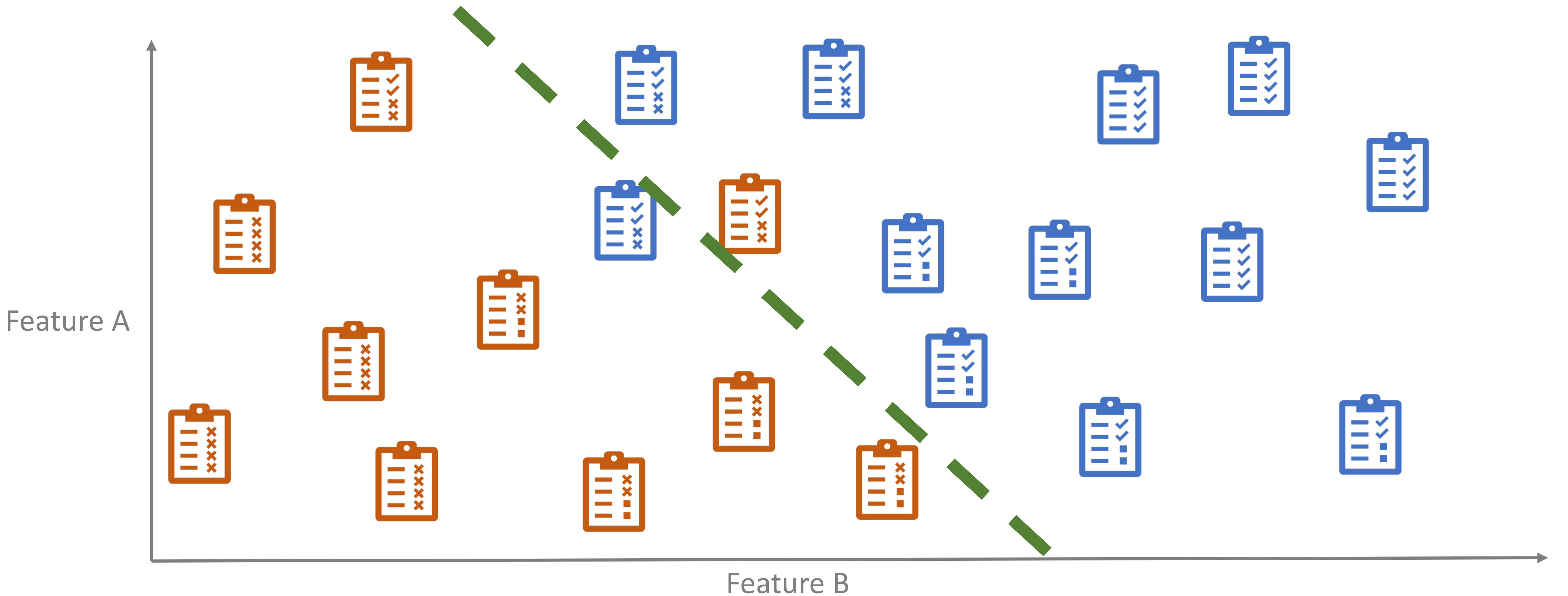
Learning from positive and unlabeled data

