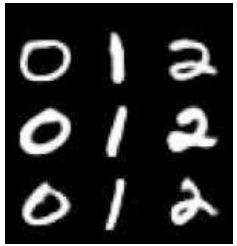
Snorkel: Rapid Training Data Creation with Weak Supervision

Michael Zhang, Jacob Klegar, Tomoka Kan

Original paper: Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2020). Snorkel: Rapid training data creation with weak supervision. The VLDB Journal, 29(2), 709-730.
Logo source: https://www.snorkel.org/

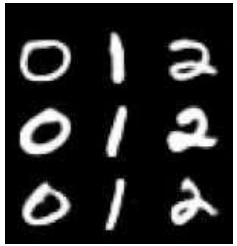
Background & Motivation: Weakly Supervised Learning

Current Supervised Learning Paradigm



Collect data (X)

Current Supervised Learning Paradigm



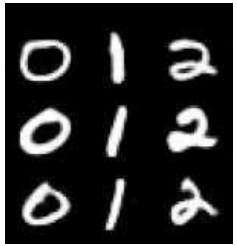
$+$

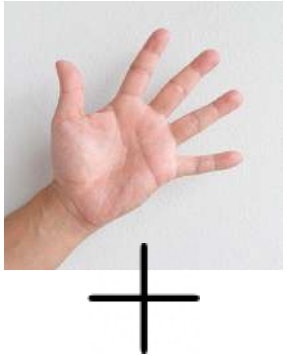
y y y y

Collect data (X)

Get ground-truth labels

Current Supervised Learning Paradigm



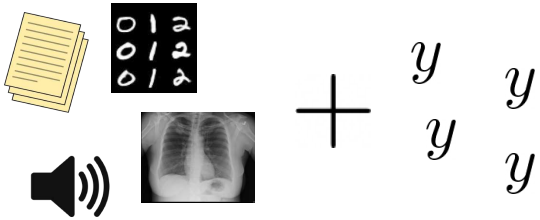


y y y y

Collect data (X)

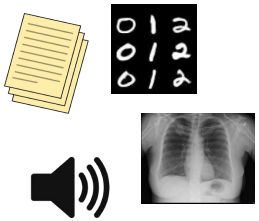
Get ground-truth labels

For Deep Learning Applications...

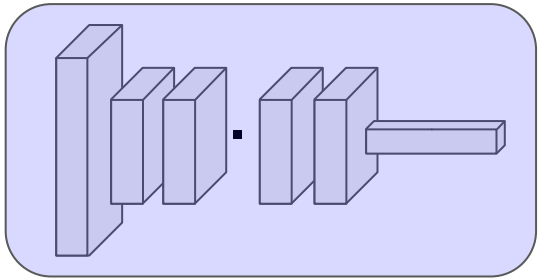


High quality, manually annotated training data

For Deep Learning Applications...



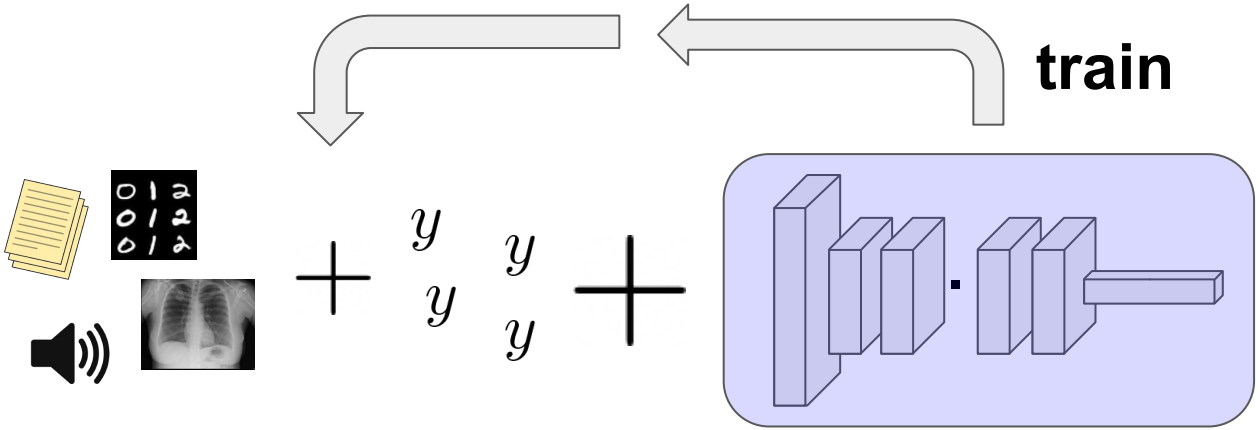
$+$ y y y y $+$



High quality, manually annotated training data

State-of-the-art deep learning model

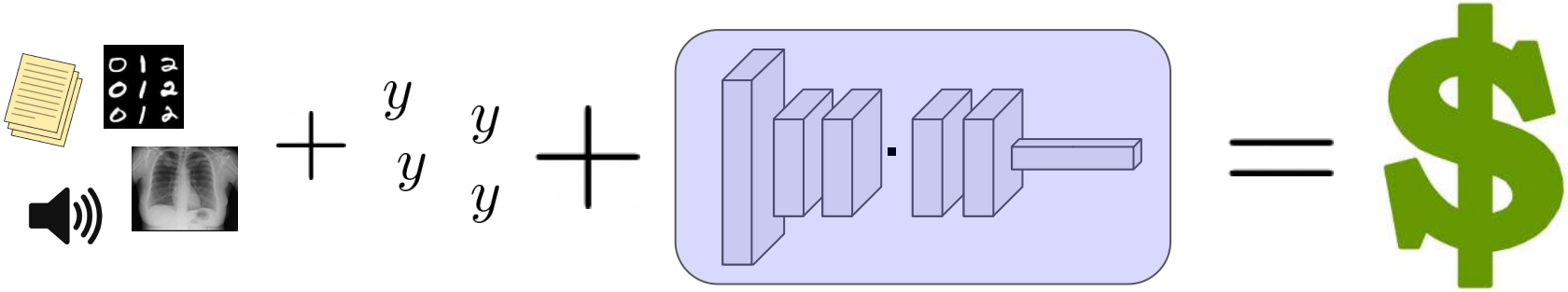
For Deep Learning Applications...



High quality, manually annotated training data

State-of-the-art deep learning model

For Deep Learning Applications...



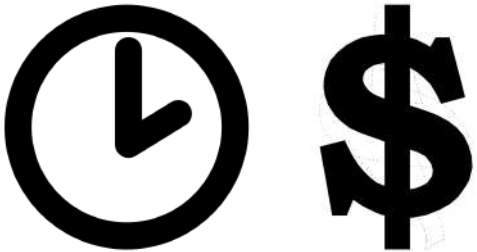
High quality, manually annotated training data

State-of-the-art deep learning model

Profit!

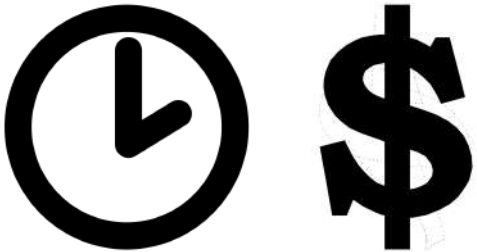
Problem: Manual Annotation is Expensive

Problem: Manual Annotation is Expensive



Slow,
Costly

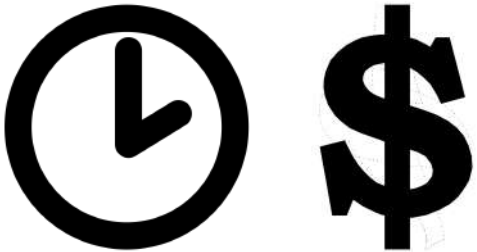
Problem: Manual Annotation is Expensive



Slow, Costly

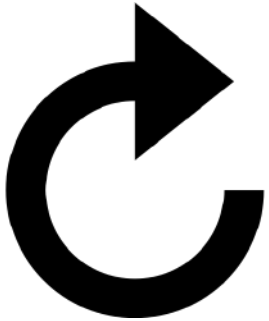
Require Experts, Error-prone

Problem: Manual Annotation is Expensive



Slow, Costly

Require Experts, Error-prone



Inflexible

Can we use *noisier* training data and still train *high-performance* models?

weakly-supervised learning

Can we use *noisier* training data and still train *high-performance* models?

Weakly Supervised Learning

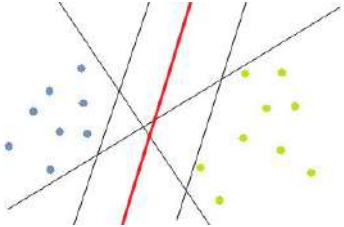
Weakly Supervised Learning

Distant Supervision

Weakly Supervised Learning

Distant Supervision

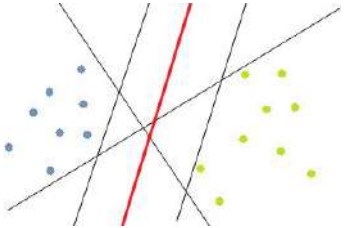
Other Models



Weakly Supervised Learning

Distant Supervision

Other Models

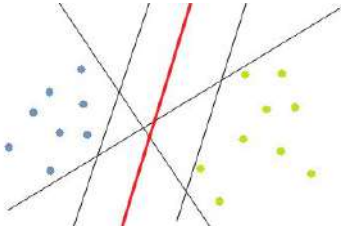


Crowdsourcing

Weakly Supervised Learning

Distant Supervision

Other Models



Crowdsourcing

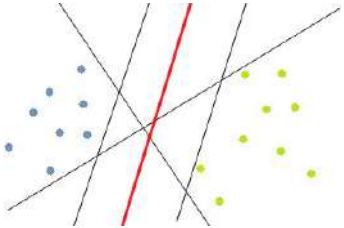
Pattern Matching

[^]*?

Weakly Supervised Learning

Distant Supervision

Other Models

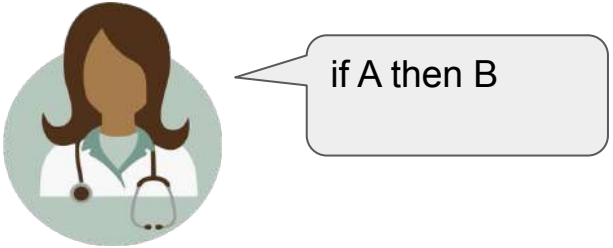


Crowdsourcing

Pattern Matching

[^]*?

Expert Heuristics



Weakly Supervised Learning

Distant Supervision Other Models Crowdsourcing

Currently done in an ad hoc manner
Snorkel tries to streamline this

Snorkel Pipeline

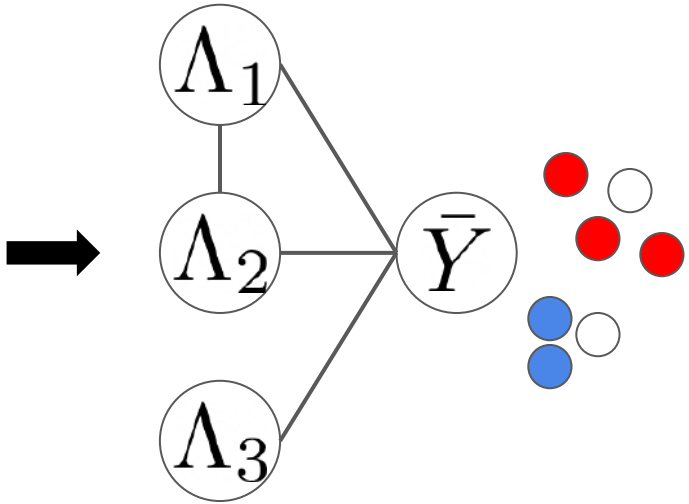
Snorkel Pipeline

```
def label_function(x, a):
    if x < a:
        return True
    elif x > 2 * a:
        return False
    else:
        return None
```

Write labeling functions

Snorkel Pipeline

```
def label_function(x, a):
    if x < a:
        return True
    elif x > 2 * a:
        return False
    else:
        return None
```

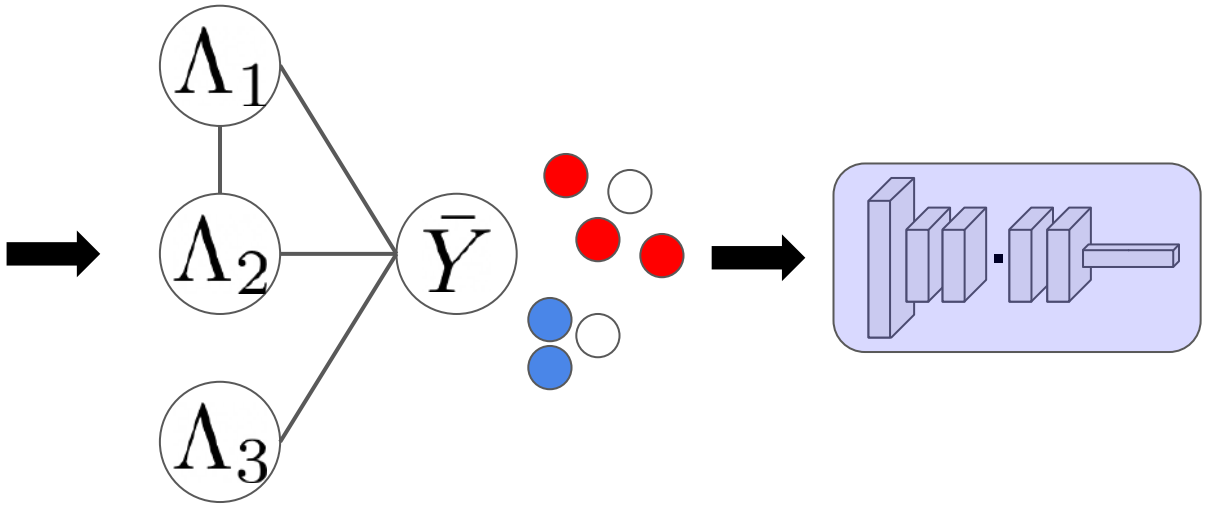


Write labeling functions

Generate probabilistic training labels

Snorkel Pipeline

```
def label_function(x, a):
    if x < a:
        return True
    elif x > 2 * a:
        return False
    else:
        return None
```



Write labeling functions

Generate probabilistic training labels

Train discriminative model

Step 1: Labeling Functions

Labeling Functions, a Key Abstraction

$$x \in \mathcal{X}$$

data point

Labeling Functions, a Key Abstraction

$$x \in \mathcal{X} \rightarrow \lambda$$

data point **labeling function**

Labeling Functions, a Key Abstraction

$$x \in \mathcal{X} \rightarrow \boxed{\lambda} \rightarrow \tilde{y} \in \mathcal{Y} \cup 0$$

data point **labeling function** **weak label**

Labeling Functions, a Key Abstraction

$$x \in \mathcal{X} \rightarrow \lambda \rightarrow \tilde{y} \in \mathcal{Y} \cup 0$$

$\{-1, 1\}$

data point **labeling function** **weak label**

Two Main Interfaces

Two Main Interfaces

```
def label_function(x, a):
    if x < a:
        return True
    elif x > 2 * a:
        return False
    else:
        return None
```

Hand-defined

Two Main Interfaces

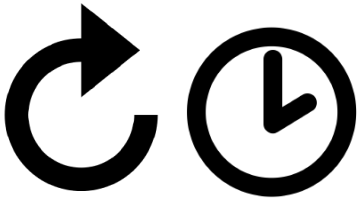
```
def label_function(x, a):
    if x < a:
        return True
    elif x > 2 * a:
        return False
    else:
        return None
```

Hand-defined

```
Ontology(ctd, [A,B,-C])

Pattern("{{a}} causes
{{b}}")

CustomFn(x:heuristic(x))
```

Declarative

Two Main Interfaces



Iterate over a year of user studies

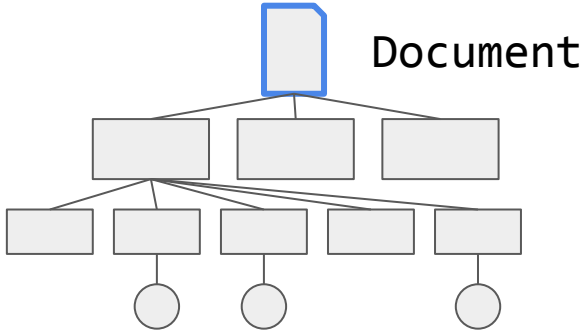
```
def label_function(x, a):
    if x < a:
        return True
    elif x > 2 * a:
        return False
    else:
        return None
```

Hand-defined

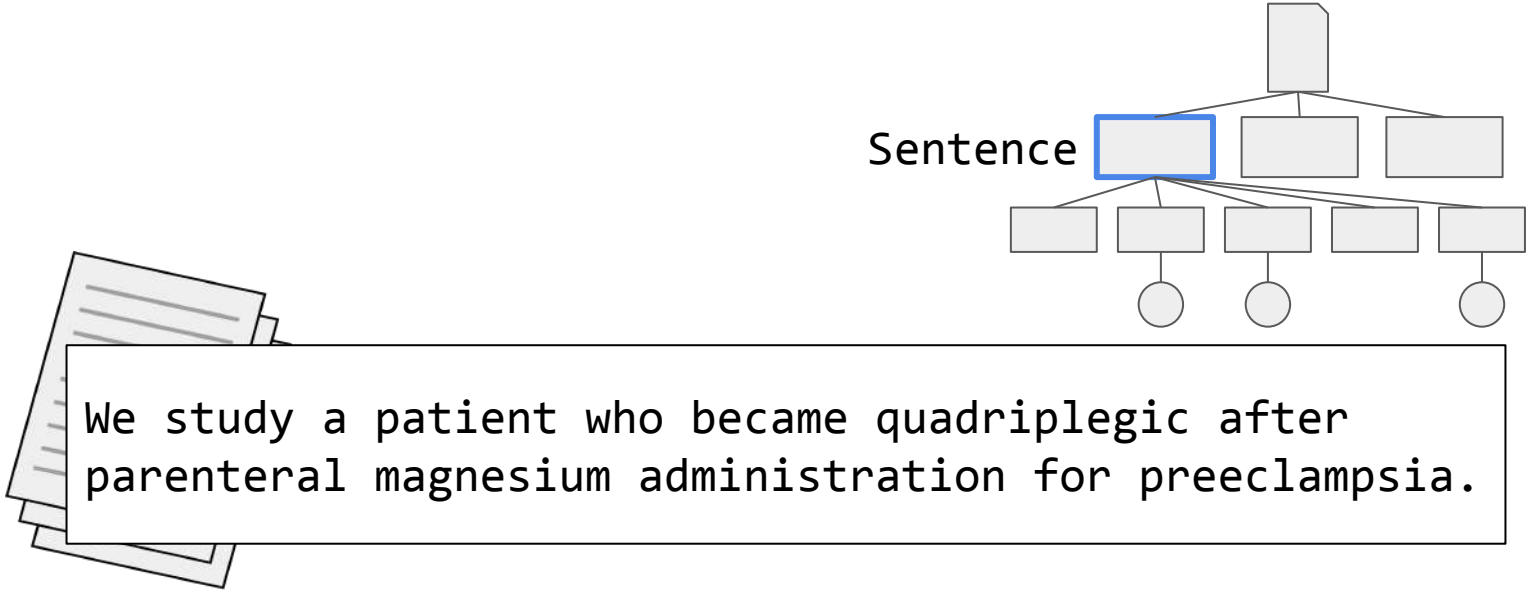
```
Ontology(ctd, [A,B,-C])

Pattern("{{a}} causes
{{b}}")

CustomFn(x:heuristic(x))
```

