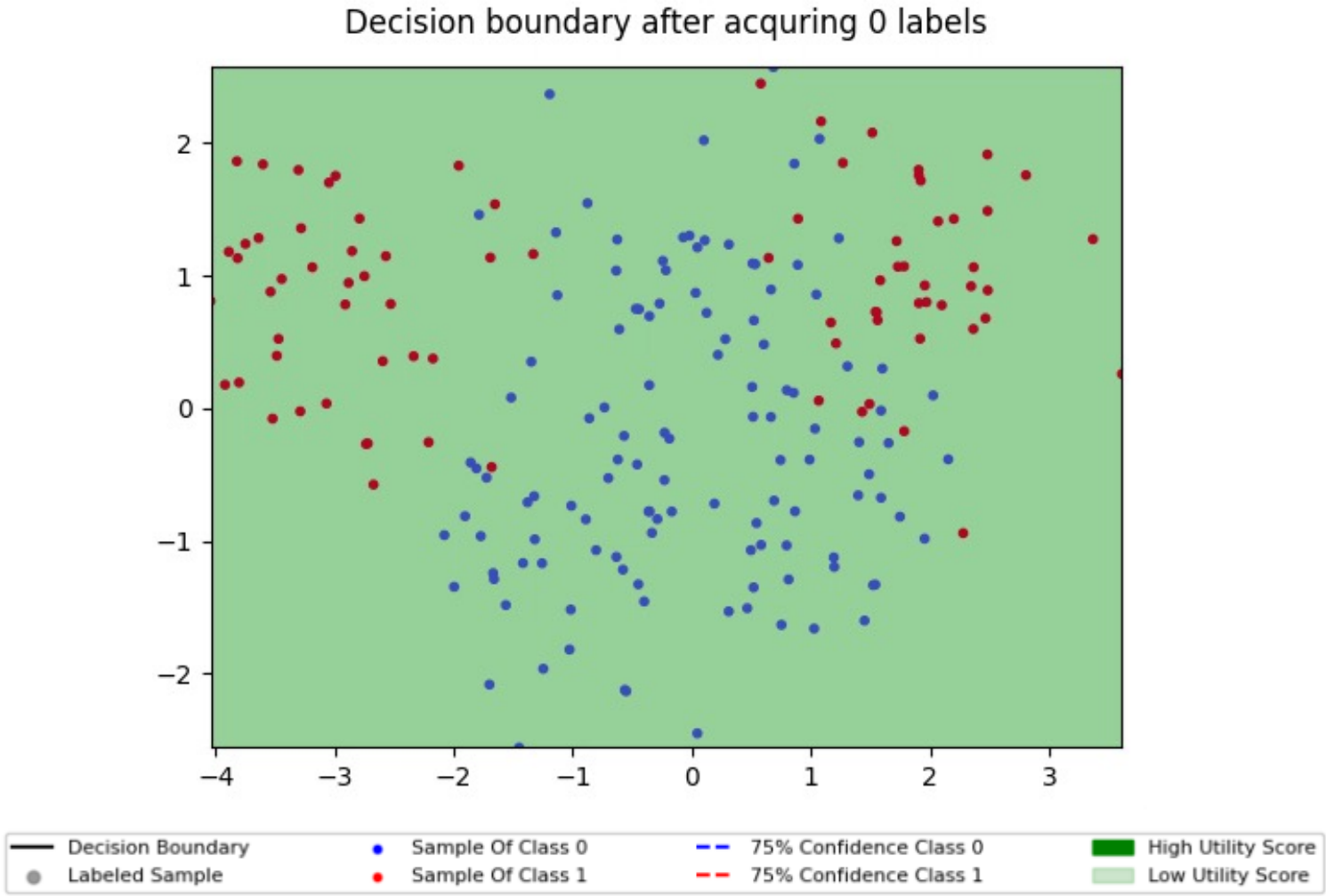


https://scikit-activeml.github.io/scikit-activeml-docs/index.html

Active Learning

Uncertainty Sampling with Entropy - Example

research



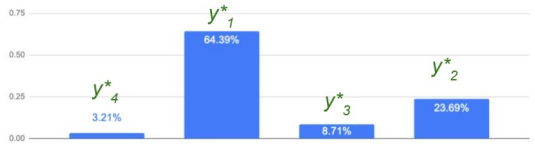
https://scikit-activeml.github.io/scikit-activeml-docs/index.html

Active Learning

Uncertainty Sampling – Cheat Sheet

research

Example Machine Learning Prediction (x):



number of labels (n) = 4

The predictions are a probability distribution (x), meaning that every prediction is between 0 and 1 and the predictions add to 1. y^*_1 is the most confident, y^*_2 is the second most confident, etc. for n predicted labels.

This example can be expressed as a NumPy array:

```
prob_dist = np.array([0.0321, 0.6439, 0.0871,
0.2369])
```

Least Confidence: *difference between the most confident prediction and 100% confidence*

$$\frac{n\,(1 - P_\theta(y^*_1 \mid x))}{n - 1}$$

```
most_conf = np.nanmax(prob_dist)
num_labels = prob_dist.size
numerator = (num_labels * (1 - most_conf))
denominator = (num_labels - 1)

least_conf = numerator / denominator
```

Margin of Confidence: difference between the top two most confident predictions

$$1 - (P_\theta(y^*_1 \mid x) - P_\theta(y^*_2 \mid x))$$

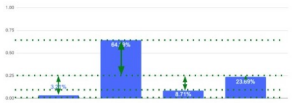
```
prob_dist[::-1].sort()
difference = (prob_dist[0] - prob_dist[1])

margin_conf = 1 - difference
```

Ratio of Confidence: ratio between the top two most confident predictions

$$\frac{P_\theta(y^*_2 \mid x)}{P_\theta(y^*_1 \mid x)}$$

```
prob_dist[::-1].sort()

ratio_conf  = (prob_dist[1] / prob_dist[0])
```

Entropy: difference between all predictions, as defined by information theory



$$\frac{-\sum_y P_\theta(y \mid x)\,\log_2 P_\theta(y \mid x))}{\log_2(n)}$$

```
prbslogs = prob_dist * np.log2(prob_dist)
numerator = 0 - np.sum(prbslogs)
denominator = np.log2(prob_dist.size)

entropy = numerator / denominator
```

https://robertmunro.com/uncertainty_sampling_example.html

Active Learning

Uncertainty Sampling - References

research fh st.pölten

- ensemble-based uncertainty sampling
  - https://aclanthology.org/D18-1318/
- entropy-based sampling
  - https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.30.6148
- margin of confidence sampling
  - https://link.springer.com/chapter/10.1007/3-540-44816-0_31
- least confidence sampling
  - https://people.cs.umass.edu/~mccallum/papers/multichoice-aaai05.pdf

Active Learning

Diversity Sampling - Example

research

- Diversity Sampling
- learn how to identify what's missing from your model
- what your model "doesn't know that it doesn't **know**" or the "**unknown unknowns**."

