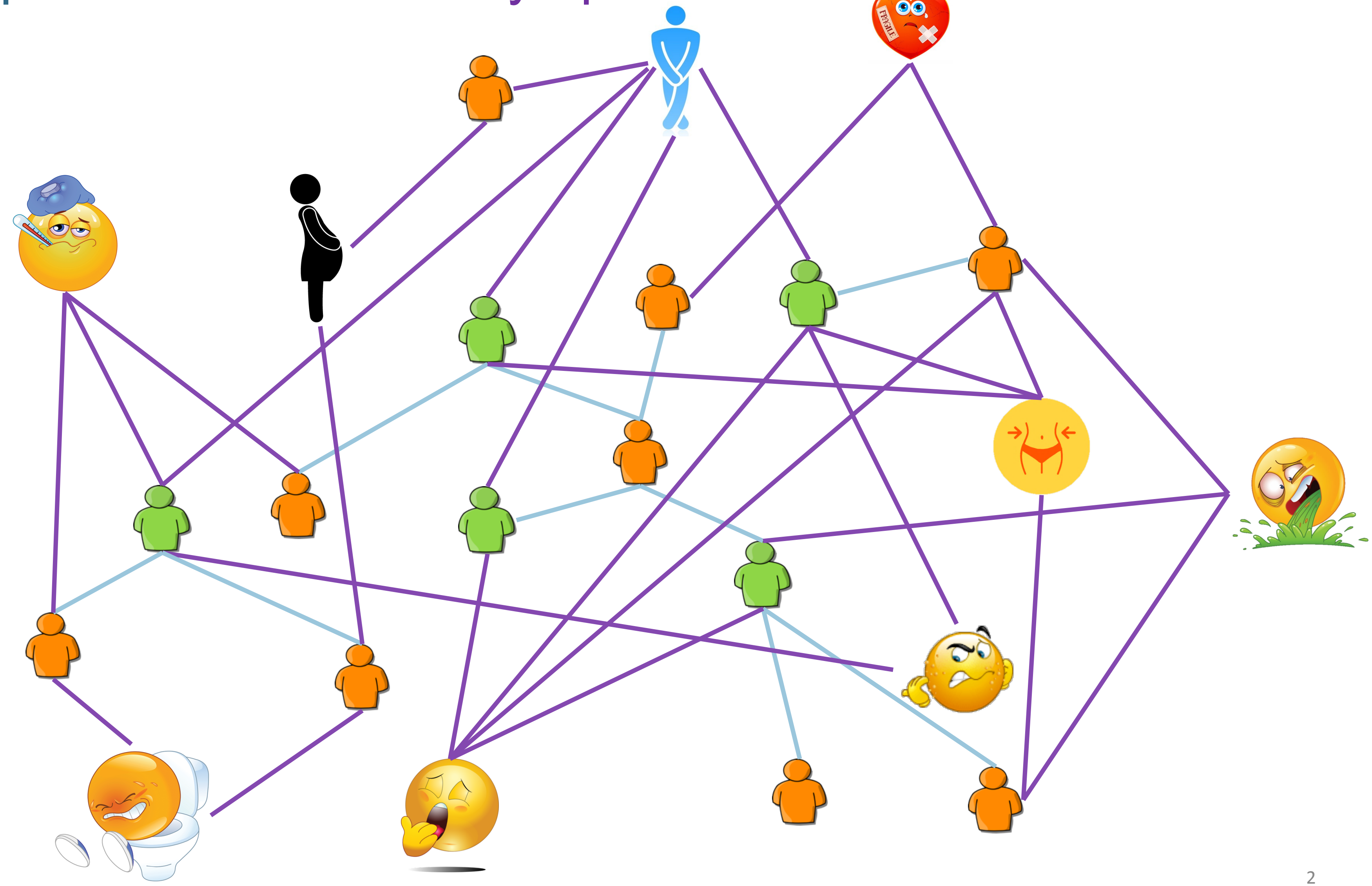
Positive and Unlabeled Relational Classification through Label Frequency Estimation

