An introduction to weakly supervised learning.
Best practices.

Kristina Khvatova

Software Developer

Softec S.p.A.

Master in Computer Science and Applied Mathematics

Saint-Petersburg State University

Master in Computer Science and Data Analysis

Milano-Bicocca University

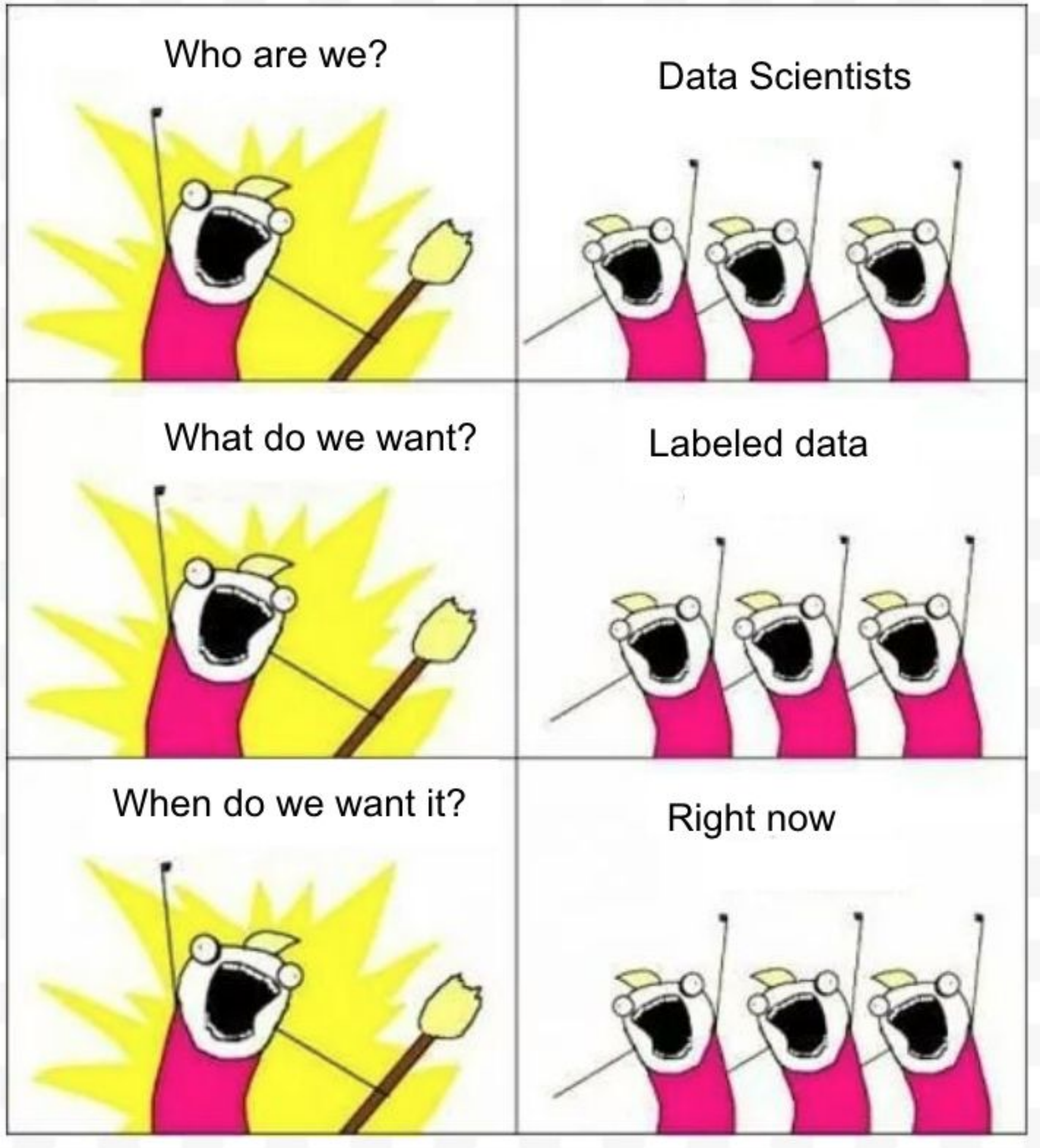
kristina.a.khvatova@gmail.com

https://www.linkedin.com/in/kristina-khvatova-a529b21

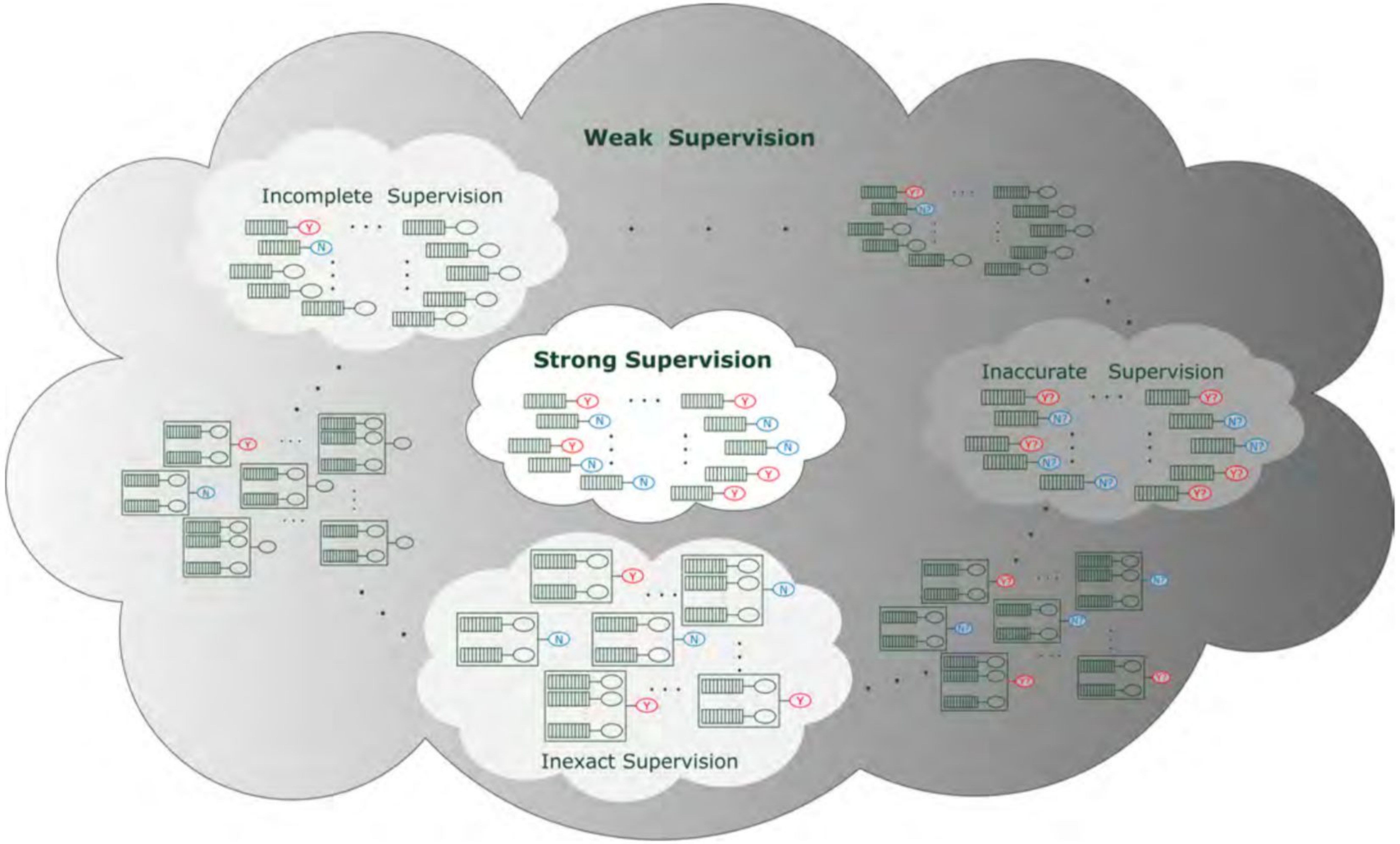
Overview

- ➢ Introduction to weak supervision
- ➢ Three types of weakly supervised learning:
    - incomplete
    - inexact
    - inaccurate
- ➢ Snorkel
- ➢ Brexit tweets classification with weak supervised learning

