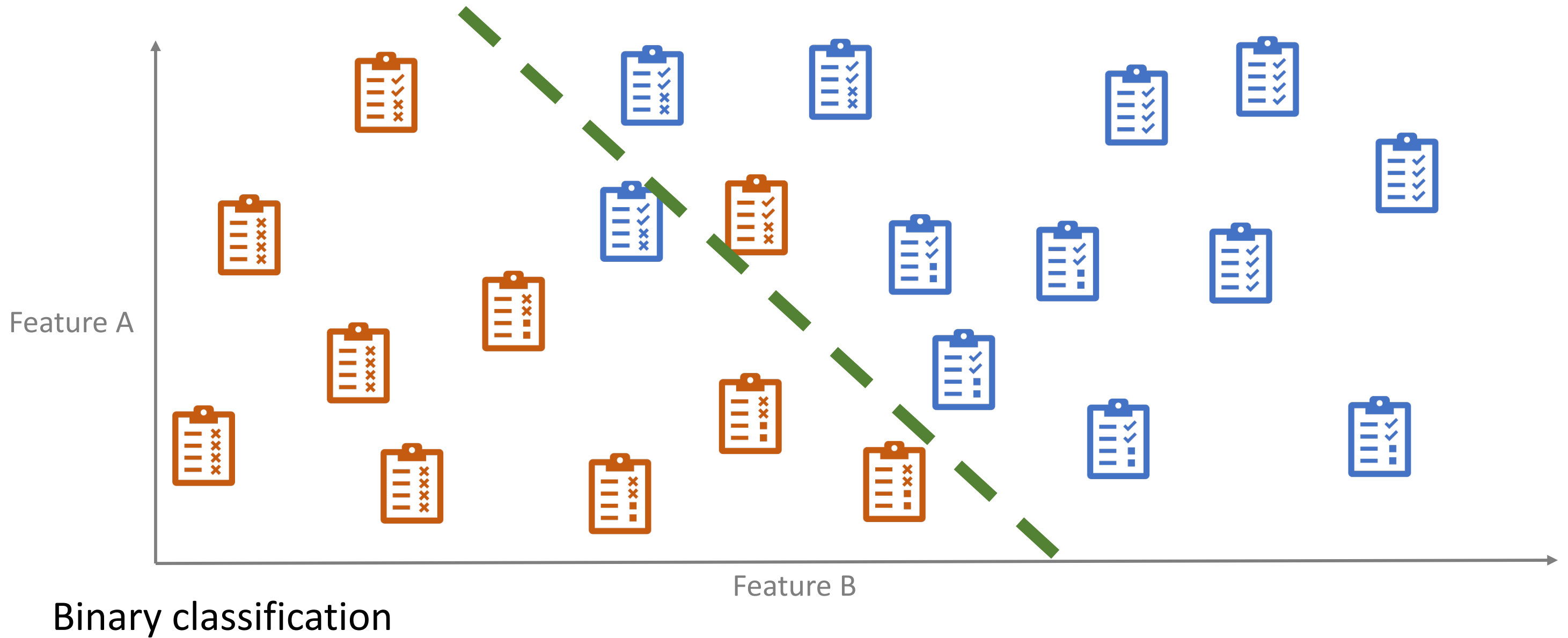


Learning from positive and
unlabeled data

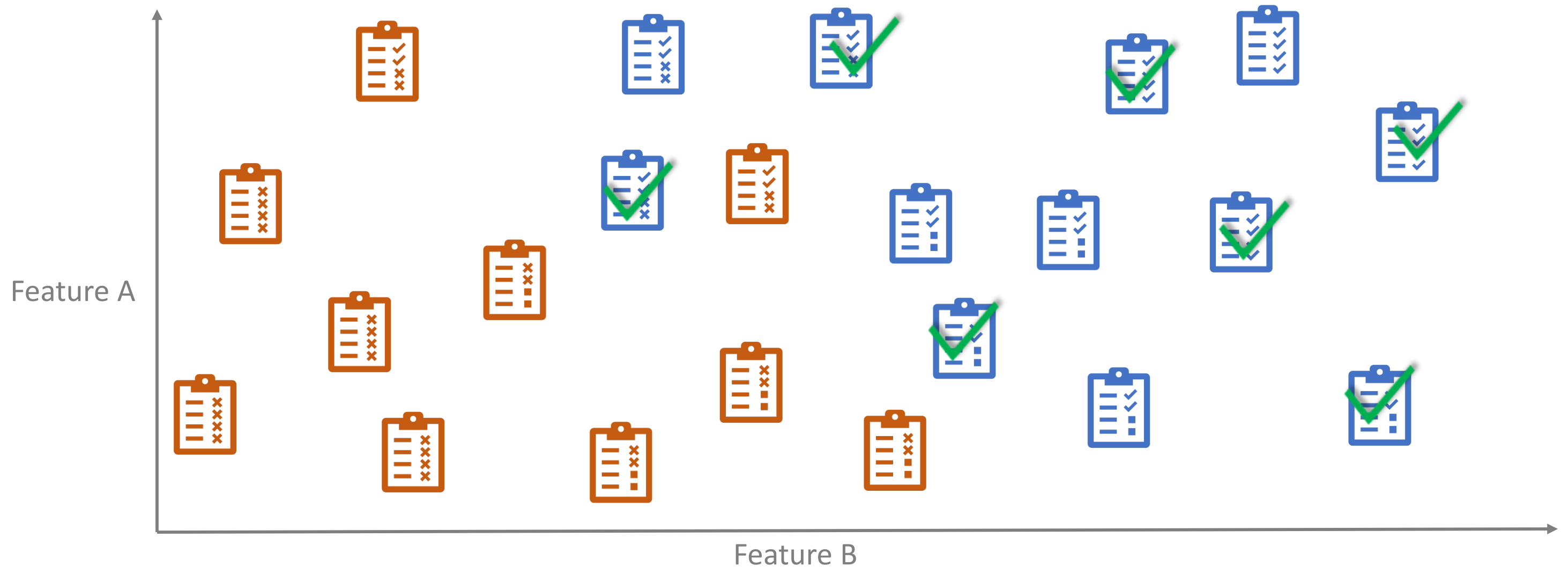
1. PU Learning and its sources

Section 7.1 in the survey paper

Learning from positive and unlabeled data



Learning from positive and unlabeled data