- For example, imagine that you build a voice assistant
  - with much broader knowledge than any one human

Active Learning

Diversity Sampling

research

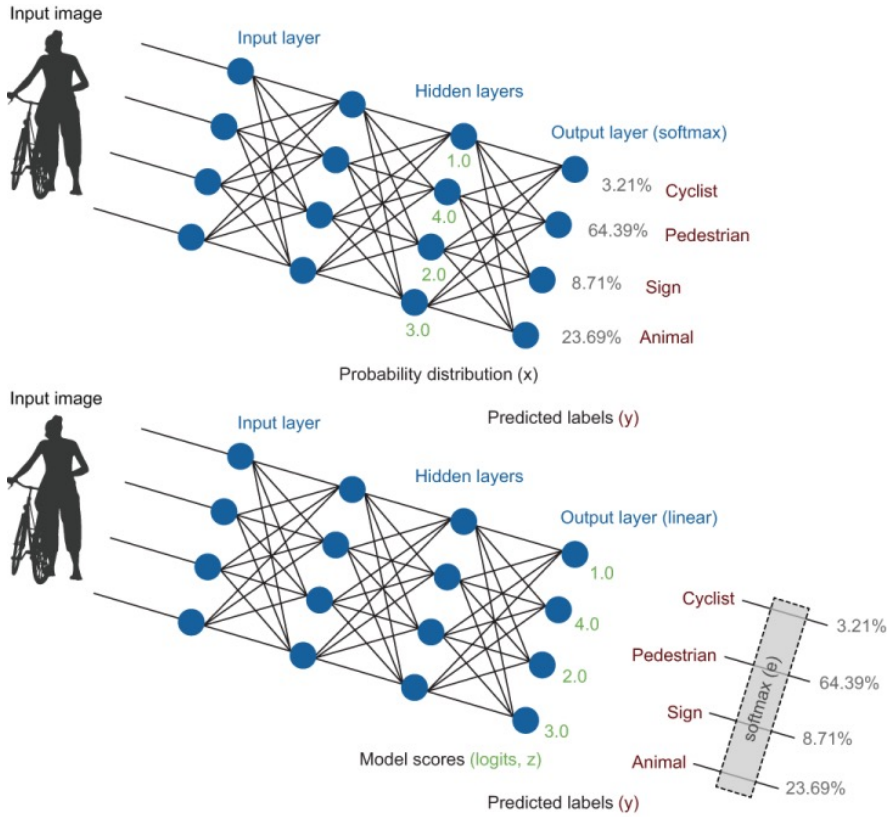
- selects items to be labeled that are maximally different from the existing training items and from one another
- four approaches to diversity sampling:
  - **Model-based outlier sampling** - determining which items are unknown to the model in its current state
  - **Cluster-based sampling** - using statistical methods independent of your model to find a diverse selection of items to label
  - **Representative sampling** - finding a sample of unlabeled items that look most like your target domain, compared with your training data
  - **Sampling for real-world diversity** - ensuring that a diverse range of real-world entities are in our training data to reduce real-world bias.

Active Learning

Diversity Sampling

research fh st.pölten

- Model-based outlier sampling
  - Model outlier in a neural model is defined as the item with the lowest activation in a given layer.
    - for our final layer, this activation is the logits
    - e.g., use validation data to rank activations
  - Limitations:
    - can generate outliers that are similar and therefore lack diversity within an active learning iteration
    - it's hard to escape some statistical biases that are inherent in the model → we may continually miss some types of outliers

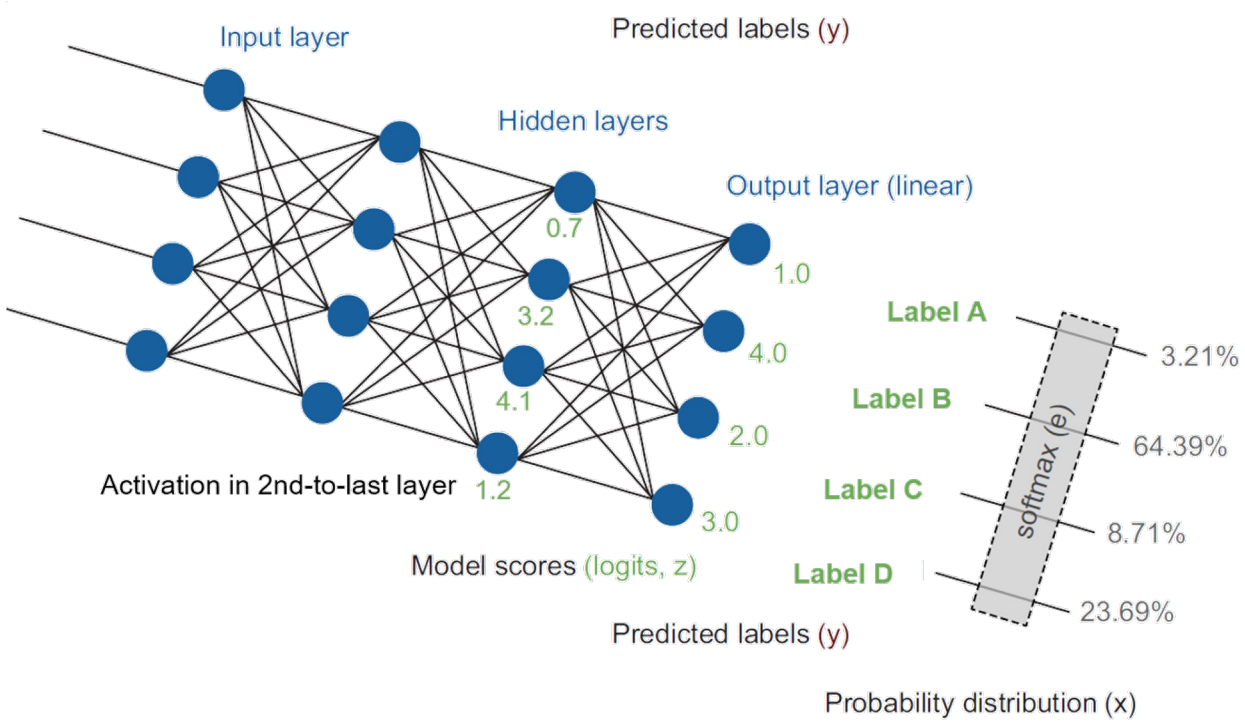


Active Learning

Diversity Sampling

research fh st.pölten

- Model-based outlier sampling
  - sampling for low activation in logits and hidden layers
- **Why?**
  - To find items that are confusing to your model because of lack of information.
  - This is different from uncertainty through conflicting information, a complementary sampling method
- **Tips:**
  - experiment with average vs max activation



https://robertmunro.com/Diversity_Sampling_Cheatsheet.pdf

Active Learning
Diversity Sampling

research

- Cluster-based sampling
  - most common method used for diversity sampling
  - can help target a diverse selection of data from the start
  - instead of sampling training data randomly to begin with, we also divide our data into a large number of clusters and sample evenly from each cluster.
  - sample from clusters in three ways:
    - Random: sampling items at random from each cluster
    - Centroids: sampling the centroids of clusters to represent the core of significant trends within our data
    - Outliers: sampling the outliers from our clustering algorithm to find potentially interesting data that might have been missed in the clusters.

