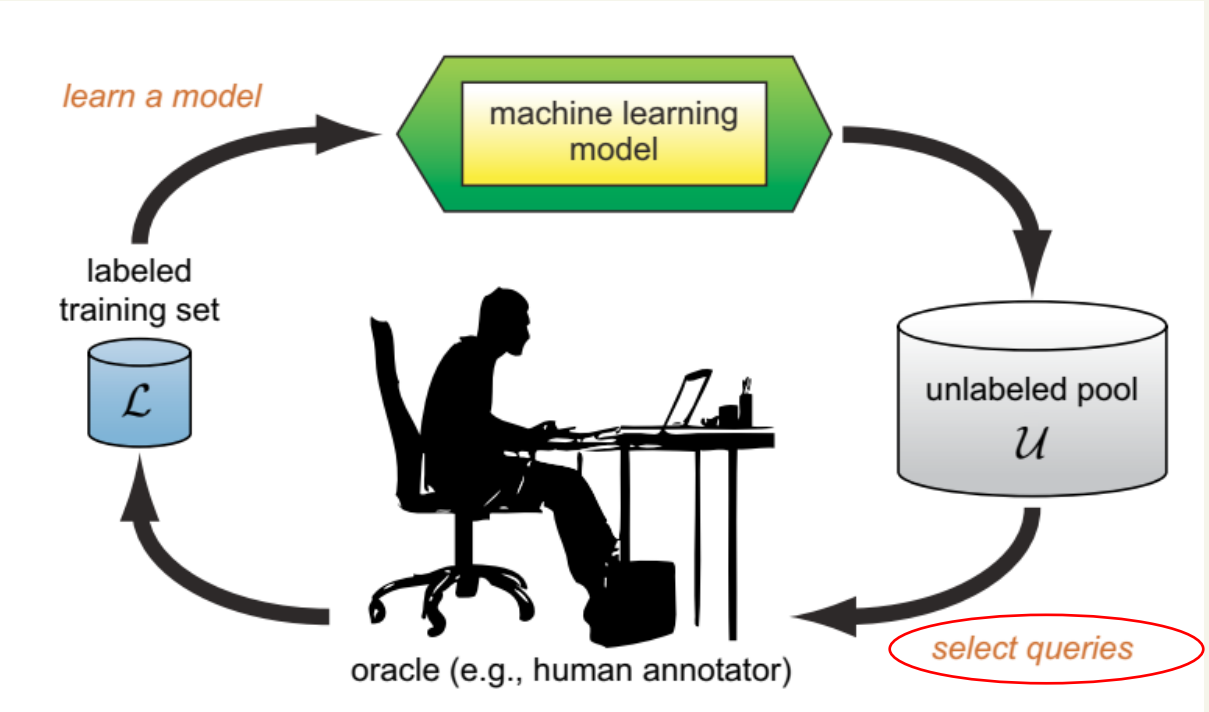
# Learning Active Learning from Data

NIPS17 Ksenia Konyushkova, Sznitman Raphael

# 目录 OUTLINE

# Background




现有的大部分工作：

$$g(\boldsymbol{x_i}) = score_i \qquad \forall x_i \in U$$

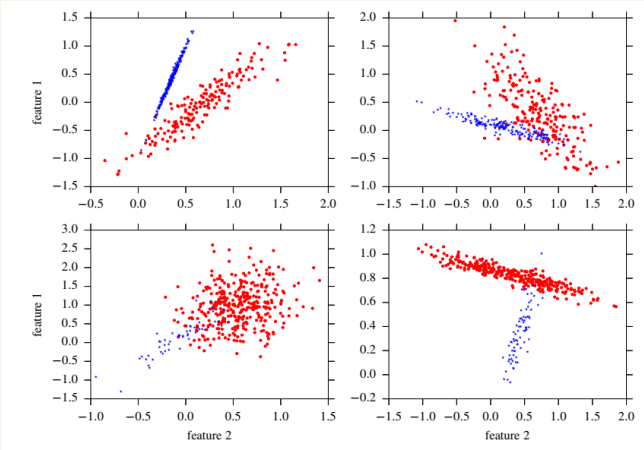
其中 $g(\cdot)$ 是一个手工设计的指标

# Introduction

训练一个回归模型来预测每个样本对模型的性能提升：

$$h(\hat{\boldsymbol{f}}, \hat{\boldsymbol{x}}_{\boldsymbol{i}}) = score_i \qquad \forall x_i \in U$$