In PU data: only a subset of the positive examples are labeled

Learning from positive and unlabeled data

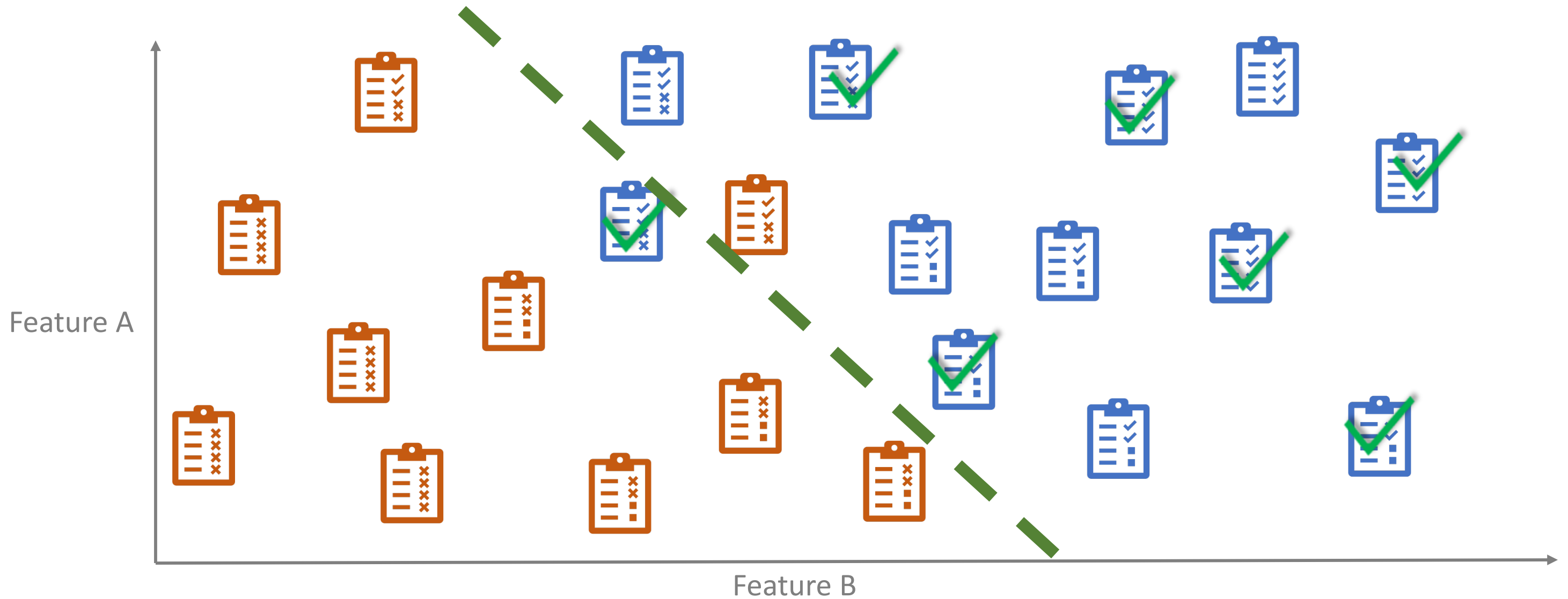


Learning from positive and unlabeled data



Now it is less clear what the decision boundary should be

Learning from positive and unlabeled data



Goal of PU Learning: Learn a good classifier from PU data