Active Learning

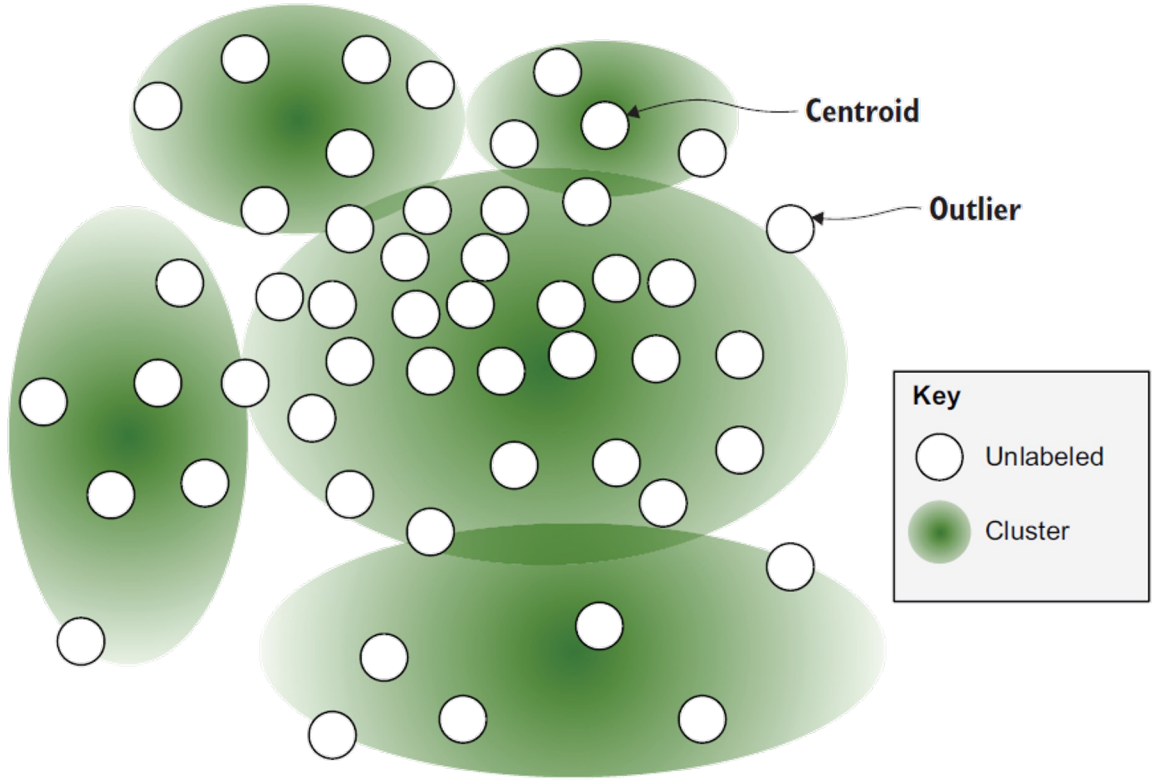
Diversity Sampling

research fh st.pölten

- Cluster-based sampling
  - using unsupervised learning to pre-segment the data

- **Why?**
  - To ensure that you are sampling data from all the meaningful trends in your data's feature-space, not just the trends that contain the most items.
  - Also to find outliers that are not part of any trend.

- **Tips:**
  - try different distance metrics and clustering algorithms

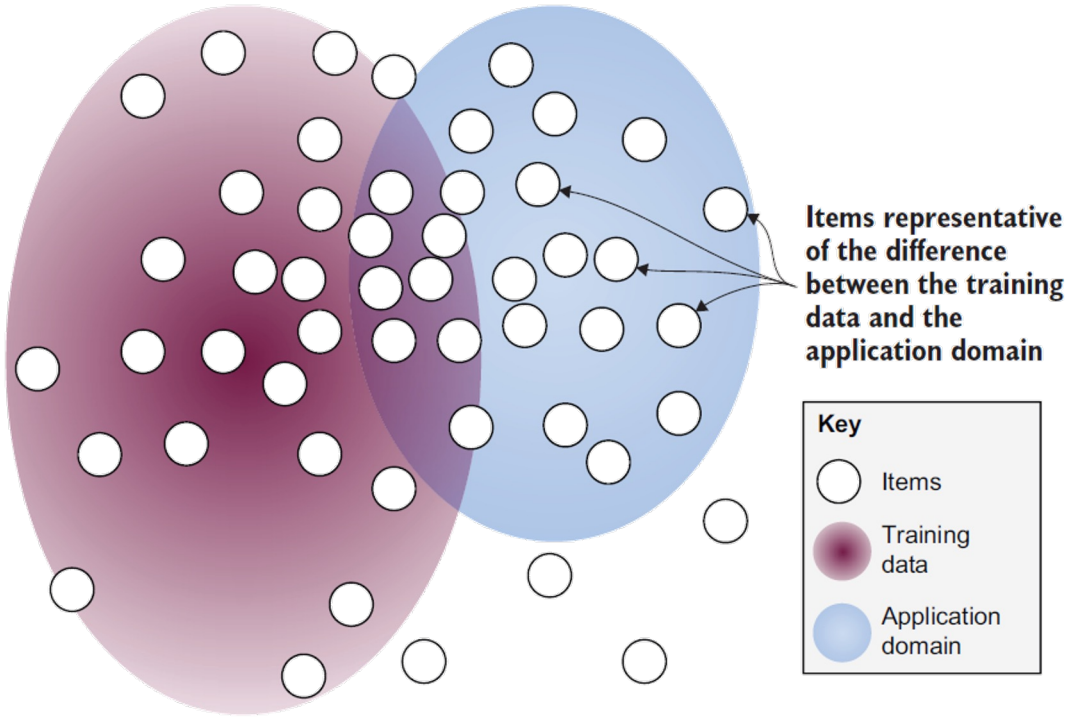


https://robertmunro.com/Diversity_Sampling_Cheatsheet.pdf

Active Learning
Diversity Sampling

research fh st.pölten

- Representative sampling
  - refers to explicitly calculating the difference between the training data and the application domain where we are deploying the model
    - what unlabeled data looks most like the domain where we are deploying our model?
    - learning what data looks most like where we are adapting it will give you good intuition about that dataset as a whole and the problems we might face