Jessa Bekker Jesse Davis

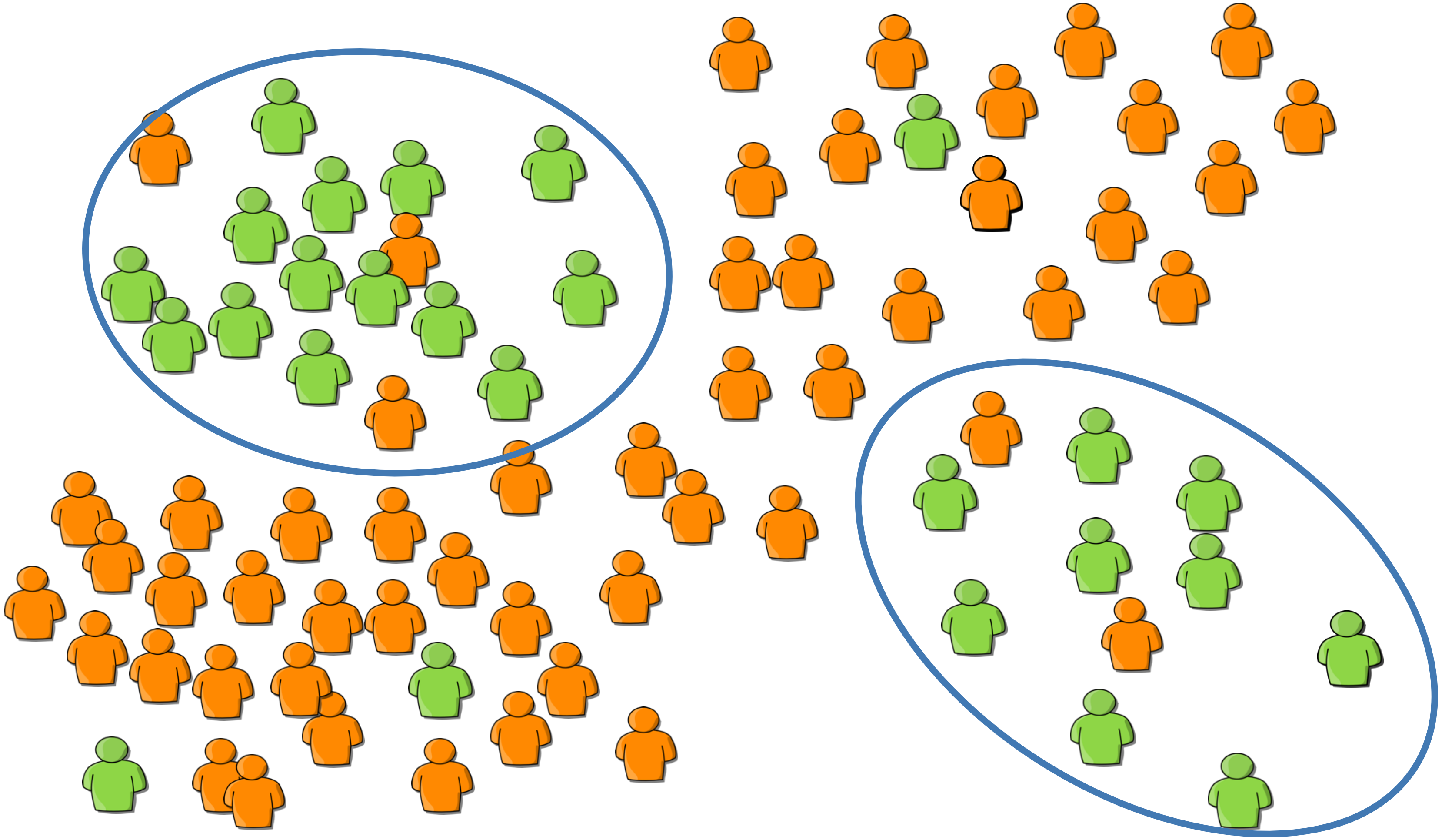
ILP 2017

Diabetes network

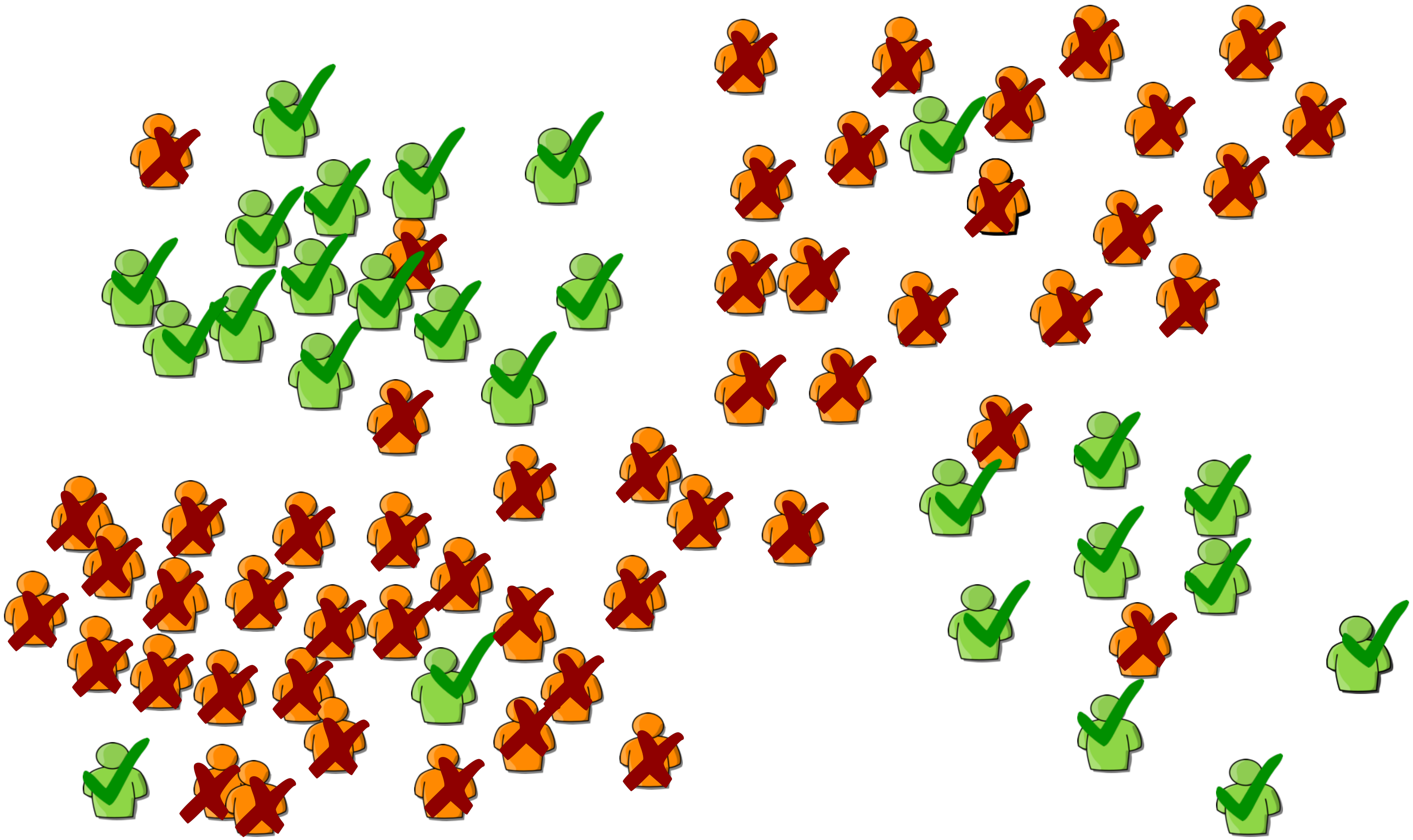
parentOf/2 and hasSymptom/2



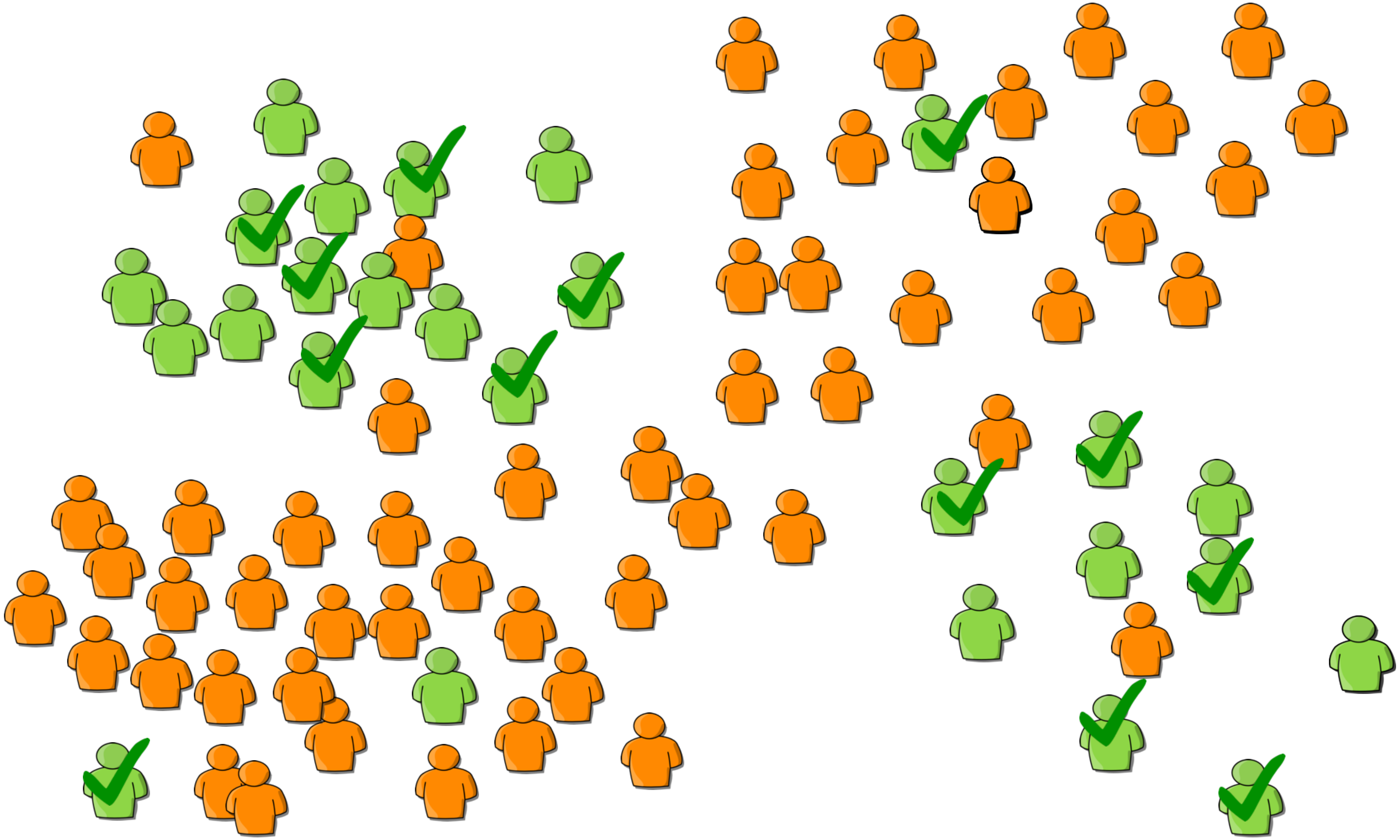
Classification



Supervised Data



Positive and Unlabeled Data

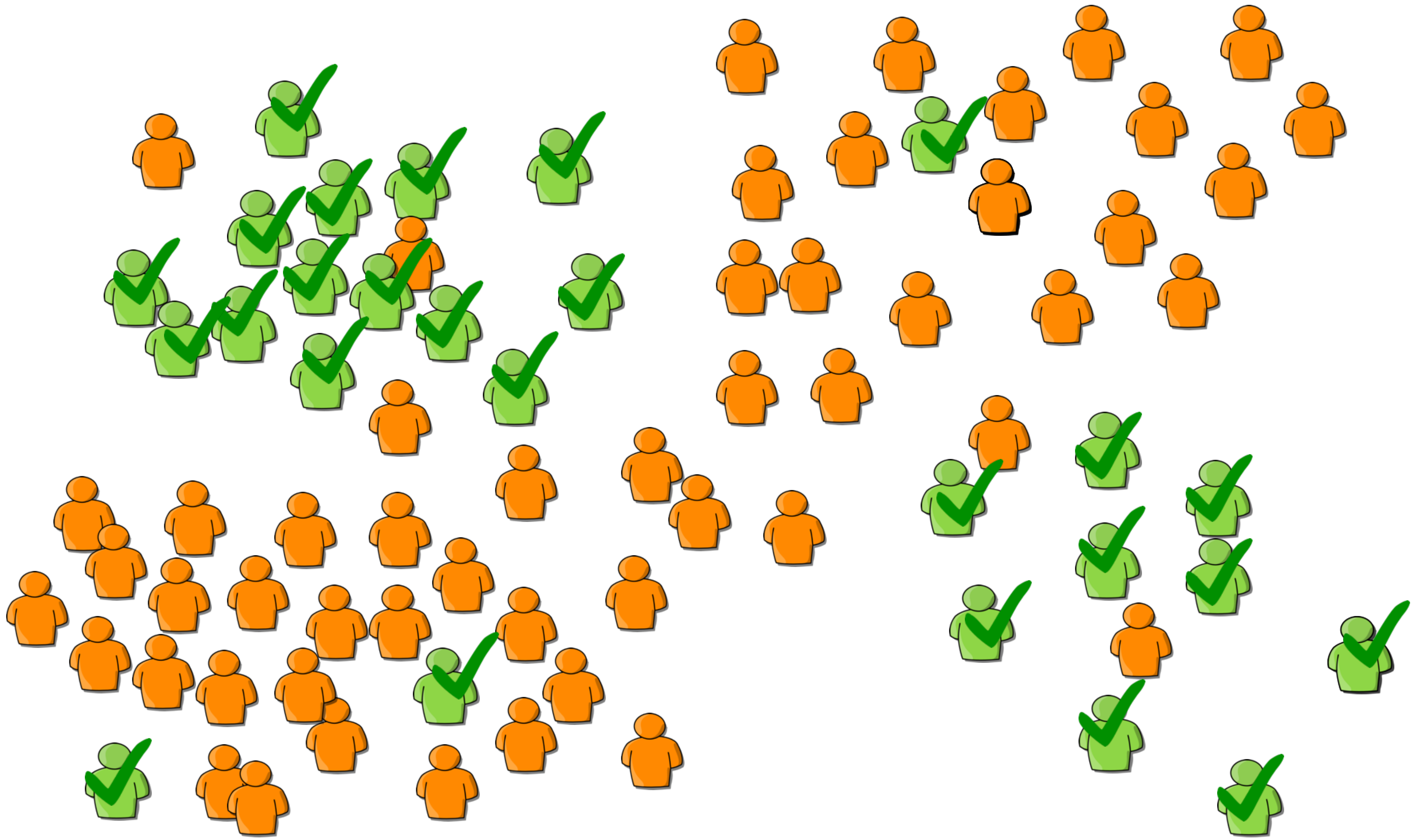