10 Sources of PU data

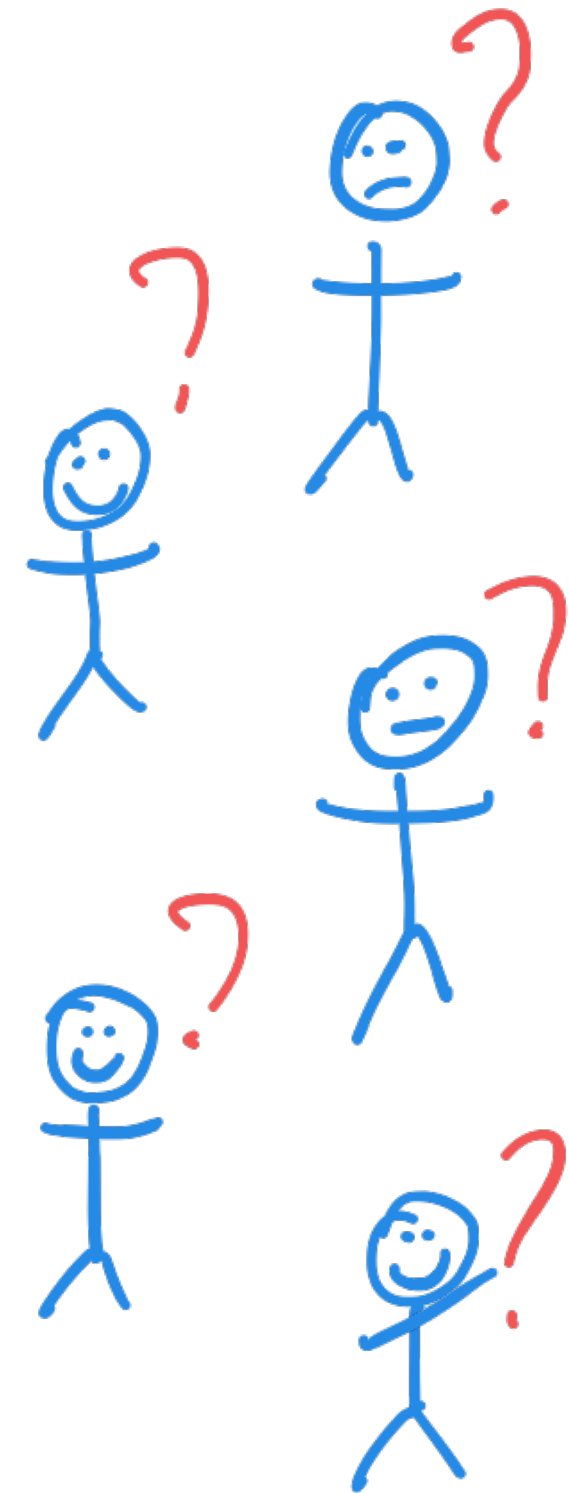
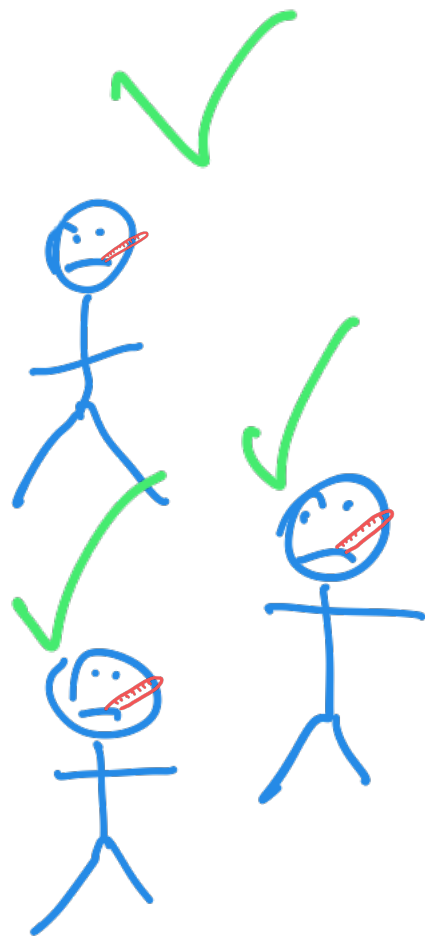
Why do we care about learning from positive and unlabeled data?

→ Because it naturally arises in many applications

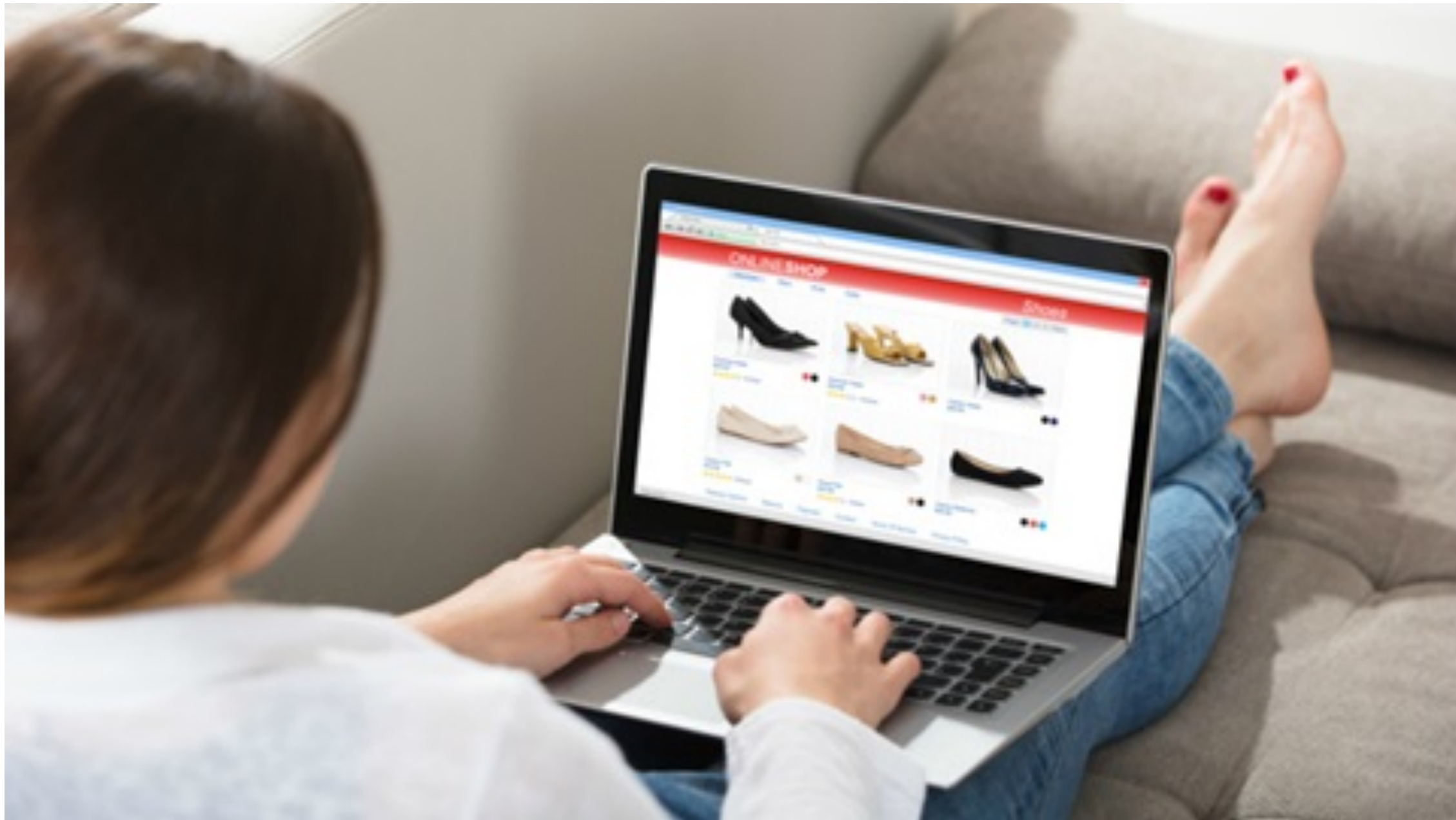
DISCLAIMER:

- This is not a complete list of sources
- The presented sources are not always strictly different.

1. *Automatic diagnosis*



2. Positive examples are easier to obtain



3. Indirect labels

KU LEUVEN 2020-2021

	Name: Jessa Bekker
	Nb: 0987654321
	Sports: Yes

Active!

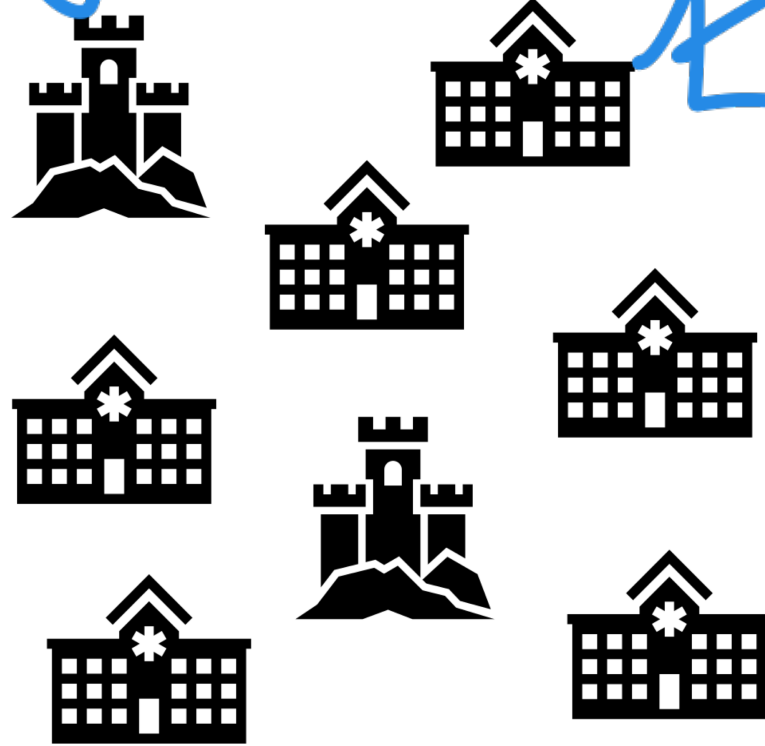
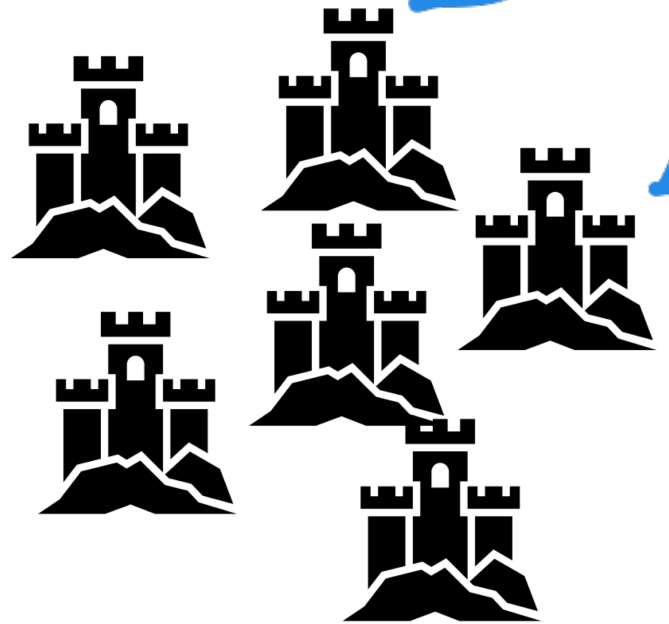
KU LEUVEN 2020-2021

	Name: Jesse Davis
	Nb: 0123456789
	Sports: No

???