Problem Definition



Weak supervision

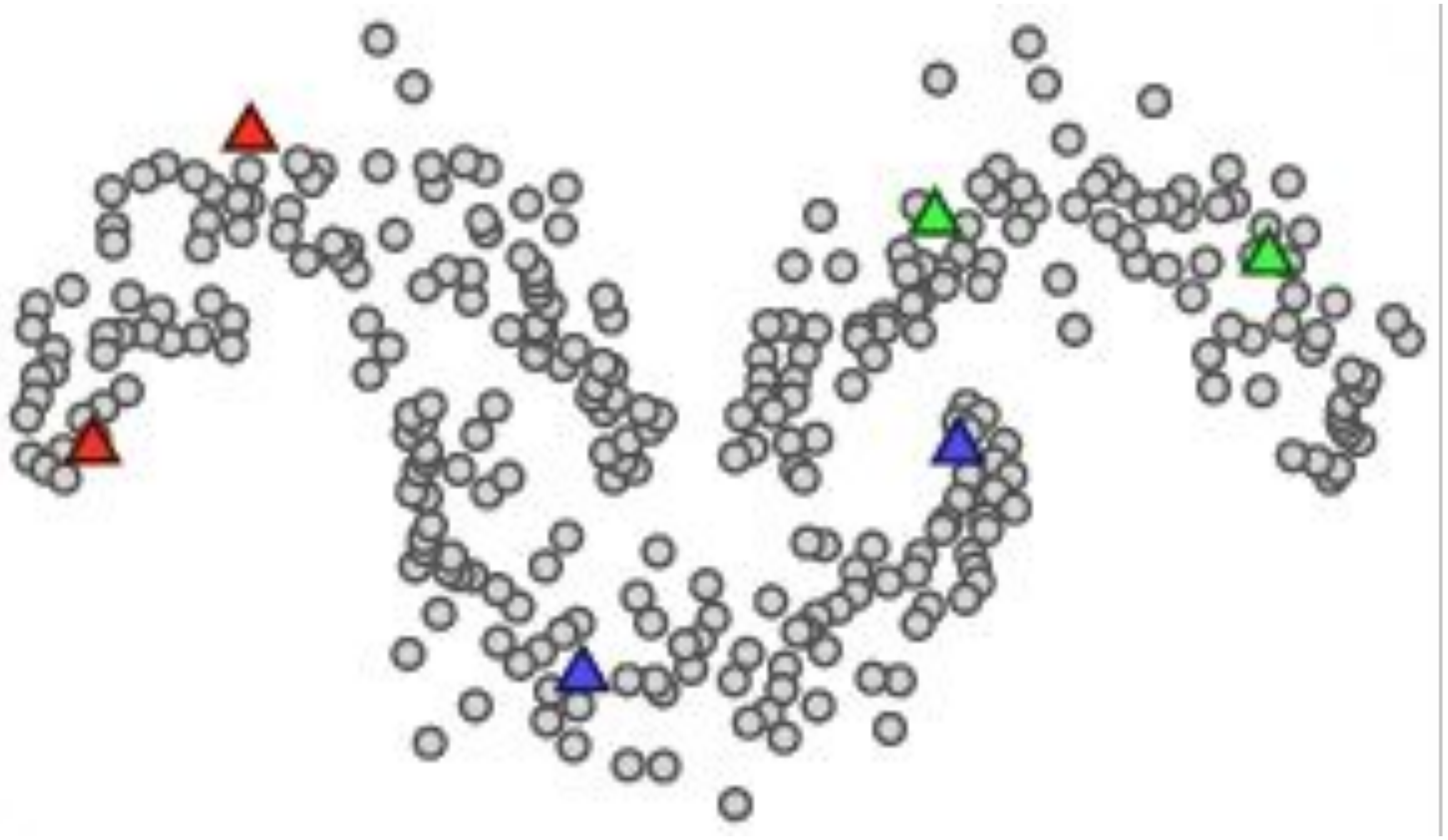


Weak supervision is the technique of building models based on new generated data.

Types:

- incomplete
- inexact
- inaccurate

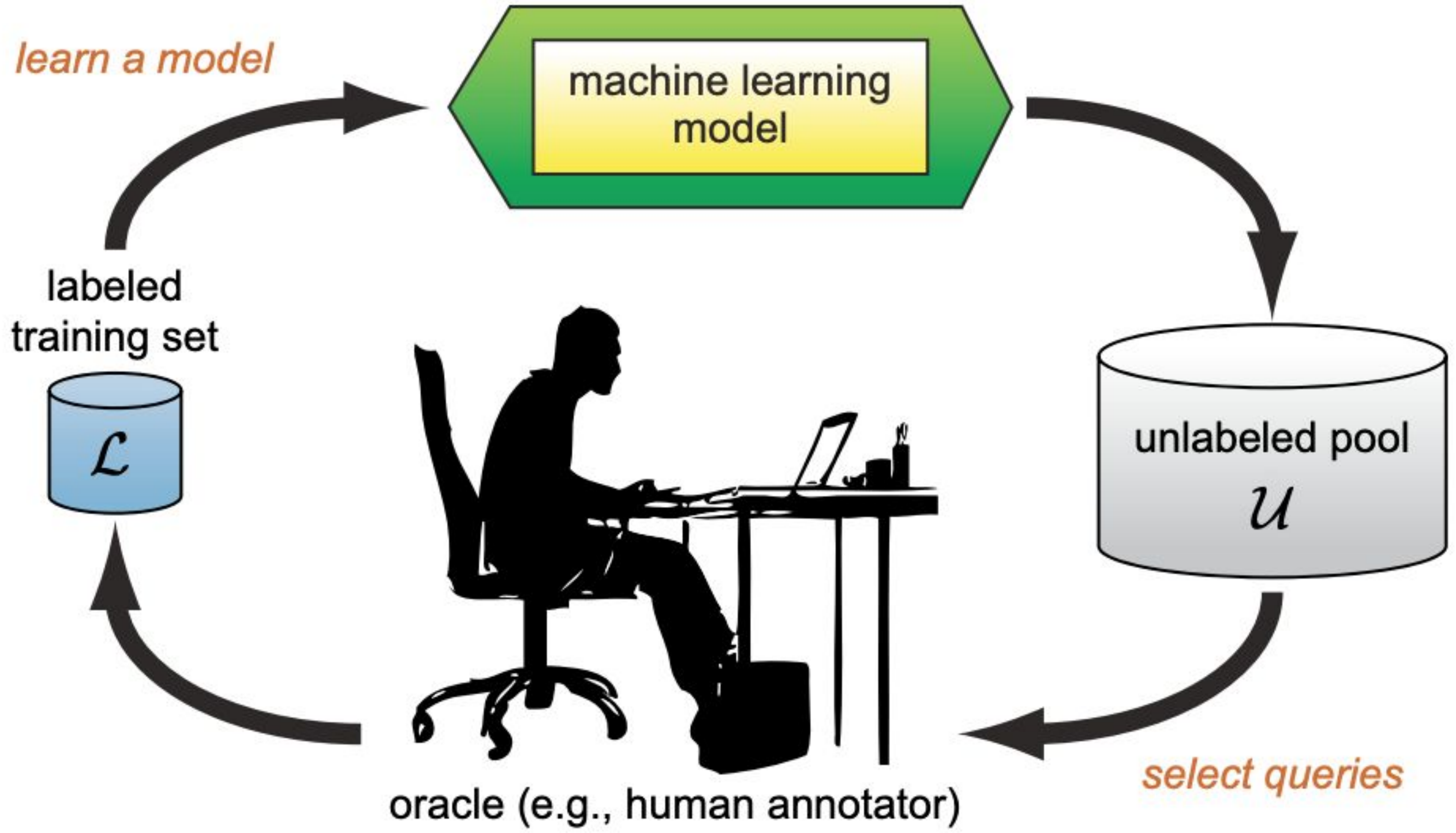
Incomplete weak supervision



- Active learning
- Semi - supervised learning

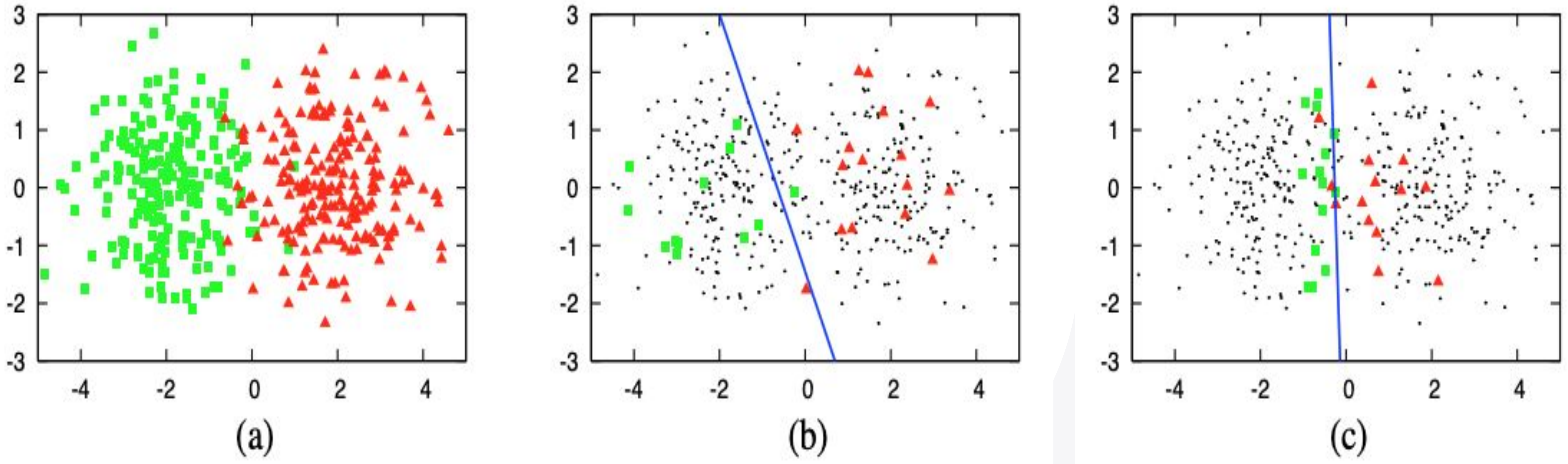
Incomplete weak supervision

Active learning



- High accuracy
- Low costs

Active learning



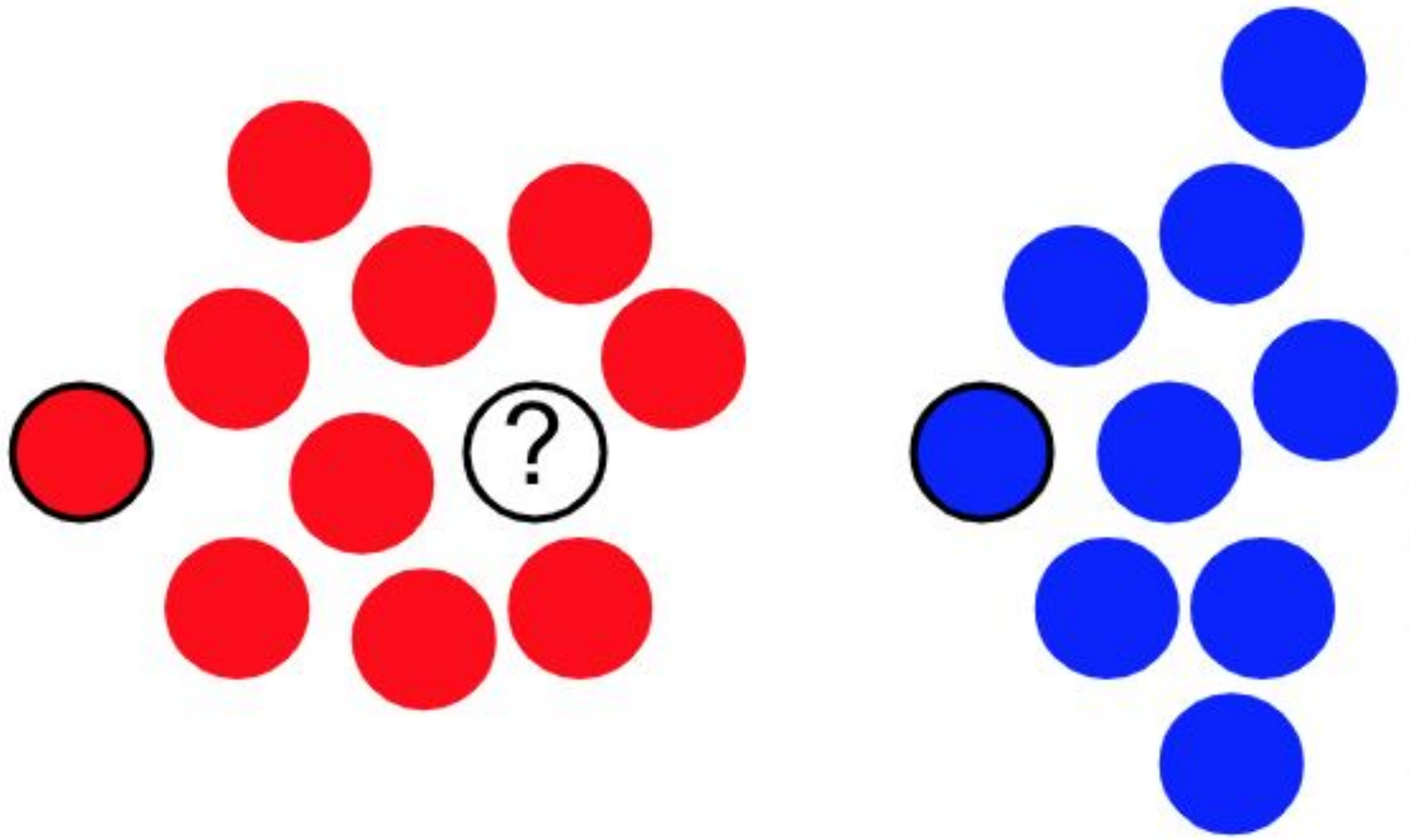
High costs for the project and high precision (90%)

Decrease costs and precision of the project (70%)

Cost of query labels is the same as in (b), but the precision is much more higher (90%) the same as (a)