Positive and Unlabeled Data: Label Frequency *c*

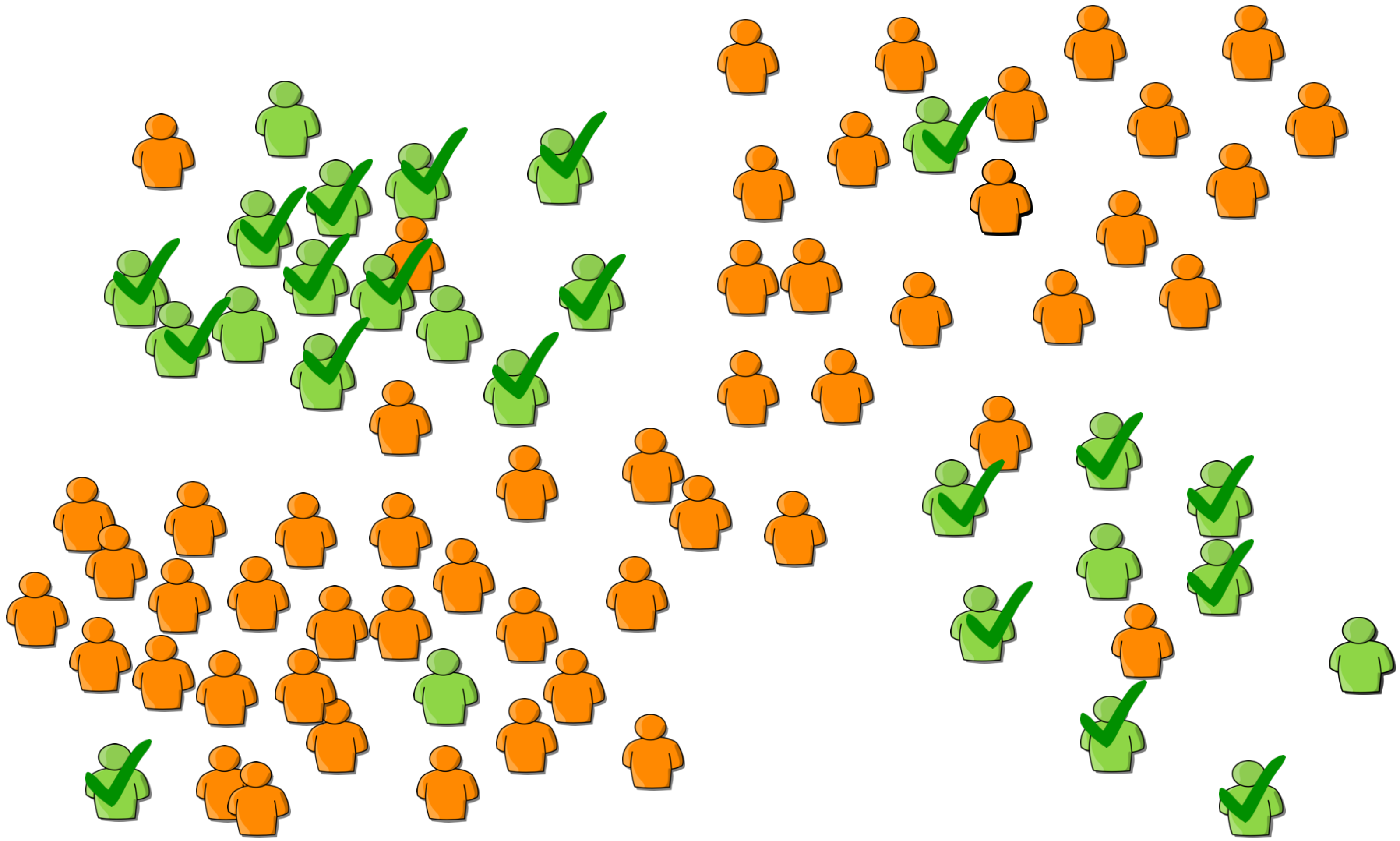
- Positive examples get labeled with constant probability c

$$
\begin{aligned}
c &= P(labeled \mid positive, facts) \\
&= P(labeled \mid positive)
\end{aligned}
$$

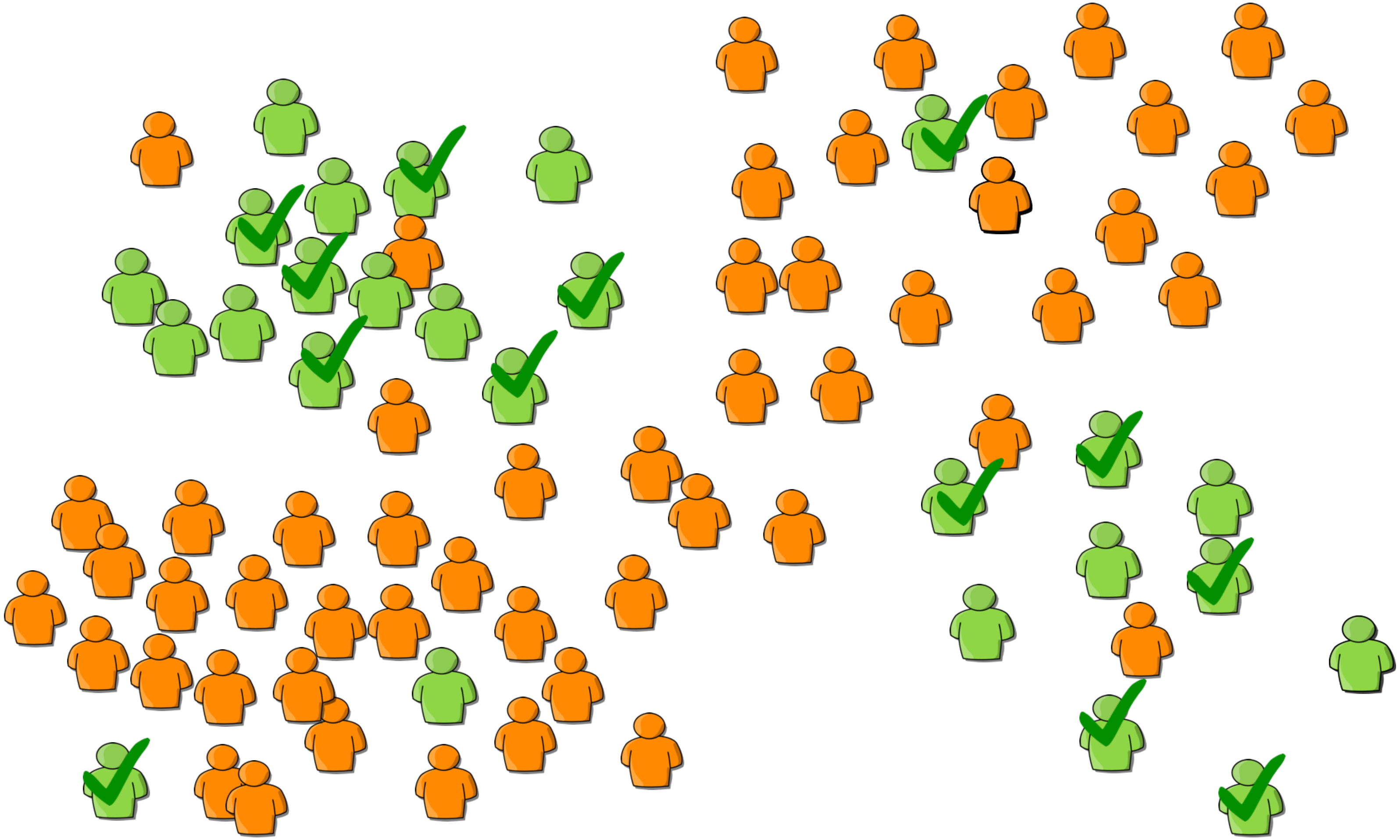
Label Frequency c = *1.0 (= Supervised data)*