Declarative

Data Model Example: Relation Extraction

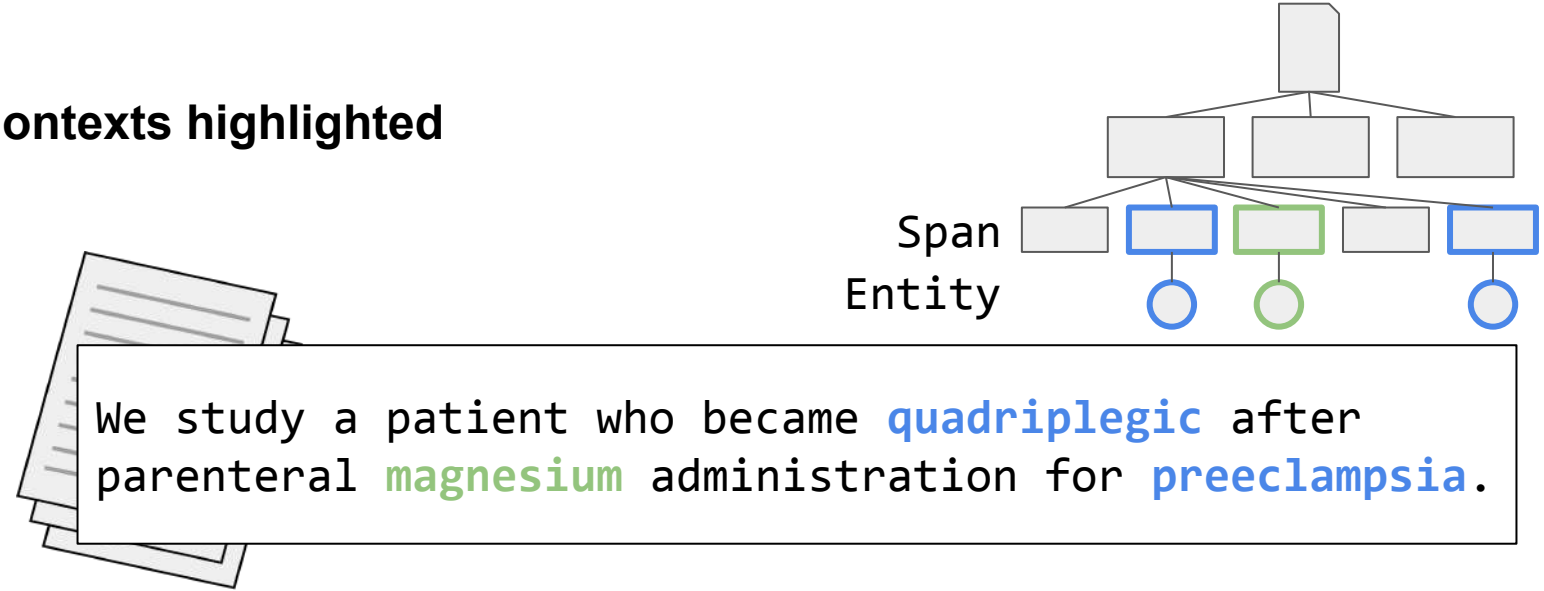


Data Model Example: Relation Extraction

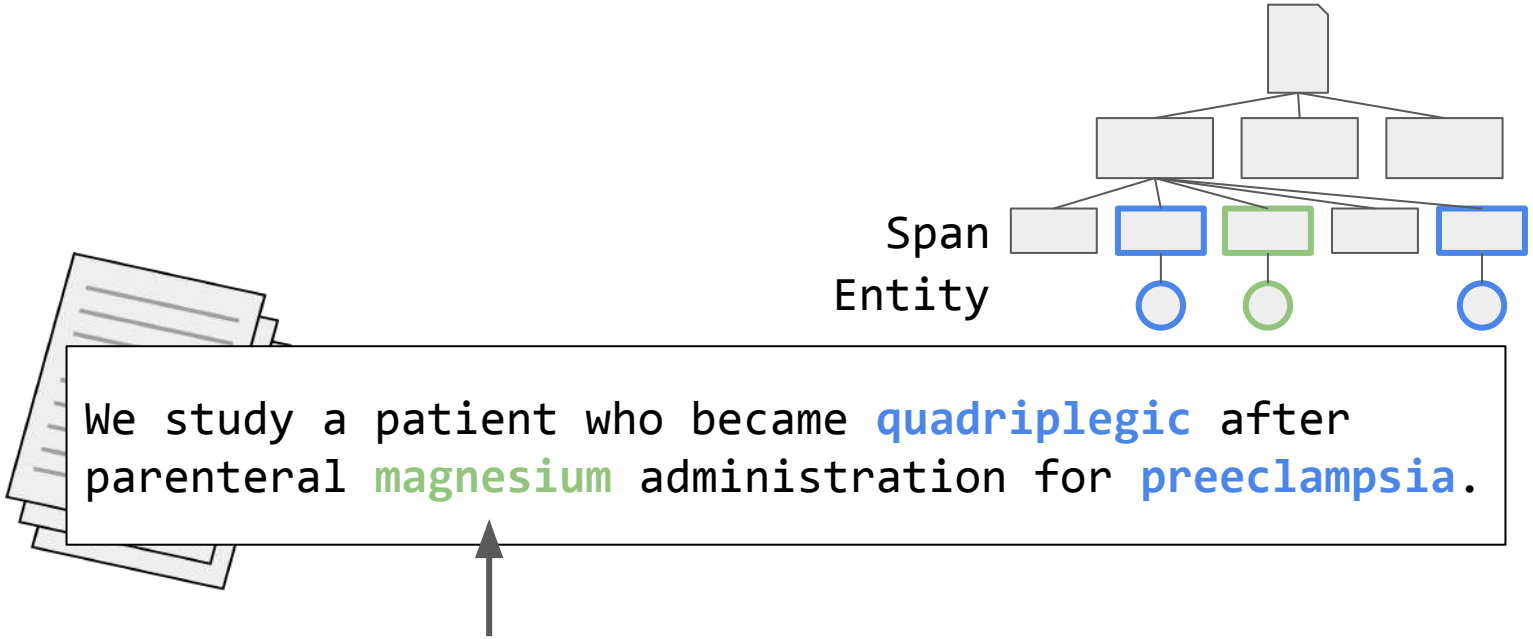


Data Model Example: Relation Extraction

Contexts highlighted

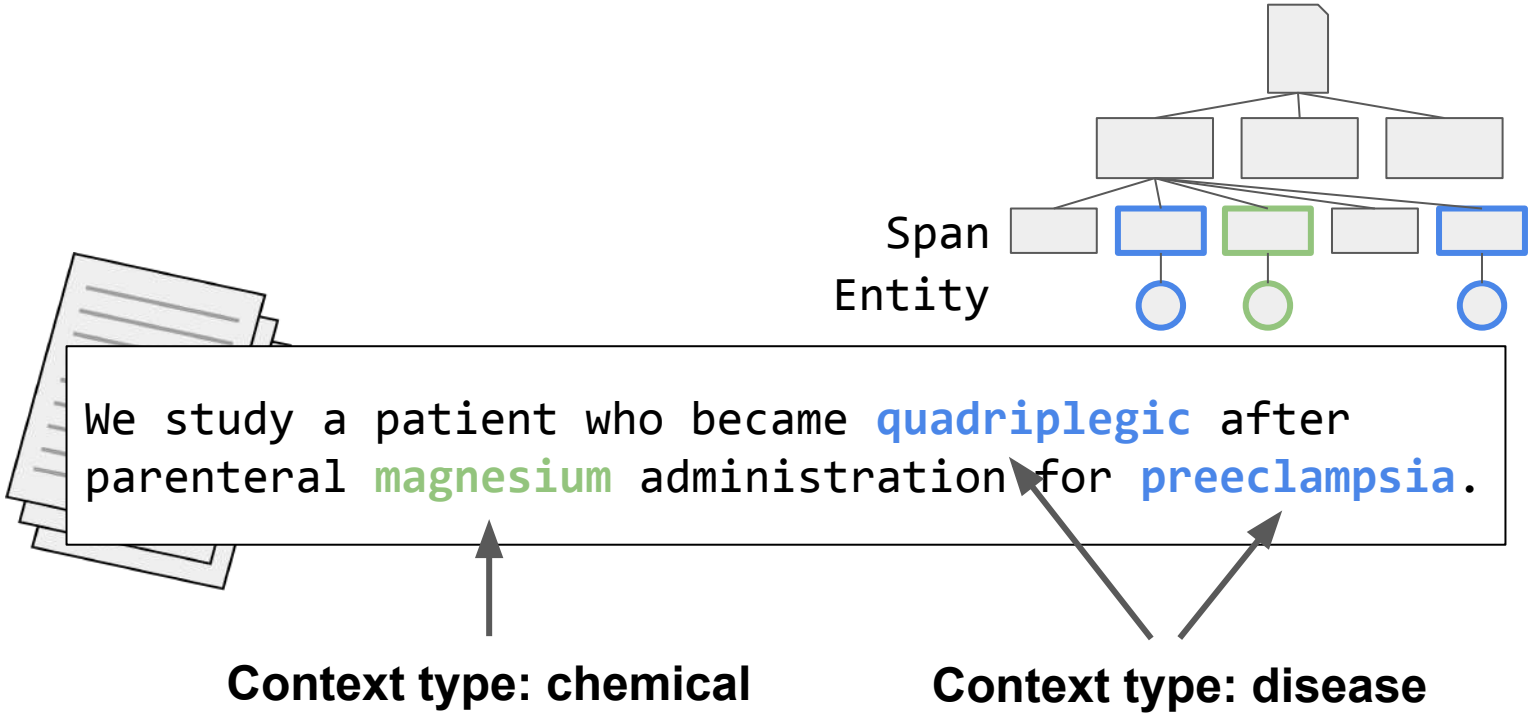


Data Model Example: Relation Extraction

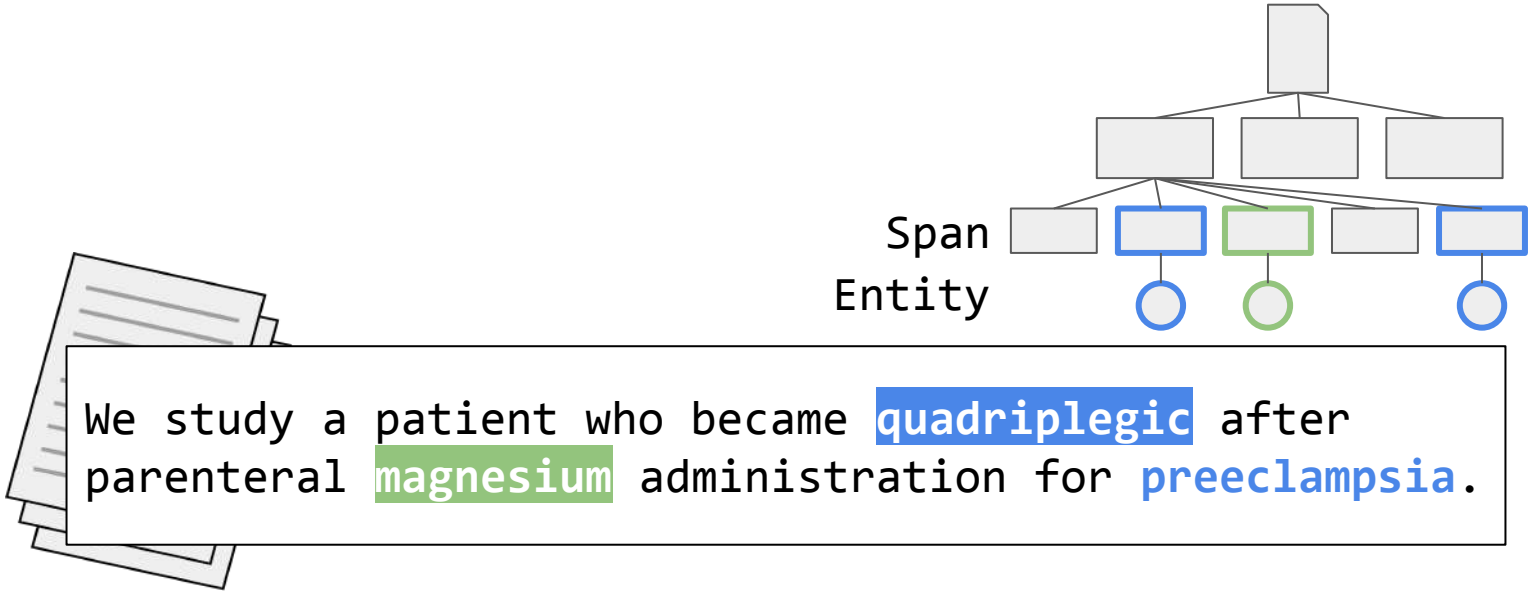


Context type: chemical

Data Model Example: Relation Extraction

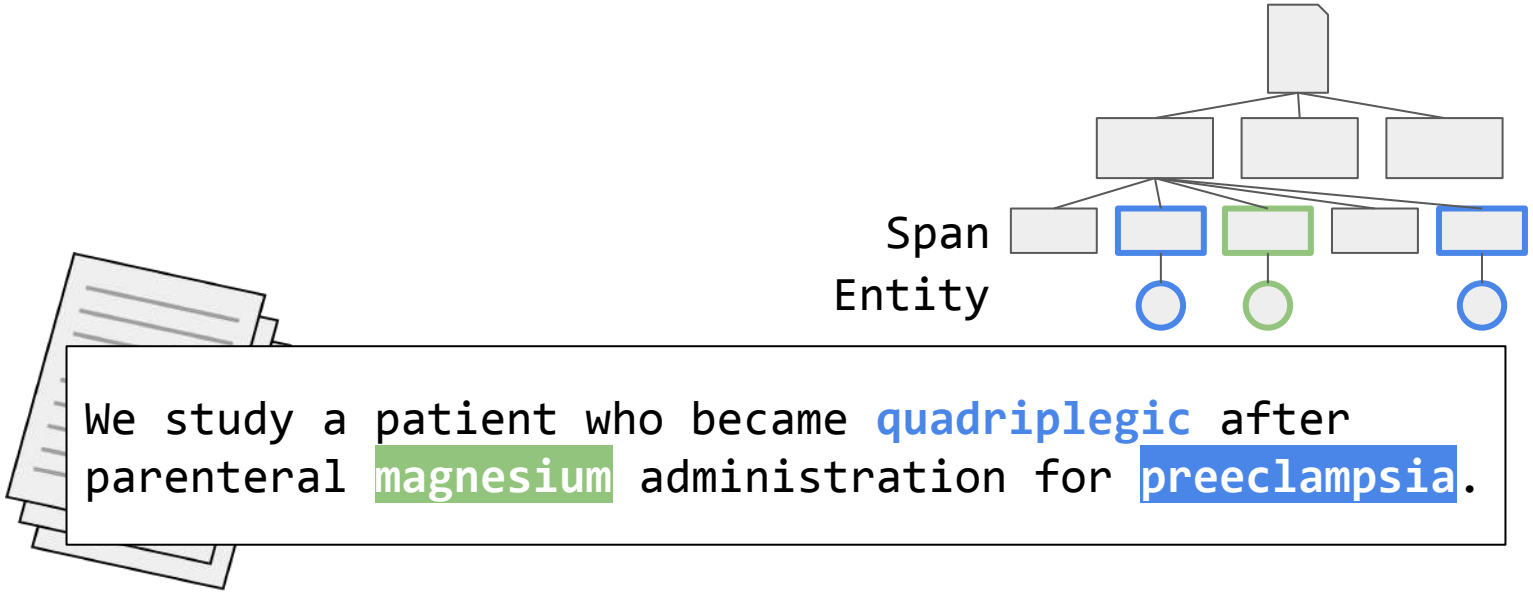


Data Model Example: Relation Extraction



Candidate 1: x_1 := Causes (magnesium, quadriplegic)