Incomplete weak supervision

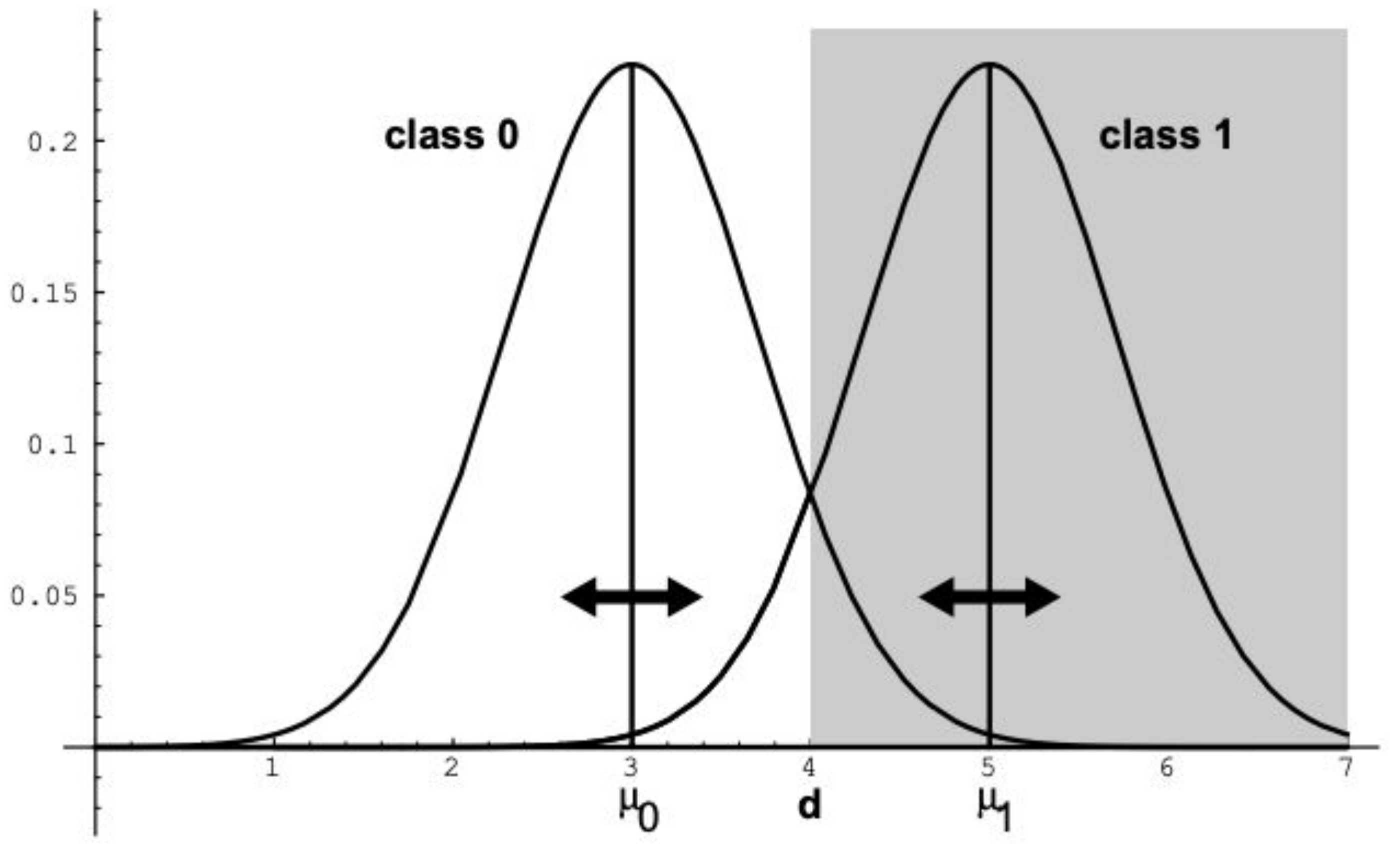
Semi-supervised learning



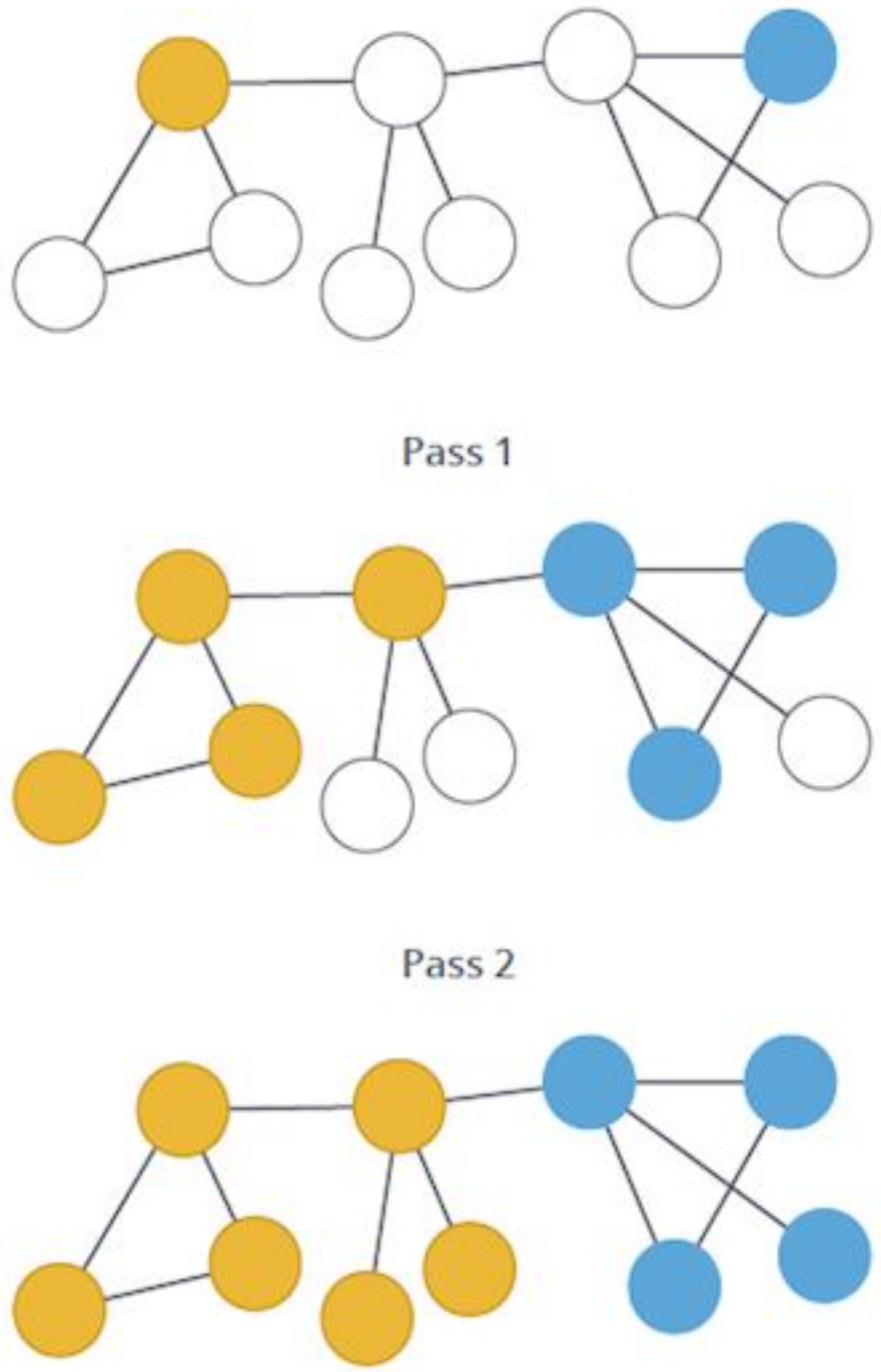
Incomplete weak supervision

Semi-supervised learning

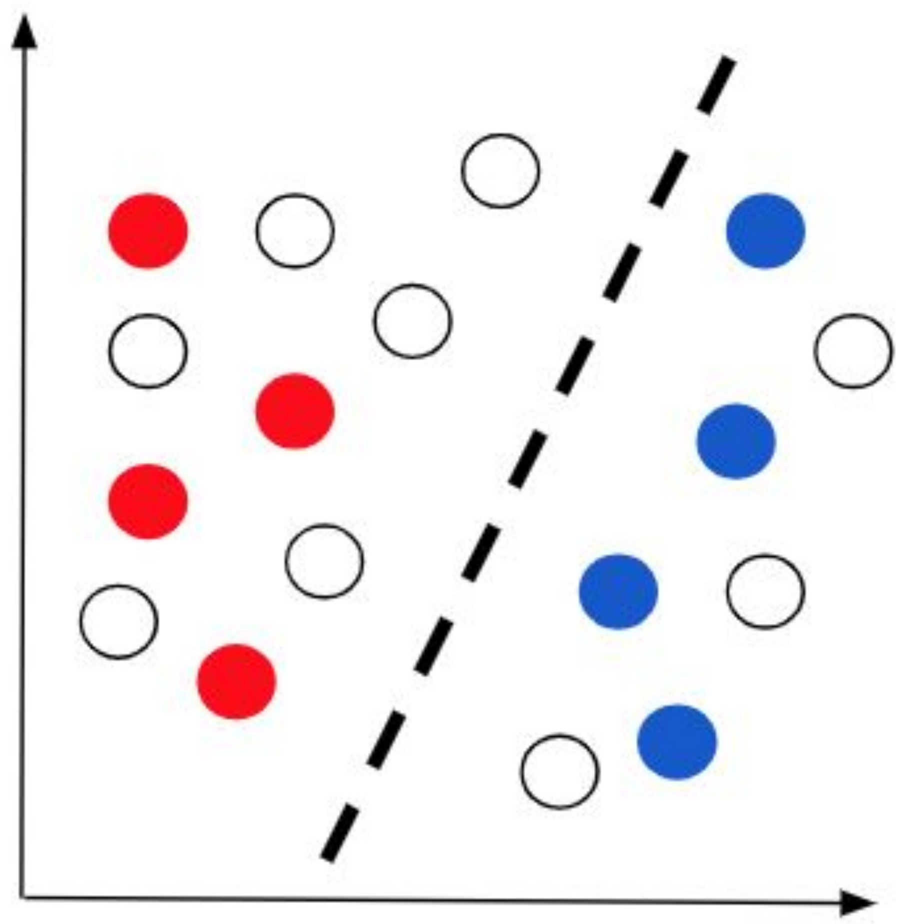
Generative models



Label propagation



TSVM



Inexact weak supervision

Single Image Surface Appearance Modeling with Self-augmented CNNs and Inexact Supervision

Wenjie Ye[1,3] Xiao Li[2,3] Yue Dong[3] Pieter Peers[4] Xin Tong[3]

[1] Tsinghua University
[2] University of Science and Technology of China
[3] Microsoft Research Asia
[4] College of William & Mary

Is object localization for free? – Weakly-supervised learning with convolutional neural networks

Léon Bottou† MSR, New York, USA
Ivan Laptev* INRIA, Paris, France
Josef Sivic* INRIA, Paris, France