Monarch, Robert Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

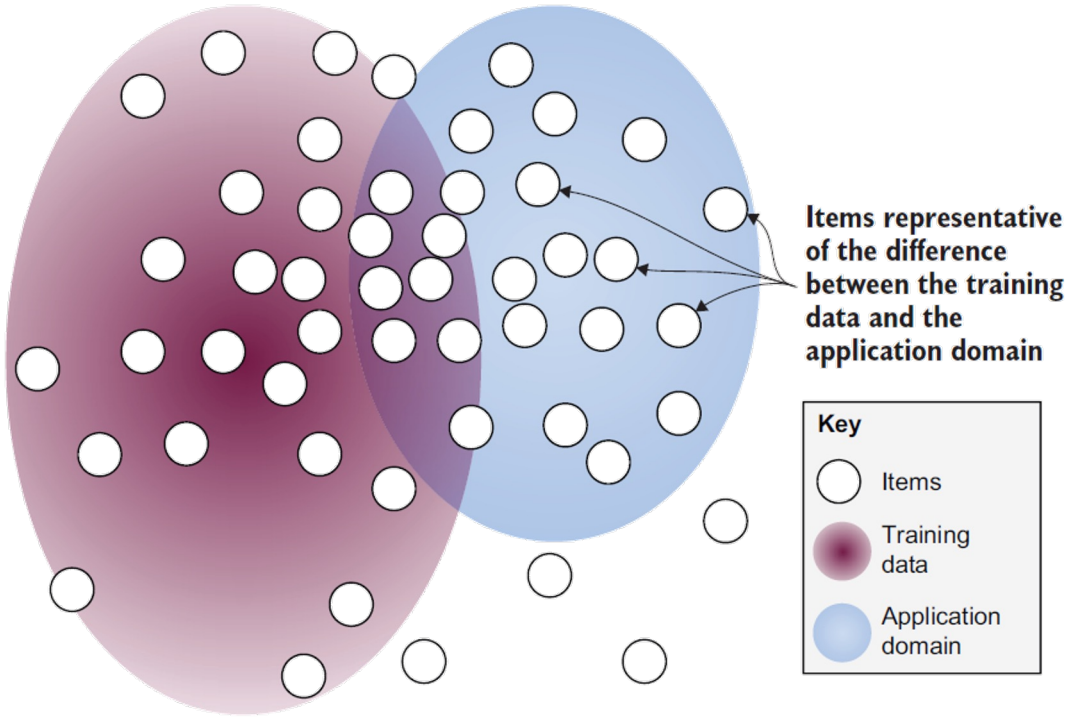
Active Learning
Diversity Sampling

research fh st.pölten

- Representative sampling
  - finding items most representative of the target domain

- **Why?**
  - When your target domain is different from your current training data, you want to sample items most representative of your target domain in order to adapt to the domain as fast as possible.

- **Tips:**
  - extendable to be adaptive within one Active Learning cycle

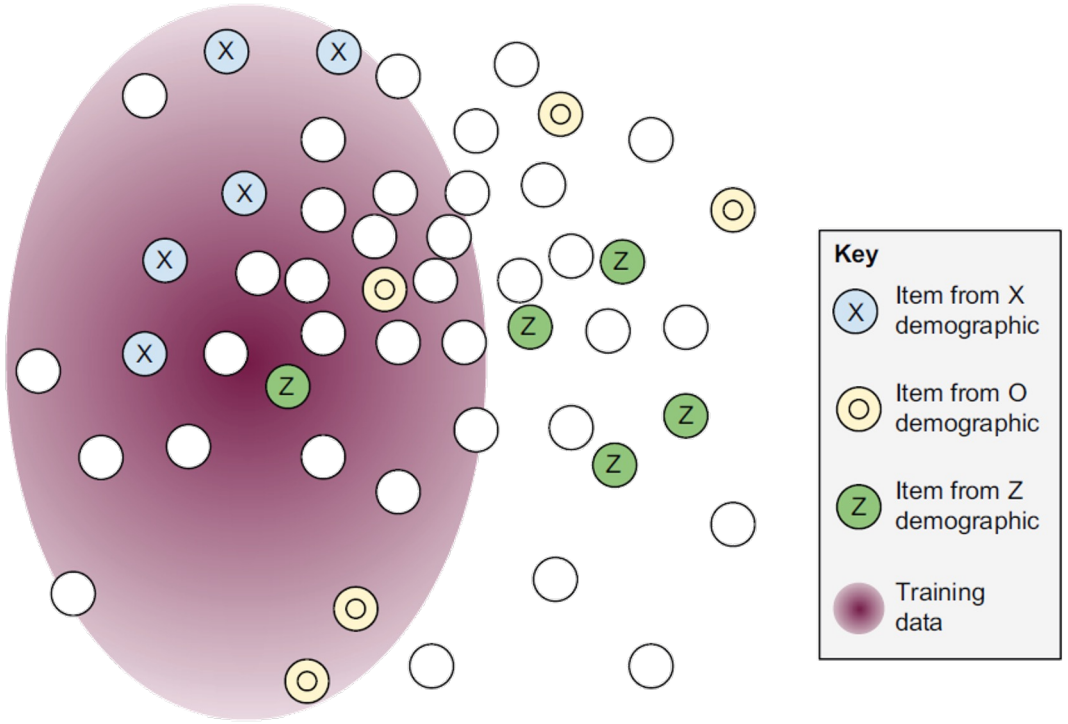


https://robertmunro.com/Diversity_Sampling_Cheatsheet.pdf

Active Learning
Diversity Sampling

research fh st.pölten

- Sampling for real-world diversity
  - ensuring that training data represents real-world diversity as fairly as possible
  - Common problems for fairness in data (e.g., for our disaster-response examples)
    - A demographic that is overrepresented in your training data but not from the same distribution as your training data (X)
    - A demographic that is from a distribution similar to the overall data distribution but not yet represented in a balanced way in the training data (O)
    - A demographic that is underrepresented in the training data in such a way that the resulting model might be worse than if random sampling were used (Z)



Monarch, Robert Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

Active Learning
Diversity Sampling

research fh st.pölten

- Sampling for real-world diversity
  - Use all Active Learning strategies to make your data as fair as possible
    1. Apply least confidence sampling for every demographic, selecting an equal number of items in each demographic where that demographic was the most-confident prediction.
    2. Apply margin of confidence sampling for every demographic, selecting an equal number of items in each demographic where that demographic was the most-confident or second-most-confident.
    3. Apply model-based outlier detection for each demographic.
    4. Apply cluster-based sampling within each demographic.

