Theoretical Analysis of Incidental Supervision

Indirectly Supervised Natural Language Processing (Part IV)

Qiang Ning
AWS AI Labs

July 2023

ACL Tutorials

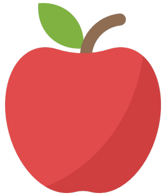
Indirectly Supervised Natural Language Processing

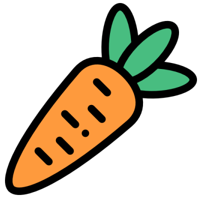
What to expect

- We pose the challenge to define a principled way to measure the benefits of these signals to a given downstream task, and the challenge to further understand why and how these signals can help reduce the complexity of the learning problem in theory.
- Main papers
  - [EMNLP'21] Foreseeing the Benefits of Incidental Supervision
  - [NeurIPS'20] Learnability with Indirect Supervision Signals

Let’s Walk Through A Toy Example

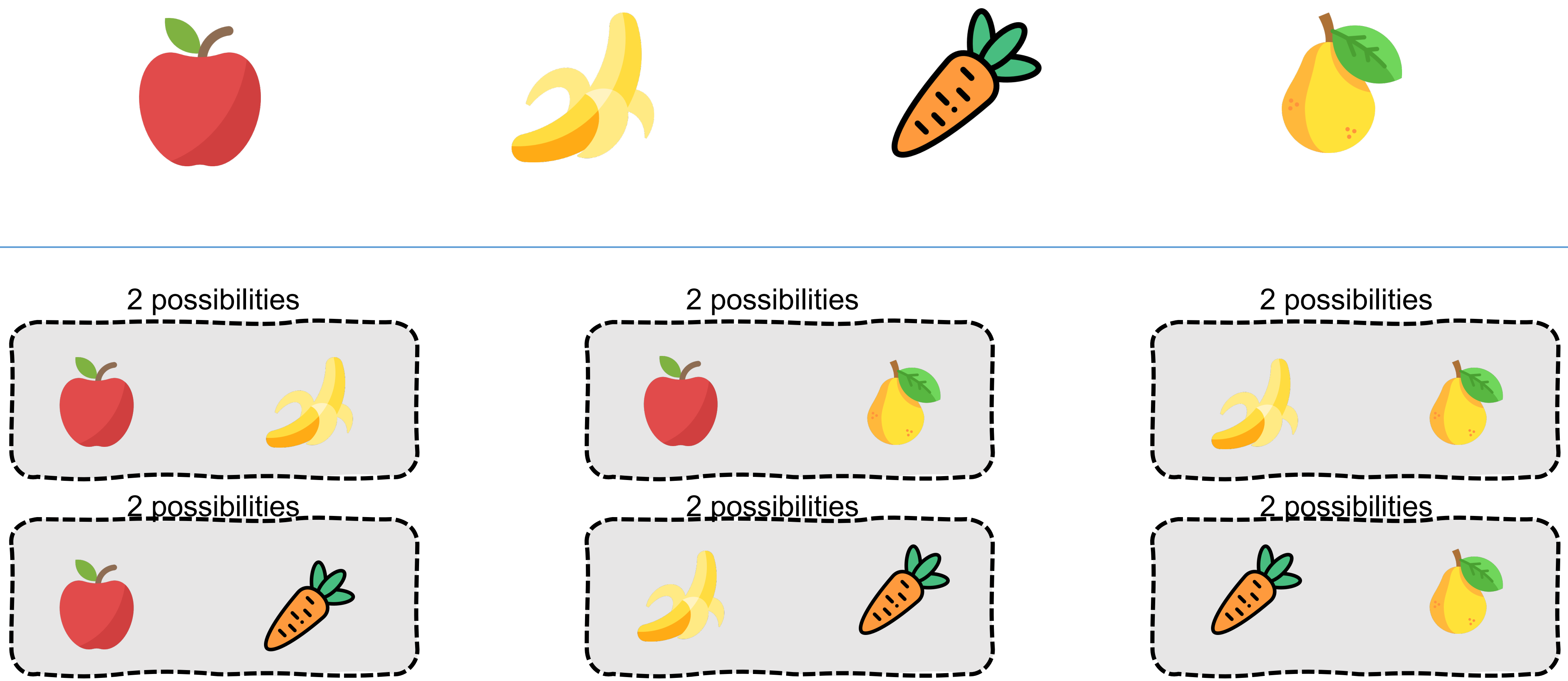
Task: Pair-wise Relationship Between Entities