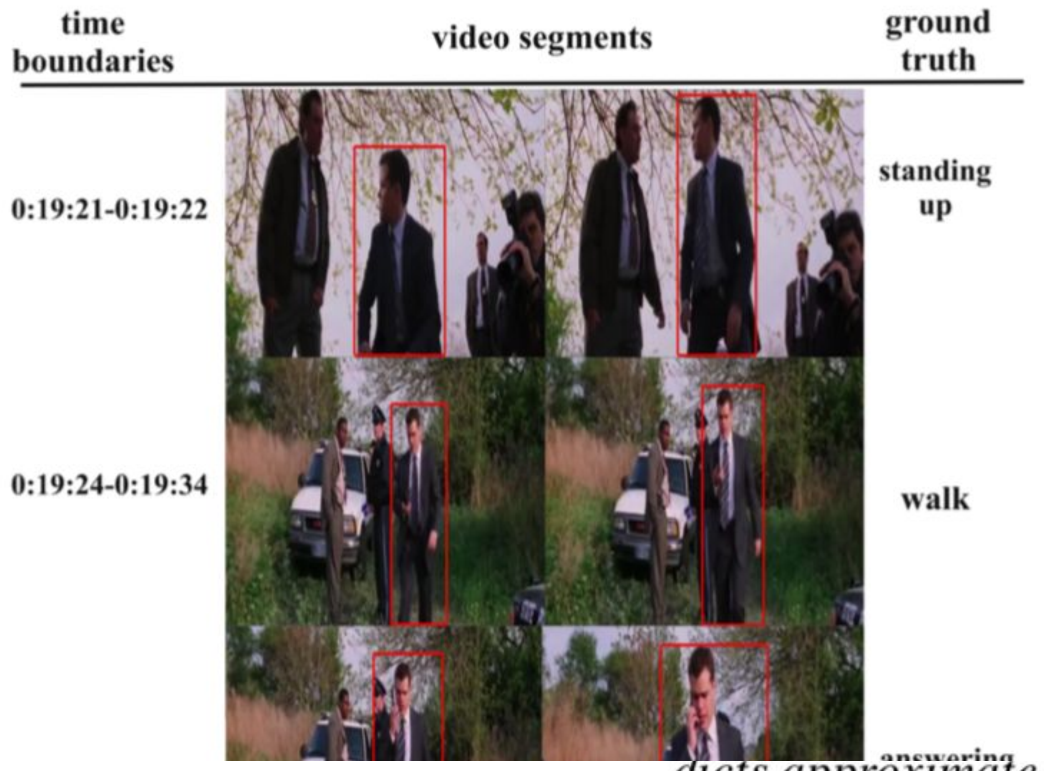
Multimodal Visual Concept Learning with Weakly Supervised Techniques

Giorgos Bouritsas, Petros Koutras, Athanasia Zlatintsi and Petros Maragos
School of E.C.E., National Technical University of Athens, Greece
gbouritsas@gmail.com, {pkoutras, nzlat, maragos}@cs.ntua.gr

[cs.CV] 4 Apr 2018

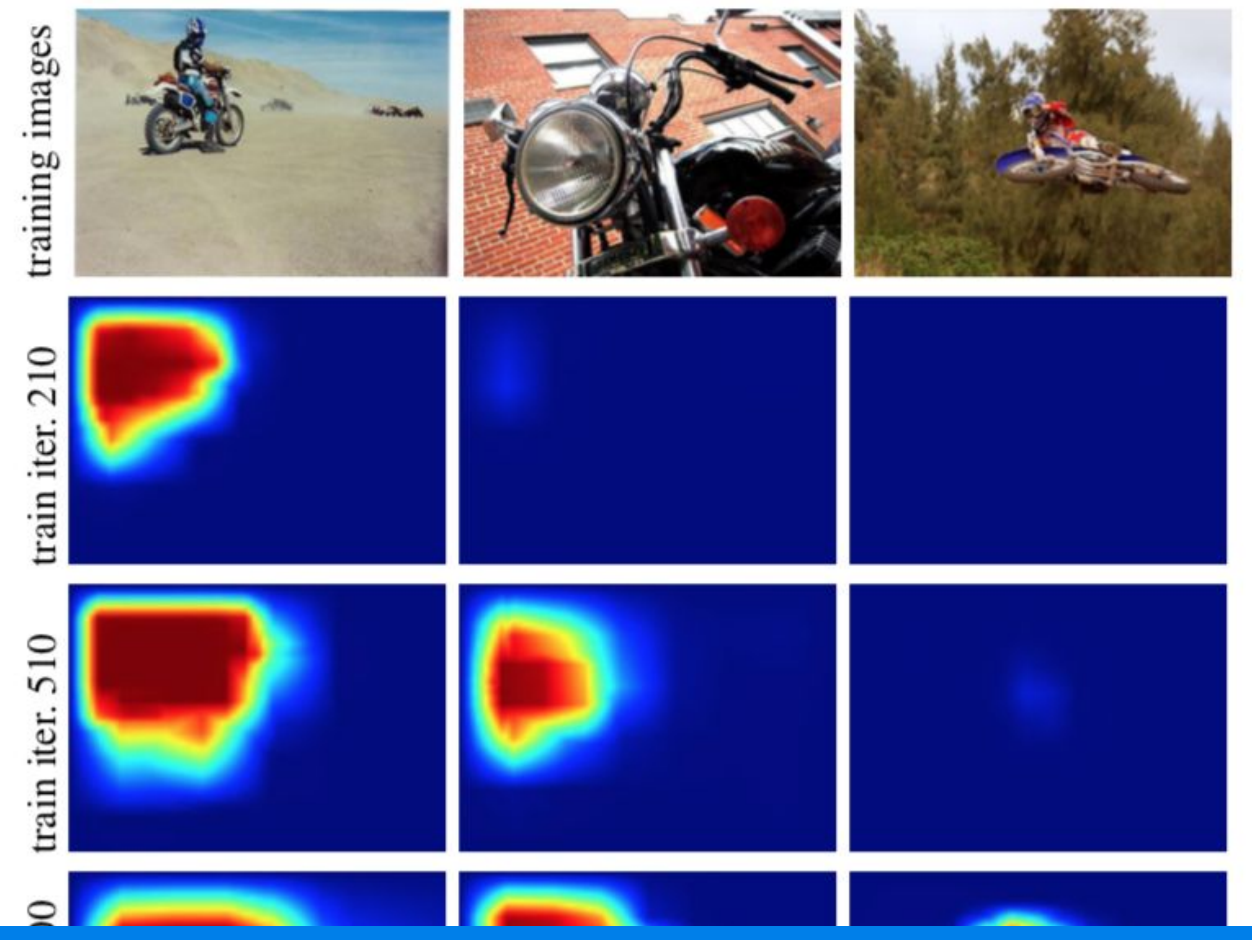
Abstract

Despite the availability of a huge amount of video data accompanied by descriptive texts, it is not always easy to exploit the information contained in natural language in order to automatically recognize video concepts. Towards this goal, in this paper we use textual cues as means of supervision, introducing two weakly supervised techniques that extend the Multiple Instance Learning (MIL) framework: the Fuzzy Sets Multiple Instance Learning (FSMIL) and the Probabilistic Labels Multiple Instance Learning (PLMIL). The former encodes the spatio-temporal imprecision of the linguistic descriptions with Fuzzy Sets, while the latter models different interpretations of each description's semantics