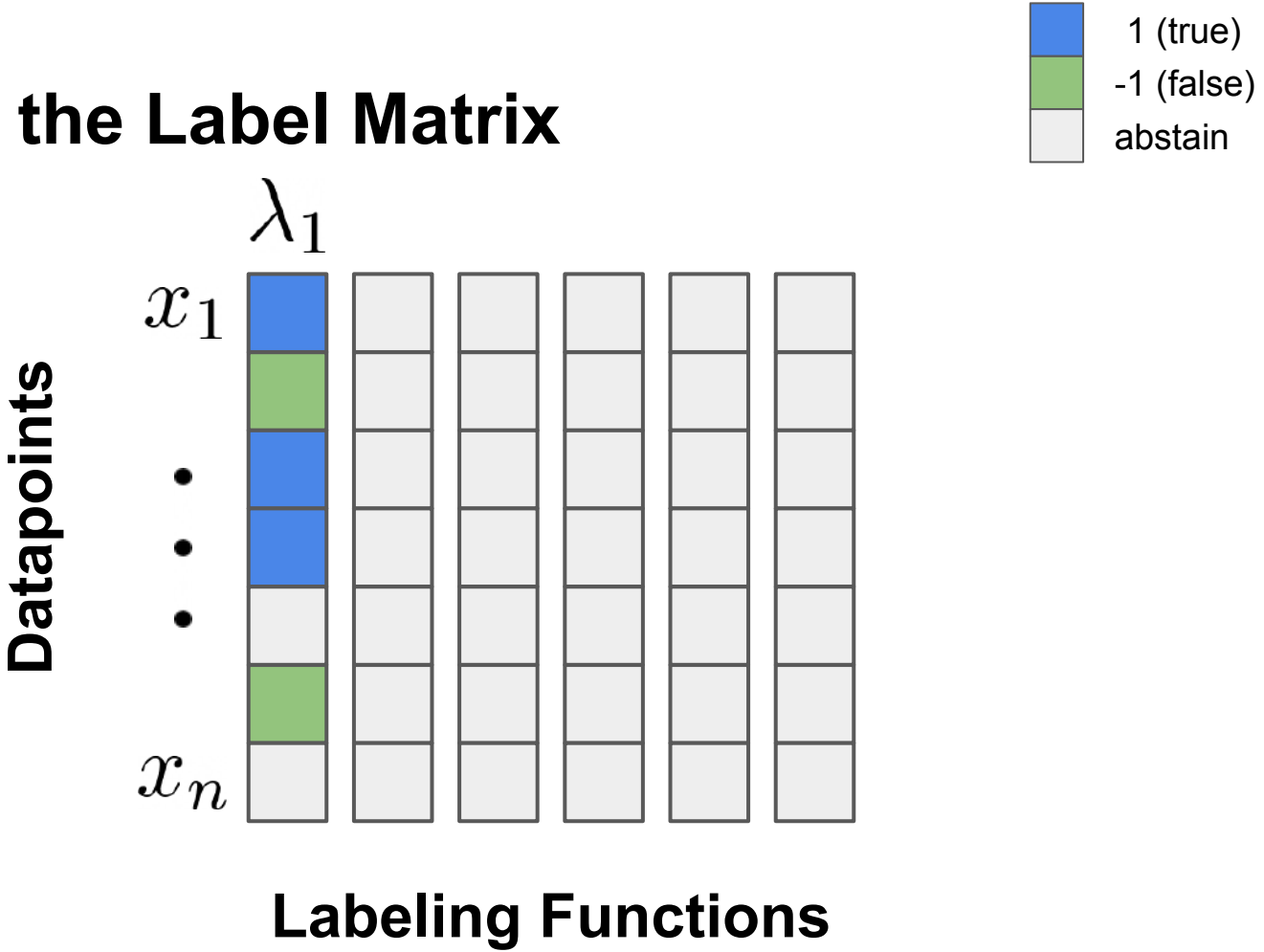
Data Model Example: Relation Extraction



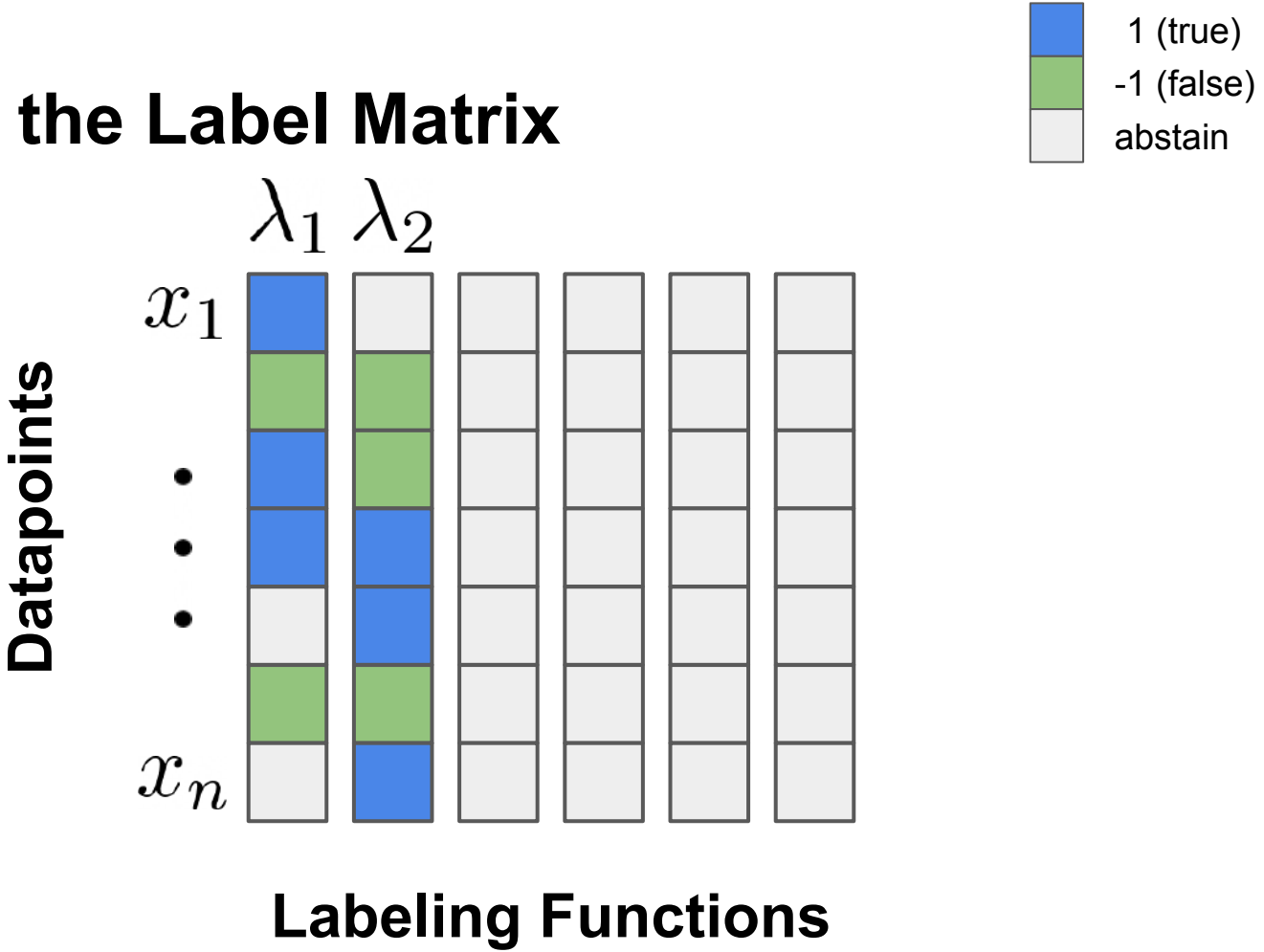
Candidate 1: x_1 := Causes (magnesium, quadriplegic)

Candidate 2: x_2 := Causes (magnesium, preeclampsia)

