

# Active Learning and Regularization

SRP Pitch by Ilia

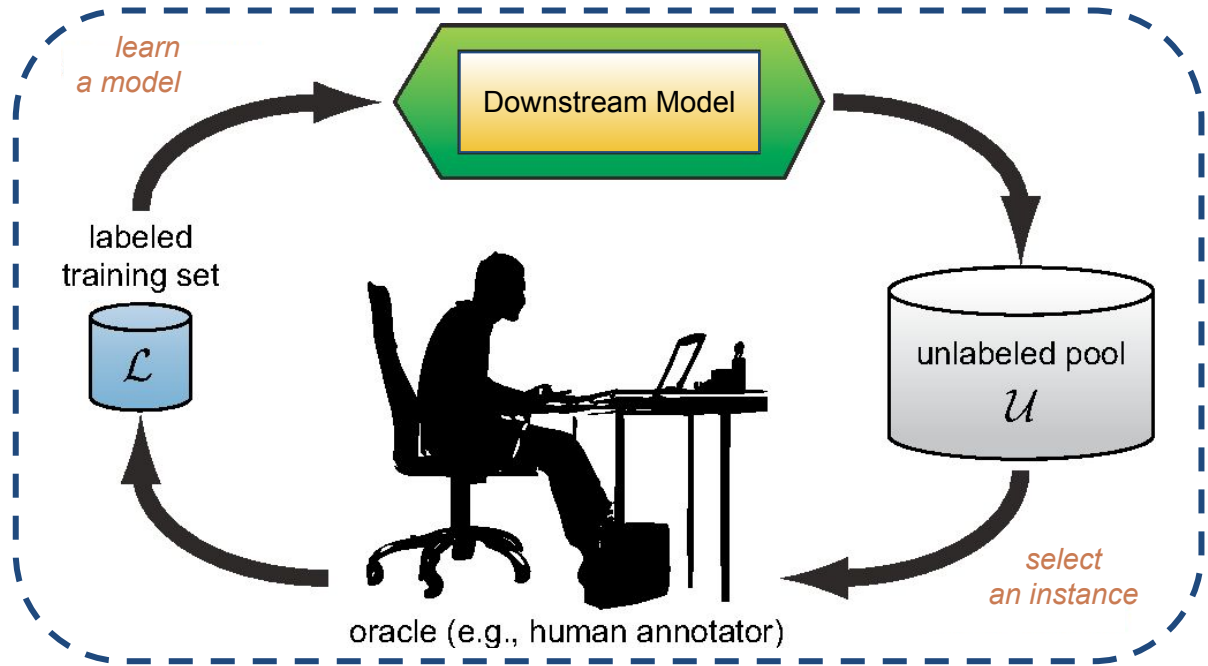
Link to the slides

Email:

koloiarov@ismll.de



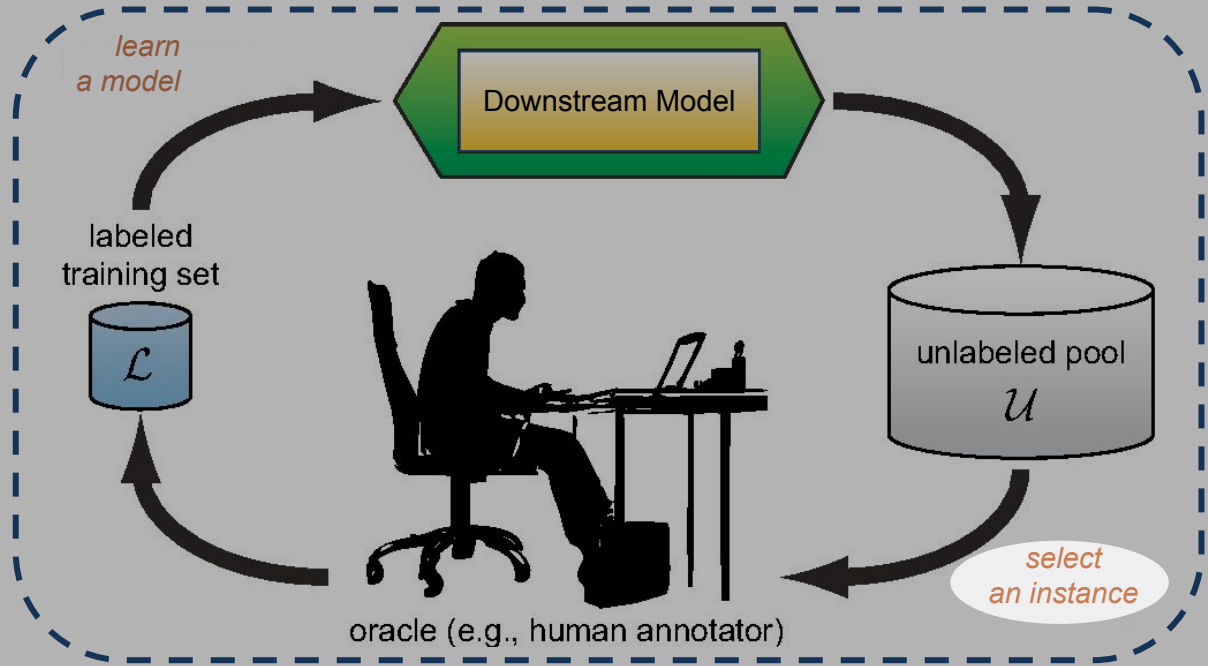
Repeat  $B$  (Budget) times



# Active Learning

High Level Visualization of pool-based AL  
[Settles, 2009]