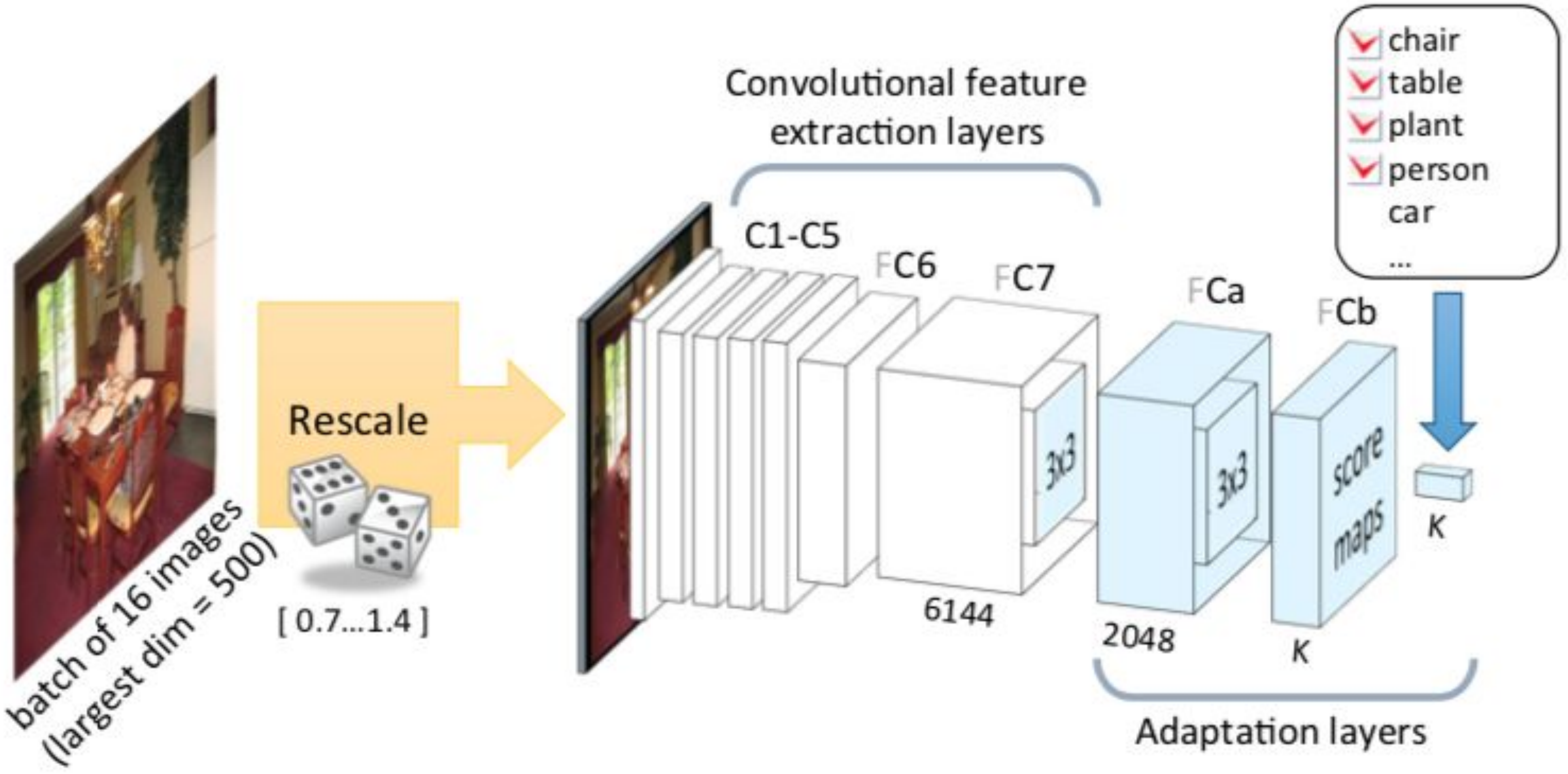


bstract

· visual object recognition typi- sets containing lots of richly an- image annotation, e.g. by object is both expensive and often sub- ·kly supervised convolutional neu- ·ect classification that relies only can learn from cluttered scenes ·s. We quantify its object classi- n prediction performance on the ·ect classes) and the much larger ·ct classes) datasets. We find that urate image-level labels, (ii) pre- dicts approximate locations (but not extents) of objects, and

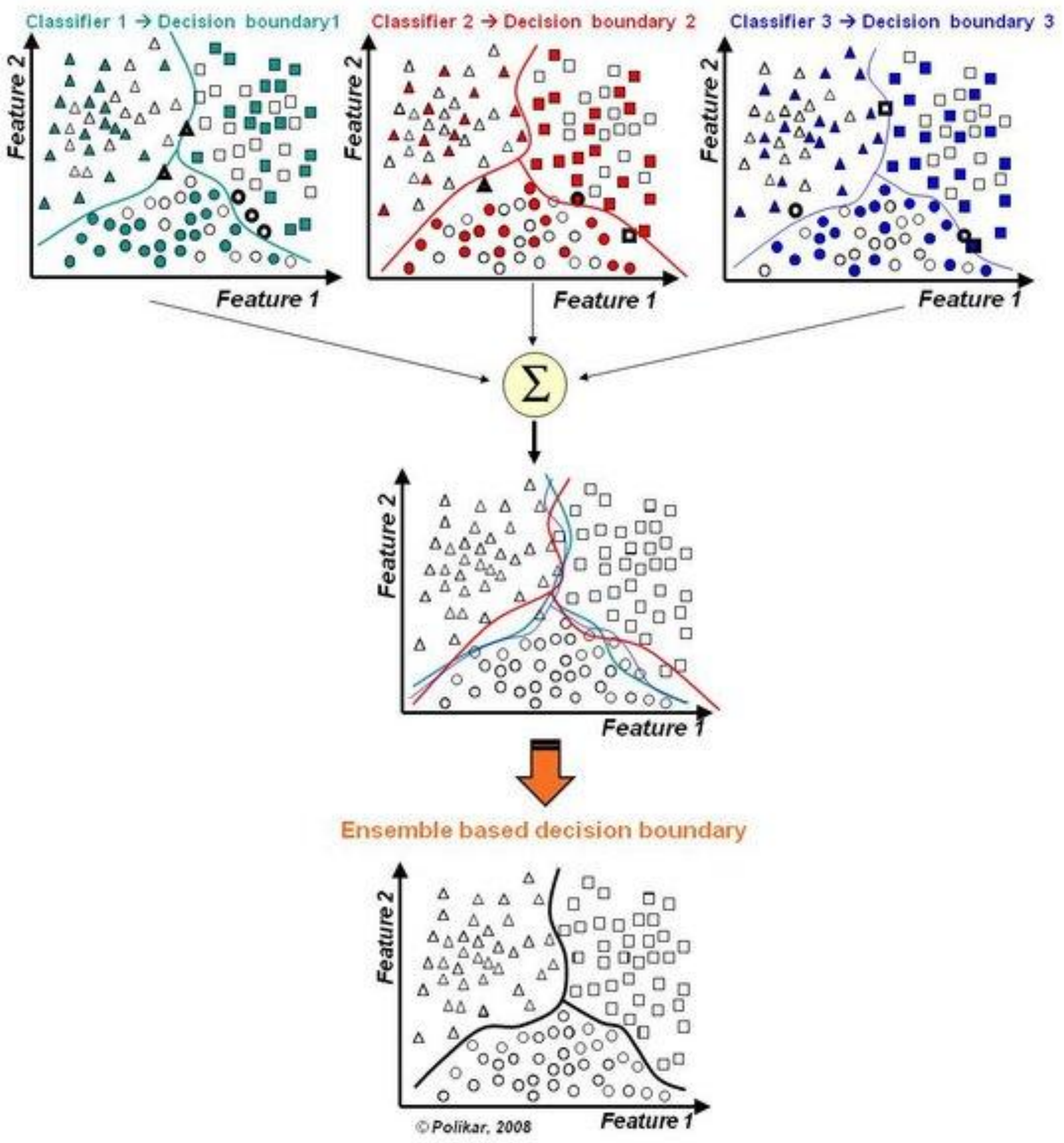


Inexact weak supervision



"Is object localization for free? – Weakly-supervised learning with convolutional neural networks." (CVPR2015)

Inaccurate weak supervision



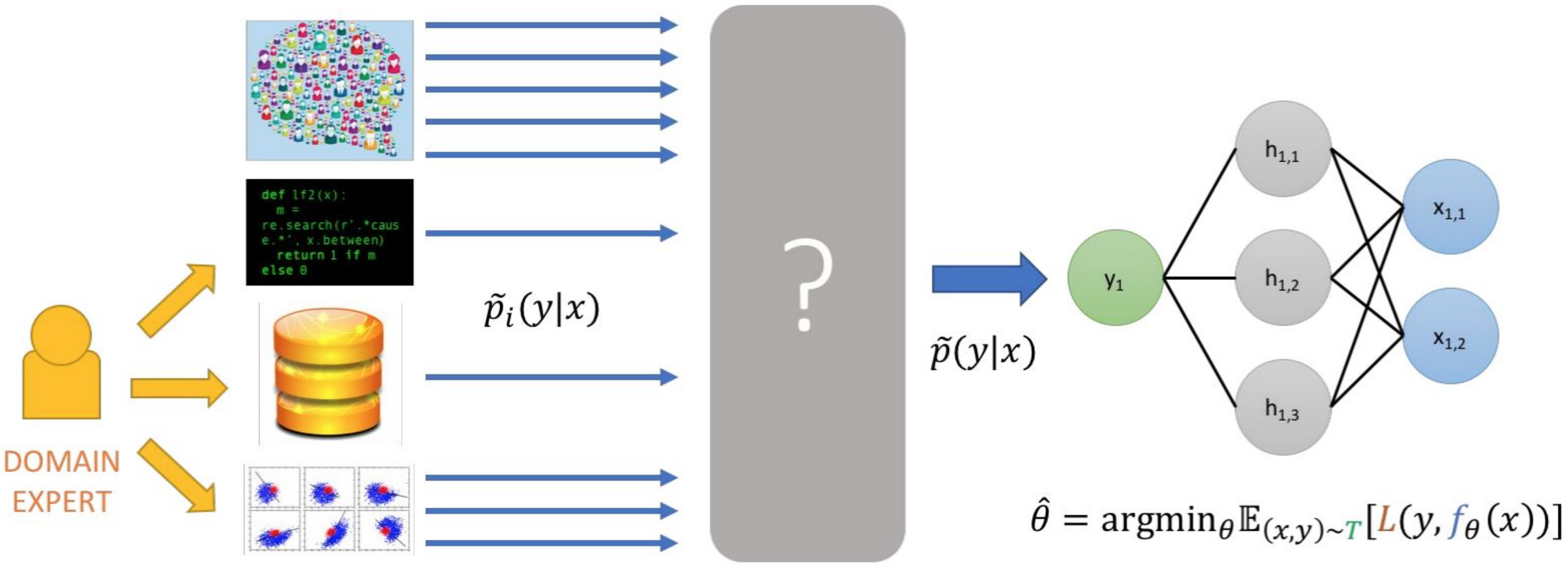
http://www.scholarpedia.org/article/Ensemble_learning

Snorkel: The System for Programmatically Building and Managing Training Data

Snorkel is a system for programmatically *building and managing* training datasets to rapidly and flexibly fuel machine learning models.