其中$\hat{\boldsymbol{f}}, \hat{\boldsymbol{x}}_{\boldsymbol{i}}$分别是与当前分类模型相关的，与未标记样本相关的特征向量



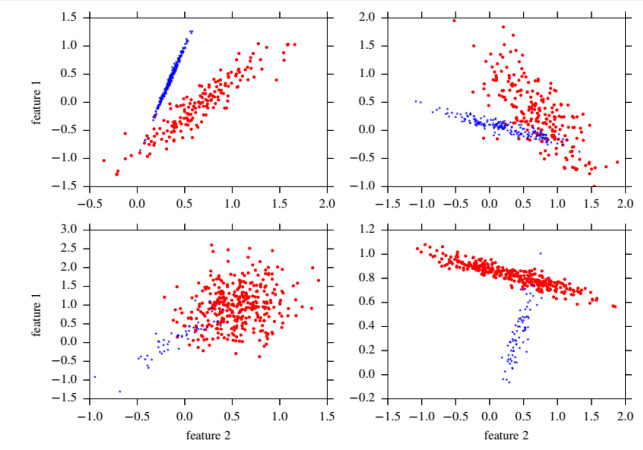

synthetic dataset (two different Gaussian distributions)

特征:
手工提取的模型与样本的特征

标记:
将样本查询后模型的性能提升

训练 →

$h(\cdot)$




Learn

Query

Real dataset

- 特征$\hat{\boldsymbol{f}}, \hat{\boldsymbol{x}}_{\boldsymbol{i}}$与domain无关

# The proposed method

获取$h(\cdot)$的训练数据:

## synthetic 2D datasets

Repeat:

随机生成均值，方差，类别比例不同的高斯分布
随机划分训练，测试集
随机初始标记点
随机采样1个点$\boldsymbol{x_i}$，提取$\hat{\boldsymbol{f}}, \hat{\boldsymbol{x}}_{\boldsymbol{i}}$，计算$score_i$，加入训练回归模型的样本集合

## Features for $h(\cdot)$

模型使用包含k颗树的随机森林

- 模型的预测值
- K个预测值的均值和方差
- 正类样本的比例
- 模型包外估计的性能
- feature importance的方差
- 树的平均深度
- 已标记的点的数量

03

# The proposed method

**Independent LAL:**

Repeat:

随机生成均值，方差，类别比例不同的高斯分布

随机划分训练，测试集

随机初始标记点

随机采样1个点$\boldsymbol{x_i}$，提取$\hat{\boldsymbol{f}}, \hat{\boldsymbol{x}}_{\boldsymbol{i}}$，计算$score_i$，加入训练回归模型的样本集合

**Iterative LAL** :

Repeat:

随机生成均值，方差，类别比例不同的高斯分布

随机划分训练，测试集

随机初始标记点固定2个，剩下的点由每一轮的策略查询得到

随机采样1个点$\boldsymbol{x_i}$，提取$\hat{\boldsymbol{f}}, \hat{\boldsymbol{x}}_{\boldsymbol{i}}$，计算$score_i$，加入训练回归模型的样本集合

# Experiment

Compared methods:

- LAL-independent 2D
- LAL-independent WS
- LAL-iterative 2D
- Random
- Uncertainty
- Kapoor(ICCV07), an algorithm that balances exploration and exploitation by incorporating mean and variance estimation of the GP classifier
- ALBE(AAAI15) , a recent example of meta-AL that adaptively uses a combination of strategies, including Us, Rs and QUIRE