Repeat  $B$  (Budget) times

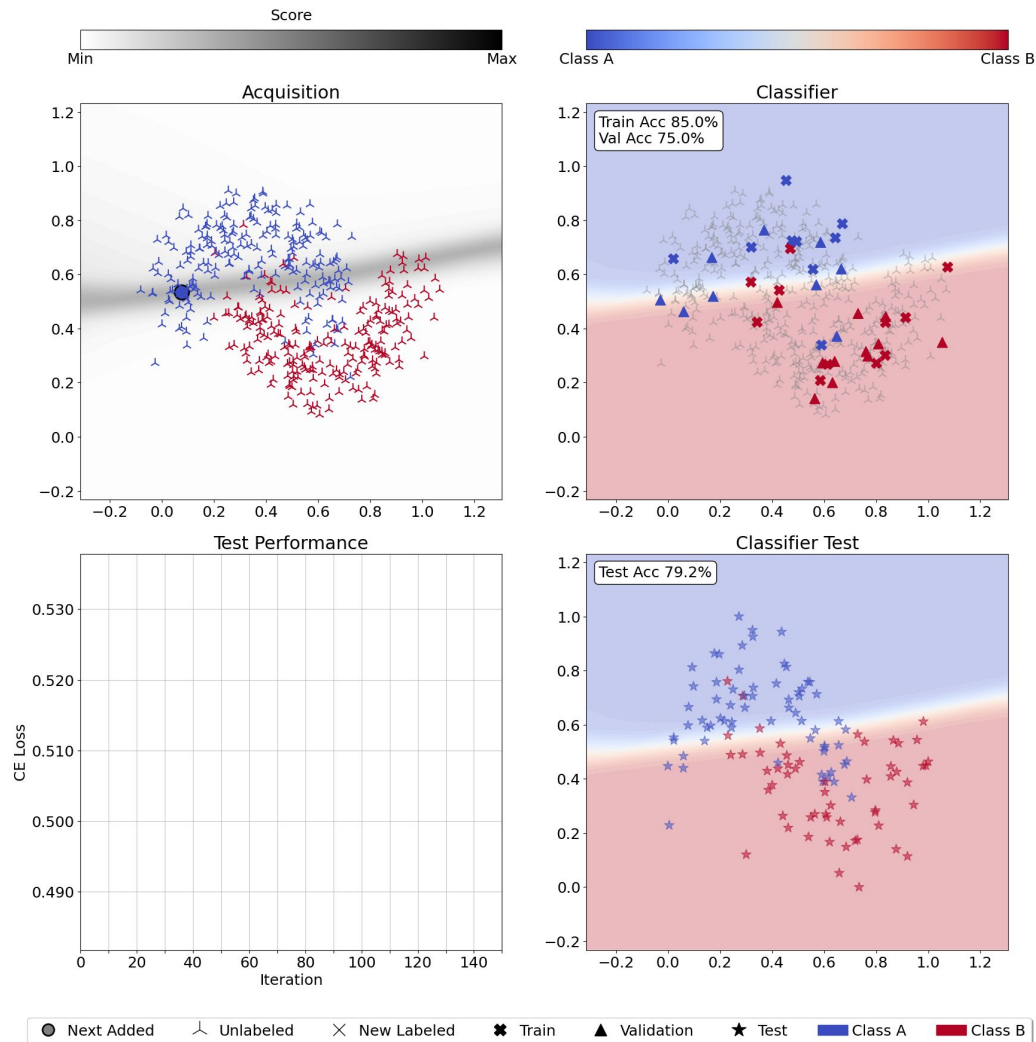


# Active Learning

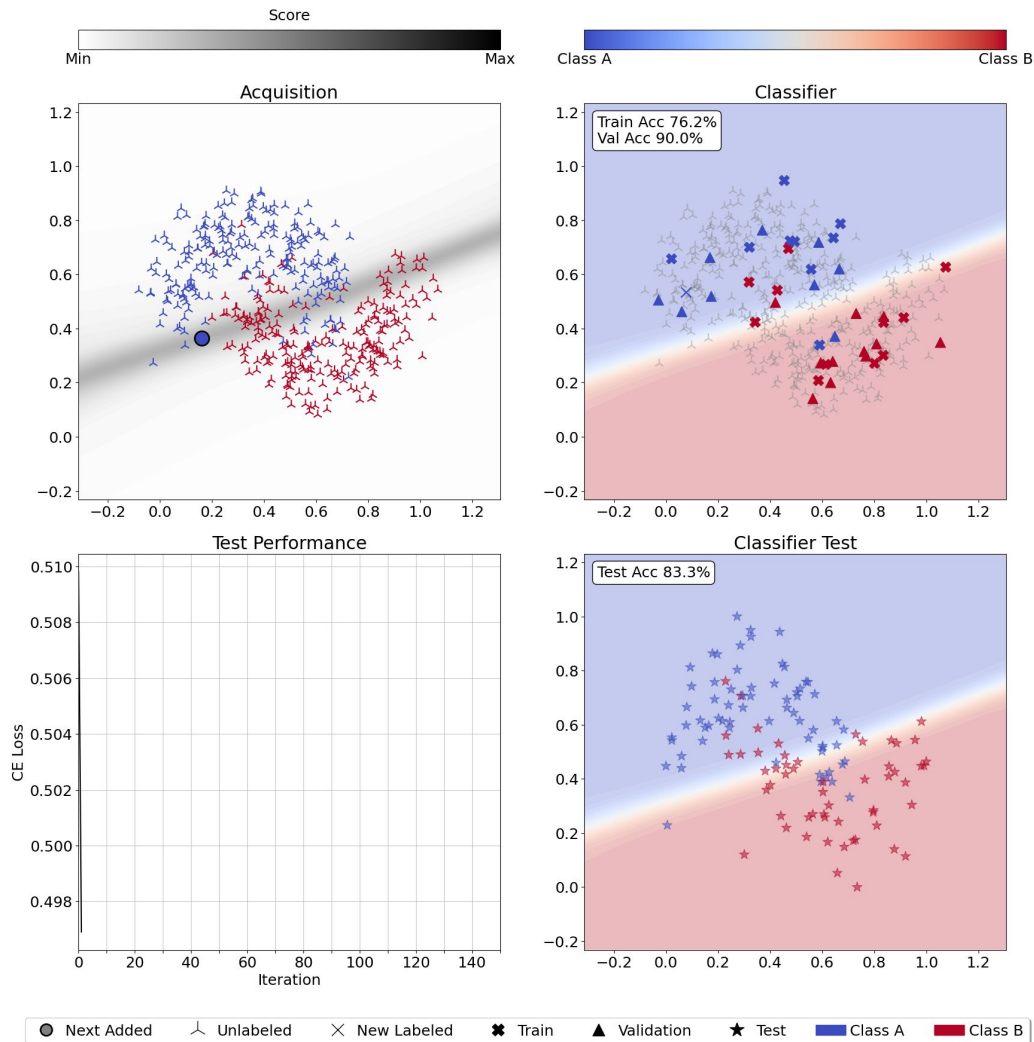
High Level Visualization of pool-based AL  
[Settles, 2009]

**Acquisition Function** — a scoring criterion  $\Phi$  used to select the most informative instance  $x^*$  from an unlabeled pool  $\mathcal{U}$  for labeling to improve a downstream model  $\mathcal{M}$  the most. If there are multiple instances with equal scores, one of them is selected randomly using a uniform distribution.