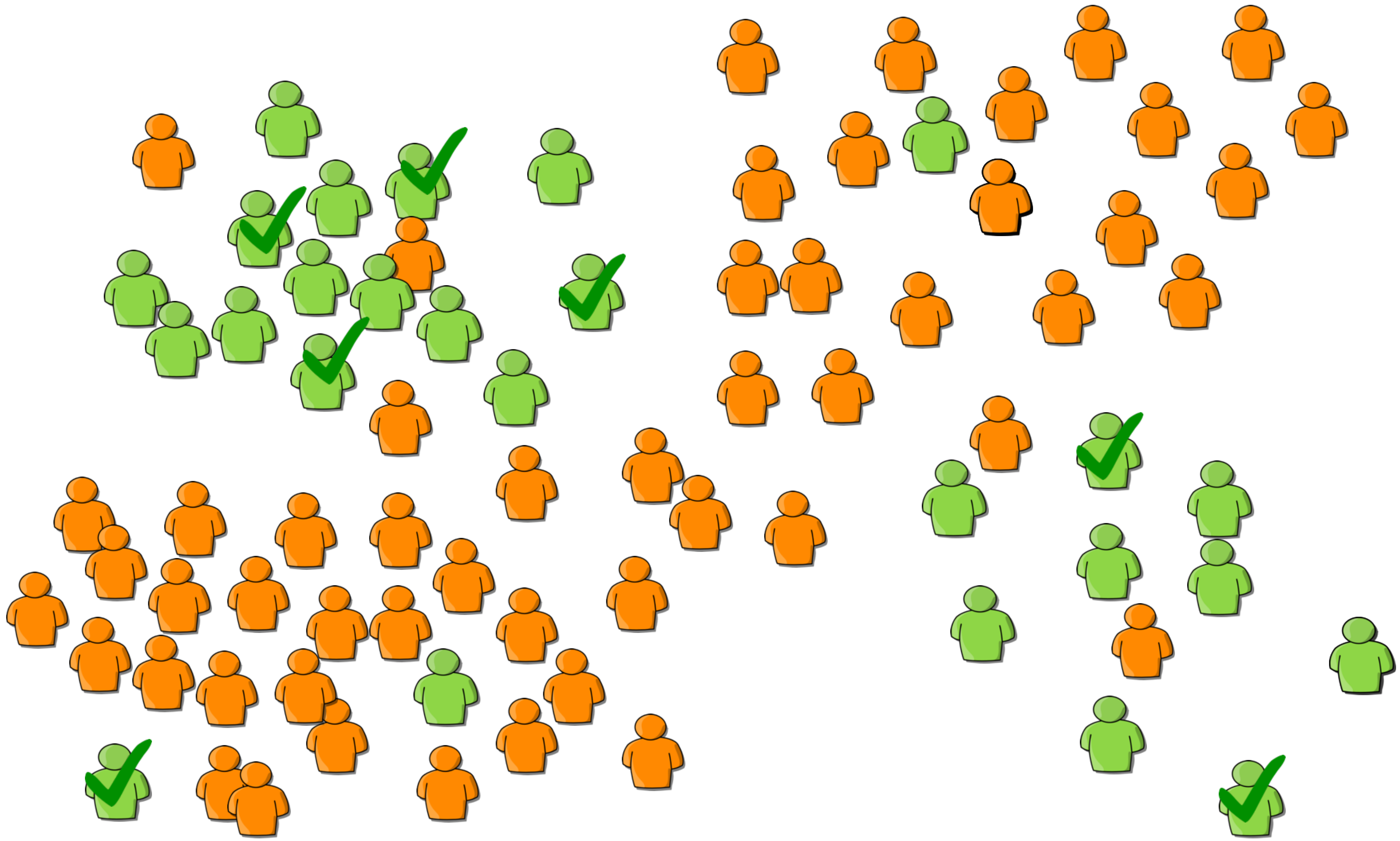
Label Frequency $c = 0.75$



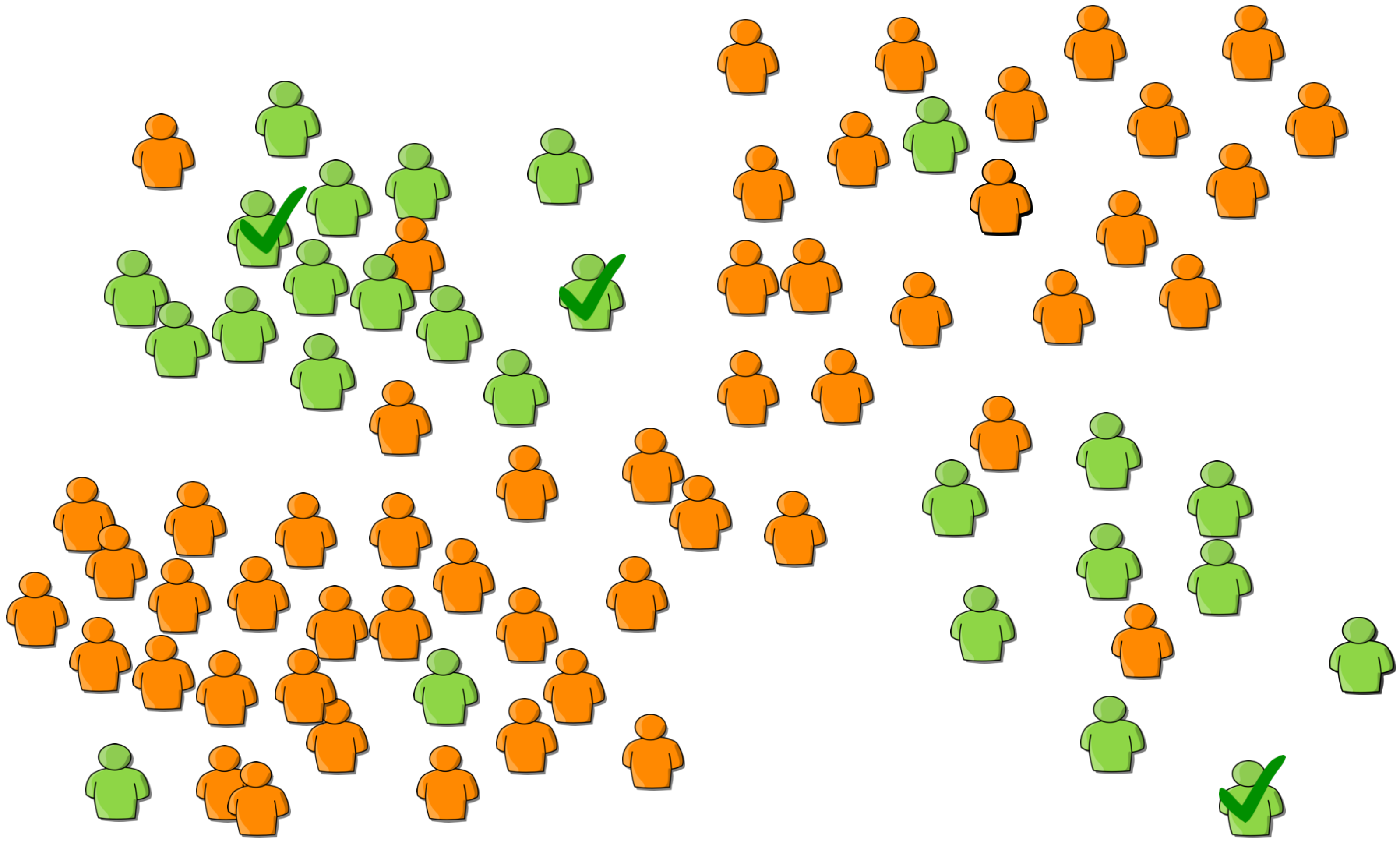
Label Frequency $c = 0.5$



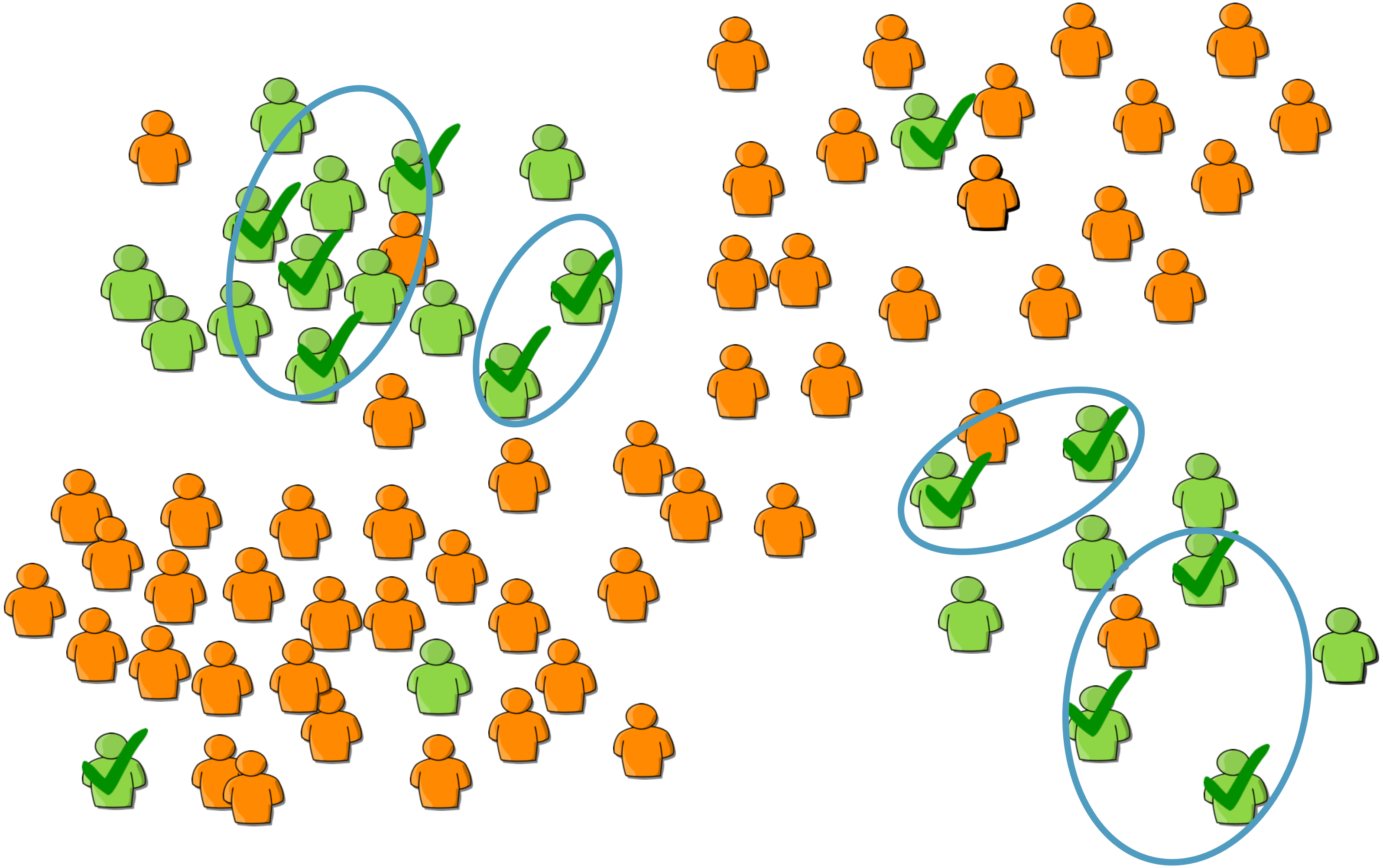
Label Frequency $c = 0.25$



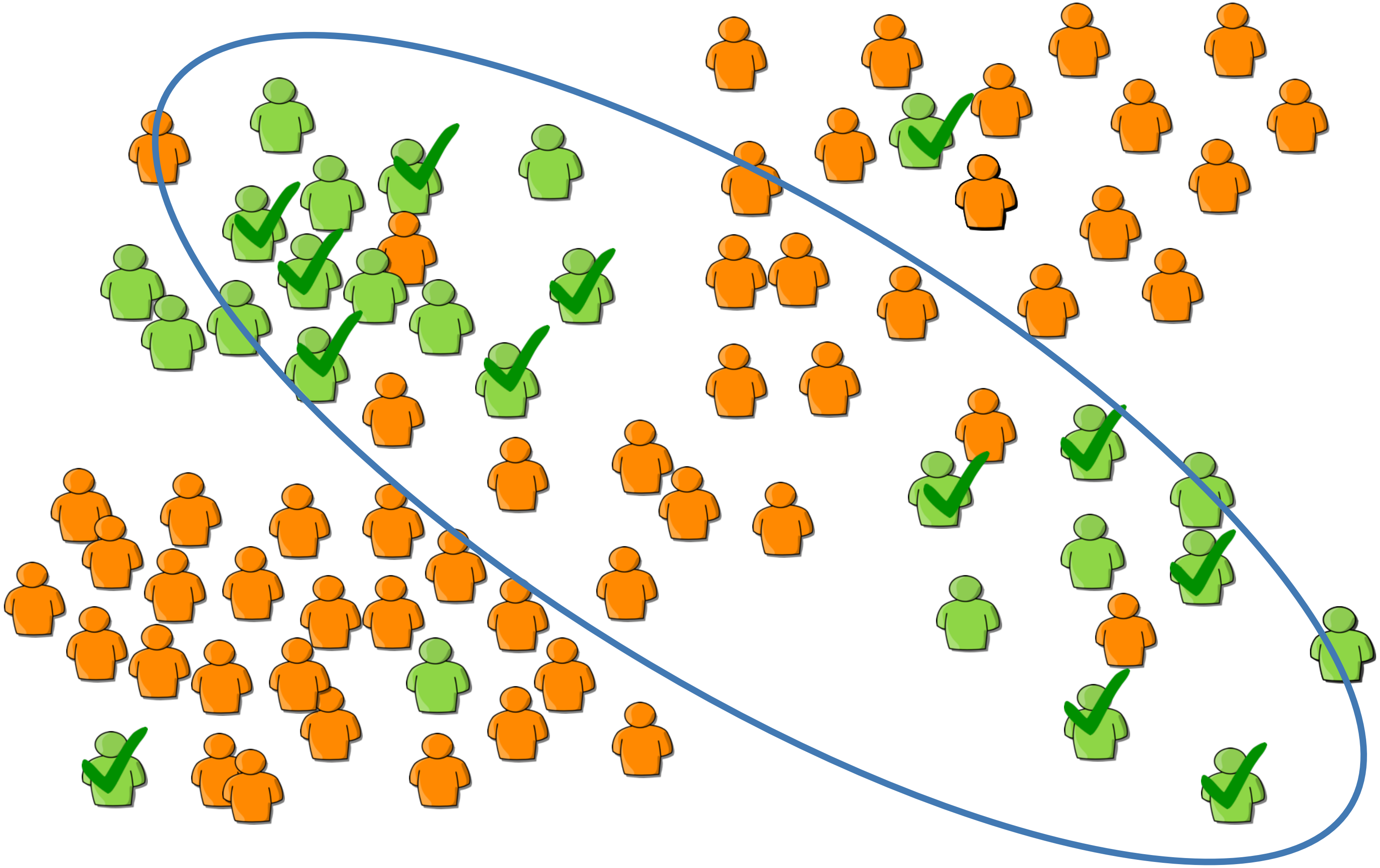
Label Frequency $c = 0.1$



Naïve Classification: Unlabeled = Negative



Common Solution: Conjunctive Concept



[Muggleton, 1996]

State of the Art in Propositional PU

Knowing the label frequency c
makes PU learning easy

[Elkan and Noto, 2008]

Using the Label Frequency *c*

- $$P(positive|facts) = \frac{P(labeled|facts)}{c}$$

Method 1: Probabilistic classifier that learns $P(labeled|facts)$

E.g. Tilde: Probabilistic Relational Decision Trees

Method 2: Adjust learning algorithm using *c*:

P=L/c and N=T-P

E.g. Aleph: adjust score function

Supervised:	Coverage = P-N
PU:	Coverage = L/*c*-(T-L/c) = 2L/c - T

How Can we Know the Label Frequency *c*?