1. PU Learning and its sources

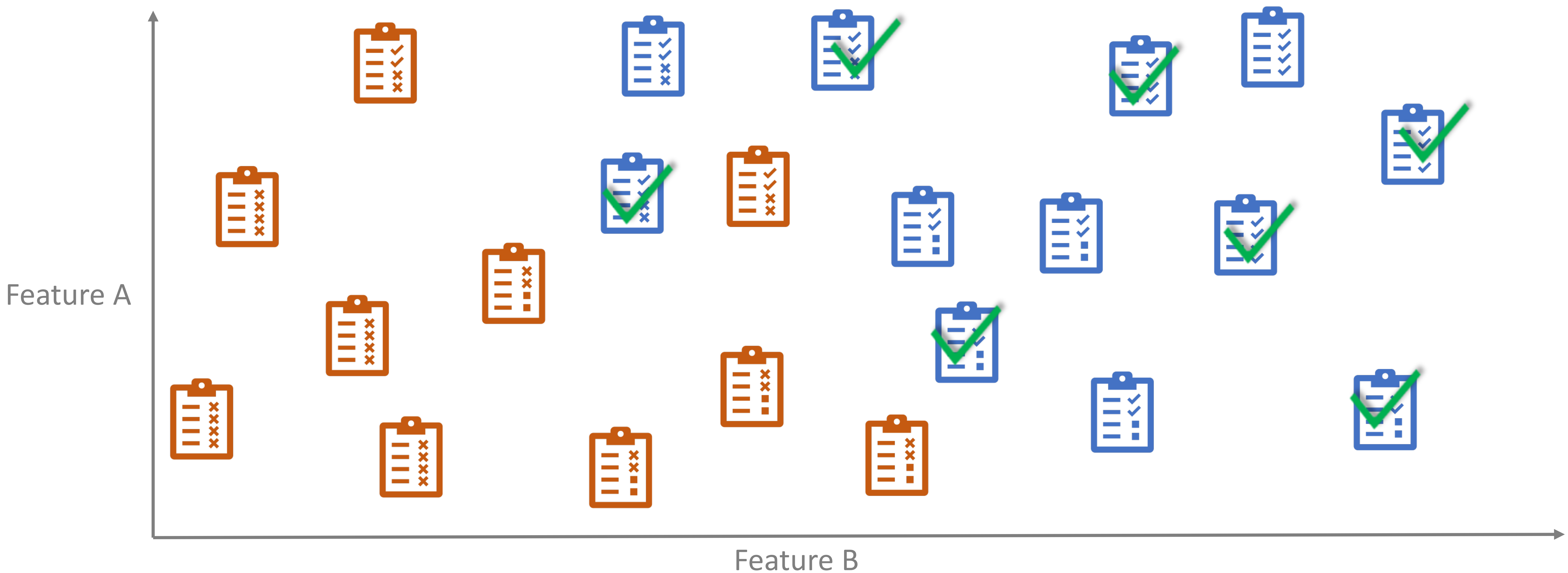
Section 7.1 in the survey paper

Learning from positive and unlabeled data



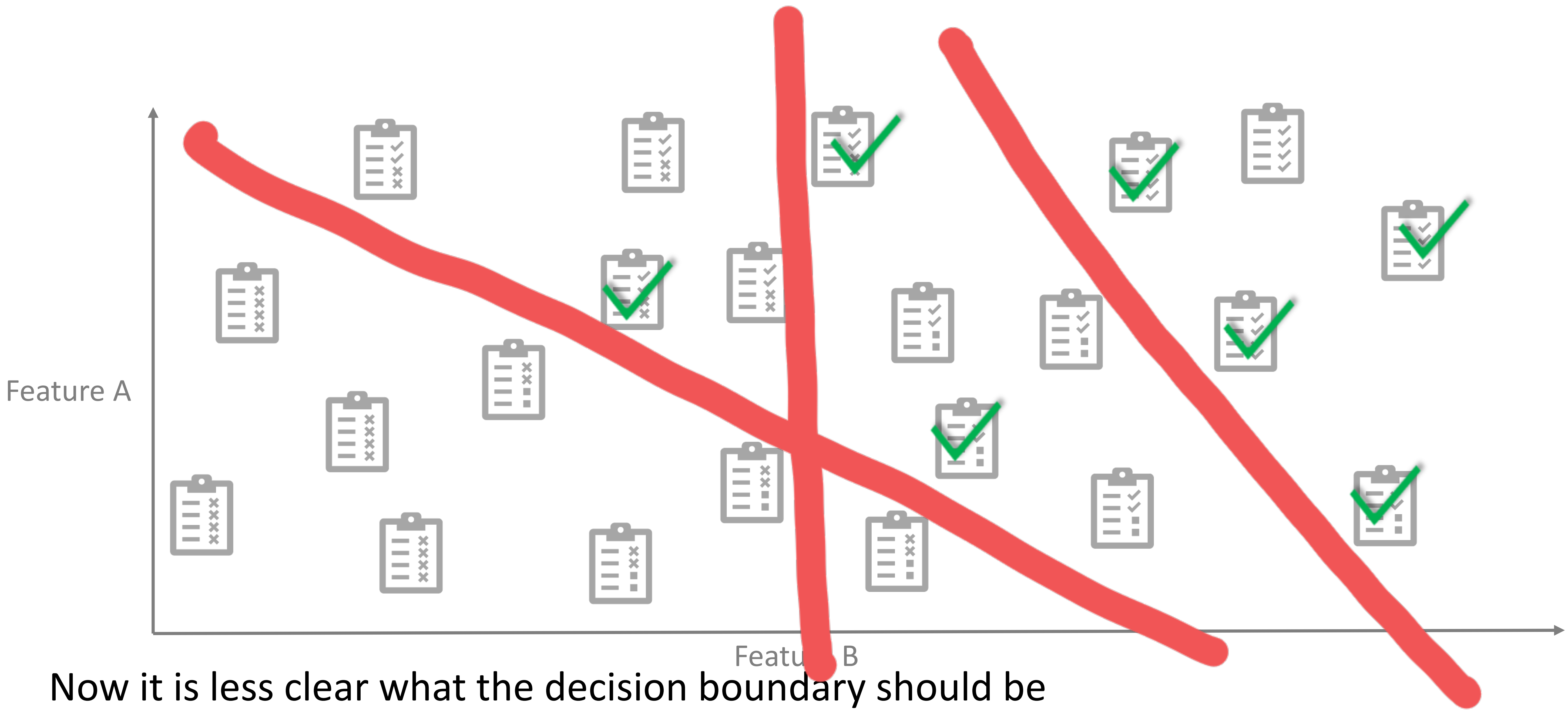
Binary classification

Learning from positive and unlabeled data



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