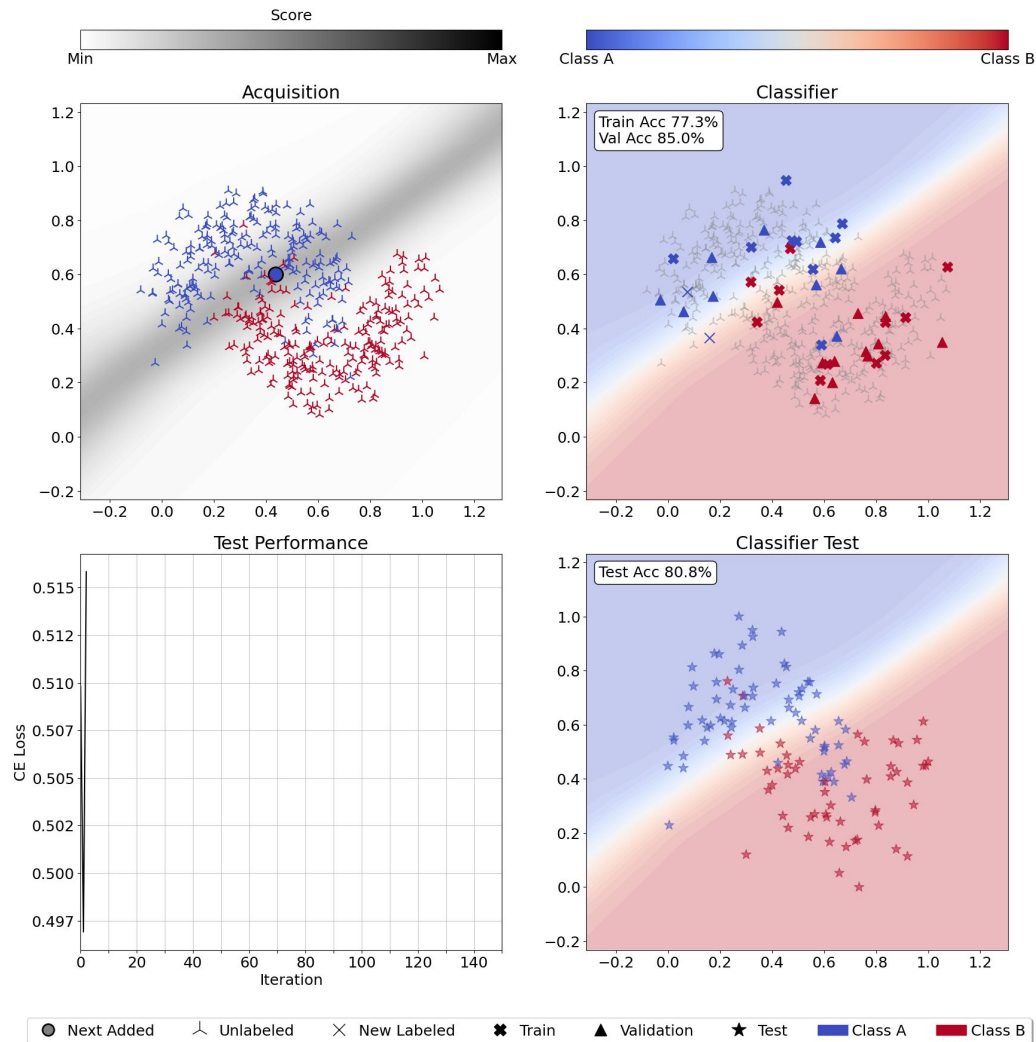
$$x^* \sim \text{Uniform}(\{x \in \mathcal{U} \mid \Phi(x, \mathcal{M}) = \max_{x \in \mathcal{U}} \Phi(x, \mathcal{M})\})$$