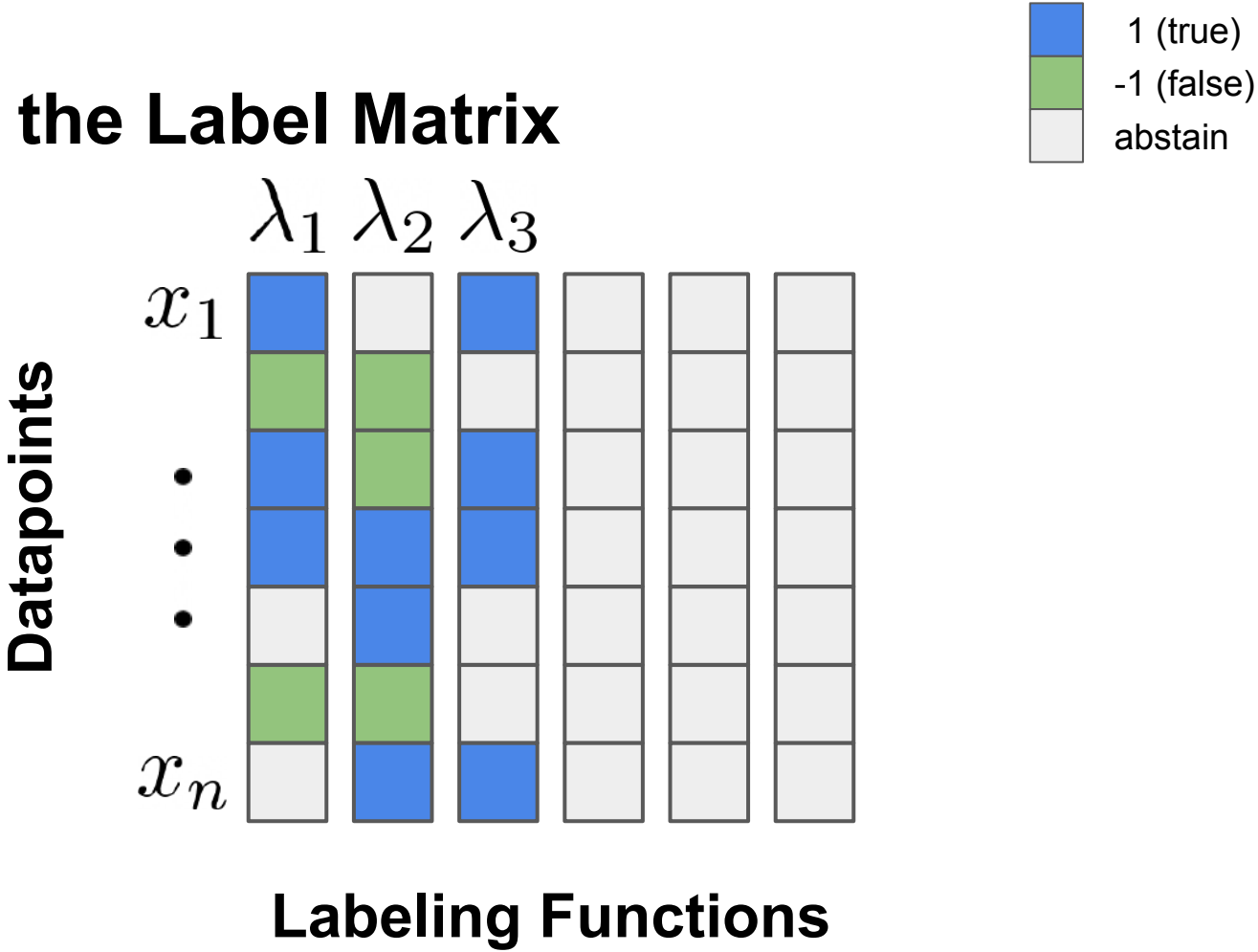
Building the Label Matrix



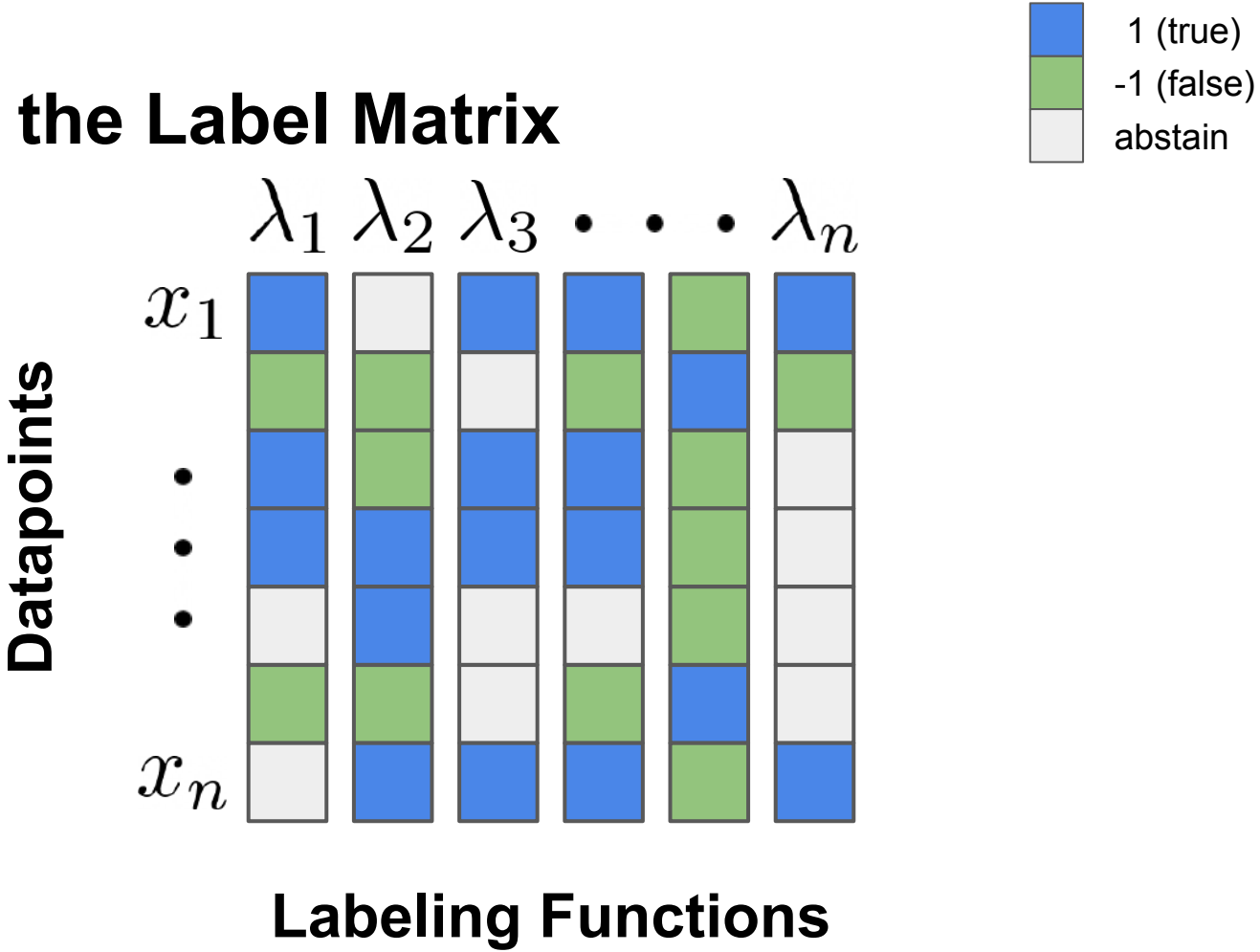
Building the Label Matrix



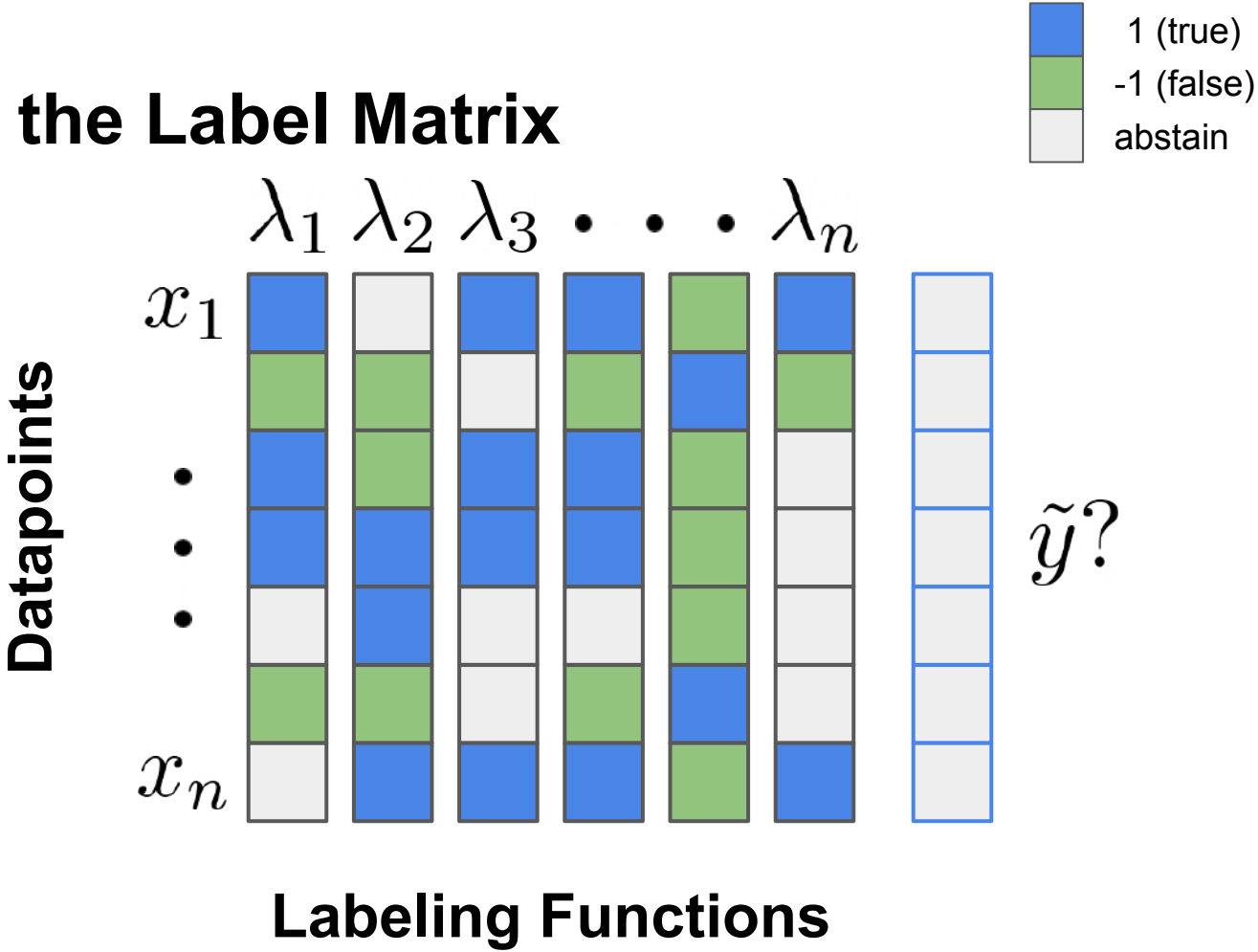
Building the Label Matrix



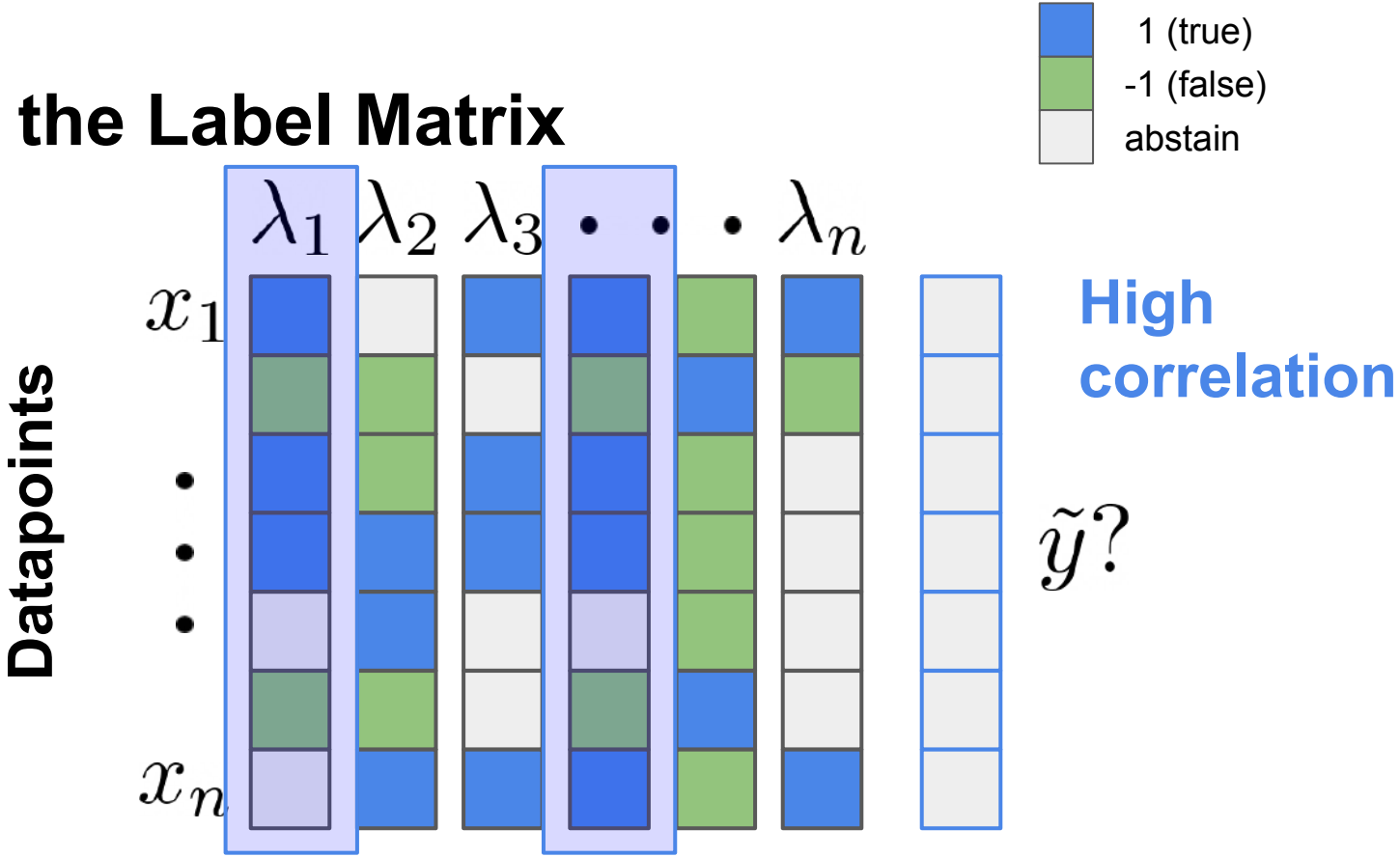
Building the Label Matrix



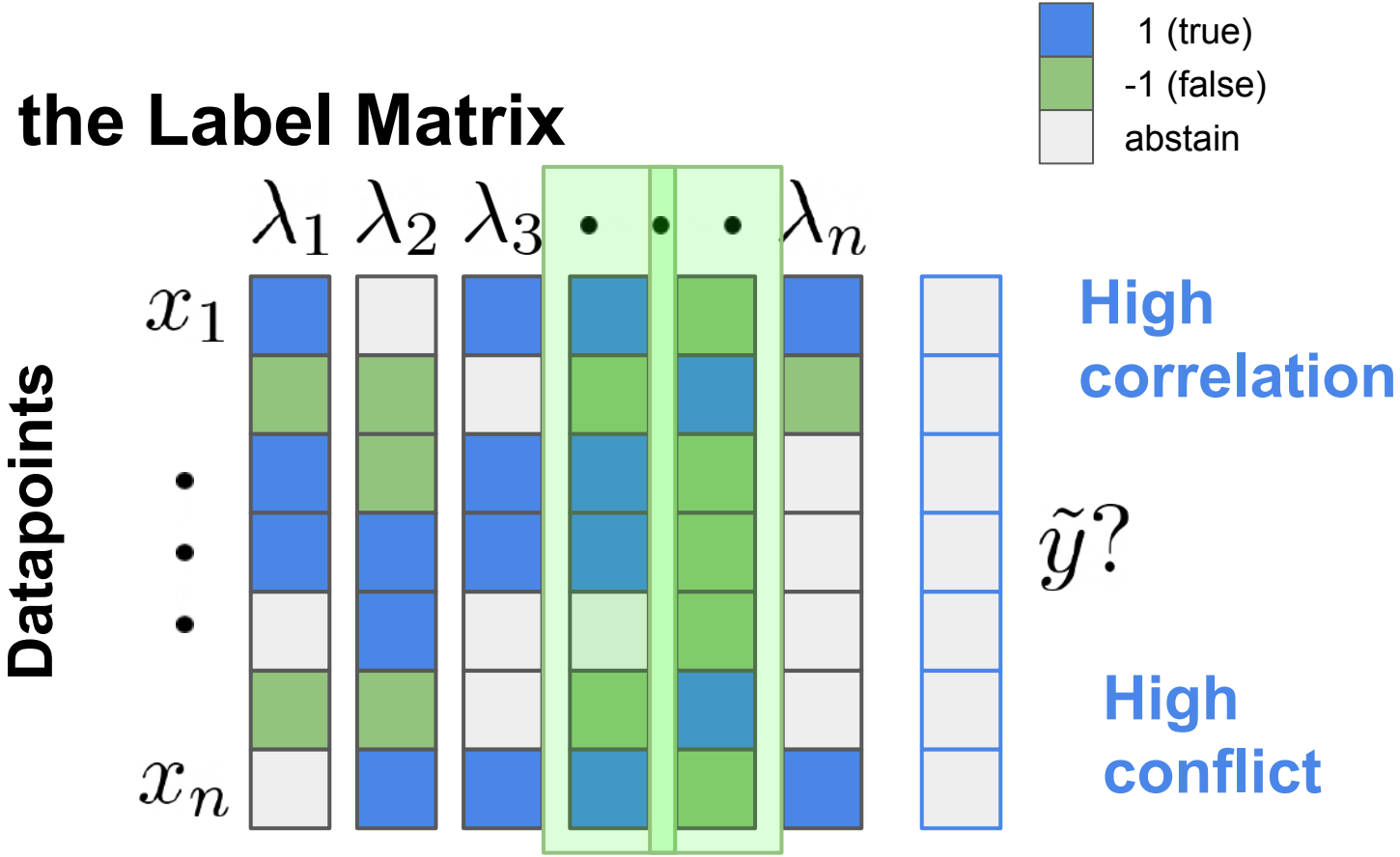
Building the Label Matrix



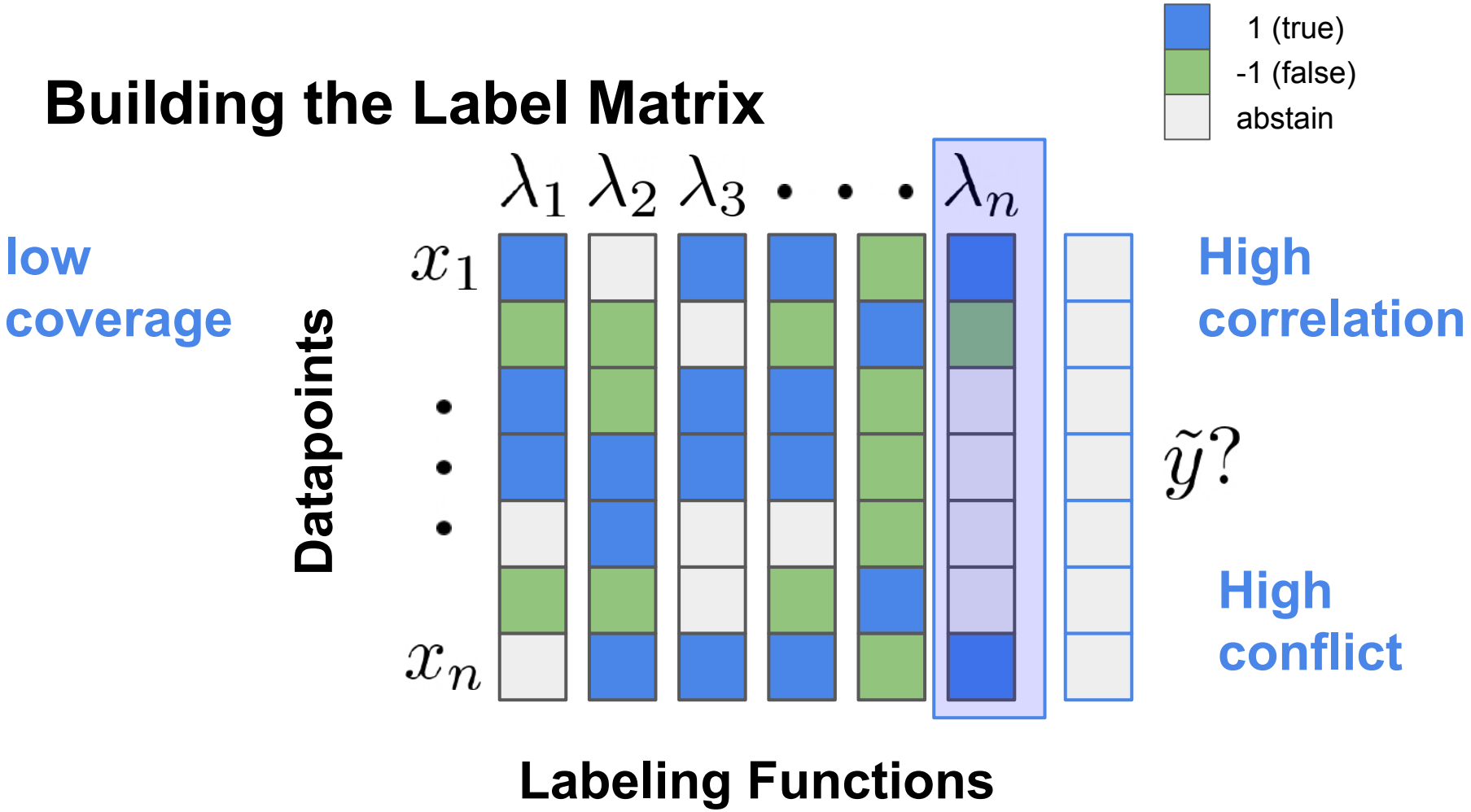
Building the Label Matrix



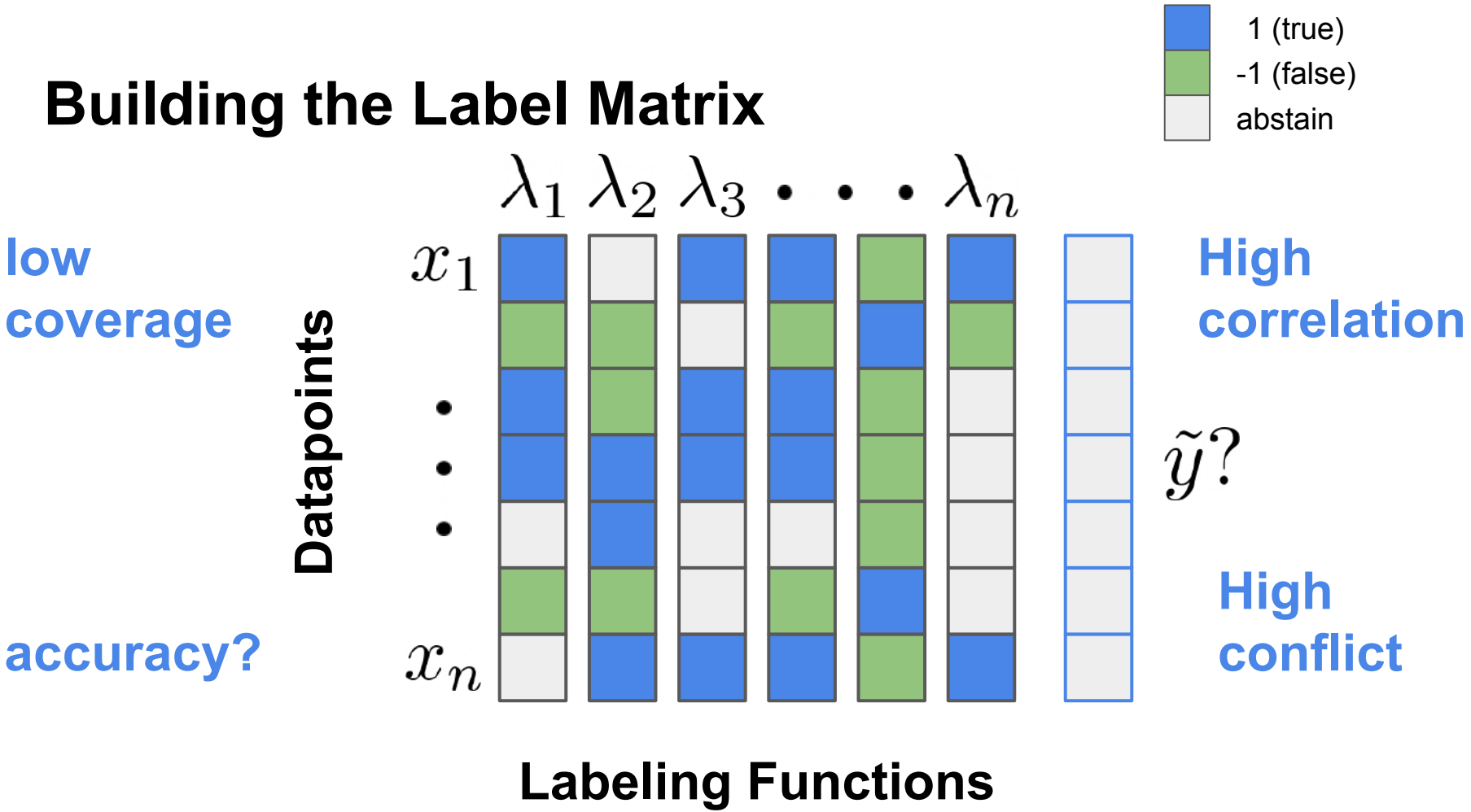
Building the Label Matrix



Building the Label Matrix

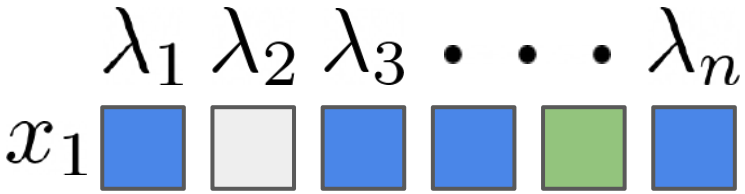


Building the Label Matrix

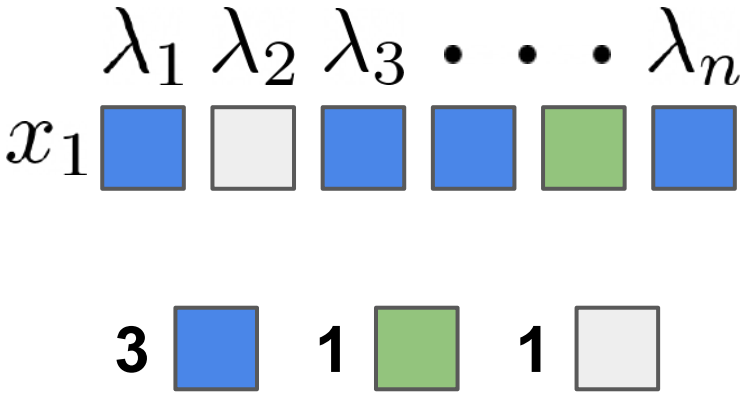


Step 2: Generative Label Model

How to aggregate labeling functions?

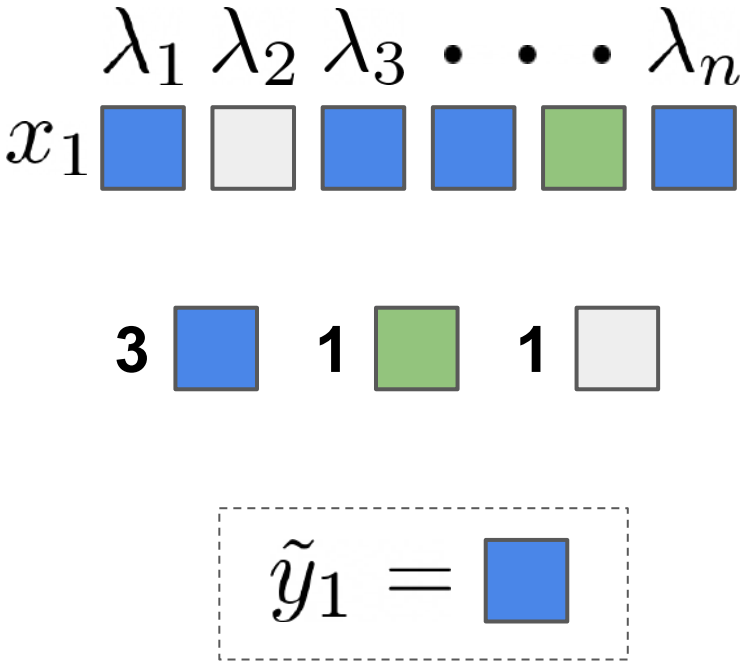


How to aggregate labeling functions?



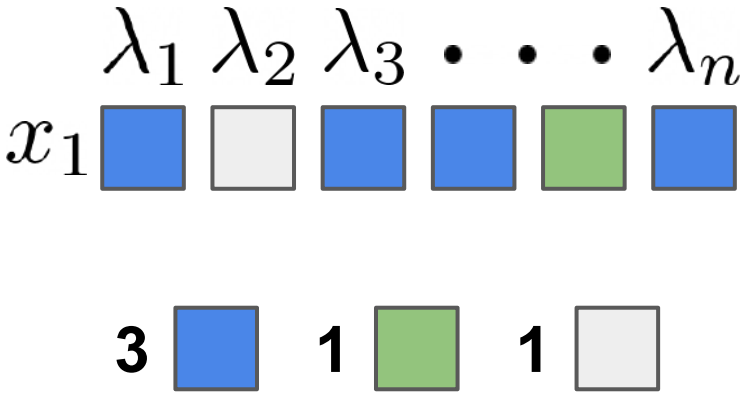
Majority Vote

How to aggregate labeling functions?



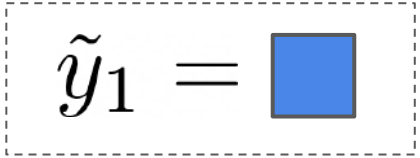
Majority Vote

How to aggregate labeling functions?

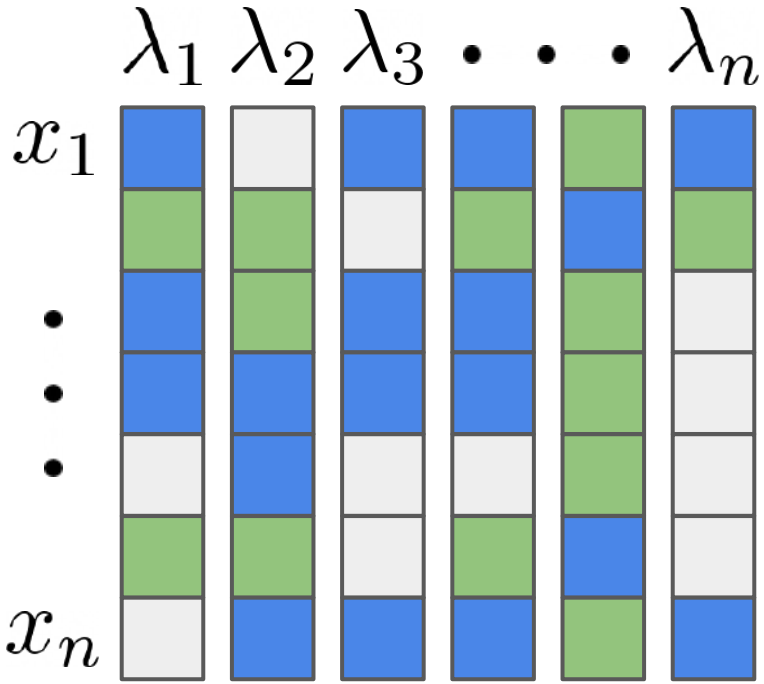


Majority Vote

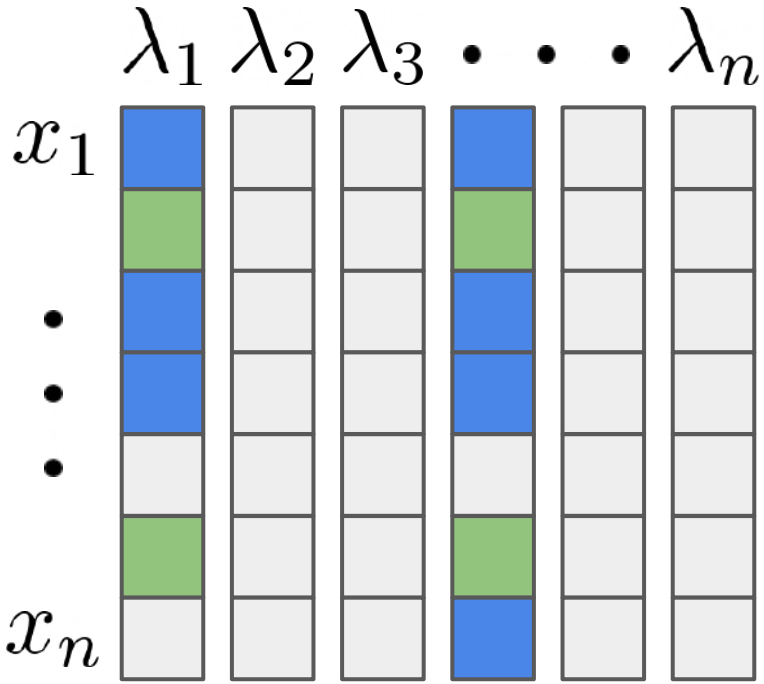
Problems?