Learning from positive and unlabeled data

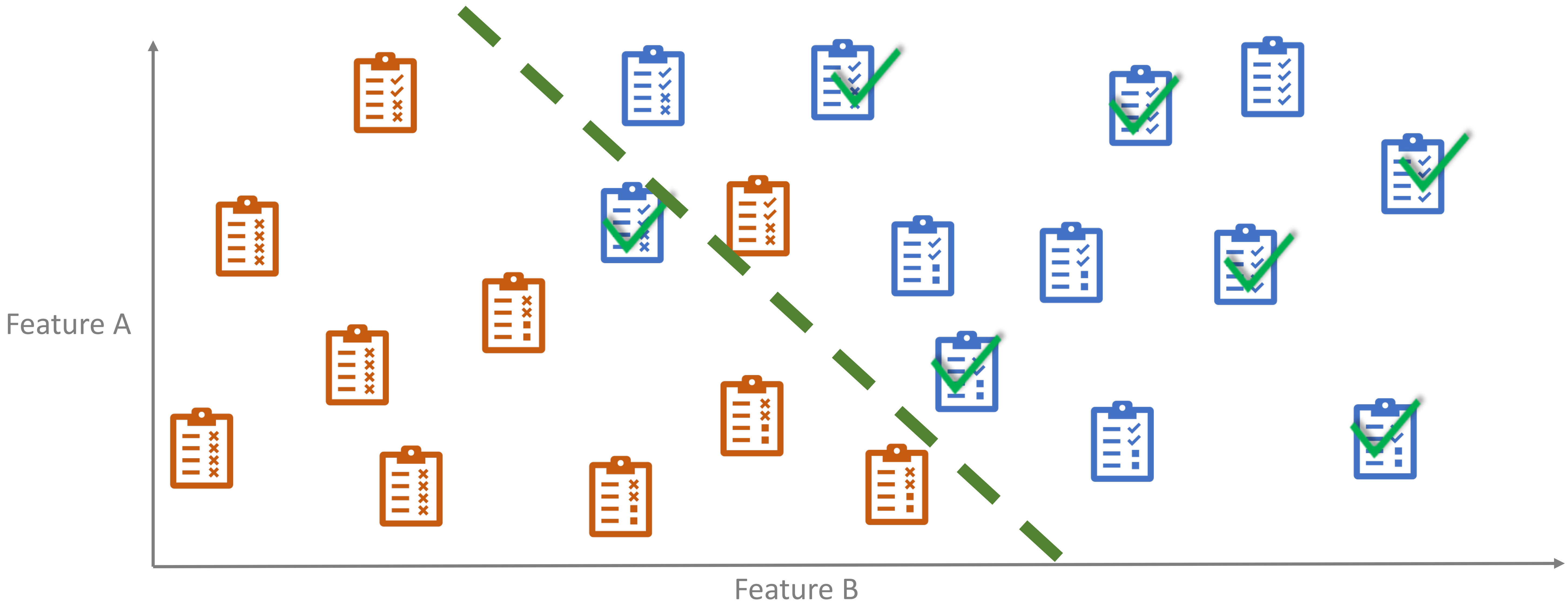


Now it is less clear what the decision boundary should be

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

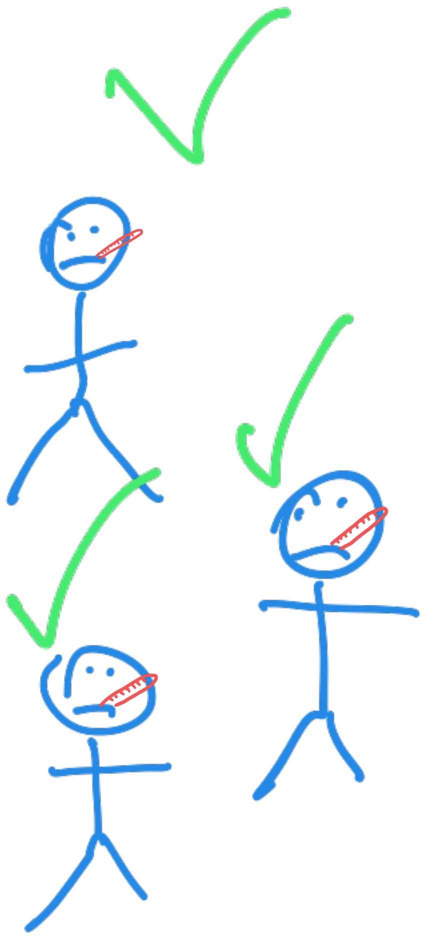
Why do we care about learning from positive and unlabeled data?