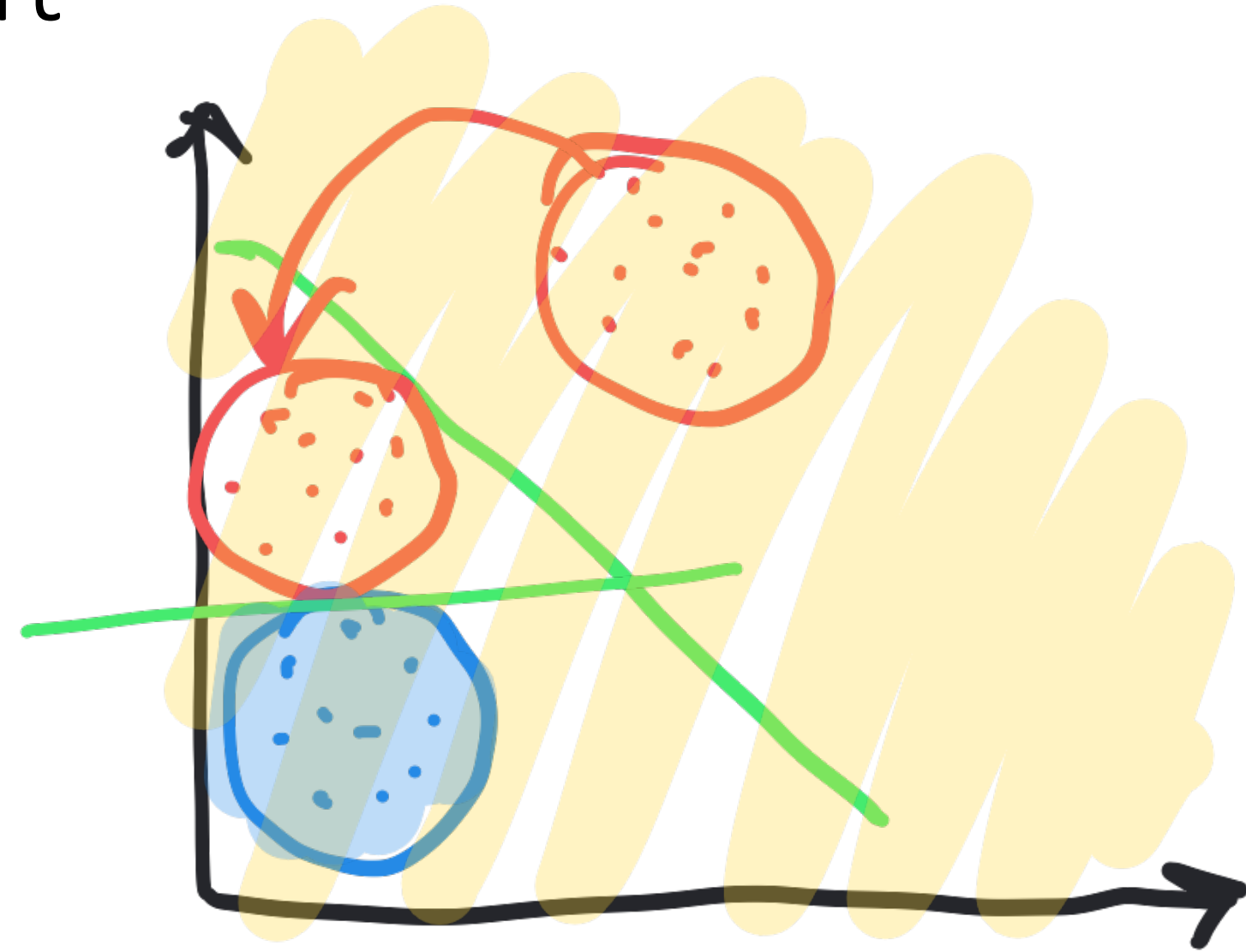
4. Case-control

P



U

5. Negative-class datashift



6. Under-reporting

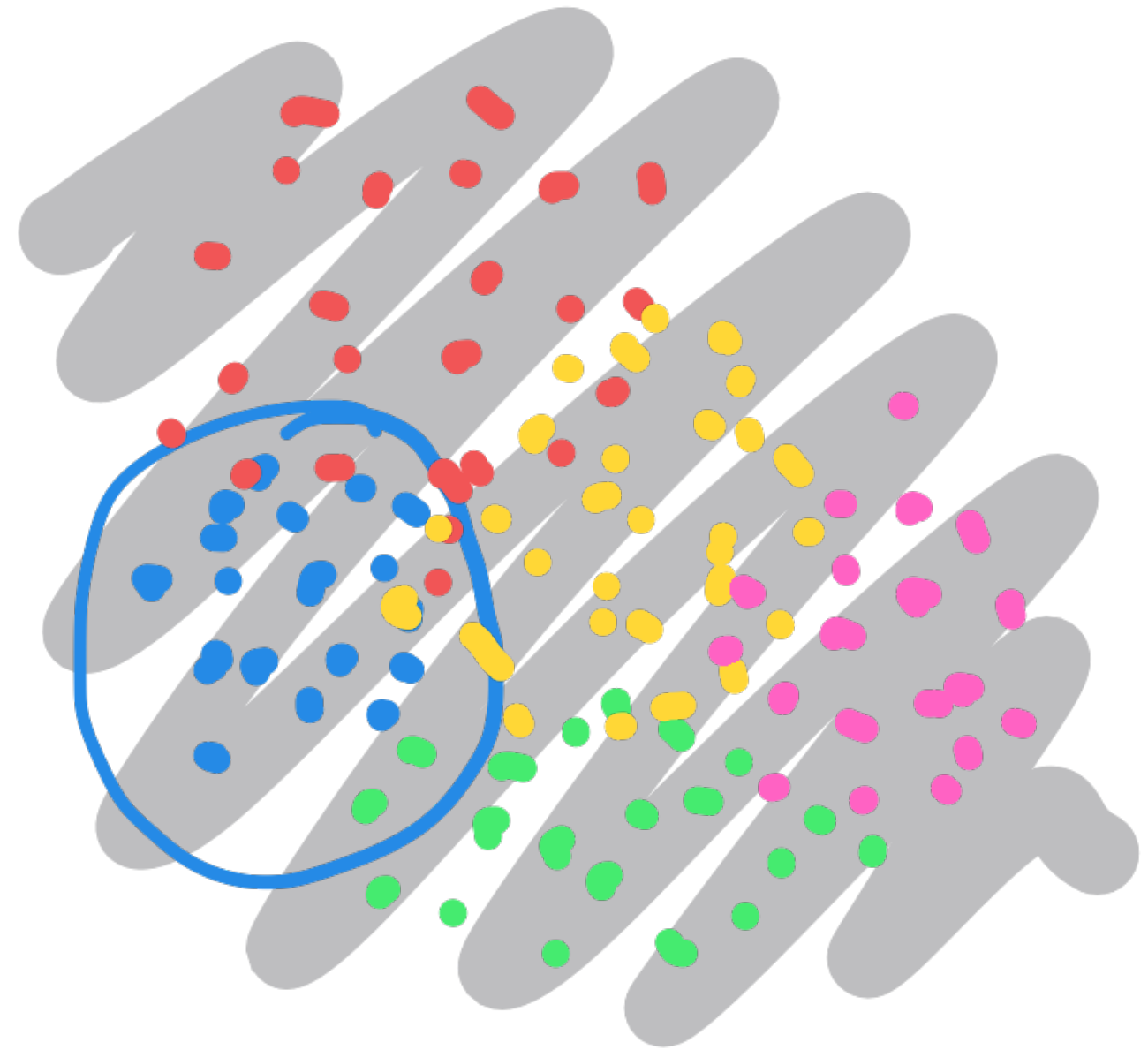


- [1] Sechidis et al. Dealing with under-reported variables: An information theoretic solution. *International Journal of Approximate Reasoning*. 2017
- [2] Gorber et al. The accuracy of self-reported smoking: A systematic review of the relationship between self-reported and cotinine- assessed smoking status. *Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco*. 2019