Active Learning

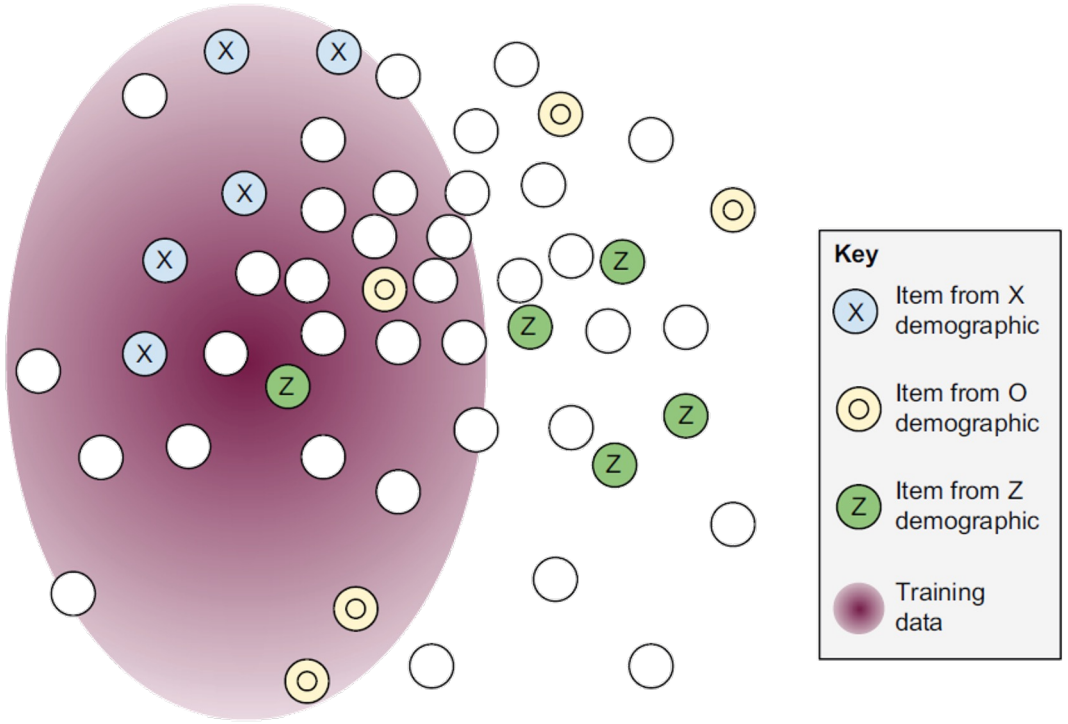
Diversity Sampling

research fh st.pölten

- Sampling for real-world diversity
  - increase fairness with data supporting real-world diversity

- **Why?**
  - So as many people can as possible take advantage of your models and you are not amplifying real-world biases.
  - Use all Active Learning strategies to make your data as fair as possible.

- **Tips:**
  - your model might not require representative data to be fair



https://robertmunro.com/Diversity_Sampling_Cheatsheet.pdf

Active Learning

Diversity Sampling - References

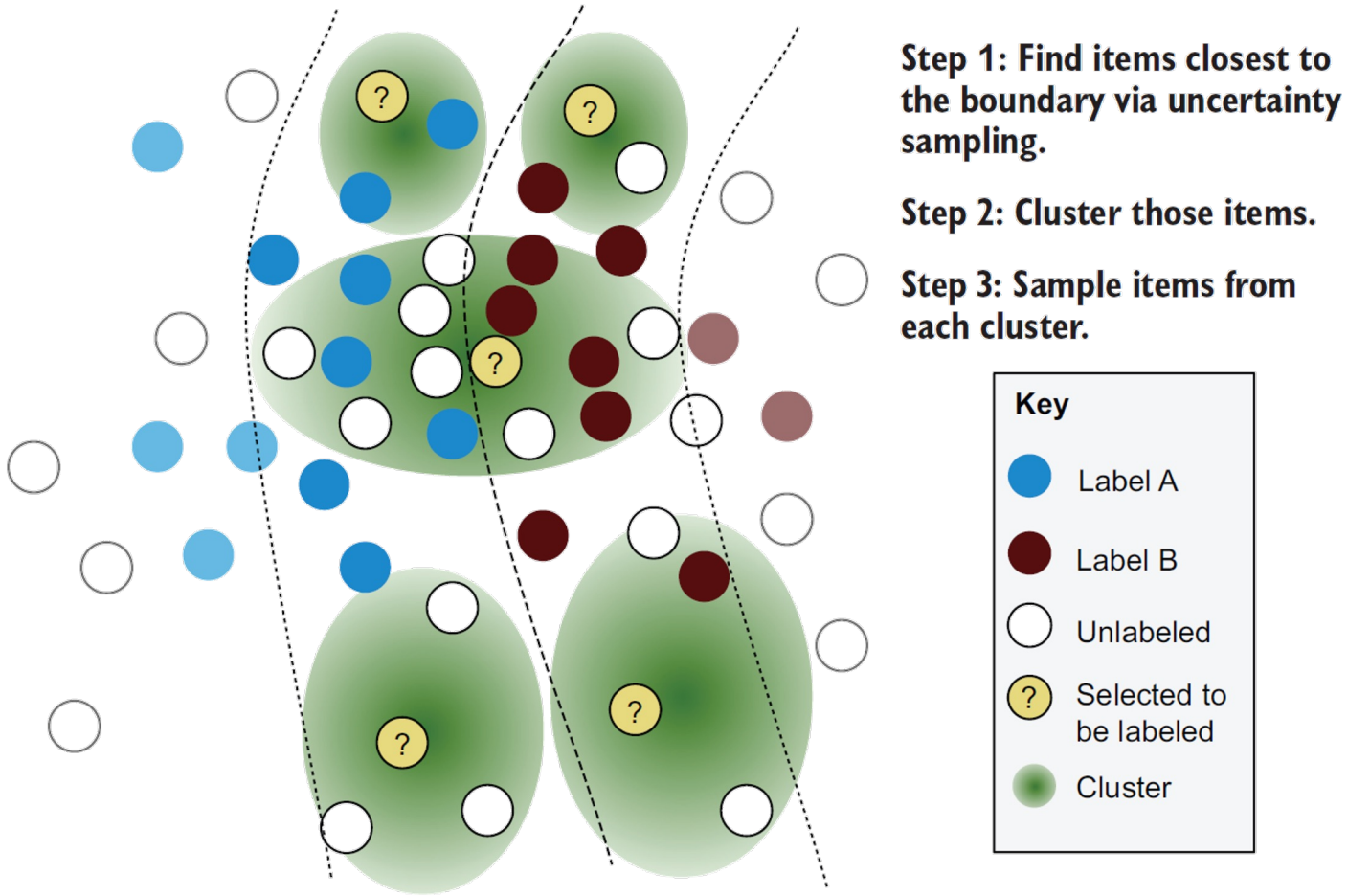
research fh st.pölten

- Cluster-based sampling
  - https://ivi.fnwi.uva.nl/isis/publications/2004/NguyenICML2004/NguyenICML2004.pdf
  - https://link.springer.com/chapter/10.1007/978-3-642-51883-6_24
- Representative sampling
  - http://www.kamalnigam.com/papers/emactive-icml98.pdf
- Sampling for real-world diversity
  - https://arxiv.org/pdf/1906.02659.pdf
  - https://aclanthology.org/P17-2009.pdf
  - https://arxiv.org/pdf/2005.14050.pdf

Active Learning

Combining least confidence and clustering-based sampling

research



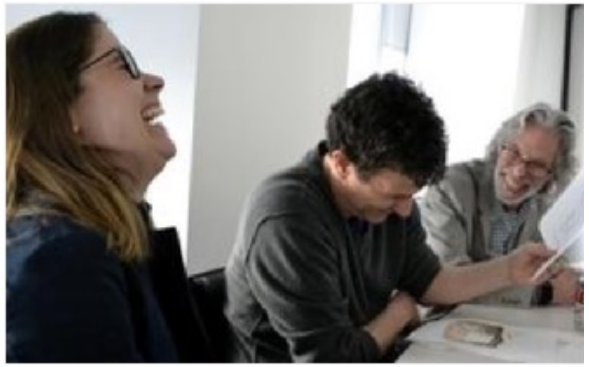
Active Learning

Real-Life Examples

research fh st.pölten

- Active learning to optimize crowdsourcing and rating in New Yorker Cartoon Caption Contest

digg



BY DOING THE EXACT OPPOSITE

How New Yorker Cartoons Could Teach Computers To Be Funny

3 diggs CNET Technology

With the help of computer scientists from the University of Wisconsin at Madison, The New Yorker for the first time is using crowdsourcing algorithms to uncover the best captions.