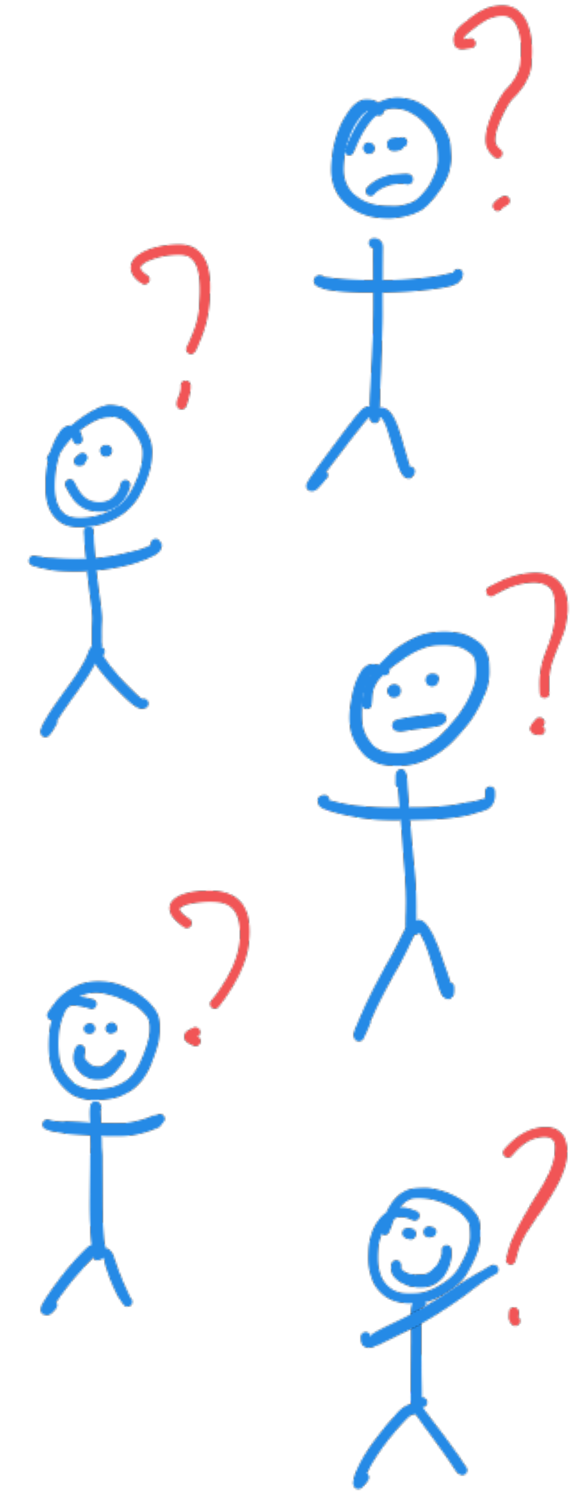
→ Because it naturally arises in many applications

DISCLAIMER:

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

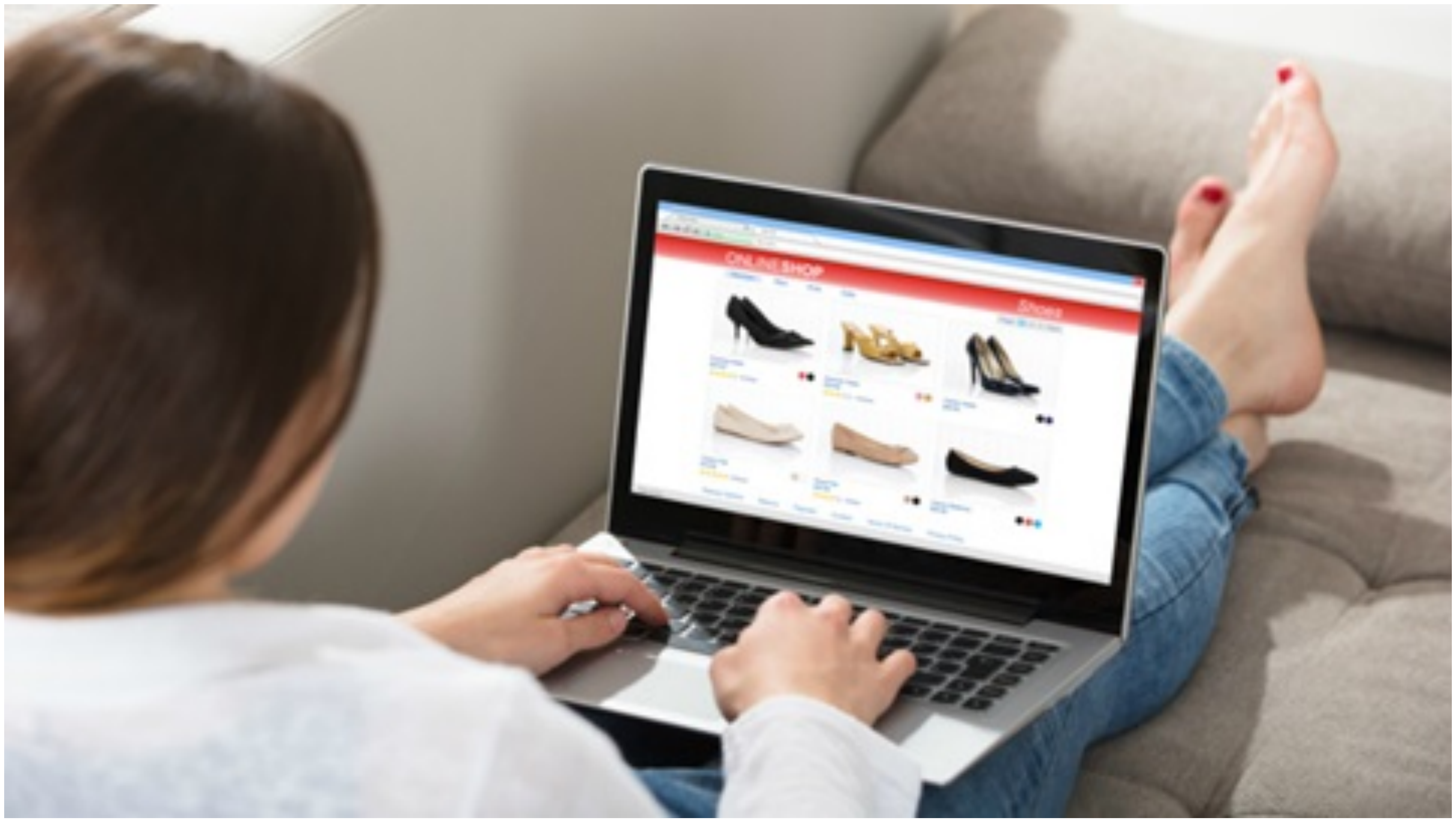
1. *Automatic diagnosis*





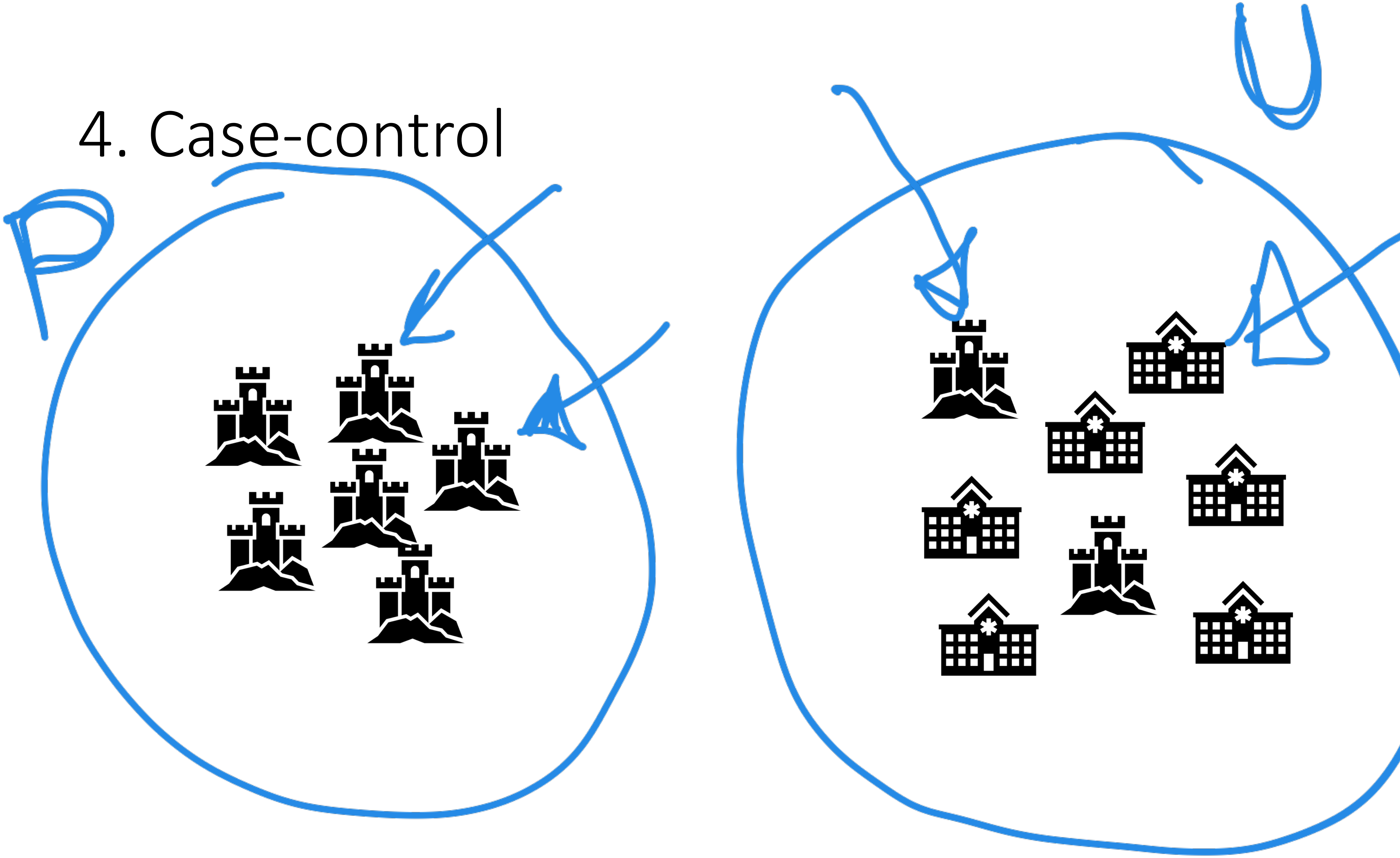
[1] Claesen et al. Building classifiers to predict the start of glucose-lowering pharmacotherapy using Belgian health expenditure data. 2015