- Data Programming with DDLite: Putting Humans in a Different Part of the Loop (**June 2016**)
- Conversational agents at IBM: Bootstrapping Conversational Agents With Weak Supervision (AAAI 2019)
- Web content & event classification at Google: Snorkel DryBell: A Case Study in Deploying Weak Supervision at Industrial Scale (SIGMOD Industry 2019), and Google AI blog post
- Business intelligence at Intel: Osprey: Non-Programmer Weak Supervision of Imbalanced Extraction Problems (SIGMOD DEEM 2019)
- Anti-semitic tweet classification w/ Snorkel + transfer learning.
- Clinical text classification: A clinical text classification paradigm using weak supervision and deep representation (BMC MIDM 2019)
- Social media text mining: Deep Text Mining of Instagram Data without Strong Supervision (ICWI 2018)
- Cardiac MRI classification with Stanford Medicine: Weakly supervised classification of rare aortic valve malformations using unlabeled cardiac MRI sequences (BioArxiv 2018)
- Medical image triaging at Stanford Radiology: Cross-Modal Data Programming for Medical Images (NeurIPS ML4H 2017)
- GWAS KBC with Stanford Genomics: A Machine-Compiled Database of Genome-Wide Association Studies (NeurIPS ML4H 2016)

Weak Supervision Formulation

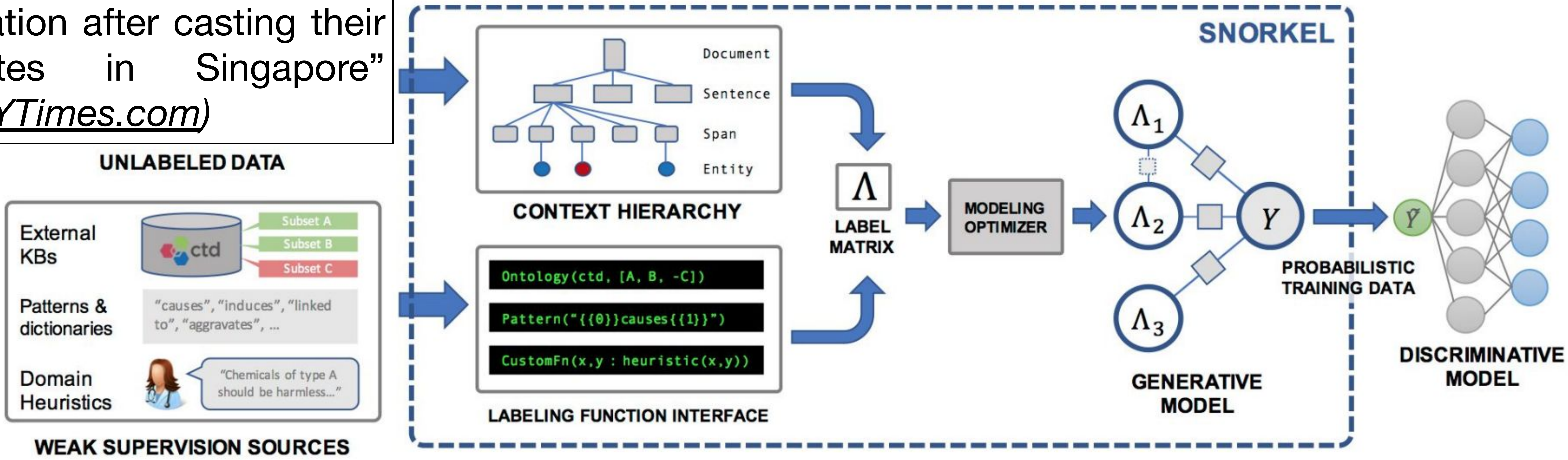


A high-level schematic of the basic weak supervision "pipeline": We start with one or more weak supervision sources: crowdsourced data, heuristic rules, distant supervision, and/or weak classifiers provided by a subject matter expert. The core technical challenge is to unify and model these sources. Then, this must be used to train the end model.

https://dawn.cs.stanford.edu/2017/07/16/weak-supervision/

Snorkel: data programming

“Prime Minister Lee Hsien Loong and his wife Ho Ching leave a polling station after casting their votes in Singapore” *(NYTimes.com)*



Demo: Step-By-Step Guide for Building a Brexit Tweet Classifier

https://github.com/HazyResearch/snorkel

https://github.com/HazyResearch/metal

Demo: Step-By-Step Guide for Building a Brexit Tweet Classifier

➔ Collecting unlabeled data: 3184

(tweets that contain #*Brexit*)

➔ Label 500 examples: 250 - ‘leave’, 250 - ‘stay’

➔ Create 5 LFs, apply on 2684 unlabeled tweets.

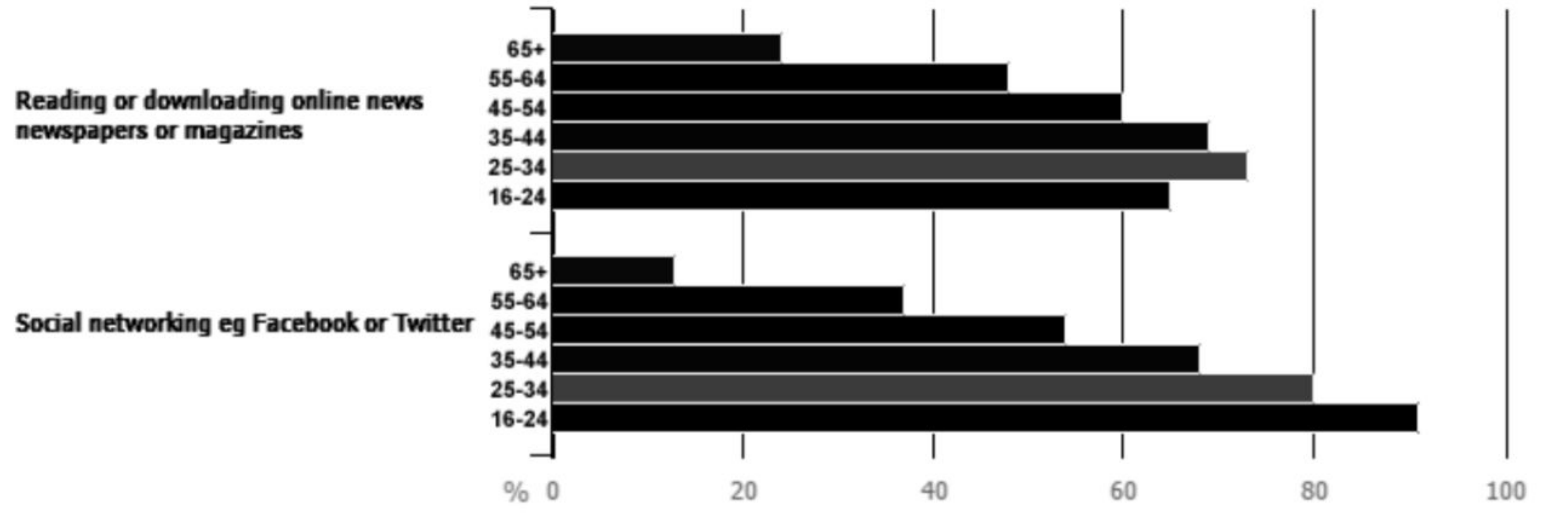


Figure 2: Percentage of online news readers and social network users by age. Source: ONS, year 2014.

Leave	Remain
euroscepticism, #beLeave, #betteroffout, #britainout, #LeaveEU, #noTTIP, #TakeControl, #VoteLeave, #VoteNO, #voteout, @end-of-europe, @leaveeuofficial, @NoThanksEU, @nothankseu, @ukleave-eu, @vote-leave	SayYes2Europe, Remain, #bremain, #betteroffin , #leadnotleave, #Remain, #Stay, #strongerin, #ukineu, #votein, #voteremain, #VoteYES, #yes2eu, #yestoeu, #SayYes2Europe,

Table 2: Sets of keywords, hashtags and mentions for assigning posts to Leave and Remain categories.