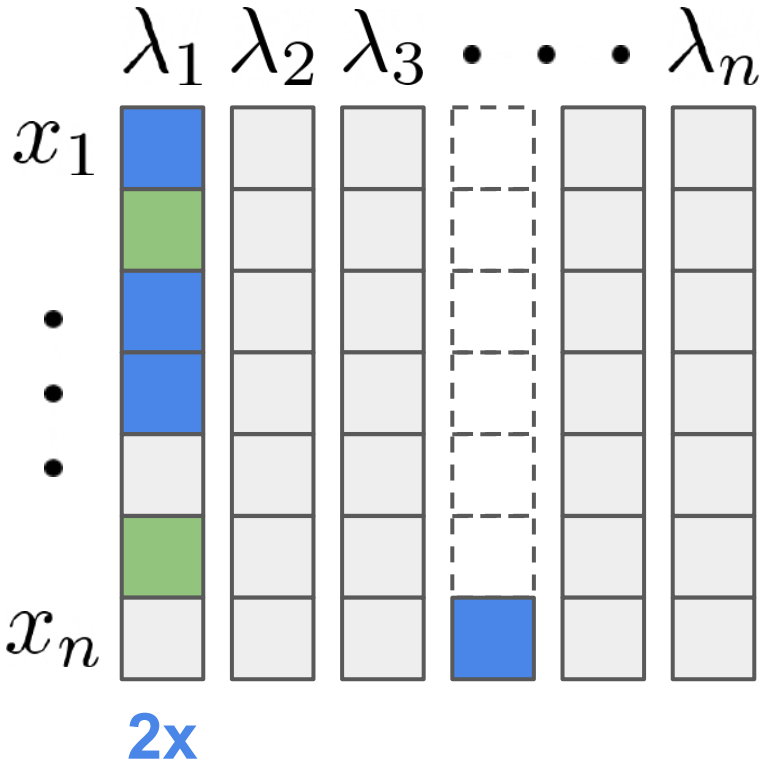
Majority Vote Issue 1: Double Counting



Majority Vote Issue 1: Double Counting

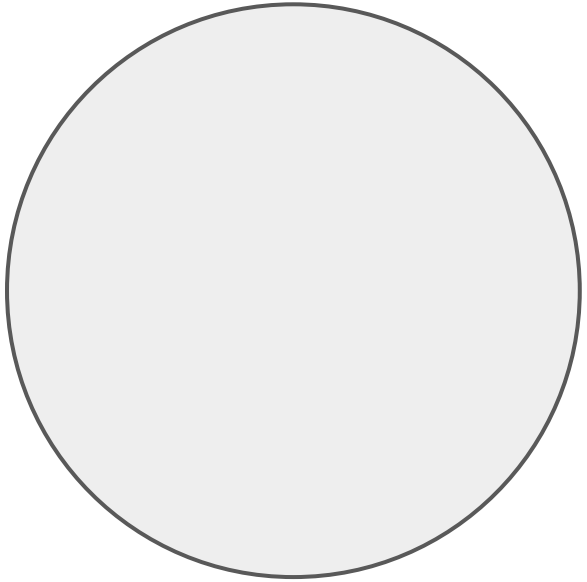


Majority Vote Issue 1: Double Counting

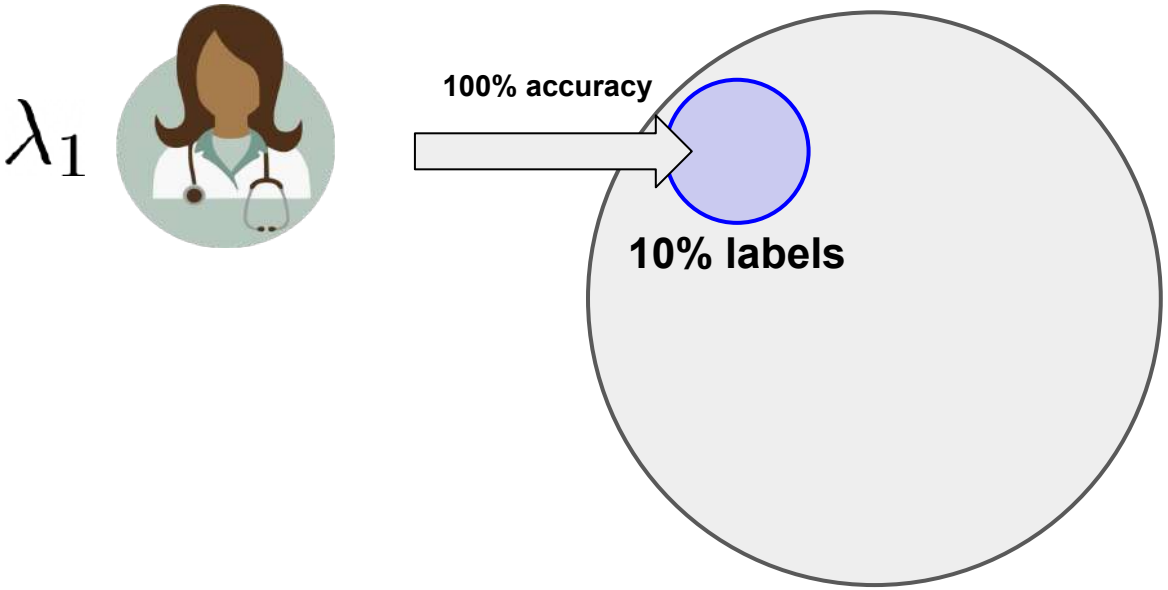


Majority Vote Issue 2: Variable Coverage, Accuracy

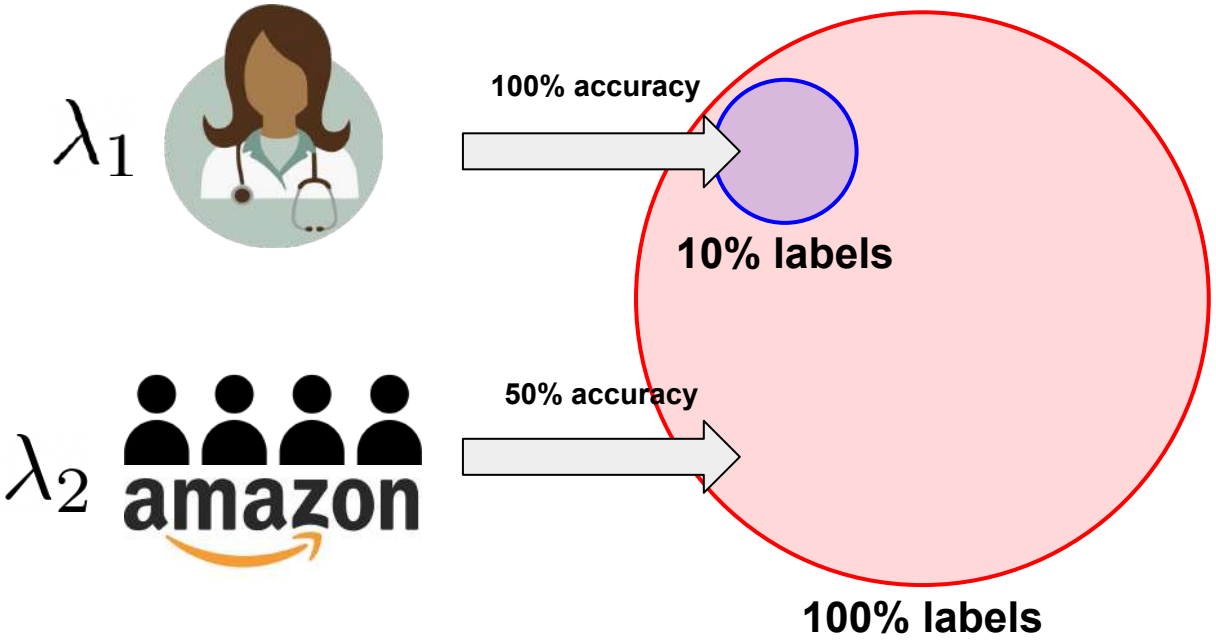
Majority Vote Issue 2: Variable Coverage, Accuracy



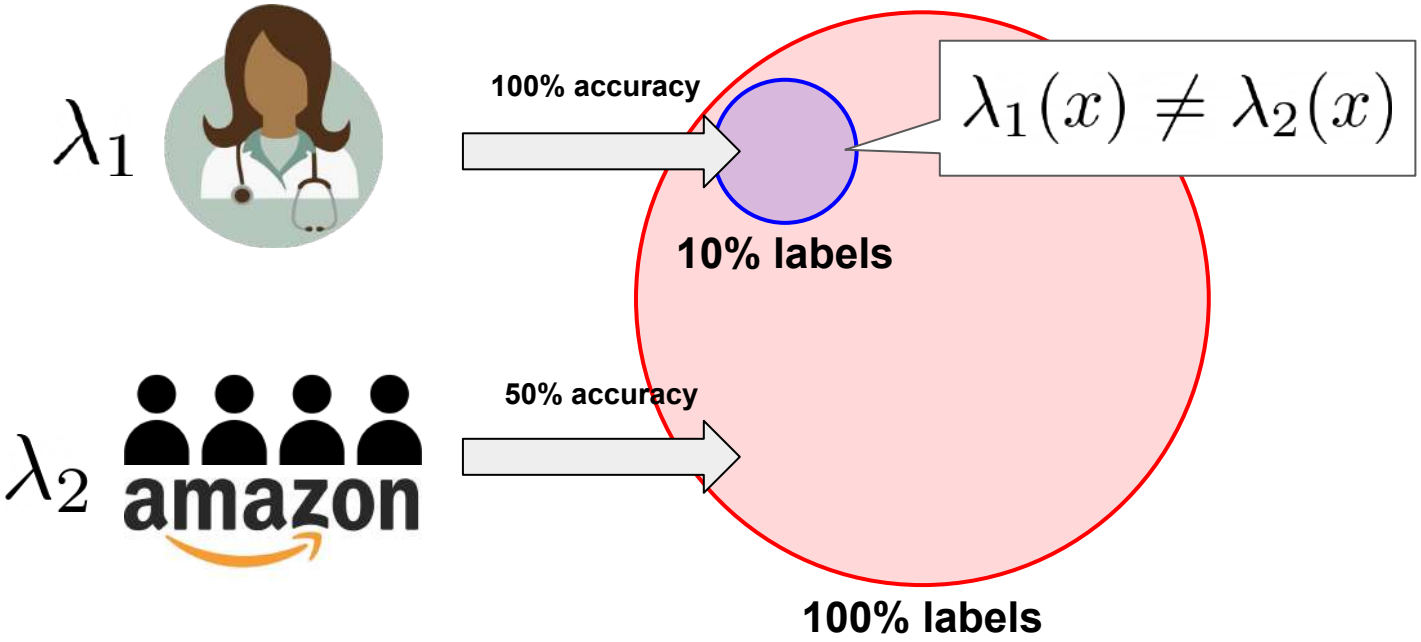
Majority Vote Issue 2: Variable Coverage, Accuracy



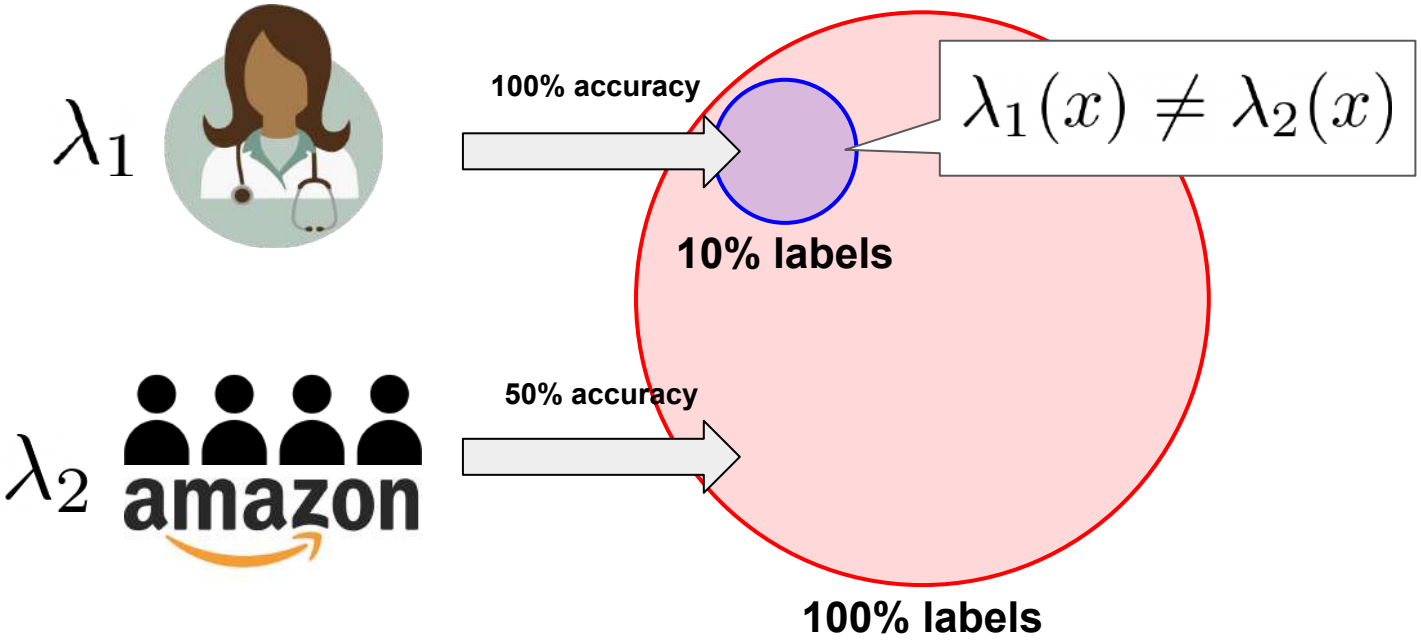
Majority Vote Issue 2: Variable Coverage, Accuracy



Majority Vote Issue 2: Variable Coverage, Accuracy

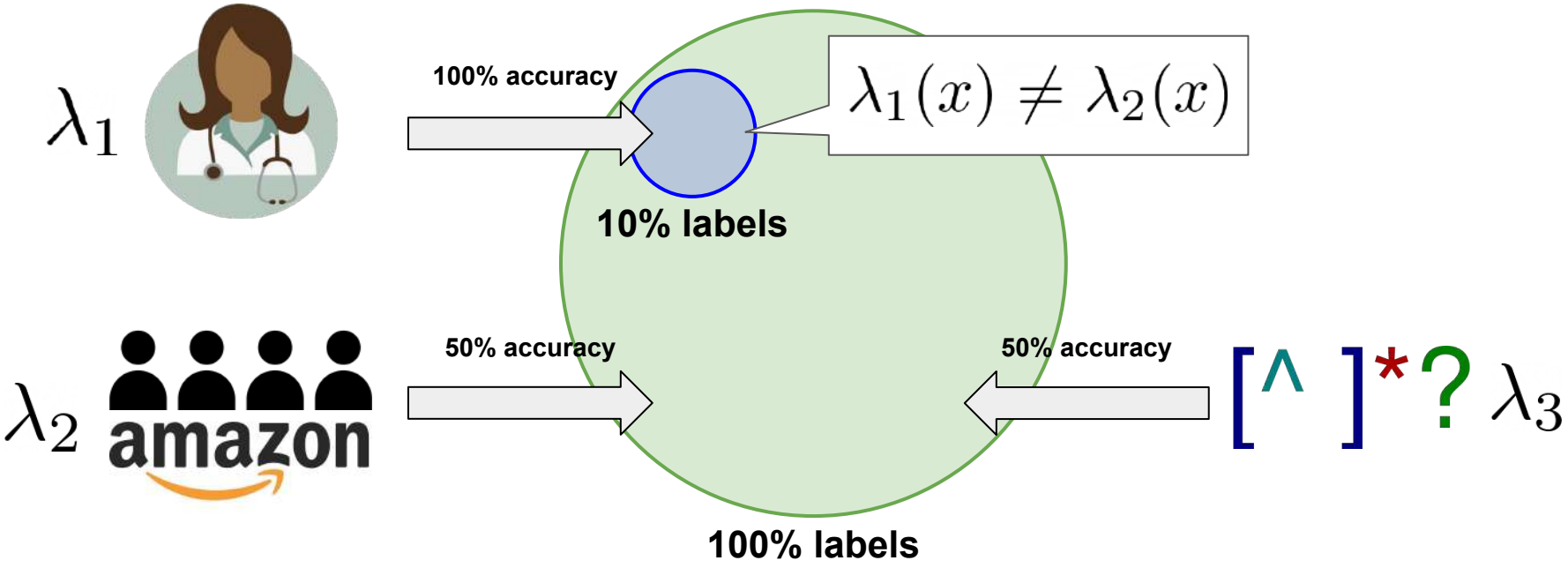


Majority Vote Issue 2: Variable Coverage, Accuracy

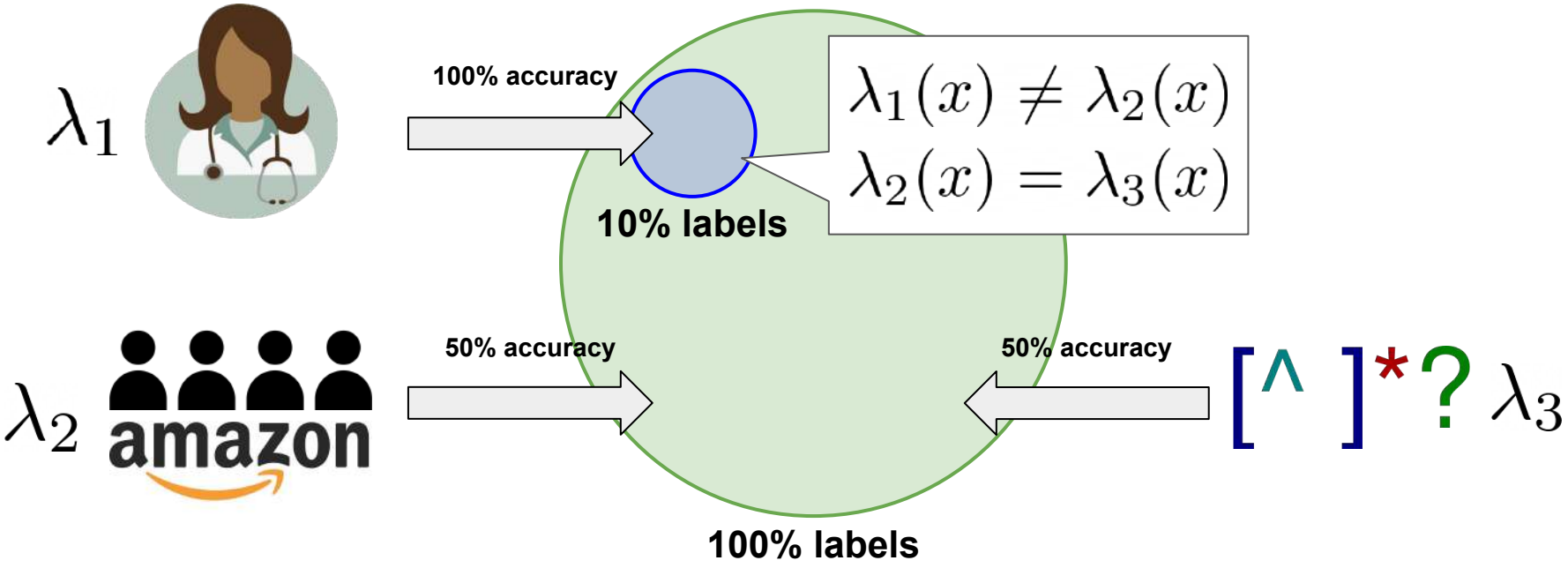


50 - 50?

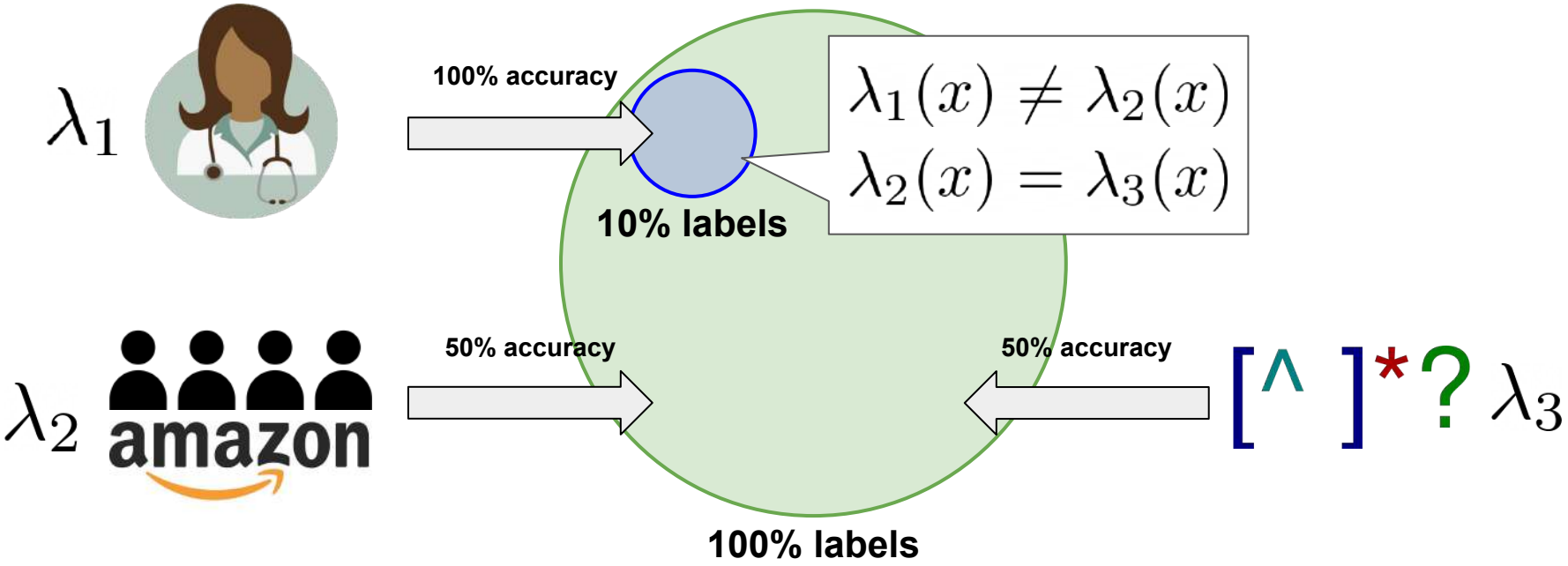
Majority Vote Issue 2: Variable Coverage, Accuracy



Majority Vote Issue 2: Variable Coverage, Accuracy



Majority Vote Issue 2: Variable Coverage, Accuracy



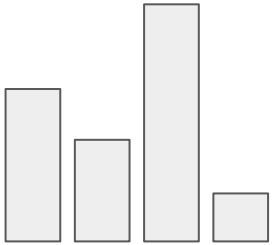
Defer to less accurate?