2. Positive examples are easier to obtain

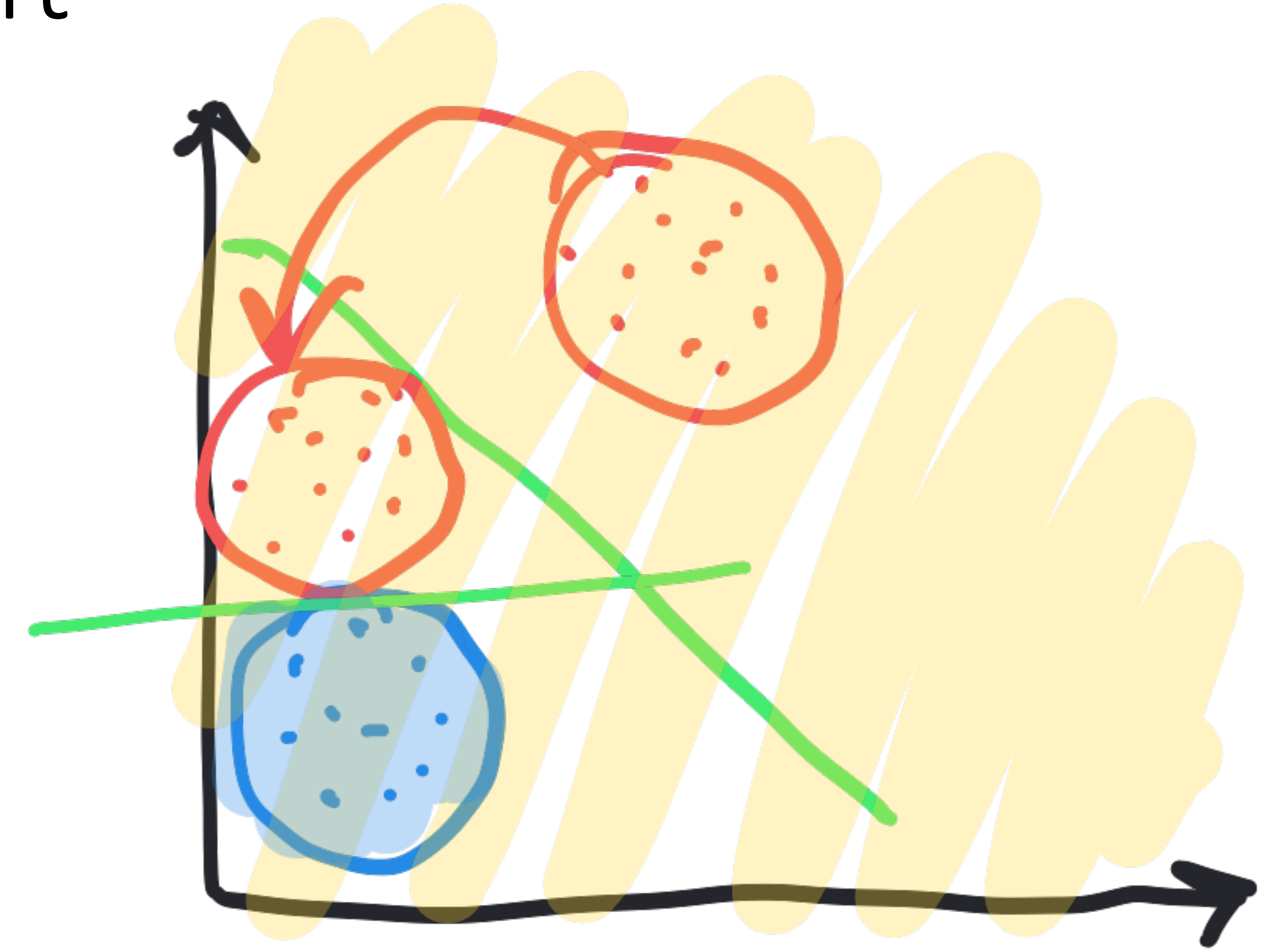


3. Indirect labels

4. Case-control



5. Negative-class datashift

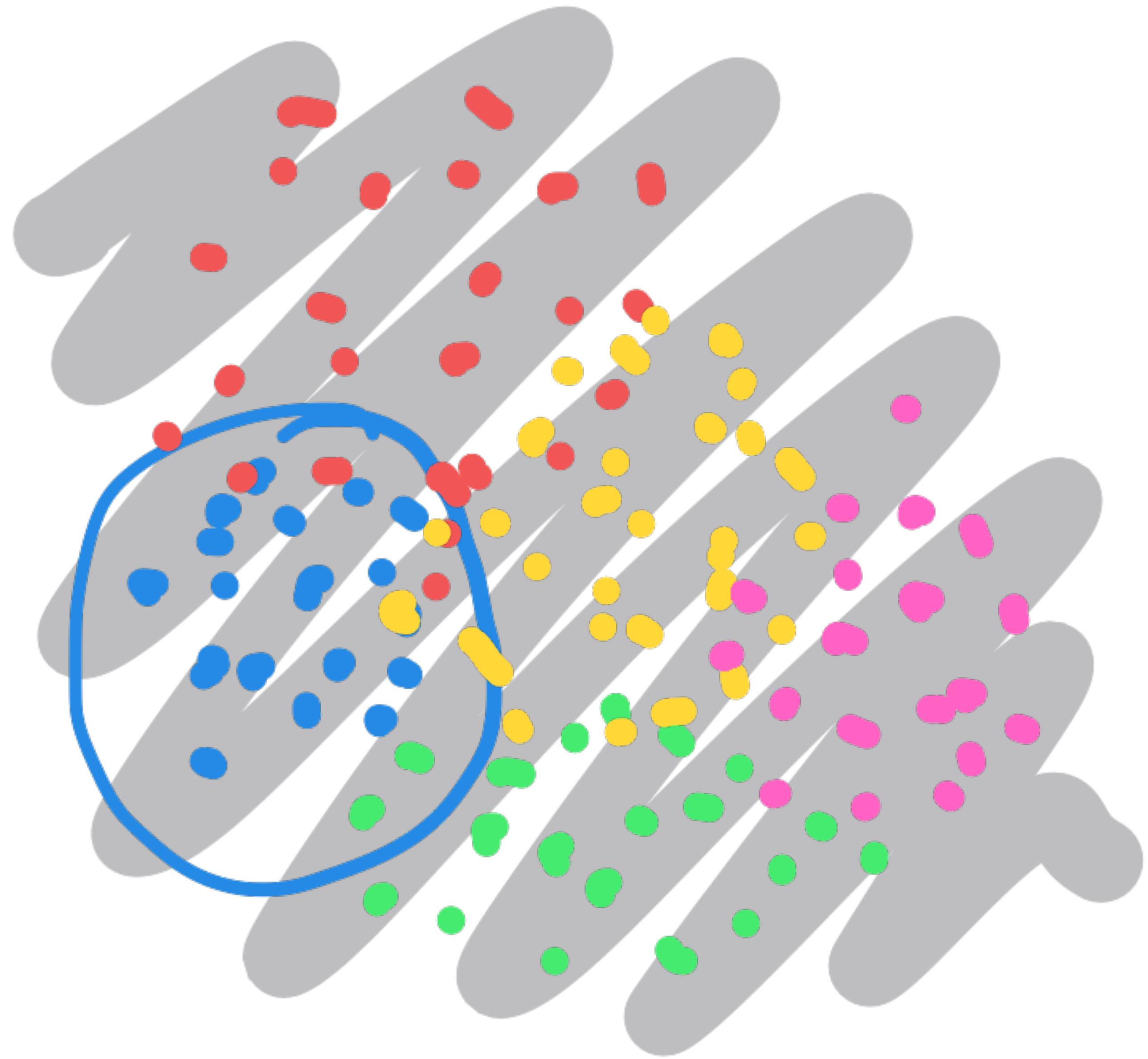


6. Under-reporting

[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

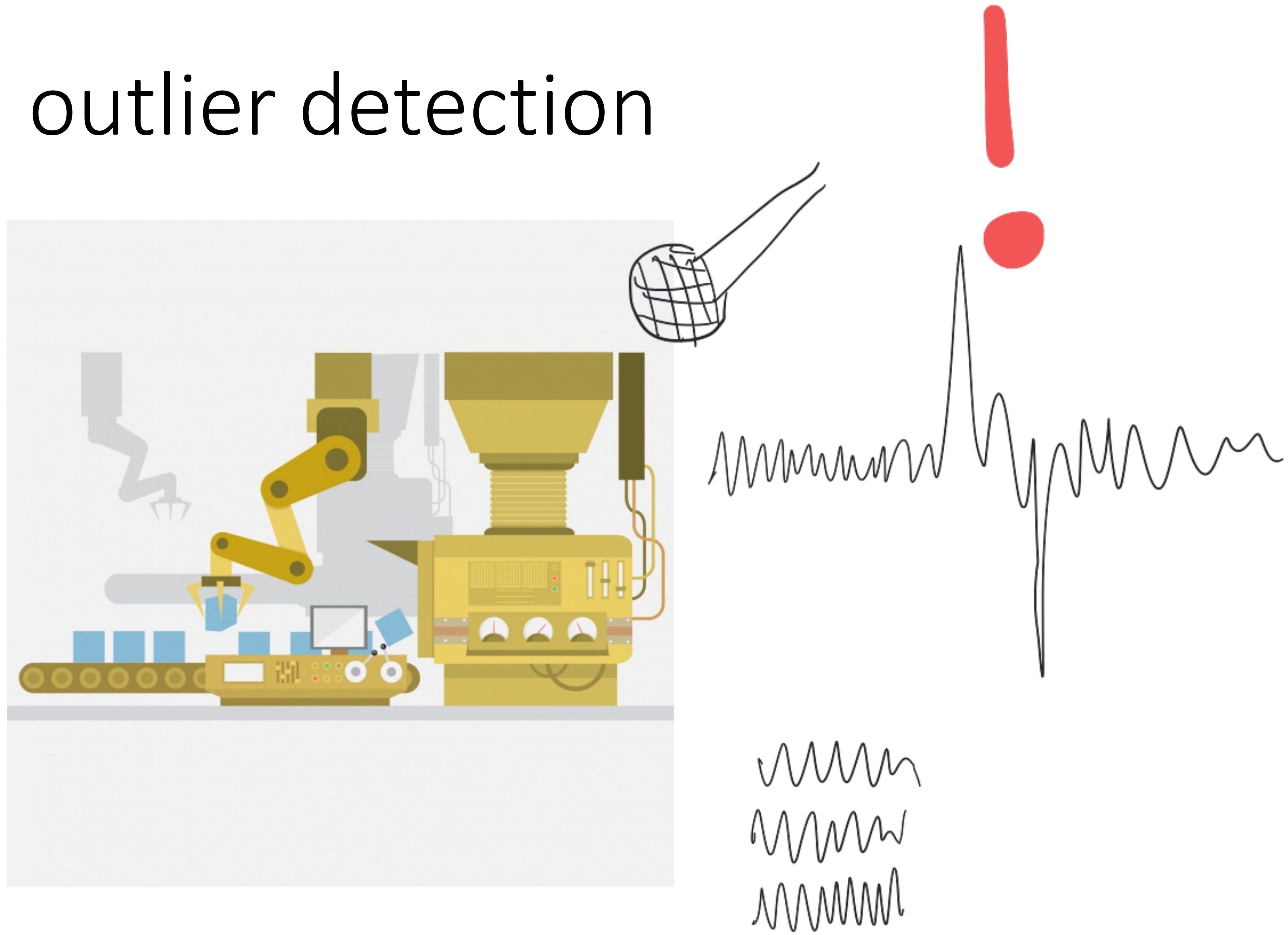