1. Domain knowledge of class proportions

2. Sample and label subset of the data

3. Estimate directly from the data

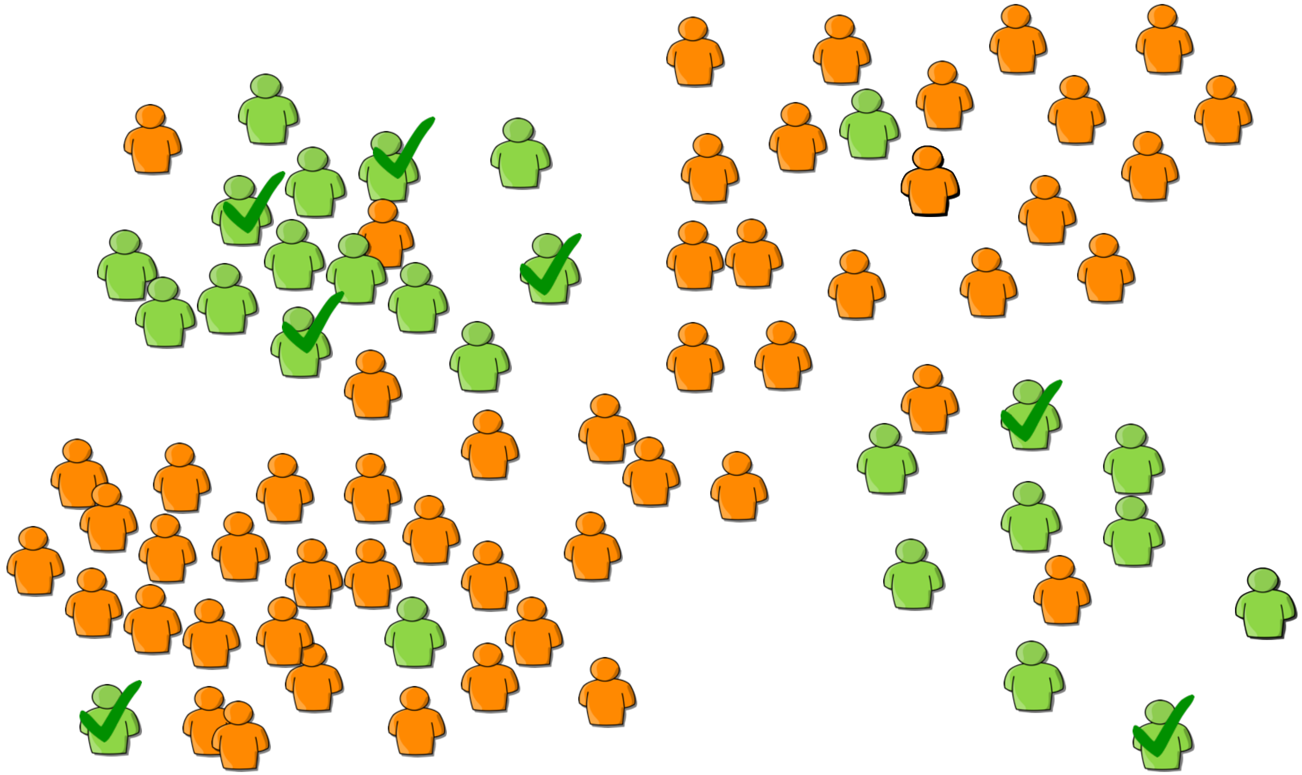
   - Only propositional methods exist
   - Recent method is adaptable for relational settings

     [Bekker&Davis, under review]

Lower bound on *c* from Data

$P \leq T \quad \Rightarrow \quad c = \frac{L}{P} \geq \frac{L}{T}$

$$T = 78$$
$$L = 7$$
$$c \geq \frac{7}{78} = 0.09$$



Estimate *c* from Data (TIcER)

- Insight 1: Data subset implies lower bound on c

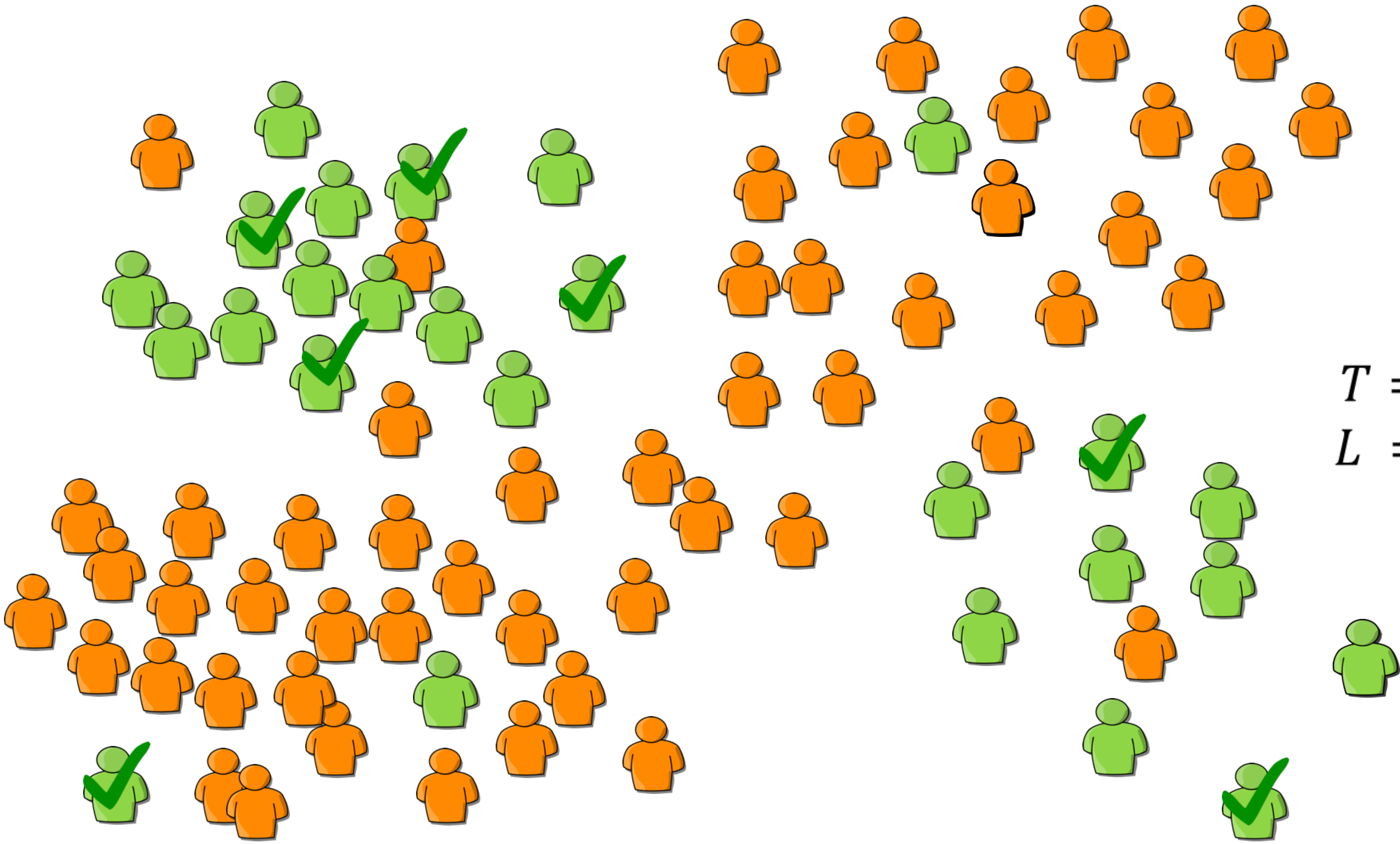
$$c \geq \frac{L}{T} \underbrace{- \varepsilon(T)}_{\text{Error term from 1-sided Chebyshev inequality}}$$