# Active Learning

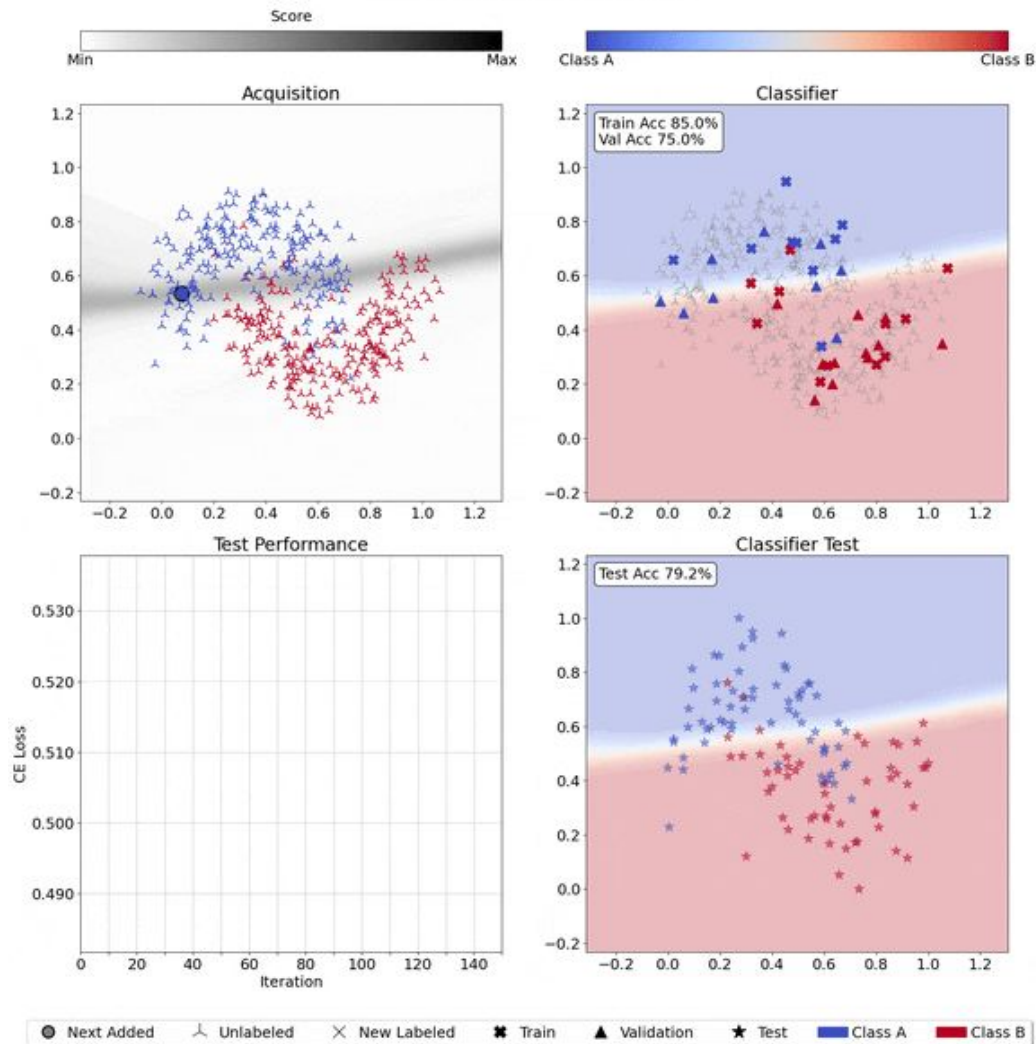


# Active Learning



# Active Learning





# Active Learning

# GitHub Repository with more examples

**ALFrameworkComparison** Public

Unpin Unwatch 1 Fork 0 Star 1

main 1 Branch 0 Tags

Go to file

Add file Code

<b>Nearpit</b>	Update README.md	88854659 · 4 months ago	259 Commits
acquisitions	Update entropy.py	4 months ago	
core	Update nl.py	4 months ago	
datasets	splice fix	4 months ago	
utilities	less verbose	4 months ago	
visualization	123	4 months ago	
.gitignore	First experiments	5 months ago	
README.md	Update README.md	4 months ago	
main.py	conventional protocol	5 months ago	
setup.py	debug splice	6 months ago	

**README**

**Please, Cease Employing Static Hyper-Parameters in Active Learning: A Comprehensive Framework Comparison**

**Acquisition Functions Visualization**

**Random Sampling Baseline**

Entropy Iter:37 Random Seed:2

Score

Min Max

Class 4 Class 8

Acquisition

Classifier

1.2 1.0

1.0 1.0

Score: 89.8%

Test Acc: 90.0%

**About**

No description, website, or topics provided.