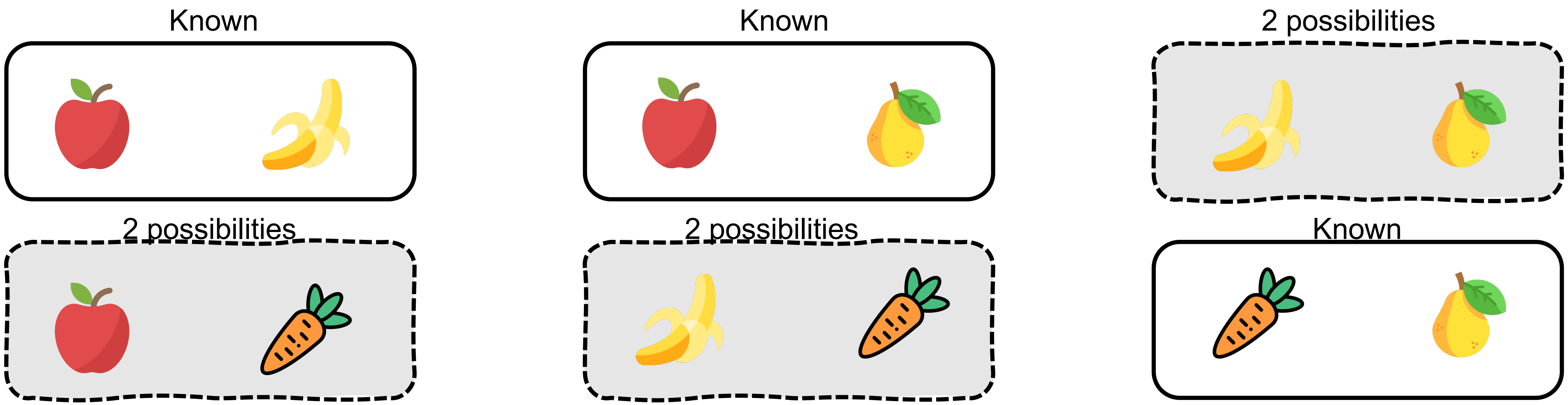

Thanks to Freepik and Flaticon for creating and sharing the icons freely.

Six Pairs of Relationships



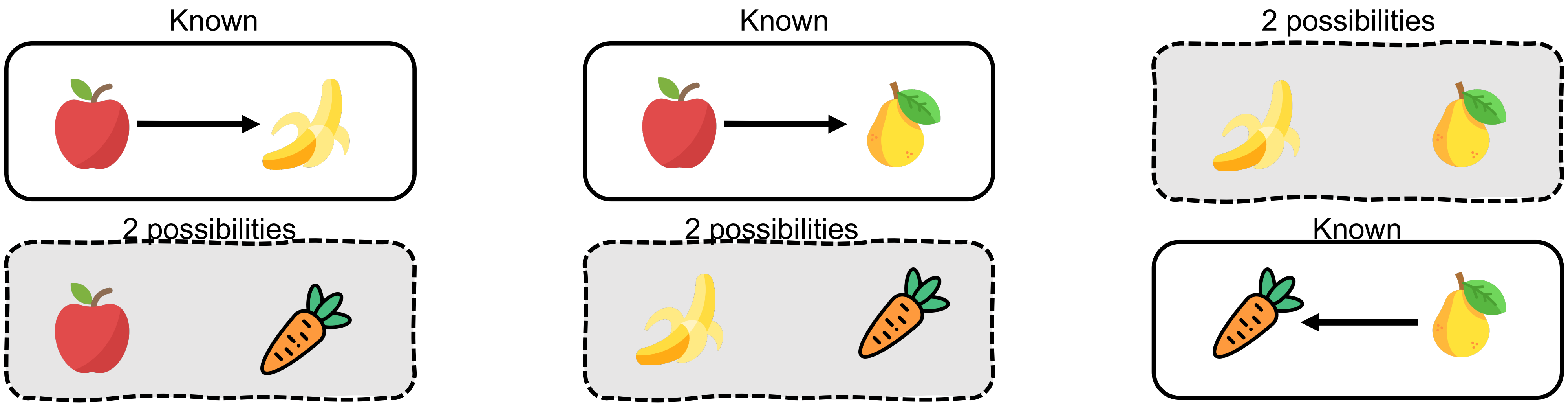
If each relation can choose from a label set of 2 labels, then there are 2^6 possibilities.