- Insight 2: Positive subsets give very tight bounds

- Insight 3: Highly labeled subsets are likely positive

Look for those through decision tree induction (Tilde)
Use subsets to tighten lower bound

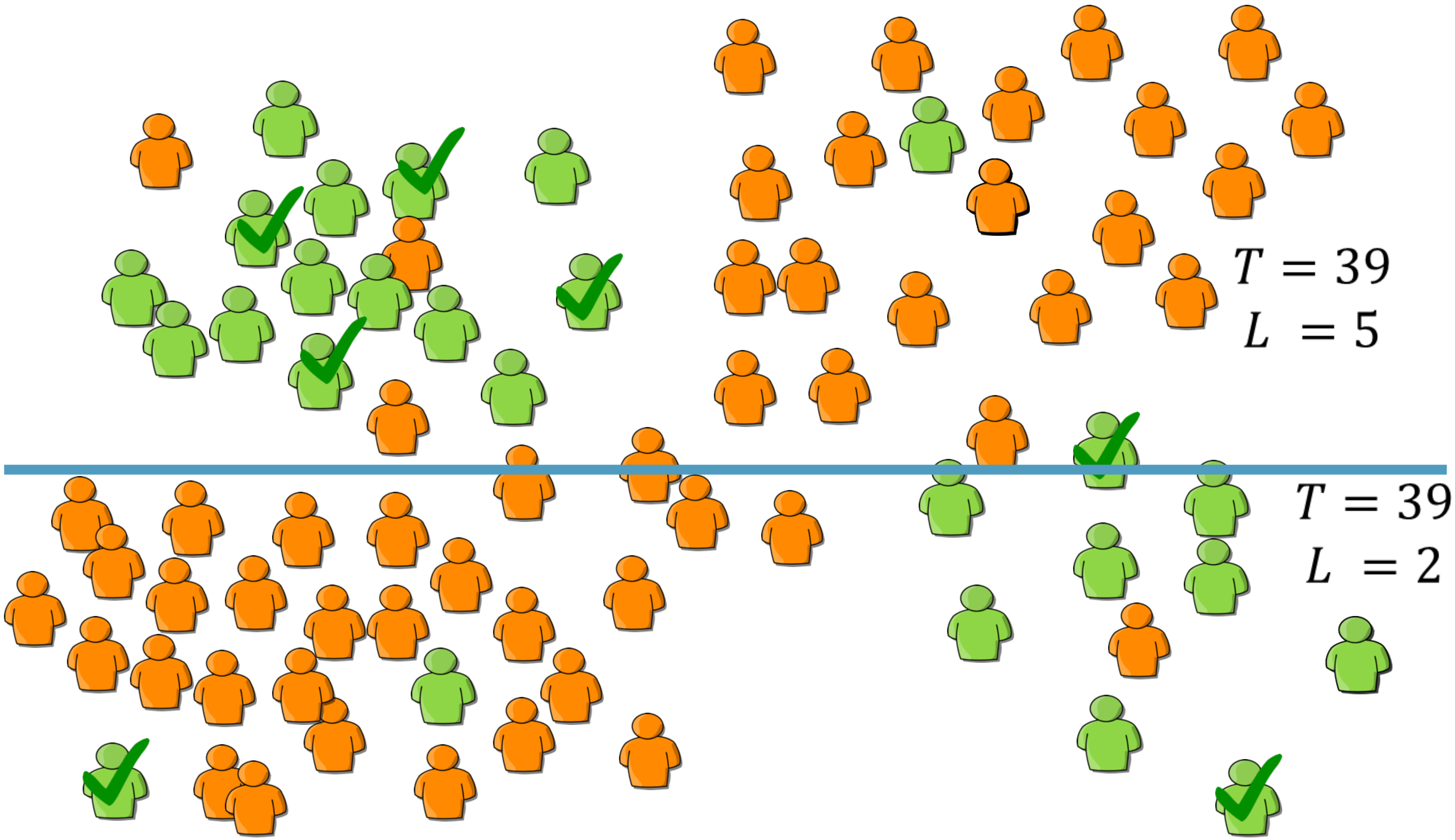
Intuition of TIcER



$$T = 78$$
$$L = 7$$

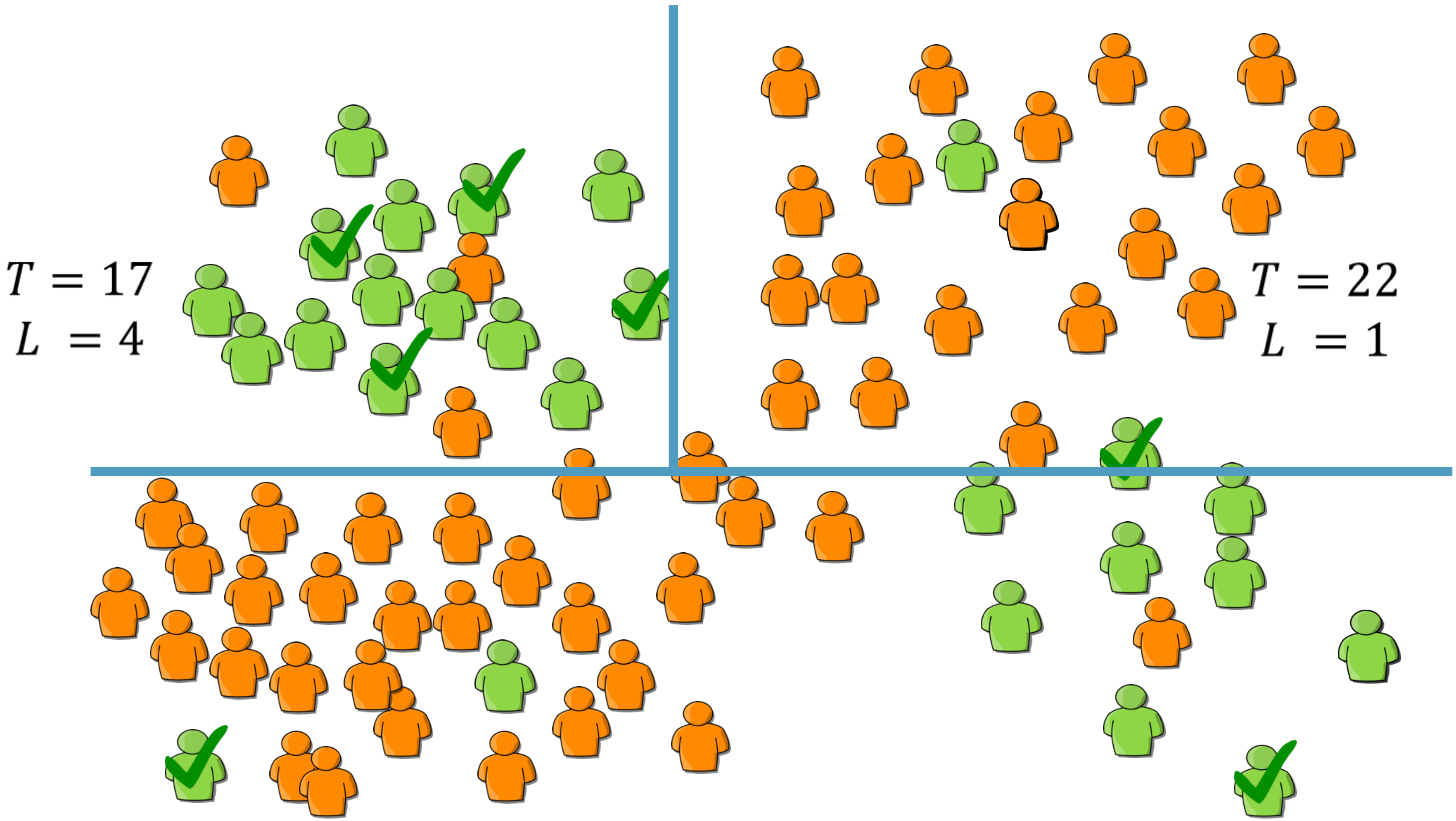
$$c \geq \frac{7}{78} - \varepsilon(78) = 0.09 - \varepsilon(78)$$

Intuition of TIcER



$$c \geq \frac{5}{39} - \varepsilon(39) = 0.13 - \varepsilon(39)$$

Intuition of TIcER



$$c \geq \frac{4}{17} - \varepsilon(17) = 0.24 - \varepsilon(17)$$

TIcER: Practical issues

Selecting subsets based on labels

$\Rightarrow$ likely to find subsets with a higher empirical label frequency.