7. One-class classification

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

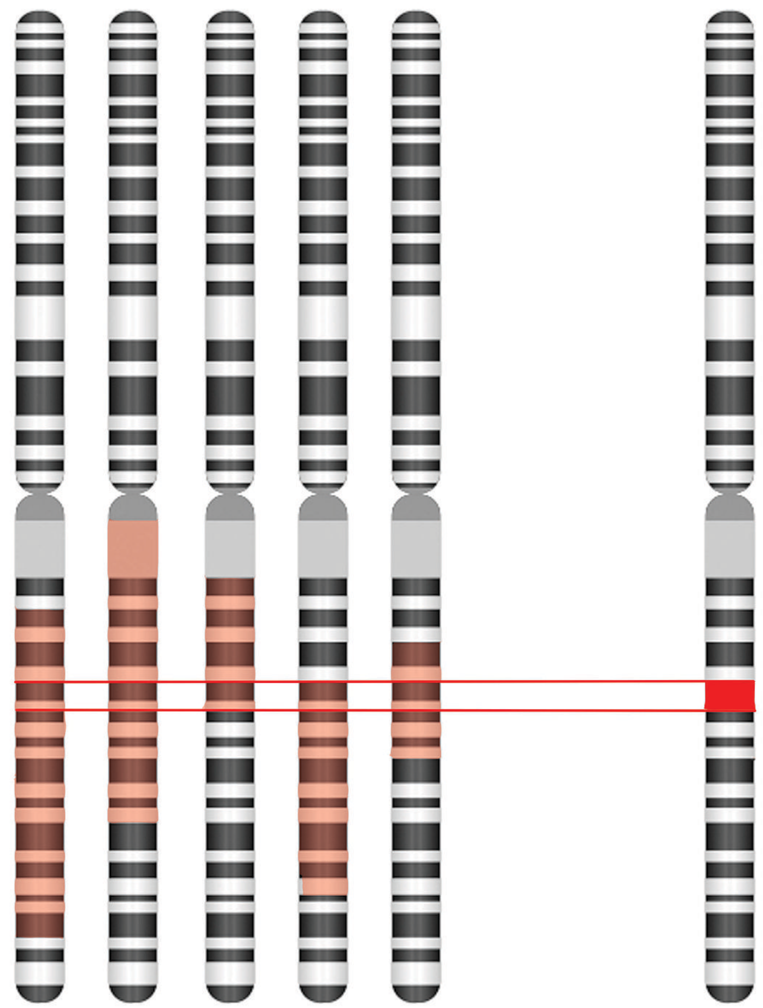


[1] Hido et al. Inlier-based outlier detection via direct density ratio estimation. In 2008 Eighth IEEE international conference on data mining. 2008
[2] Smola et al. Relative novelty detection. IJCAI. 2009
[3] Blanchard et al. Semi-supervised novelty detection. JMLR. 2010

9. Knowledge base completion

[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



[1] Mordelet & Vert. ProDiGe: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. 2011

Image from Esherma1 on Wikipedia

Up next...