7. One-class classification

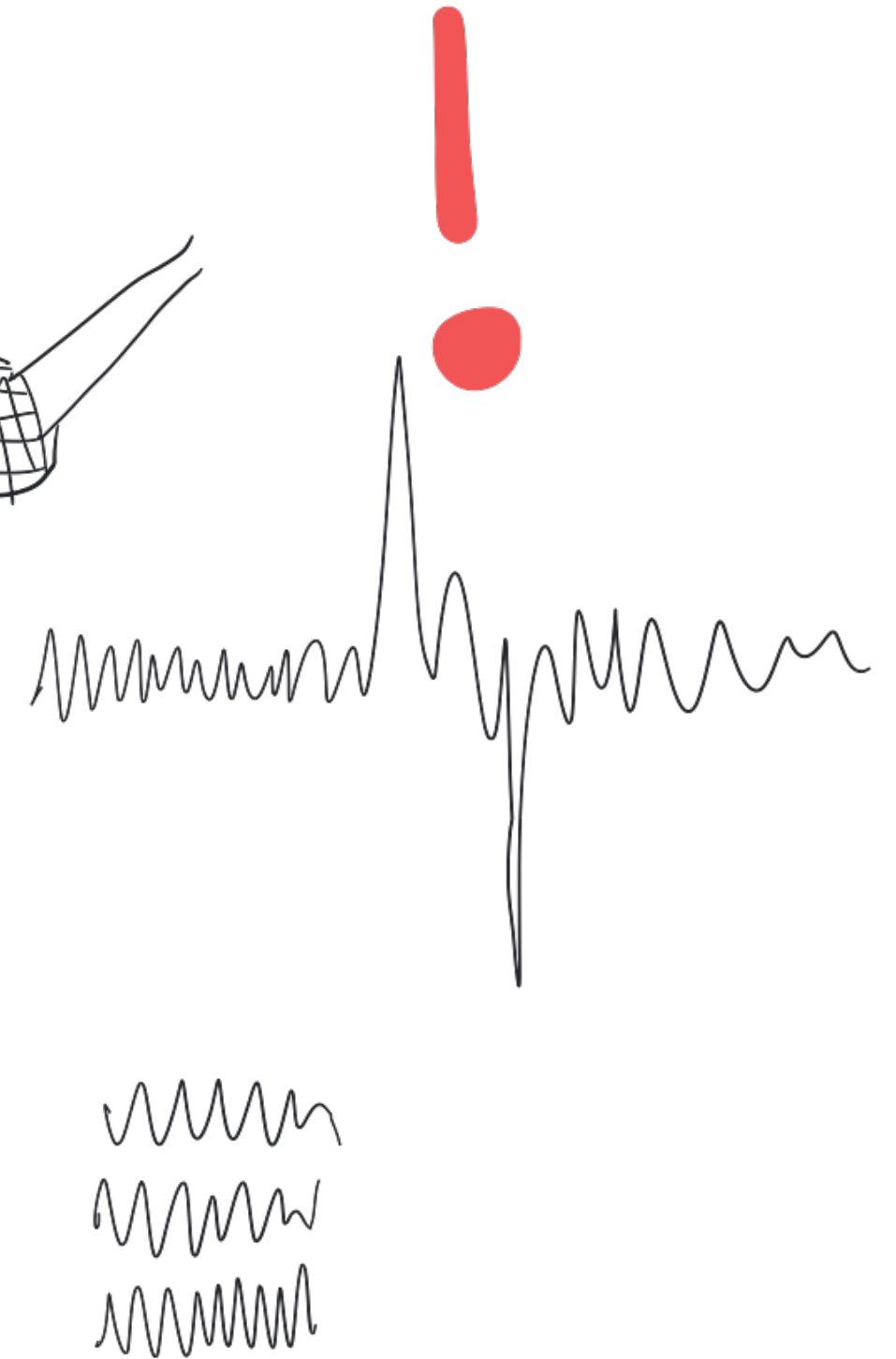
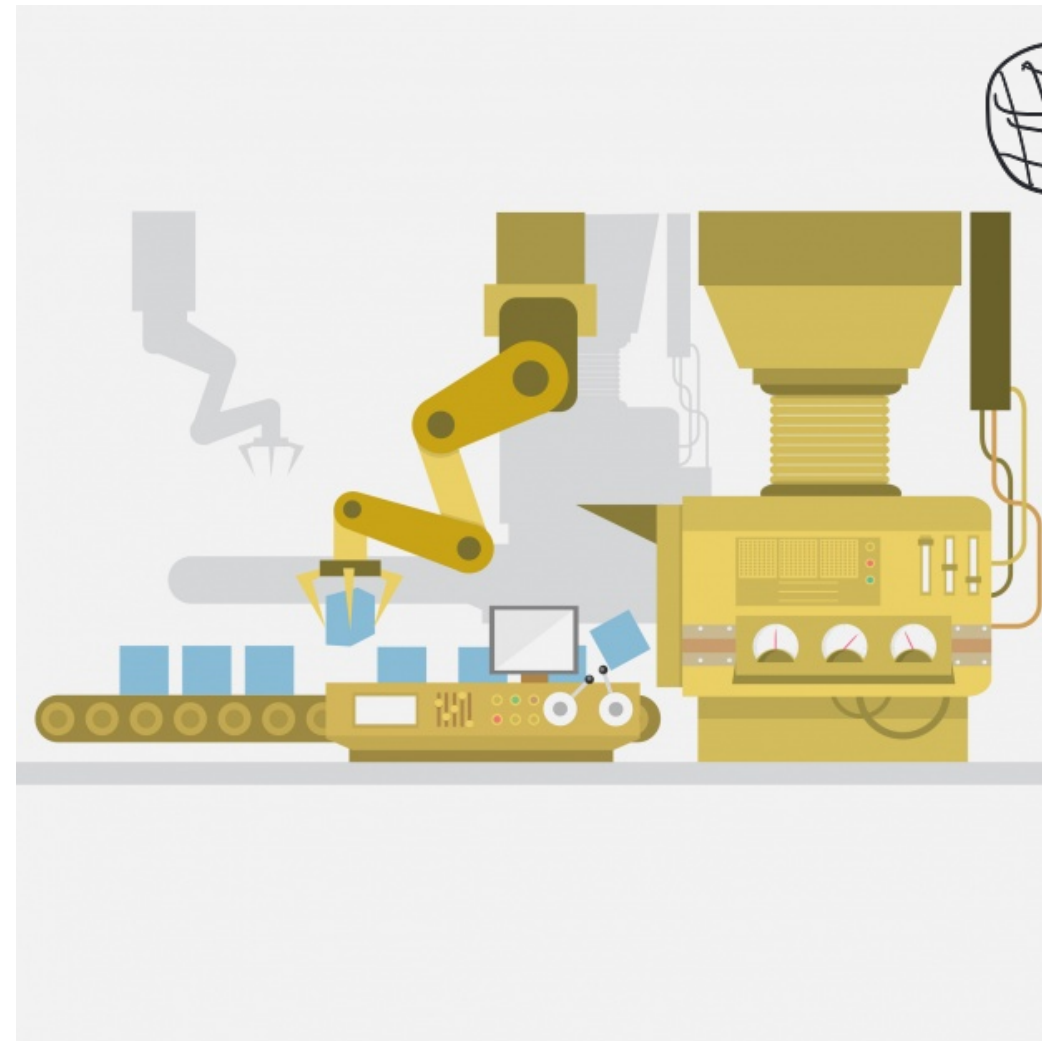


Picture from [1]



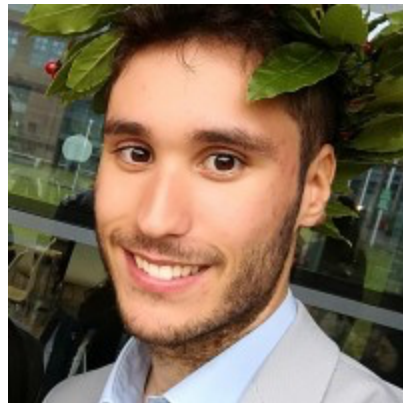
[1] Li et al. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. IEEE Transactions on Geoscience and Remote Sensing. 2011

8. Inlier-based outlier detection



- [1] Hido et al. Inlier-based outlier detection via direct density ratio estimation. In 2008 Eighth IEEE international conference on data mining. 2008
- [2] Smola et al. Relative novelty detection. IJCAI. 2009
- [3] Blanchard et al. Semi-supervised novelty detection. JMLR. 2010