Active Learning

Real-Life Examples

research

- Actively learning user's beer preferences

Active Learning

Learning by Doing

research

- https://github.com/rmunro/pytorch_active_learning

- Diversity Sampling
- https://github.com/rmunro/pytorch_active_learning/blob/master/diversity_sampling.py

- Uncertainty Sampling
- https://github.com/rmunro/pytorch_active_learning/blob/master/uncertainty_sampling.py

BREAK

Hands on

research

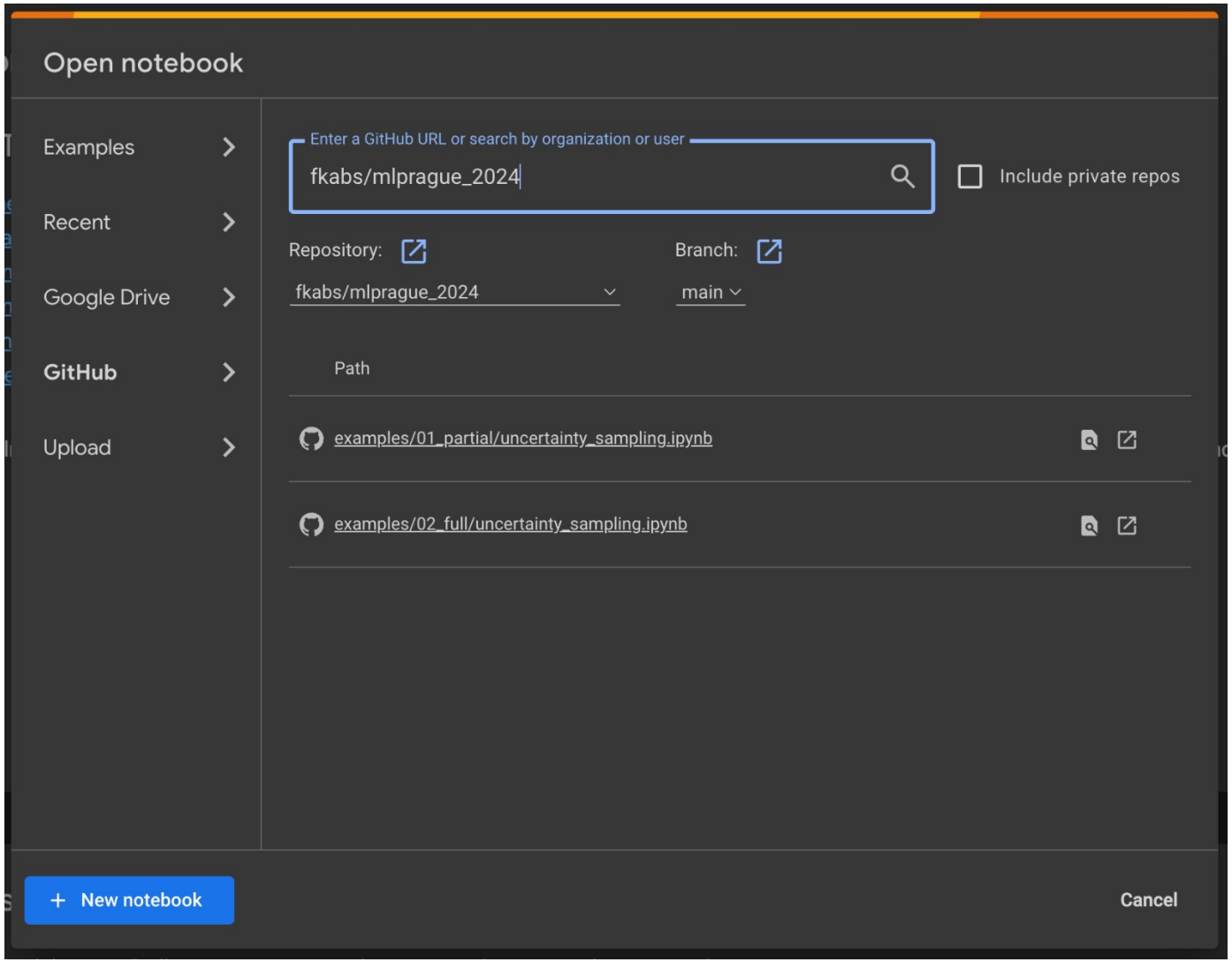
- https://github.com/fkabs/mlprague_2024

- We suggest, you open a new Google Colab Notebook and open a new notebook from GitHub:
- Simply search for

„ fkabs/mlprague_2024 “

Hands on

research



Road ahead and other ideas

research

- Efficient Supervised Learning
  - One-Shot (one labeled example)
  - Zero-Shot (absolutely no labeled data)
  - Few-Shot Learning (limited number of labeled examples)

- Semi-Supervised Learning approaches
  - Self-training
  - Co-training

Semi-Supervised Learning

Road ahead and other ideas

Semi-Supervised Learning

research

fh st.pölten

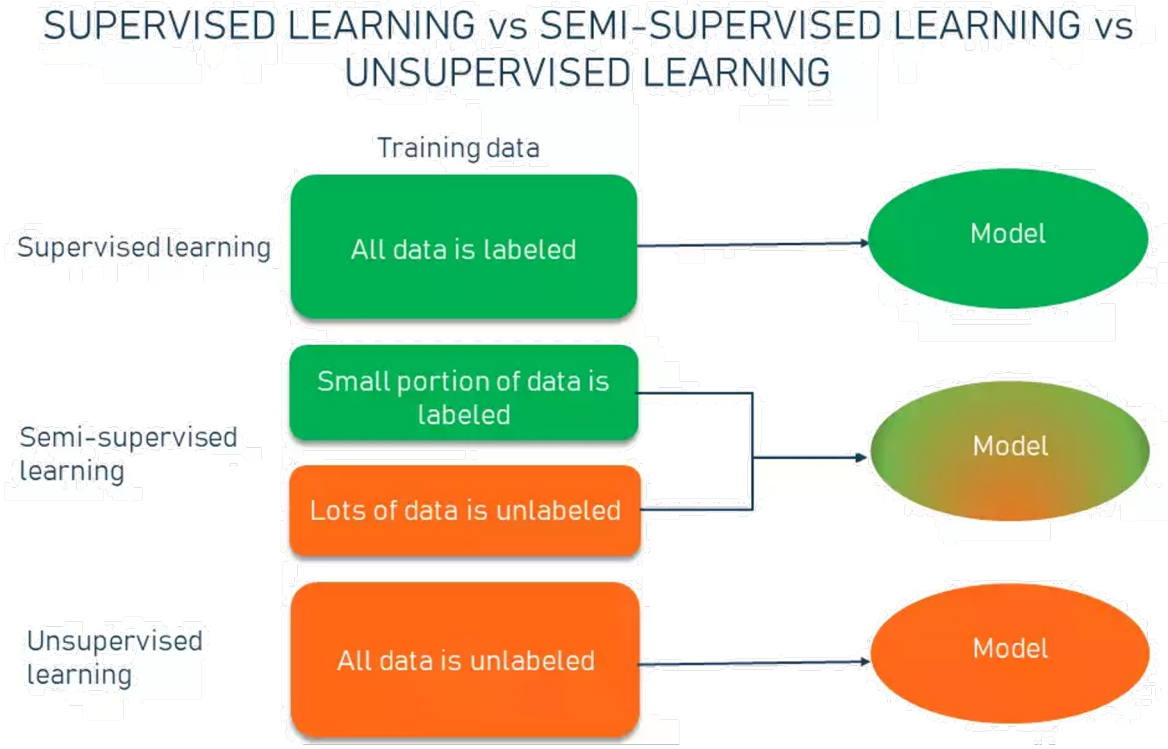
- middle ground between supervised learning and unsupervised learning
- model learns from both labeled and unlabeled data to improve its performance
- leverages the availability of a small amount of labeled data along with a larger amount of unlabeled data to achieve better generalization and accuracy compared to purely supervised learning methods
- valuable in scenarios where labeled data is scarce or expensive to obtain
  - model can benefit from the wealth of information present in the unlabeled instances, leading to improved performance and higher accuracy
  - performance gain may vary depending on the quality and relevance of the unlabeled data.