Alternative: Modeling Latent Accuracies

Unweighted Majority Vote

Alternative: Modeling Latent Accuracies

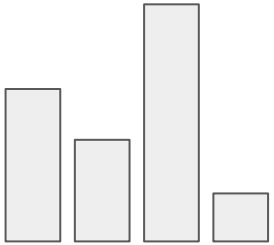
Unweighted Majority Vote



Weighted by Accuracies

Alternative: Modeling Latent Accuracies

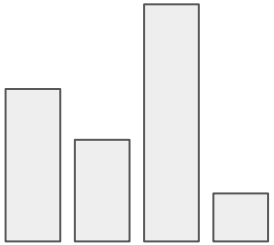
Unweighted Majority Vote



Weighted by Accuracies

Alternative: Modeling Latent Accuracies

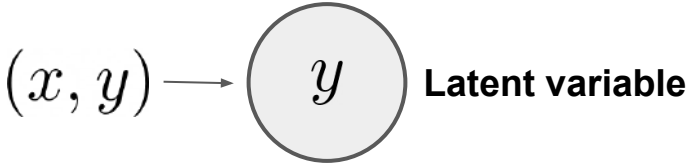
Unweighted Majority Vote



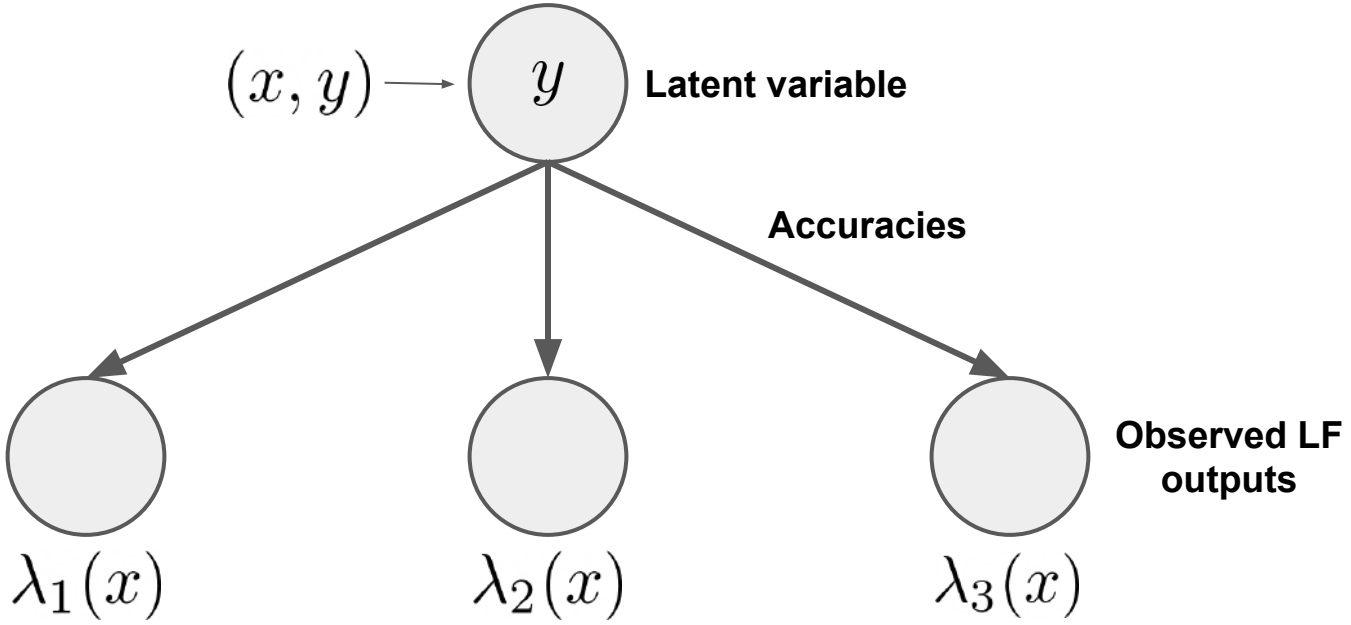
Weighted by Accuracies

How to model accuracies without ground-truth labels?

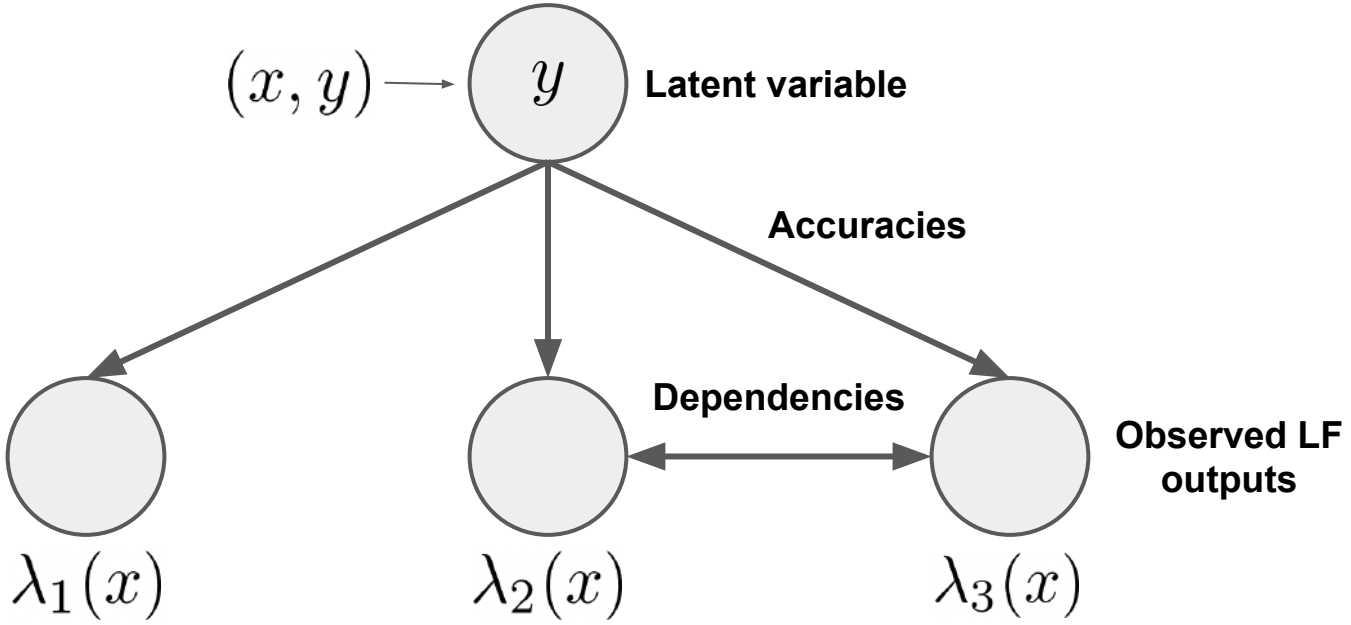
Generative Label Model



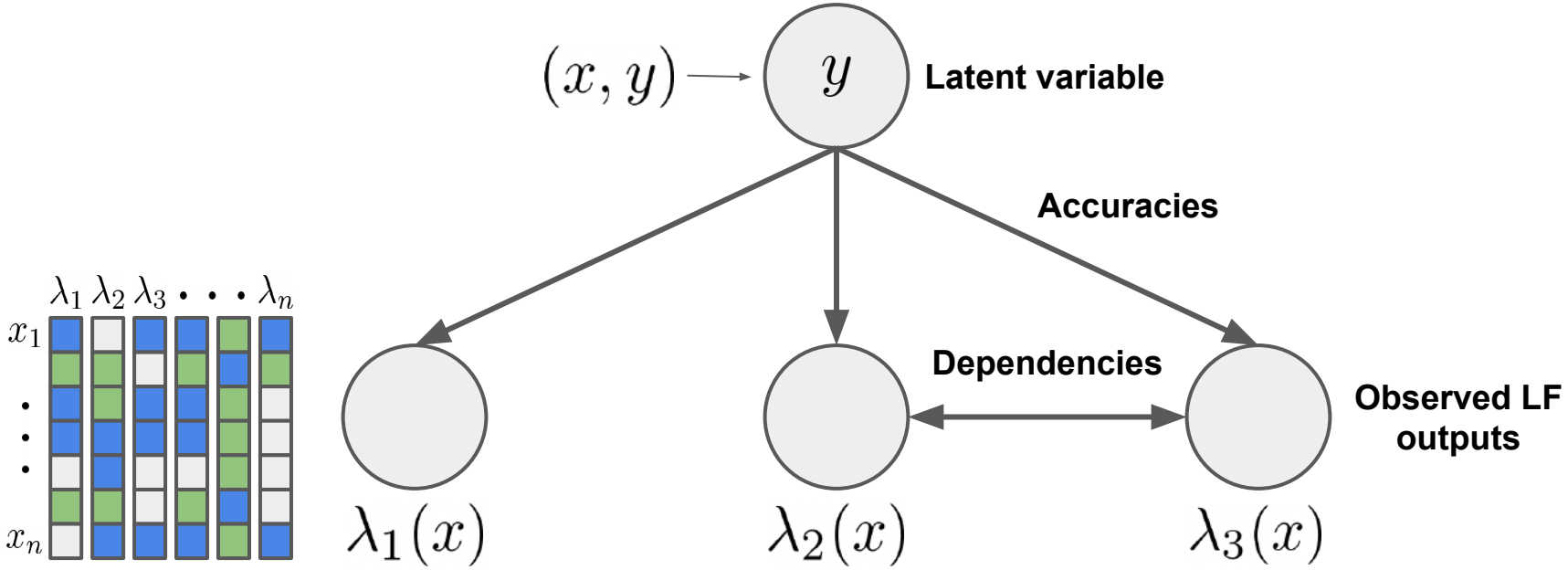
Generative Label Model



Generative Label Model



Generative Label Model

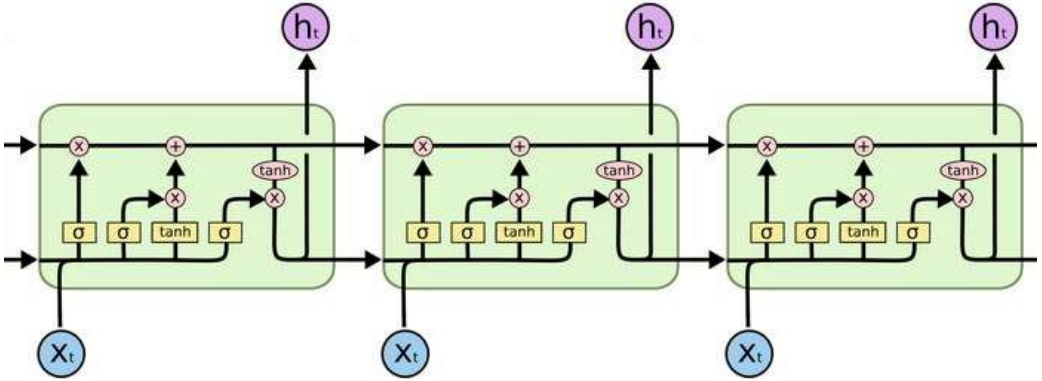


Try to maximize probability of LF outputs given generative model

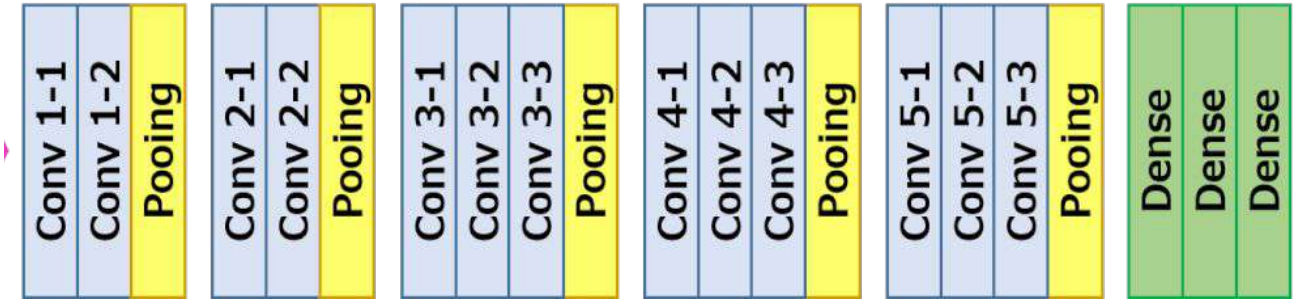
Step 3: Discriminative End Model

Training a Discriminative Model

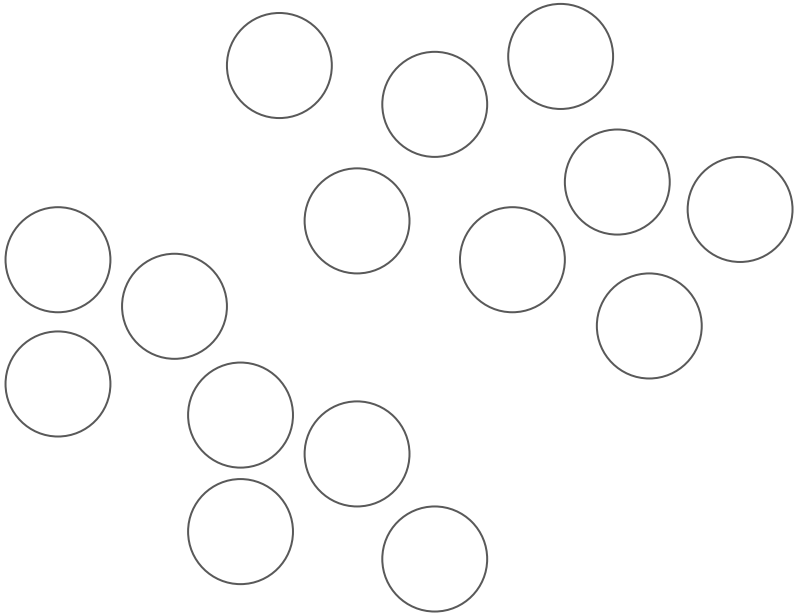
LSTM



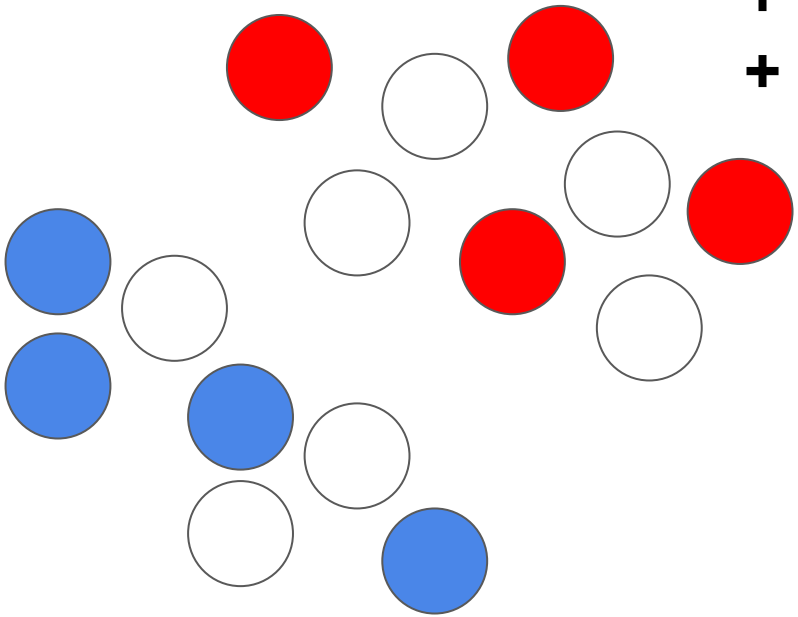
CNN



Why Add A Discriminative Model?



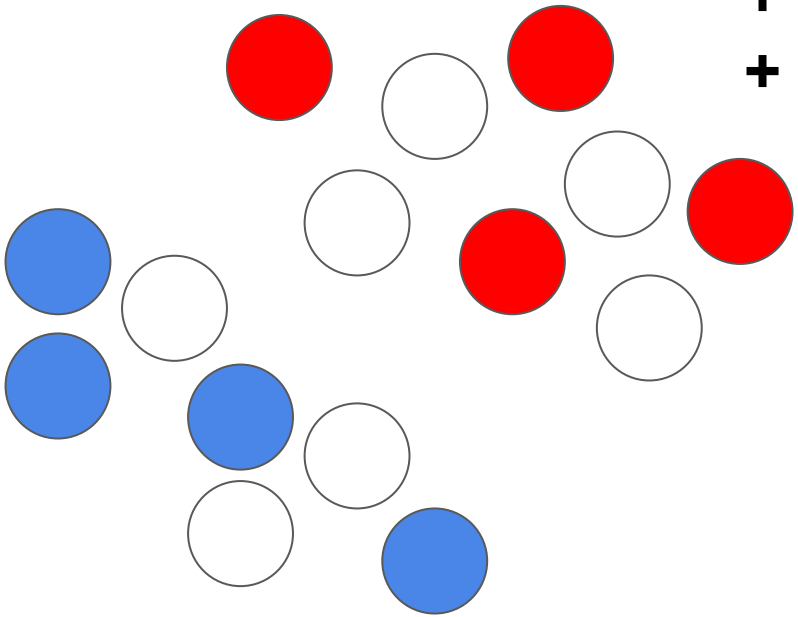
Why Add A Discriminative Model?



+ labeling functions
+ generative model

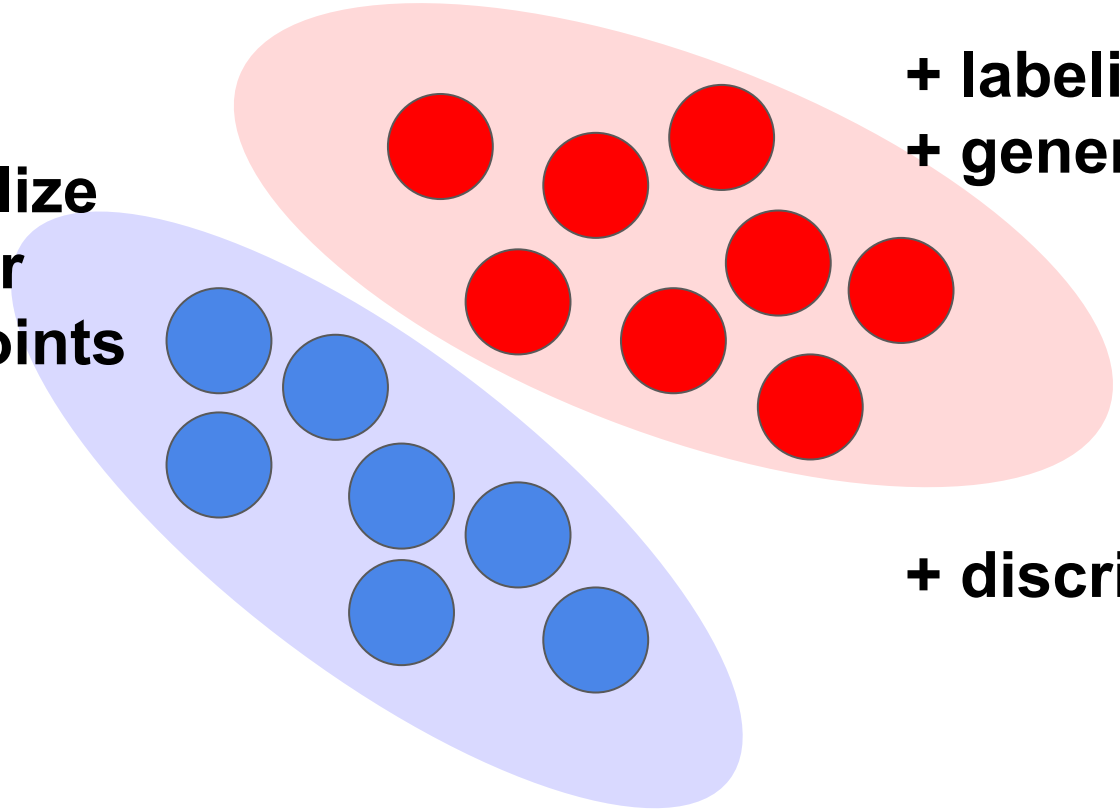
Why Add A Discriminative Model?

still
potentially
sparse



+ labeling functions
+ generative model

Why Add A Discriminative Model?