9. Knowledge base completion



affiliated



Machine Learning (2020) 109:719–760
<https://doi.org/10.1007/s10994-020-05877-5>

Learning from positive and unlabeled data: a survey
Jessa Bekker¹ · Jesse Davis¹

Received: 12 November 2018 / Revised: 18 December 2019 / Accepted: 1 March 2020 /
Published online: 2 April 2020
© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2020

Abstract
Learning from positive and unlabeled data or PU learning is the setting where a learner only has access to positive examples and unlabeled data. The assumption is that the unlabeled data can contain both positive and negative examples. This setting has attracted increasing interest within the machine learning literature as this type of data naturally arises in applications such as medical diagnosis and knowledge base completion. This article provides a survey of the current state of the art in PU learning. It proposes seven key research questions that commonly arise in this field and provides a broad overview of how



author

Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data

Jessa Bekker^{{0000-0003-1928-7374} (✉)}, Pieter Robberechts<sup>{0000-0002-3734-0047}, and Jesse Davis^{0000-0002-3748-9263}
Dept of Computer Science, KU Leuven, Belgium
first.lastname@cs.kuleuven.be</sup>

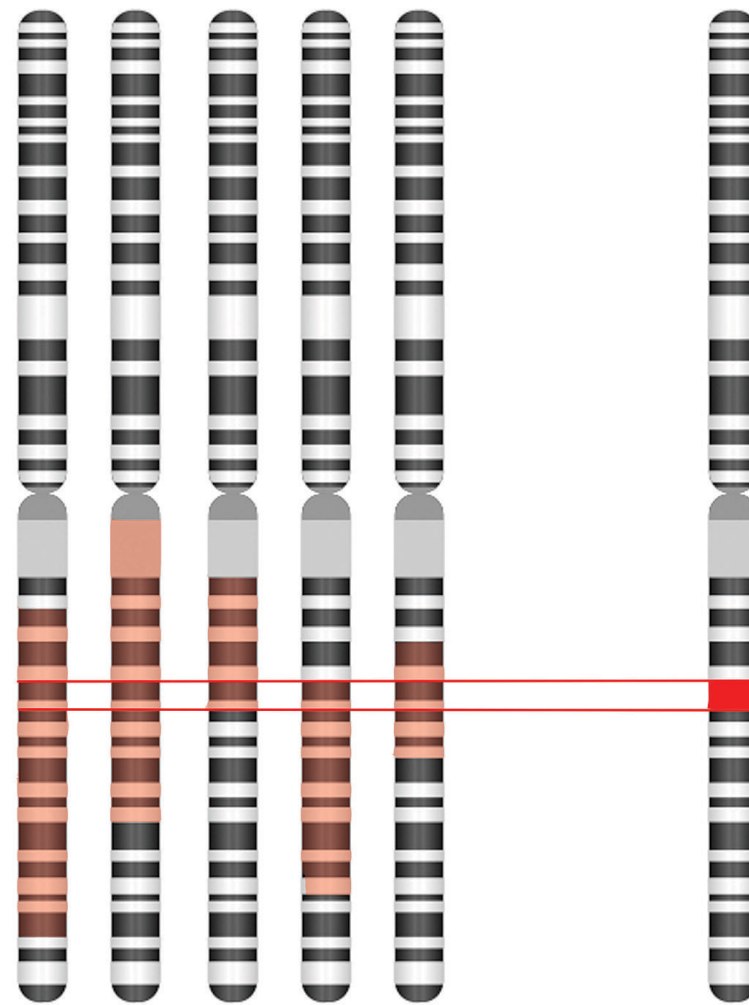
Abstract. Most positive and unlabeled data is subject to selection biases. The labeled examples can, for example, be selected from the positive set because they are easier to obtain or more obviously positive. This paper investigates how learning can be enabled in this setting. We propose and theoretically analyze an empirical-risk-based method for incorporating the labeling mechanism. Additionally, we investigate under which assumptions learning is possible when the labeling mechanism is not fully understood and propose a practical method to enable this. Our

Class Prior Estimation in Active Positive and Unlabeled Learning
Lorenzo Perini, Vincent Vercryssen and Jesse Davis
DTAI Research group, KU Leuven, Belgium
lorenzo.perini@kuleuven.be, vincent.vercryssen@kuleuven.be, jesse.davis@kuleuven.be

Abstract
Estimating the proportion of positive examples (i.e., the class prior) from positive and unlabeled (PU) data is an important task that facilitates learning a classifier from such data. In this paper, we explore how to tackle this problem when the observed labels were acquired via active learning. This introduces the challenge that the observed labels were not selected completely at random, which is the process to only unlabeled data, and positive labels are gradually acquired using an active learning strategy. In anomaly detection applications, for instance, one often starts with a completely unlabeled dataset and labels are acquired via active learning because the labeling process is costly [Vercryssen et al., 2018]. Because anomalies are rare events and not well-understood, the user almost always ends up labeling only normal examples (i.e., examples from the positive class). Another class of problems where active learning would only return positive labels arises when measuring the interestingness

[1] Galárraga et al. Fast rule mining in ontological knowledge bases with AMIE+. The International Journal on Very Large Data Bases. 2015
[2] Neelakantan et al. Compositional vector space models for knowledge base completion. ACL | IJCNLP. 2015

10. Identification



Up next...