Road ahead and other ideas

Semi-Supervised Learning - concept

research fh st. pölten

- semi-supervised learning (SSL) uses a small portion of labeled data and lots of unlabeled data to train a predictive model
- Two main counterparts:
- Supervised learning
  - training a machine learning model using the labeled dataset.
  - Limitations: slow → label training examples
- Unsupervised learning
  - A model tries to mine hidden patterns, differences, and similarities in unlabeled data by itself, without human supervision
  - data points are grouped into clusters based on similarities
  - Limitations: limited area of applications and less accurate results

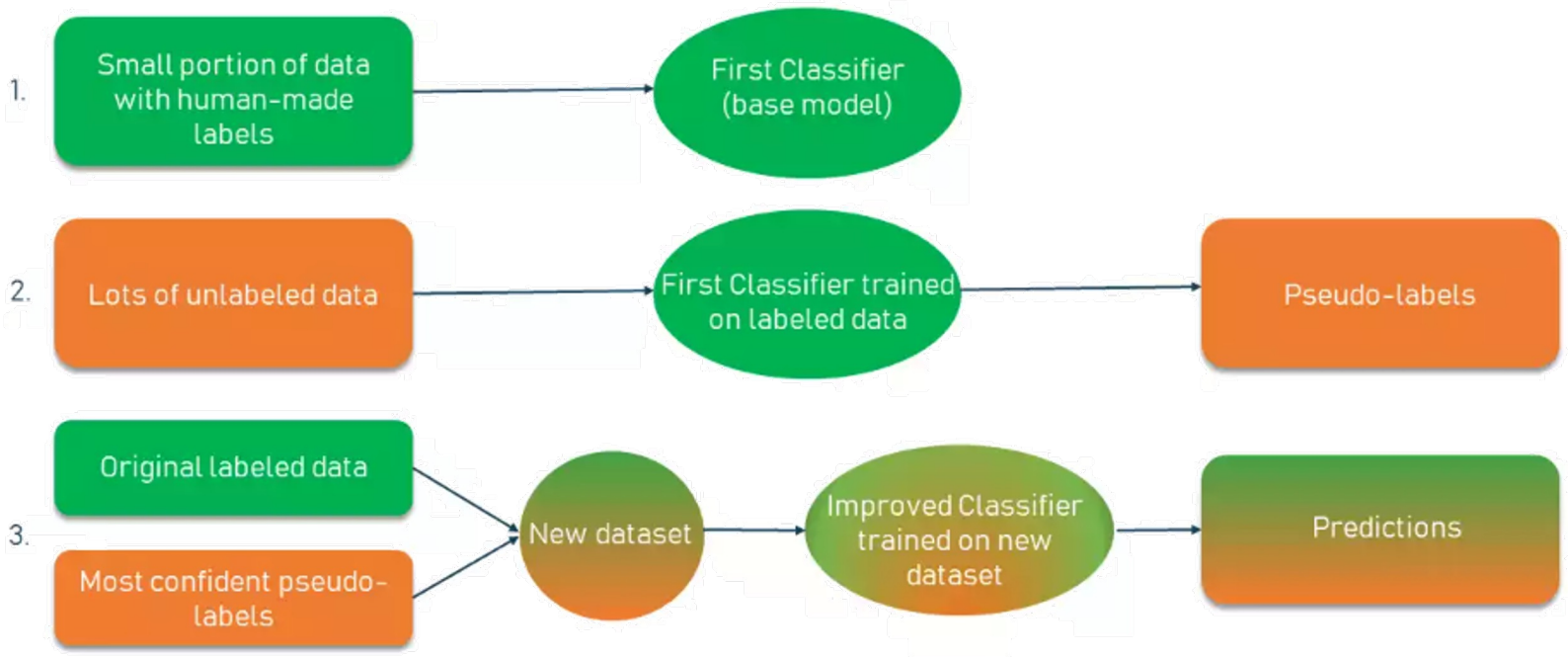


Road ahead and other ideas

Semi-Supervised Learning - Self-Training approach

research fh st.pölten

- simplest examples is self-training
- procedure in which we can take any supervised method for classification or regression and modify it to work in a semi-supervised manner
  - taking advantage of labeled and unlabeled data