+ labeling functions
+ generative model

generalize
to other
data points

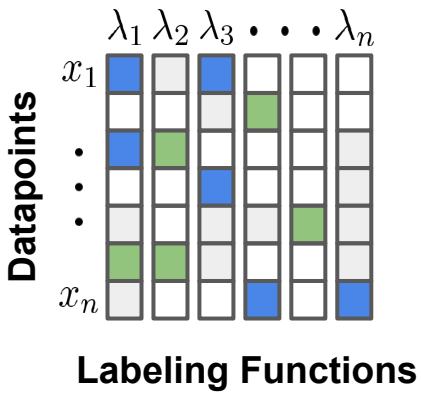
+ discriminative model

Label Model Tradeoffs

Modeling Label Function Accuracies

Unweighted Majority Vote or More Complicated Model?

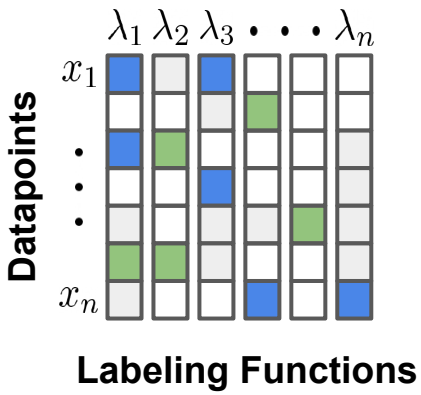
Modeling Label Function Accuracies



Low Density

Modeling Label Function Accuracies

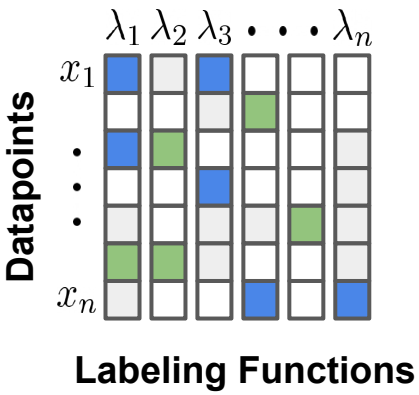
Little Disagreement

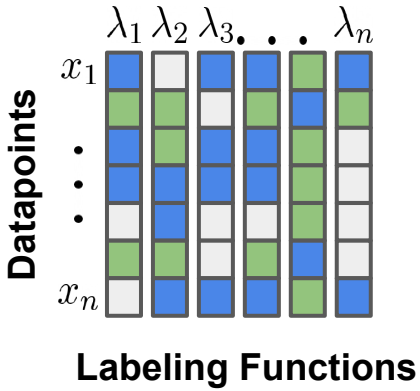


Low Density

Modeling Label Function Accuracies

Little Disagreement



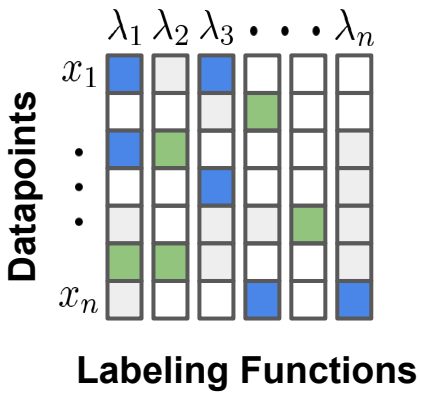


Low Density

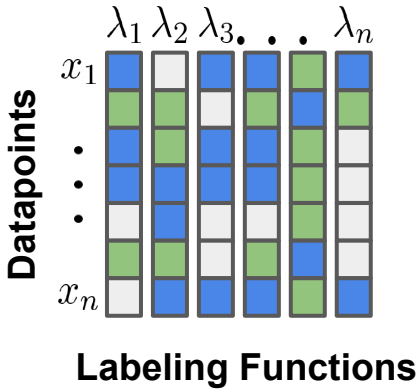
High Density

Modeling Label Function Accuracies

Little Disagreement



Consensus Reached

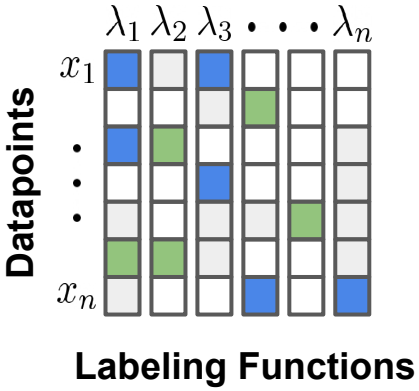


Low Density

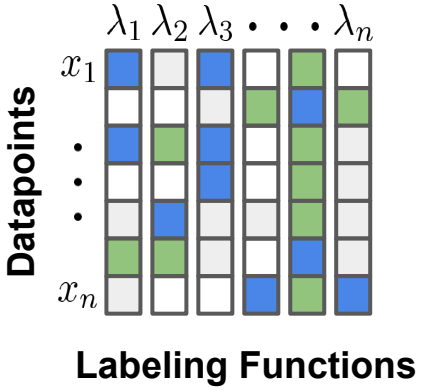
High Density

Modeling Label Function Accuracies

Little Disagreement

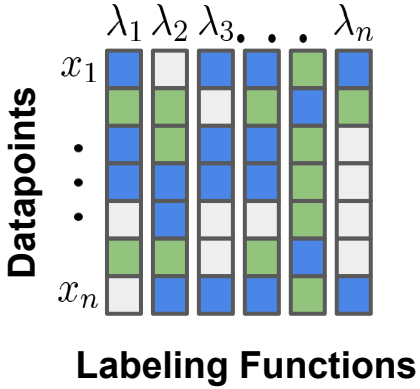


Low Density



Medium Density

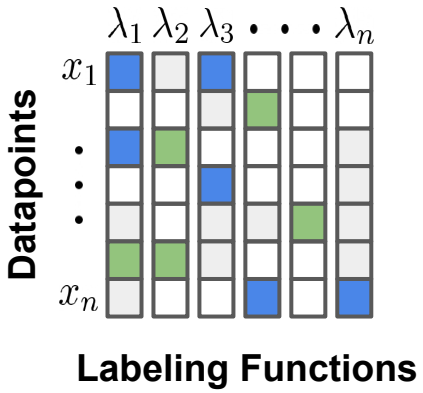
Consensus Reached



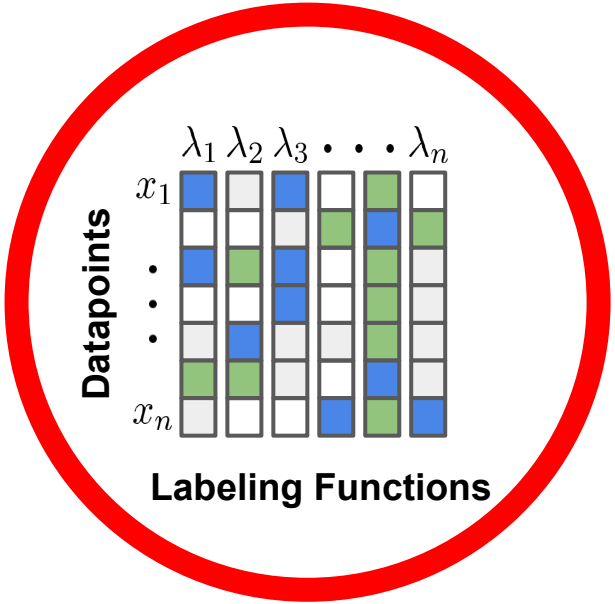
High Density

Modeling Label Function Accuracies

Little Disagreement

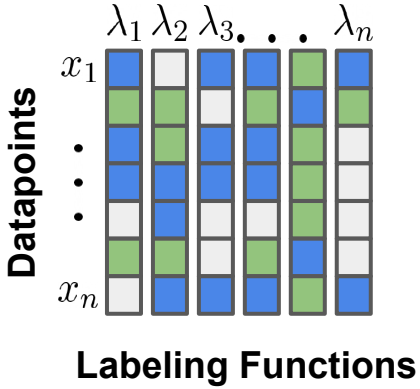


Low Density



Medium Density

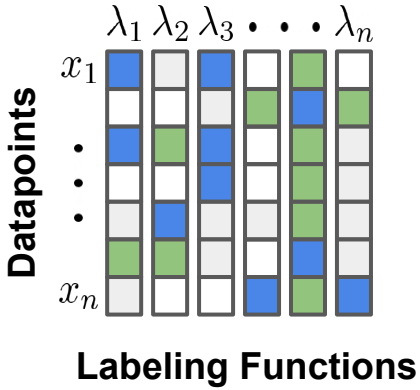
Consensus Reached



High Density

Modeling Label Function Accuracies

Modeling Helps!

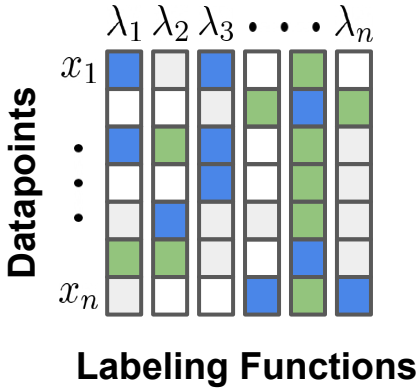


What is medium density?

Medium Density

Modeling Label Function Accuracies

Modeling Helps!



Medium Density

What is medium density?

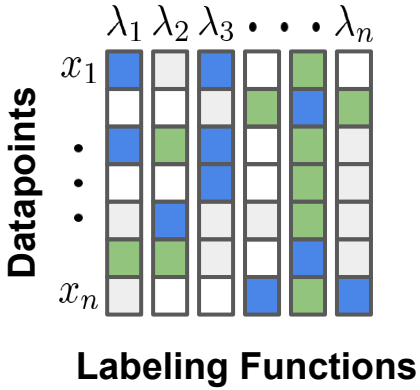
How many labels per datapoint?

or

Could a weighted vote flip an incorrect unweighted majority vote?

Modeling Label Function Accuracies

Modeling Helps!



Medium Density

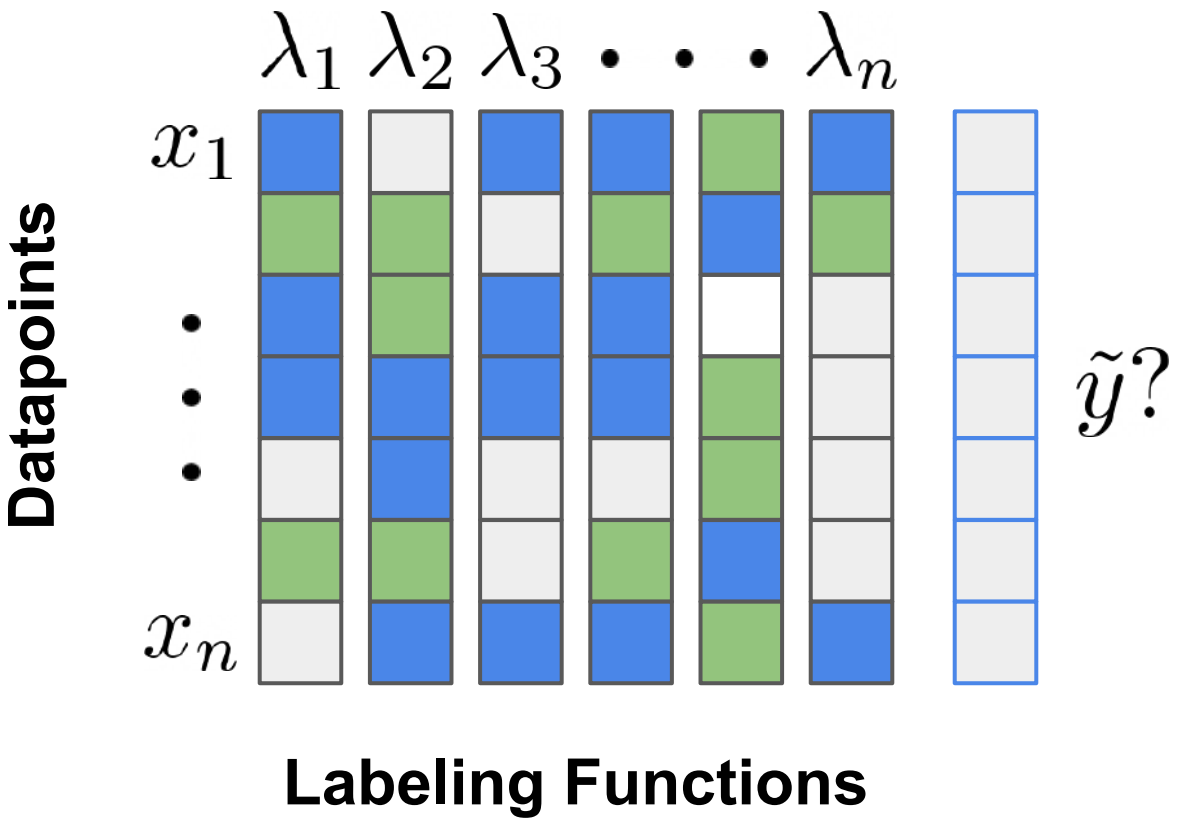
What is medium density?

How many labels per datapoint?

or

Could a weighted vote flip an incorrect unweighted majority vote?