Readme

Activity

1 star

1 watching

0 forks

**Releases**

No releases published

[Create a new release](#)

**Packages**

No packages published

[Publish your first package](#)

**Languages**

Python 100.0%

**Suggested workflows**

Based on your tech stack

**Publish Python Package** Configure

Publish a Python Package to PyPI on release.

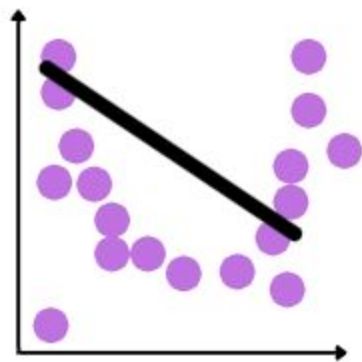
**Python application** Configure

Create and test a Python application.

**Python Package** Configure

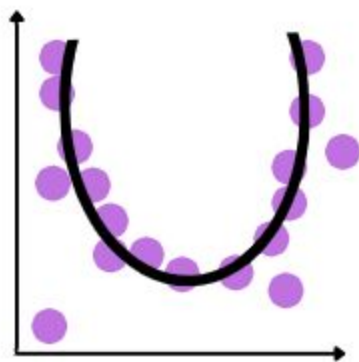


<https://github.com/Nearpit/ALFrameworkComparison>

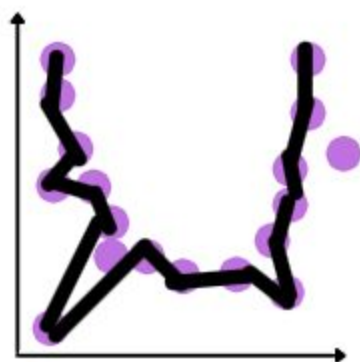


Underfitting

(model is too simple)



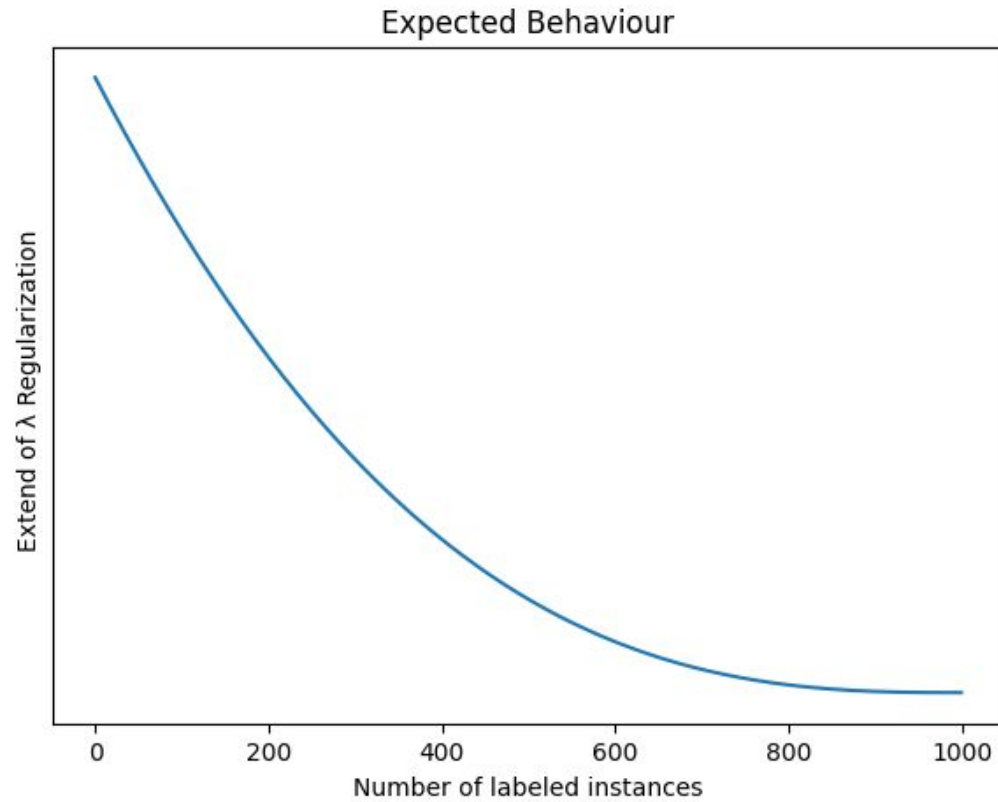
Optimal



Overfitting

(model is too complex and captures even noise in the data)