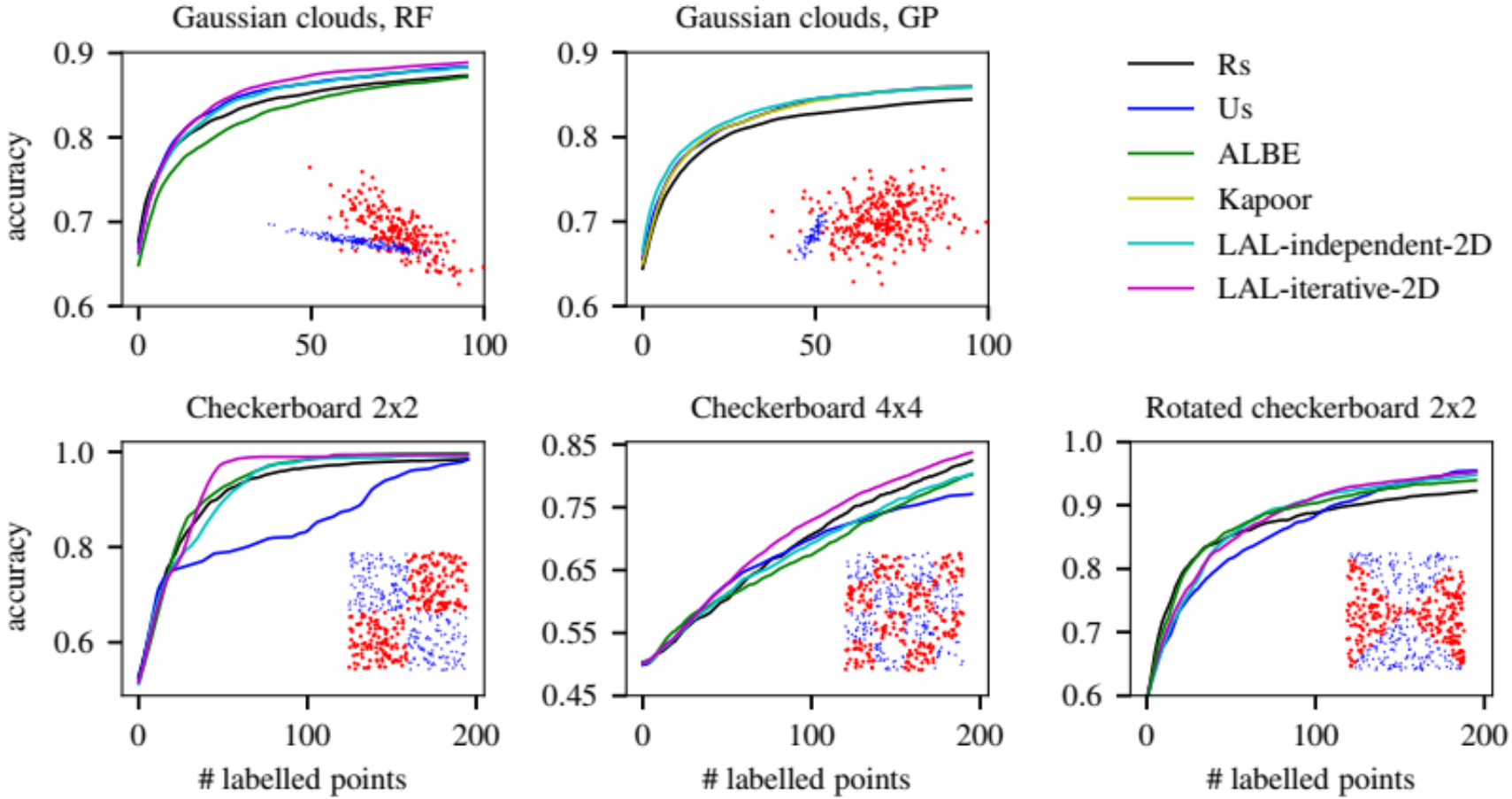



Figure 2: Experiments on the synthetic data. Top row: RF and GP on 2 Gaussian clouds. Bottom row from left to right: experiments on *Checkerboard* $2 \times 2$, *Checkerboard* $4 \times 4$, and *Rotated Checkerboard* $2 \times 2$ datasets.