Solution:

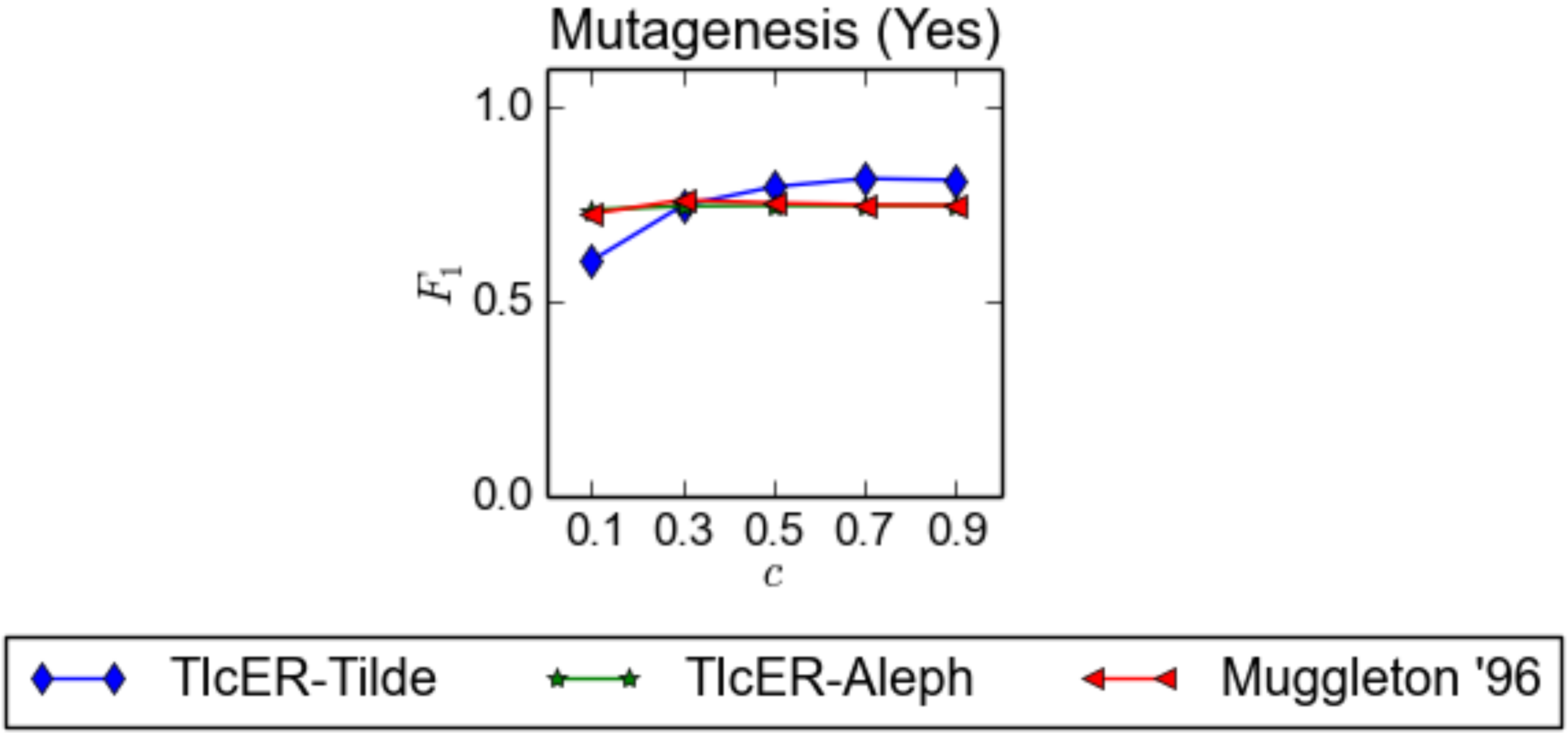
Different datasets for tree induction and c estimation

$\sim$ k-fold cross validation

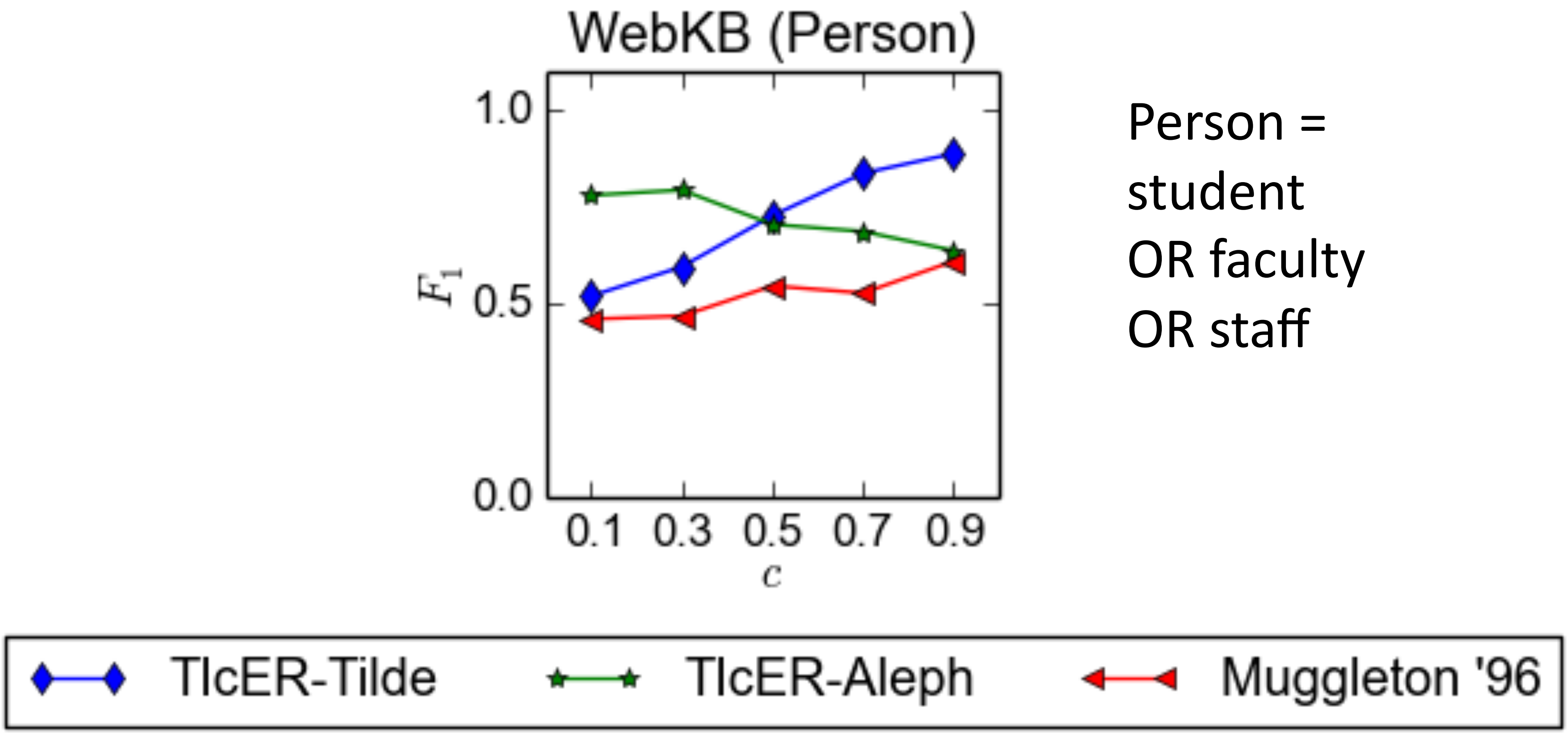
Experimental results

- Estimate *c* from subsets found with Tilde
- use *c* to adjust 1) Tilde and 2) Aleph
- Compare with [Muggleton, 1996]

Experimental results



Experimental results



Person = student OR faculty OR staff

Conclusion

- Knowing the label frequency makes PU learning easier
- Our method is capable of learning disjunctive concepts

References

- Muggleton, Stephen. Learning from positive data. ILP, 1996.
- Elkan, Charles, and Noto, Keith. Learning classifiers from only positive and unlabeled data. KDD, 2008.
- Bekker, Jessa, and Davis, Jesse. Estimating the Class Prior in Positive and Unlabeled Data through Decision Tree Induction. Under review.

Questions?