“Predicting Brexit:Classifying Agreement is Better than Sentiment and Pollsters”

Demo: Step-By-Step Guide for Building a Brexit Tweet Classifier

Safer In #EU? No! No! No! Terrorists want the UK to STAY Remember 7/7 Paris #EUreferendum #VoteLeave

#Liverpool have broke the #Spanish dominance in Europe... #English #football says Yes We Belong in #Europe! #Stay #strongerin

Tweet

```
COMMON_HASHTAG_VOTE_LEAVE = r"(?i)VoteLeave|VoteNO|VoteOUT"

# def function
def most_common_hashtag_leave(tweet_text):
    return 'leave' if re.search(COMMON_HASHTAG_VOTE_LEAVE, tweet_text) else 0

COMMON_HASHTAG_VOTE_STAY = r"(?i)StrongerIN|VoteYES|VoteIN"

# def function
def most_common_hashtag_stay(tweet_text):
    return 'stay' if re.search(COMMON_HASHTAG_VOTE_STAY, tweet_text) else 0
```

Label functions

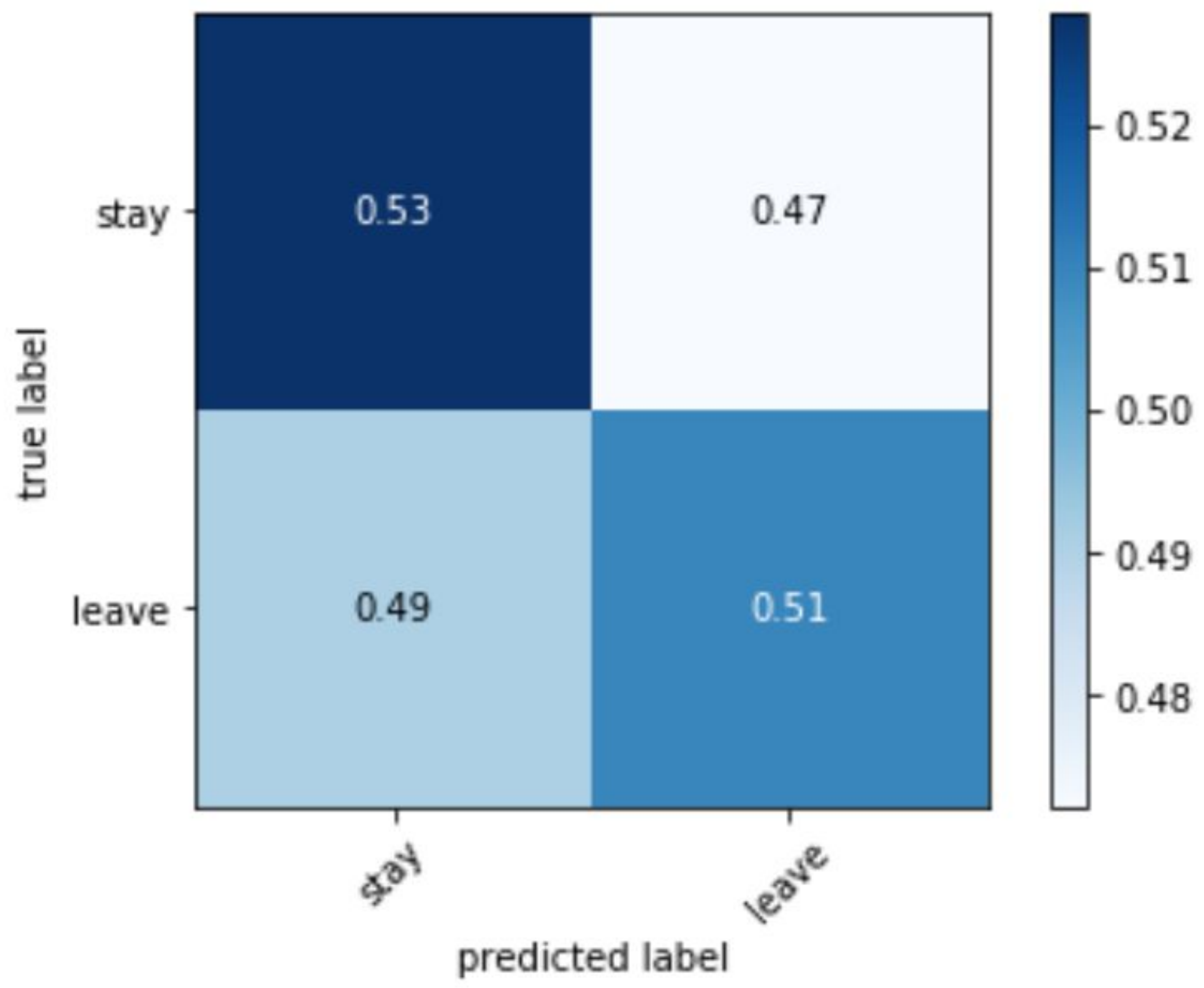
@StrongerIn so if we stay in eu that means we get more zero hours contracts and employers can say 'we dont need to now, fuck off' 🇬🇧 #TakeControl #VoteLeave

"Predicting Brexit:Classifying Agreement is Better than Sentiment and Pollsters"

Result: Brexit Tweet Classifier

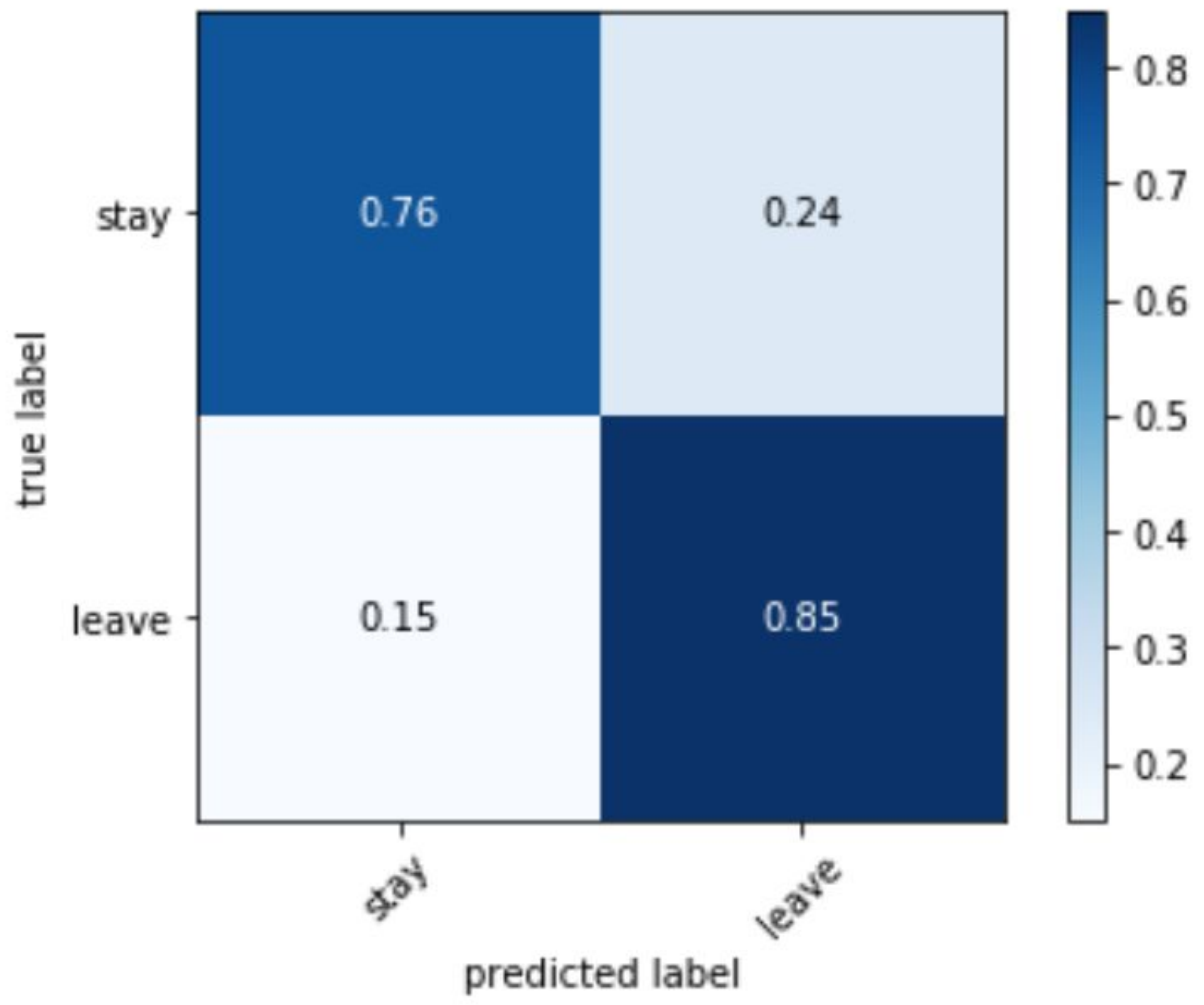
Tweet Classifier on 500 labeled examples

LR ACCURACY: 0.52



Tweet Classifier with Snorkel

LR ACCURACY: 0.78



Summary

- ➢ Weak supervision
  - incomplete
  - inexact
  - inaccurate
- ➢ Snorkel and Snorkel metal
- ➢ Demo application: *Brexit Tweet Classifier*