Modeling Label Function Correlations

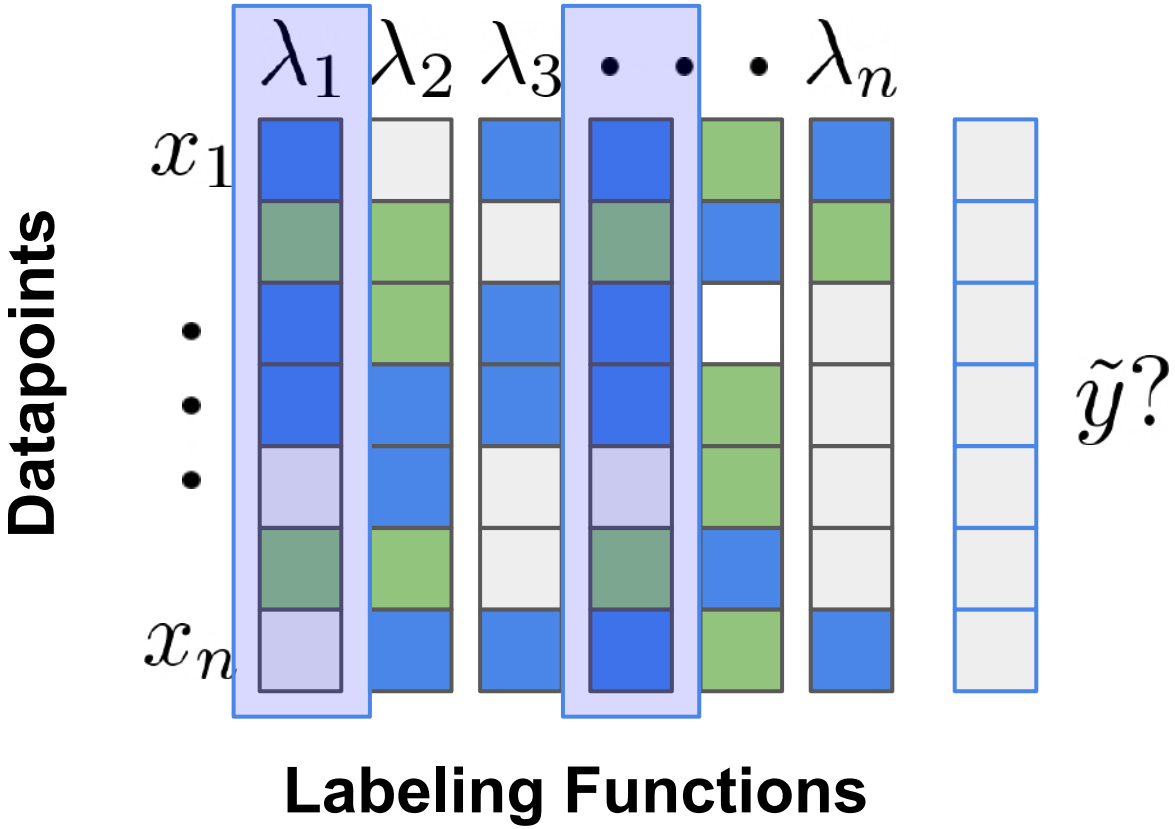


ε = 1

Correlations: 0

Predictive Performance: 20

Modeling Label Function Correlations

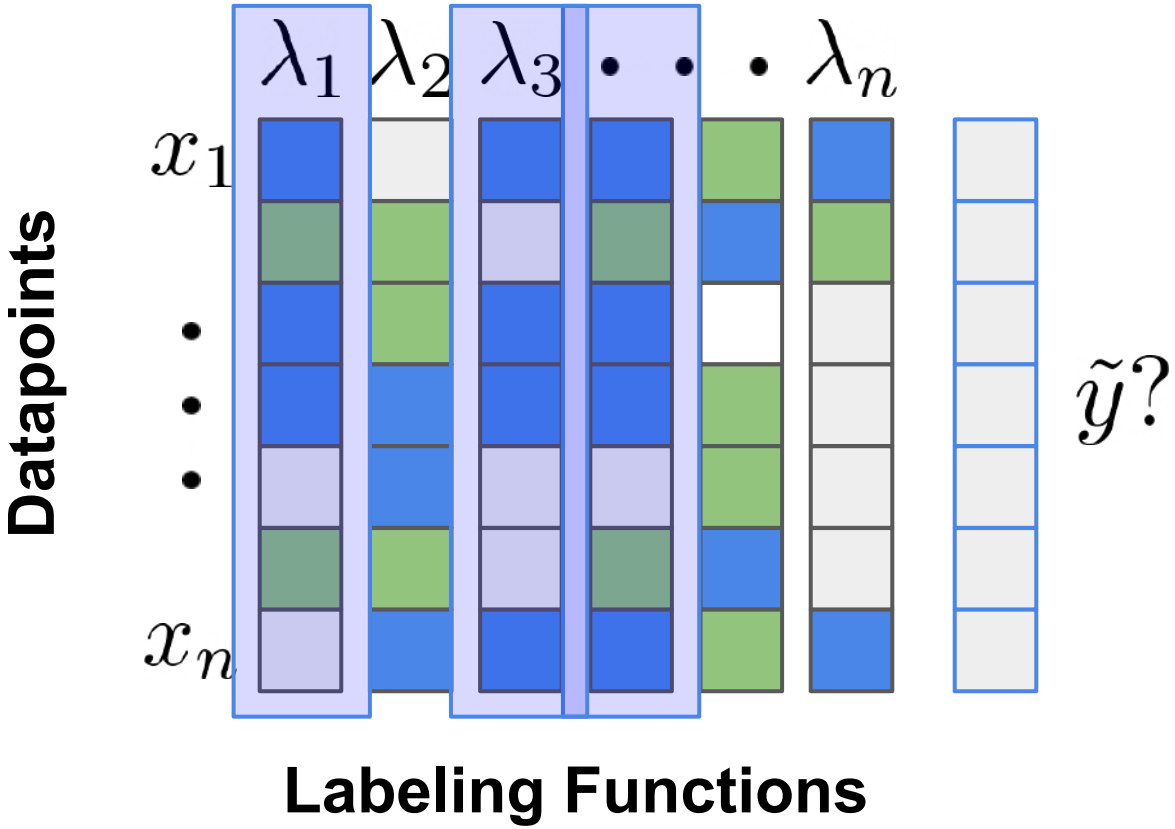


ε = 0.85

Correlations: 1

Predictive Performance: 40

Modeling Label Function Correlations

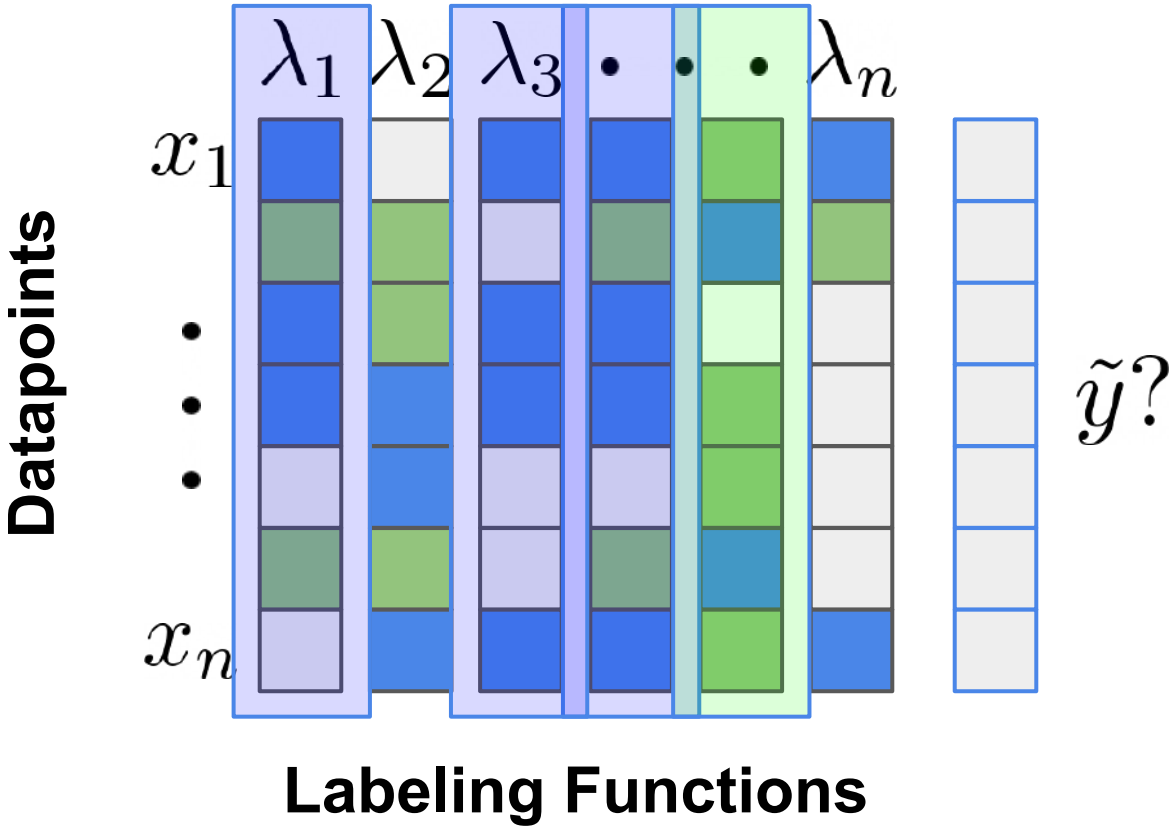


ε = 0.7

Correlations: 2

Predictive Performance: 60

Modeling Label Function Correlations

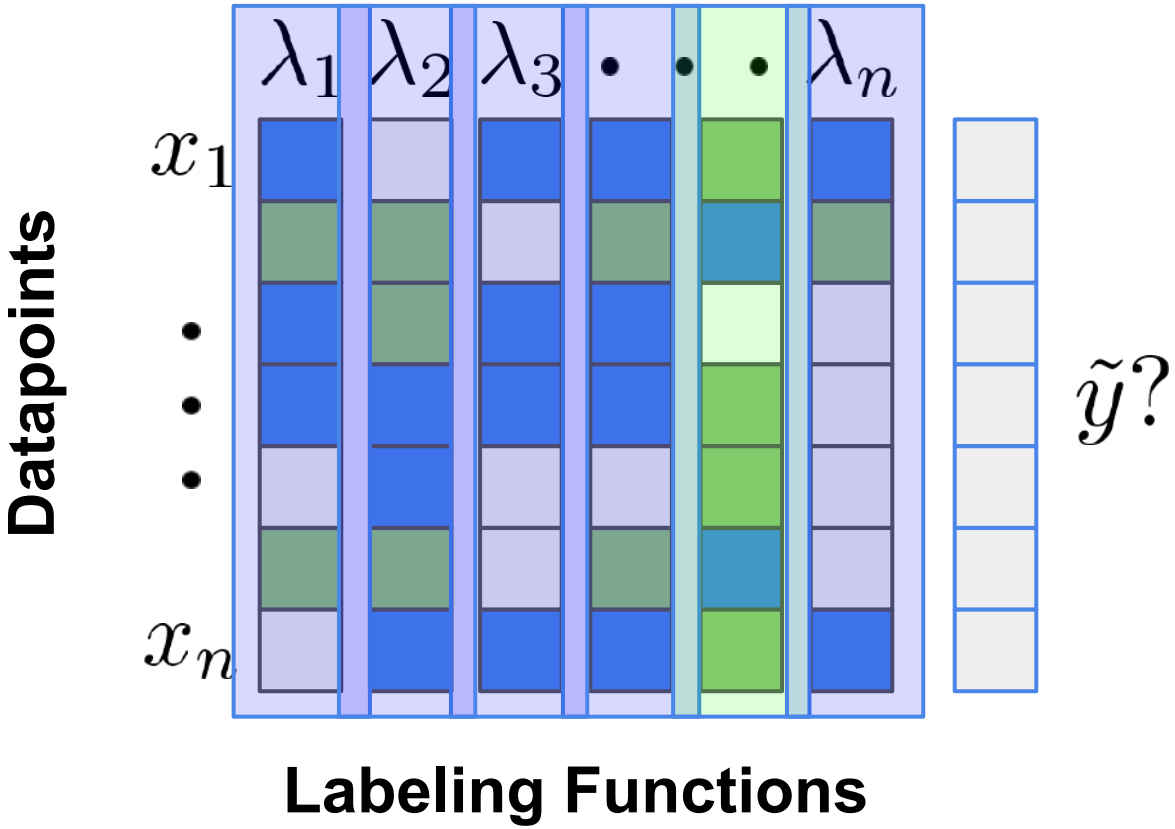


ε = 0.55

Correlations: 3

Predictive Performance: 80

Modeling Label Function Correlations

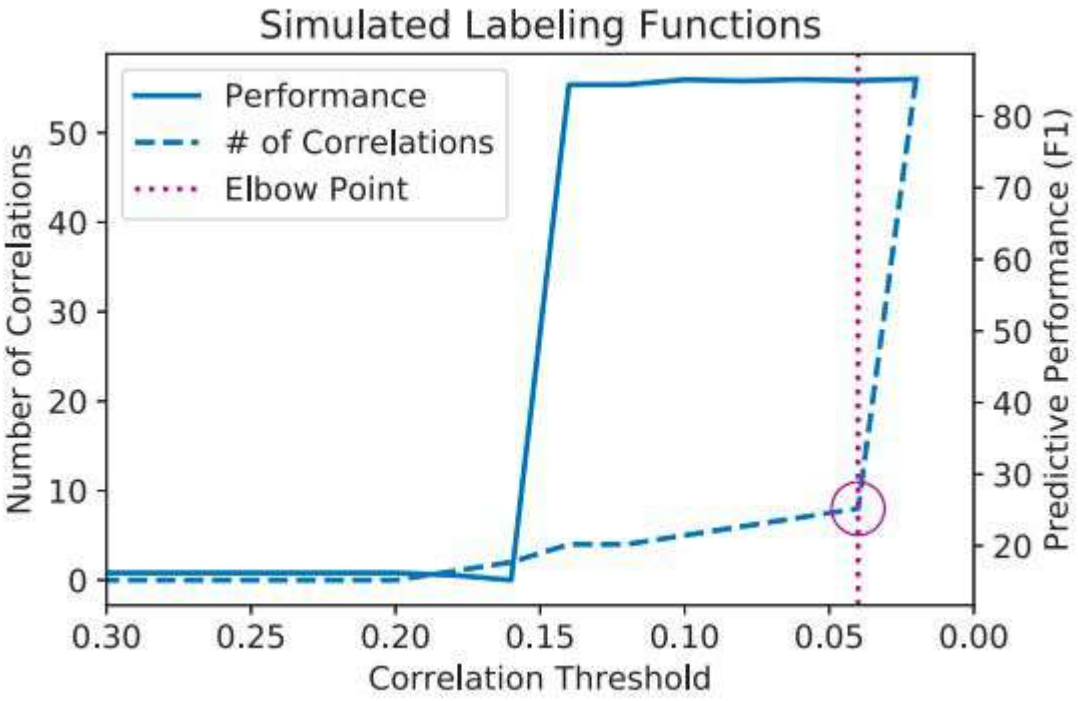


ε = 0.4

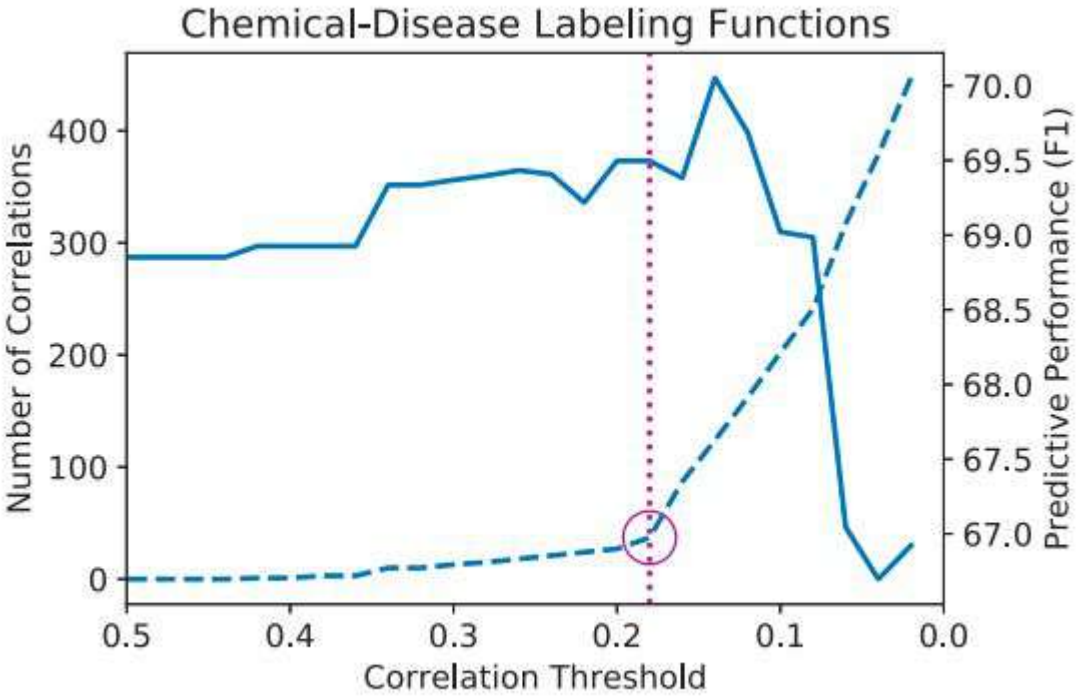
Correlations: 15

Predictive Performance: 81

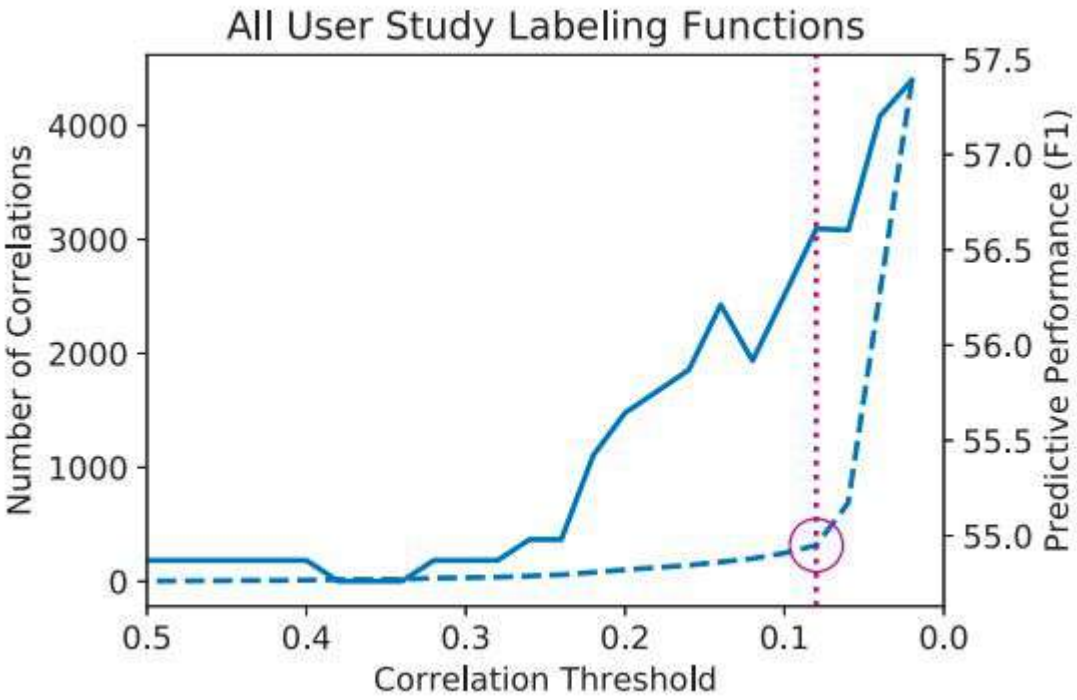
Modeling Label Function Correlations



Modeling Label Function Correlations

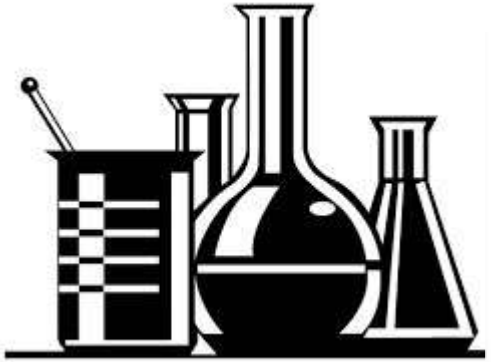


Modeling Label Function Correlations



Results

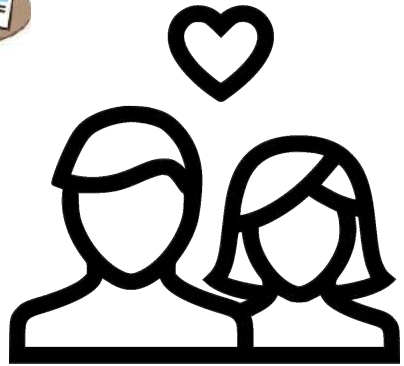
Relation Extraction from Text



Scientific Articles

Electronic Health Records

Chemical-Disease relations



Spouses

Relation Extraction from Text

	Distant Supervision			Snorkel (Gen.)				Snorkel (Disc.)				Hand Supervision		
Task	P	R	**F1**	P	R	**F1**	**Lift**	P	R	**F1**	**Lift**	P	R	**F1**
Chem	11.2	41.2	17.6	78.6	21.6	33.8	+16.2	87.0	39.2	54.1	+36.5	-	-	-
EHR	81.4	64.8	72.2	77.1	72.9	74.9	+2.7	80.2	82.6	81.4	+9.2	-	-	-
CDR	25.5	34.8	29.4	52.3	30.4	38.5	+9.1	38.8	54.3	45.3	+15.9	39.9	58.1	47.3
Spouses	9.9	34.8	15.4	53.5	62.1	57.4	+42.0	48.4	61.6	54.2	+38.8	47.8	62.5	54.2

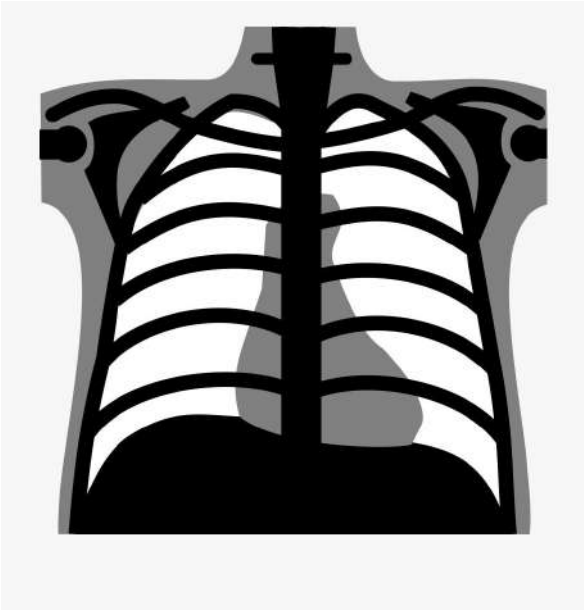
Snorkel exceeds models trained by distant supervision by an average of 132%

Relation Extraction from Text

	Distant Supervision			Snorkel (Gen.)				Snorkel (Disc.)				Hand Supervision		
Task	P	R	**F1**	P	R	**F1**	**Lift**	P	R	**F1**	**Lift**	P	R	**F1**
Chem	11.2	41.2	17.6	78.6	21.6	33.8	+16.2	87.0	39.2	54.1	+36.5	-	-	-
EHR	81.4	64.8	72.2	77.1	72.9	74.9	+2.7	80.2	82.6	81.4	+9.2	-	-	-
CDR	25.5	34.8	29.4	52.3	30.4	38.5	+9.1	38.8	54.3	45.3	+15.9	39.9	58.1	47.3
Spouses	9.9	34.8	15.4	53.5	62.1	57.4	+42.0	48.4	61.6	54.2	+38.8	47.8	62.5	54.2

Snorkel comes within 2.11% of the F1 score of hand supervision

Cross-Modal Experiments



Abnormality Detection in Lung Radiographs

Crowdsourcing

Cross-Modal Experiments

Task	Snorkel (Disc.)	Hand Supervision
Radiology (AUC)	72.0	76.2
Crowd (Acc)	65.6	68.8

Snorkel comes within 5.08% of of
the accuracy of hand supervision

Effect of Generative Modelling

Task	Disc. Model on Unweighted LFs	Disc. Model	Lift
Chem	48.6	54.1	+5.5
EHR	80.9	81.4	+0.5
CDR	42.0	45.3	+3.3
Spouses	52.8	54.2	+1.4
Crowd (Acc)	62.5	65.6	+3.1
Rad. (AUC)	67.0	72.0	+5.0

Improvement of 5.81% on average

Impact of Labeling Functions

LF Type	P	R	F1	Lift
Text Patterns	42.3	42.4	42.3	
+ Distant Supervision	37.5	54.1	44.3	+2.0
+ Structure-based	38.8	54.3	45.3	+1.0

Adding different types of labeling functions improves predictive performance

User Study

How quickly can users learn to write labelling functions?

Is writing labelling functions more time-efficient than hand-labelling data?

Future Directions?