# Regularization



Current Hypothesis

## Desirable Result/Goal

Build a new acquisition function that relies on regularization

# Preliminary work – Get aware of the problem's domain

1. **ANY PAPER YOU find suitable** – keep in mind, it's your research experience
2. Overall AL – [Settles, 2009]
3. AL Training protocol – [Munjal et al., 2022]
4. Acquisition families:
  - a. Ensembles – [Beluch et al., 2018]
  - b. Uncertainty-based:
    - i. Entropy/Margin/etc... – [Settles 2009]
    - ii. BALD&DBAL – [Gal et al., 2017]
  - c. Geometrical-based:
    - i. Core-set – [Sener and Savarese, 2018]
    - ii. TypiClust – [Hacohen et al., 2022]

# Data&Model Domain:

## Path A:

1. Computer Vision **Embeddings** (raw data is computationally intractable on our cluster)
  - a. CIFAR10/CIFAR100
2. Models
  - a. Embedding retrieval
    - i. ResNet/DinoV2
  - b. Predictions
    - i. Simple Linear Layer
3. Pros
  - a. Much simpler to implement
  - b. More generic approach
4. Cons
  - a. Many pitfalls since I have never implemented it myself
  - b. Less applicative in the real-world scenario

## Path B:

1. Raw Tabular Data
  - a. Spline, DNA, Synthetic Datasets, etc...
2. Models
  - a. MLP
  - b. ResNet [He et al., 2015]
  - c. FT-Transformer [Gorishny et al., 2023]  
SAINT [Somepalli et al., 2021]
  - d. [Optional] XGBoost [Chen and Guestrin 2016]
3. Pros
  - a. You'll get aware more about the State of the Art in the Tabular domain
  - b. I will help you with it (my research domain)
4. Cons
  - a. More challenging topic
  - b. Much more coding

# Data&Model Domain:

Path C:

1. Any domain **YOU** are intrinsically connected to
  - a. Up to you
2. Any model **YOU** find fascinating and great performing
  - a. Up to you
3. Pros
  - a. Huge amount of experience(i.e. you will learn a lot)
  - b. You'll be ready for any Industrial/PhD task
4. Cons
  - a. HUMONGOUS amount of pitfalls
  - b. It requires a lot of time investment



## Desirable Result/Goal

Build a new acquisition function that relies on regularization

\*and have some fun

# Reminder:

Don't forget to choose your topic preferences on  
Monday–Tuesday

# Literature

- Beluch, William H., Tim Genewein, Andreas Nurnberger, and Jan M. Kohler. “The Power of Ensembles for Active Learning in Image Classification.” In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9368–77, 2018. <https://doi.org/10.1109/CVPR.2018.00976>.
- Gal, Yarin, Riashat Islam, and Zoubin Ghahramani. “Deep Bayesian Active Learning with Image Data.” arXiv, March 8, 2017. <https://doi.org/10.48550/arXiv.1703.02910>.
- Gorishniy, Yury, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. “Revisiting Deep Learning Models for Tabular Data.” arXiv, November 10, 2021. <https://doi.org/10.48550/arXiv.2106.11959>.
- Hacohen, Guy, Avihu Dekel, and Daphna Weinshall. “Active Learning on a Budget: Opposite Strategies Suit High and Low Budgets.” arXiv, June 16, 2022. <https://doi.org/10.48550/arXiv.2202.02794>.
- Munjal, Prateek, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. “Towards Robust and Reproducible Active Learning Using Neural Networks.” arXiv, June 15, 2022. <https://doi.org/10.48550/arXiv.2002.09564>.
- Sener, Ozan, and Silvio Savarese. “Active Learning for Convolutional Neural Networks: A Core-Set Approach.” arXiv, June 1, 2018. <https://doi.org/10.48550/arXiv.1708.00489>.
- Settles, Burr. “Active Learning Literature Survey.” Technical Report. University of Wisconsin-Madison Department of Computer Sciences, 2009. <https://minds.wisconsin.edu/handle/1793/60660>.
- Somepalli, Gowthami, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. “SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training.” arXiv, June 2, 2021. <https://doi.org/10.48550/arXiv.2106.01342>.