Efficient supervised learning

One-Shot, Zero-Shot and Few-Shot Learning

Road ahead and other ideas

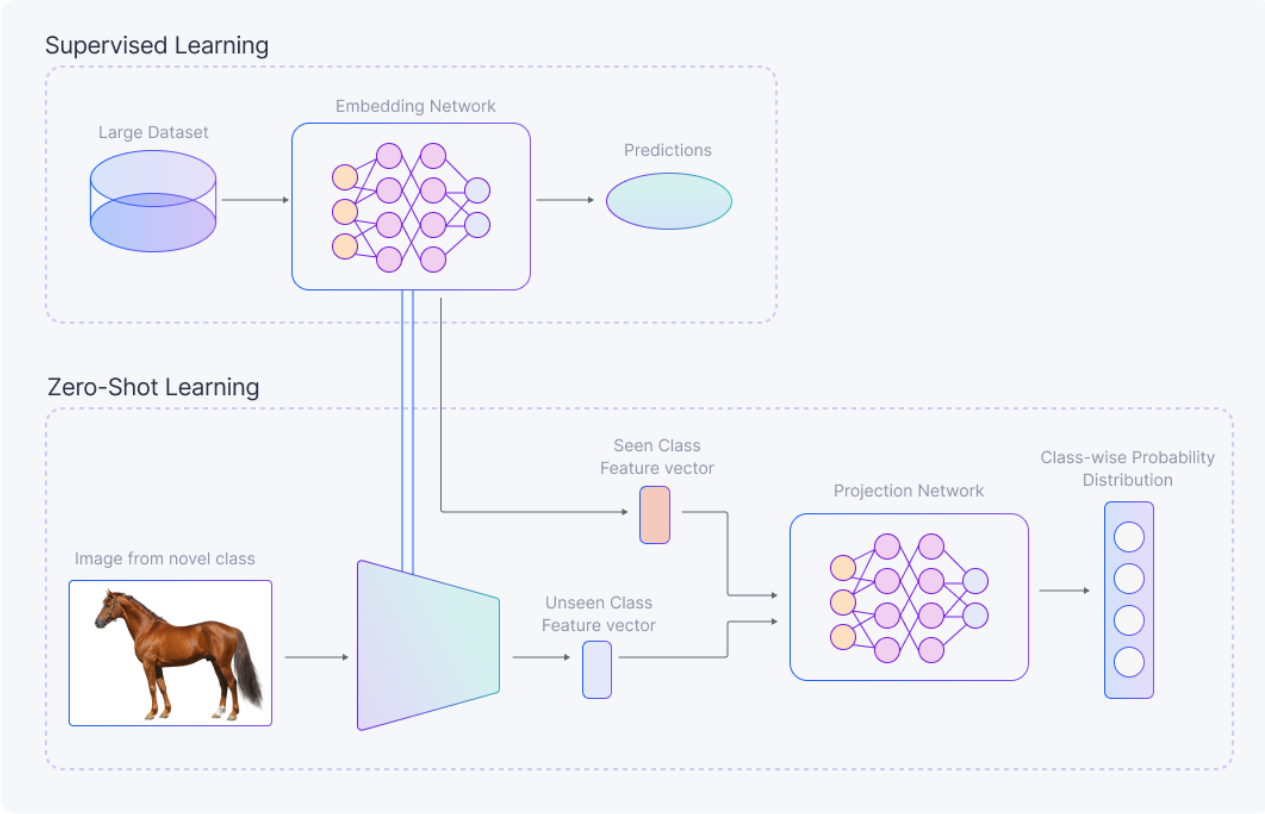
Efficient supervised learning

research

- address the problem of learning from limited or no labeled data
- **One-shot learning**
  - each new class has one labeled example
  - make predictions for the new classes based on this single example
- **Few-shot learning**
  - **l**imited number of labeled examples for each new class.
  - make predictions for new classes based on just a few examples of labeled data
- **Zero-shot learning**
  - absolutely no labeled data available for new classes
  - Subfield of Transfer Learning
  - make predictions about new classes by using prior knowledge about the relationships that exist between classes it already knows
    - large language models (LLMs) like ChatGPT, prior knowledge is likely include semantic similarities

Road ahead and other ideas

Zero-Shot Learning - Process

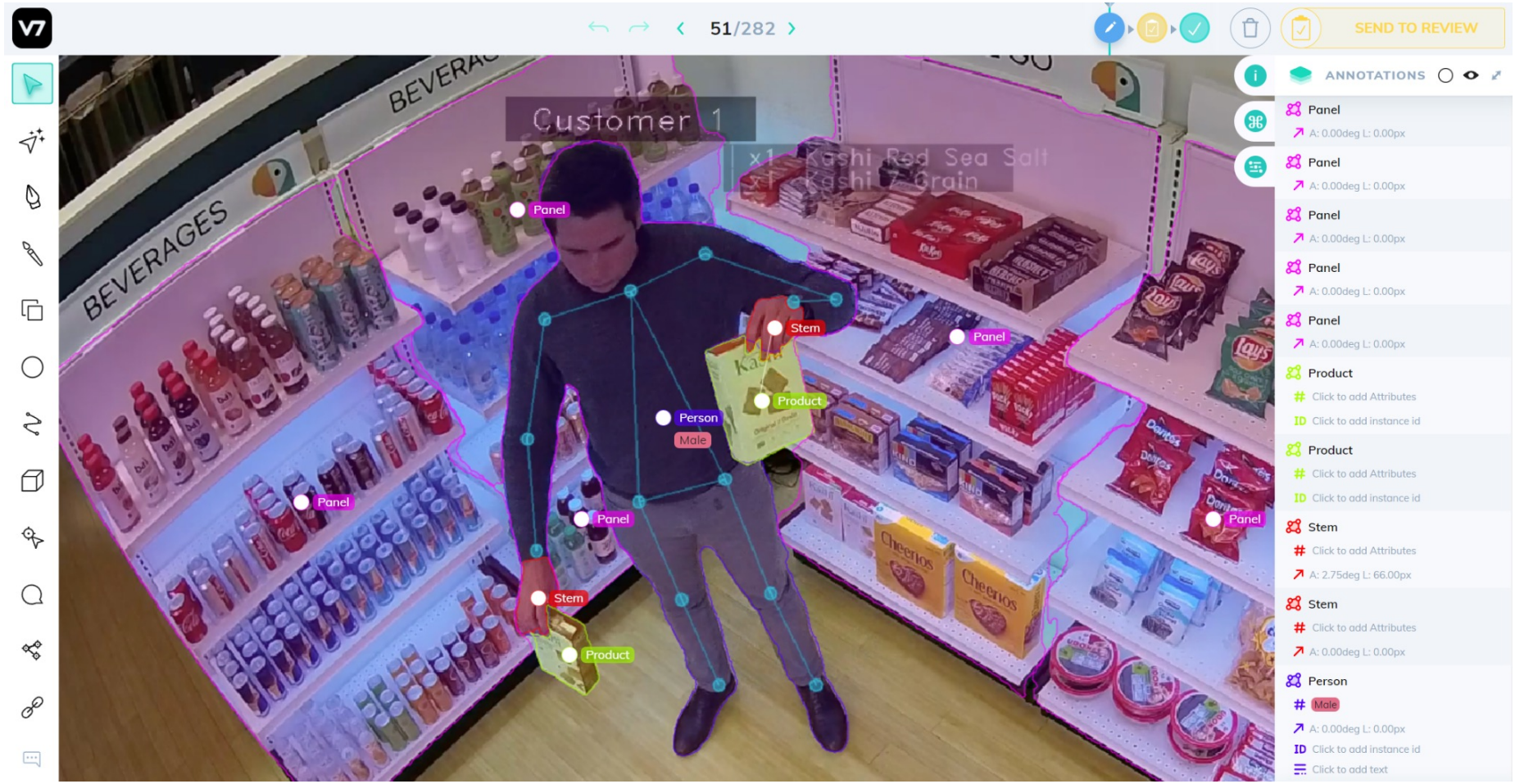


https://www.v7labs.com/blog/zero-shot-learning-guide

Road ahead and other ideas

Zero-Shot Learning - Applications

research



Thank you!

Fabian Kovac
fabian.kovac@fhstp.ac.at

Oliver Eigner
oliver.eigner@fhstp.ac.at

Workshop Material
https://github.com/fkabs/mlprague_2024