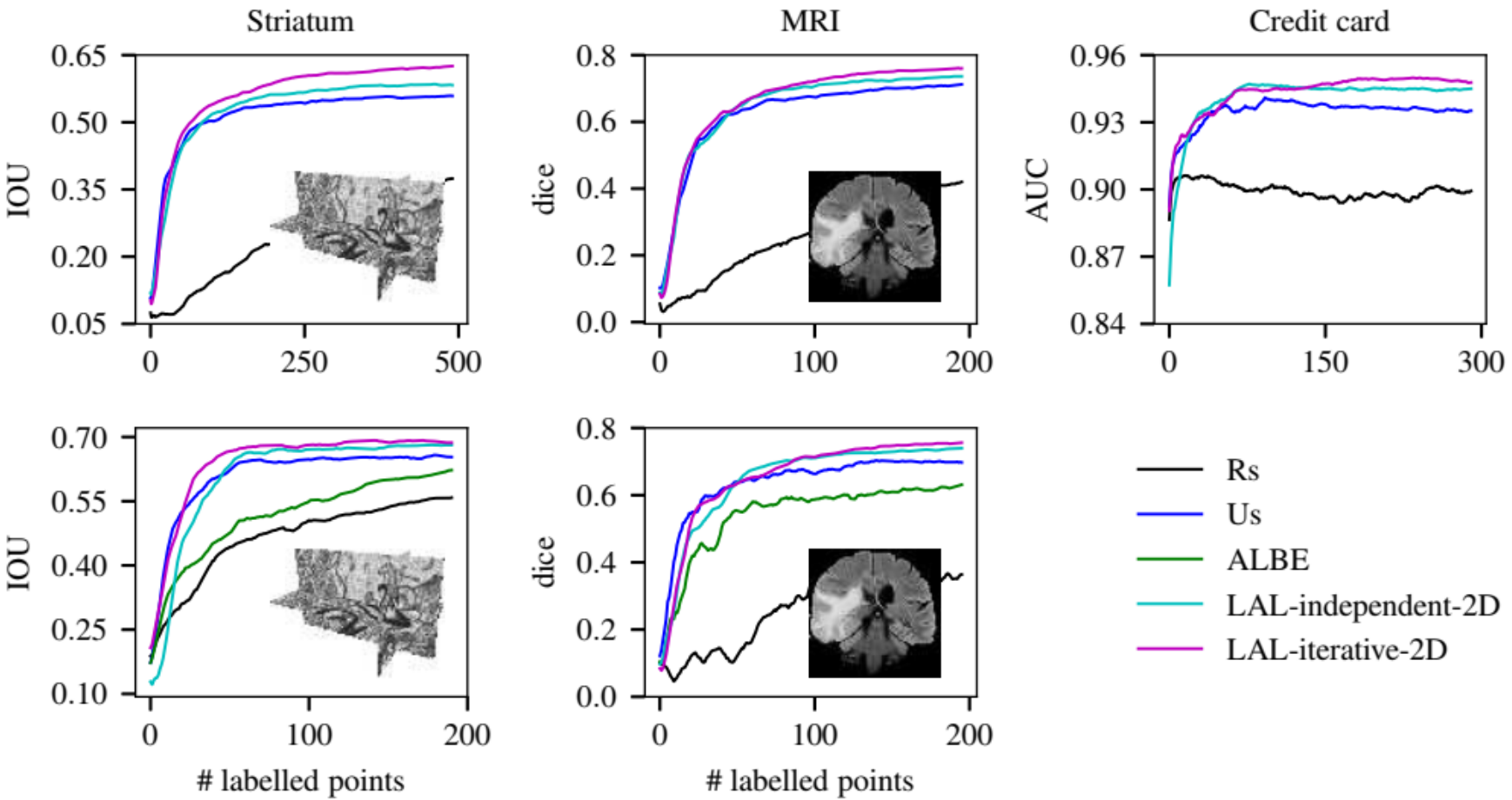
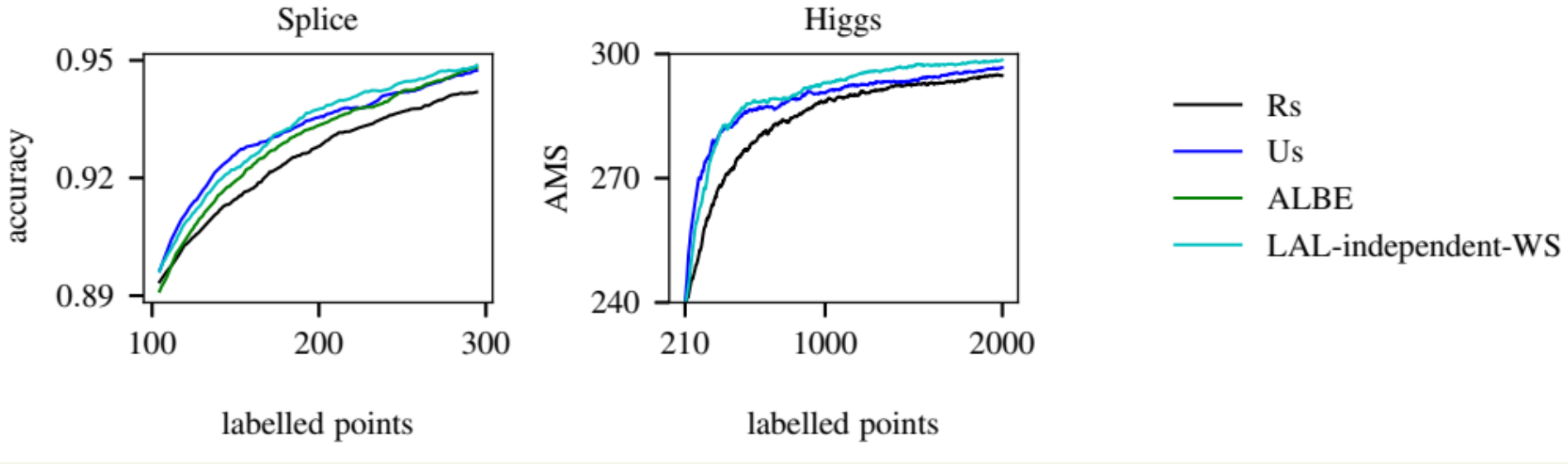



Figure 3: Experiments on real data. Top row: IOU for *Striatum*, dice score for *MRI* and AUC for *Credit card* as a function of a number of labeled points. Bottom row: Comparison with **ALBE** on the *Striatum mini* and *MRI mini* datasets.


Splice
0.95
0.92
0.89
accuracy
100
200
300
labelled points
Higgs
300
270
240
AMS
210
1000
2000
labelled points
Rs
Us
ALBE
LAL-independent-WS

# 问题:

1. 在2维数据集上训练的回归模型与现有指标差别不大
2. Warm Start的设置下需要较多初始标记样本来训练回归模型

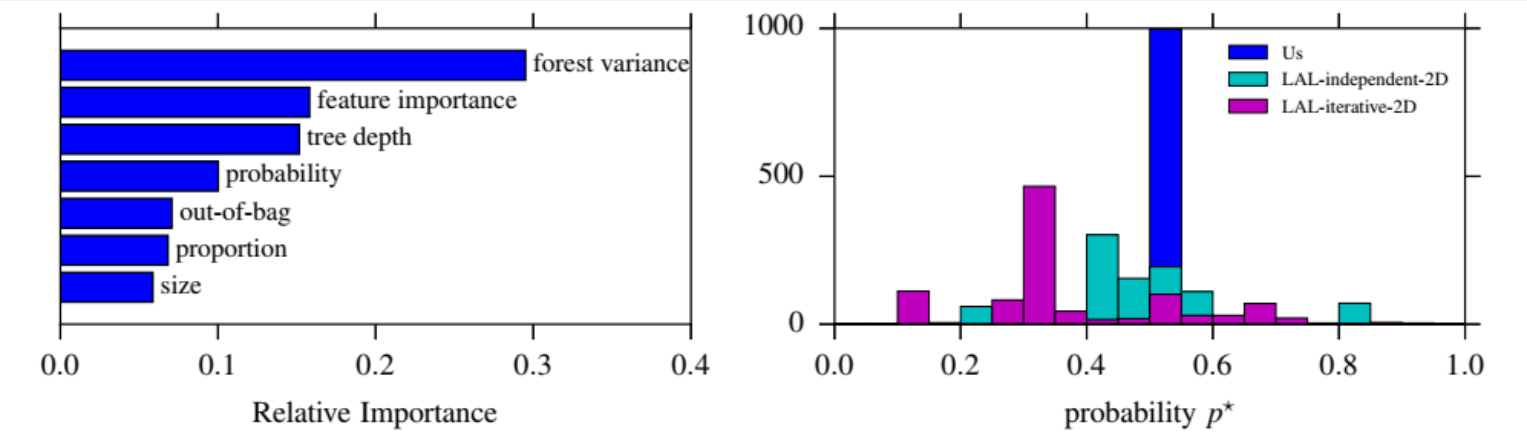



Figure 5: Left: feature importances of the RF regressor representing LALITERATIVE strategy. Right: histograms of the selected probability for different AL strategies in experiments with *MRI* dataset.

# THANKS