Six Pairs of Relationships



Suppose that we already know the label for 3 pairs of them. The total number of possibilities is reduced from $2^6=64$ to $2^3=8$. In other words, we still know nothing about the remaining 3 pairs of relationships.

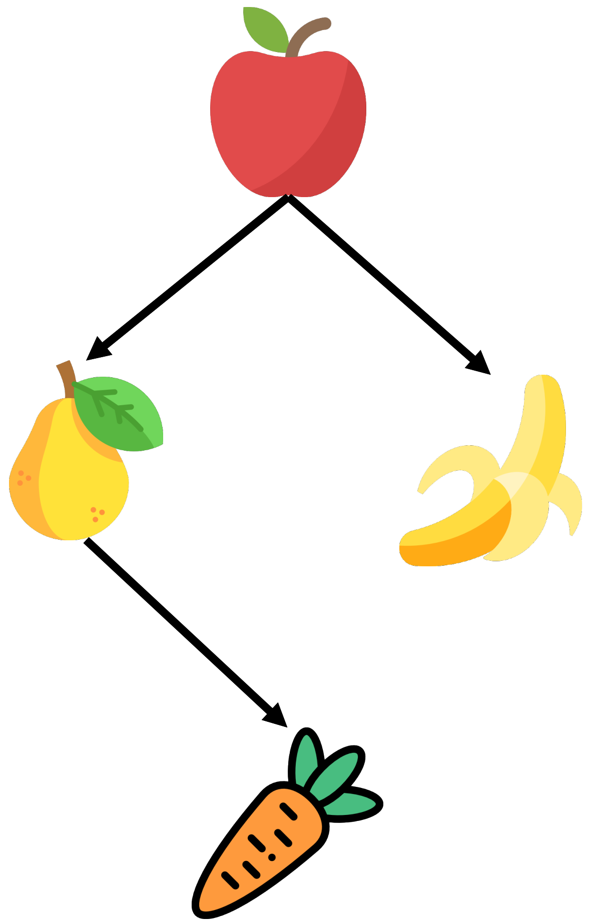
Introducing A Structure Among the Entities



Now, assume that we learn more information about the problem!
(1) the pair-wise relation between entities is an "order relation"
(2) all of the entities create a Directed Acyclic Graph (DAG)

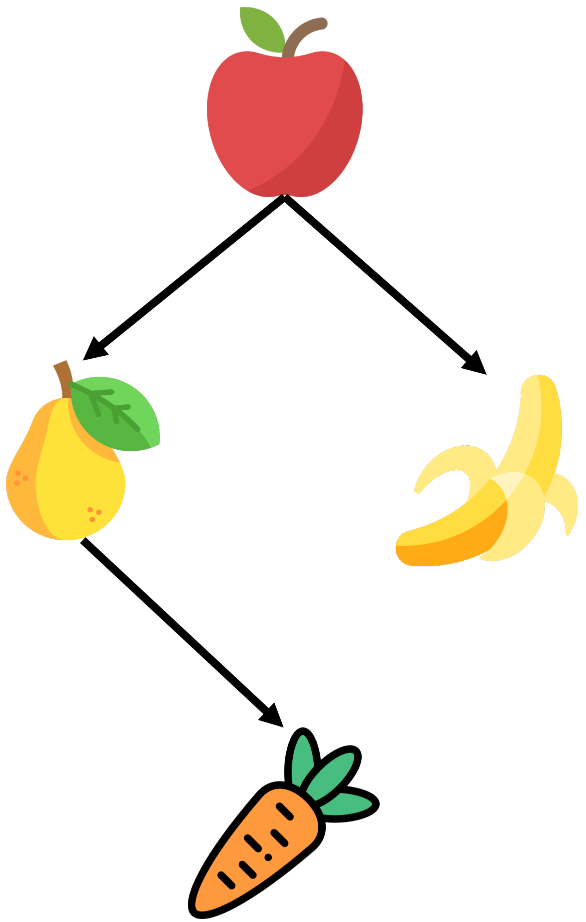
Introducing A Structure Among the Entities

Now with 3 known edges, we have a "partial order."

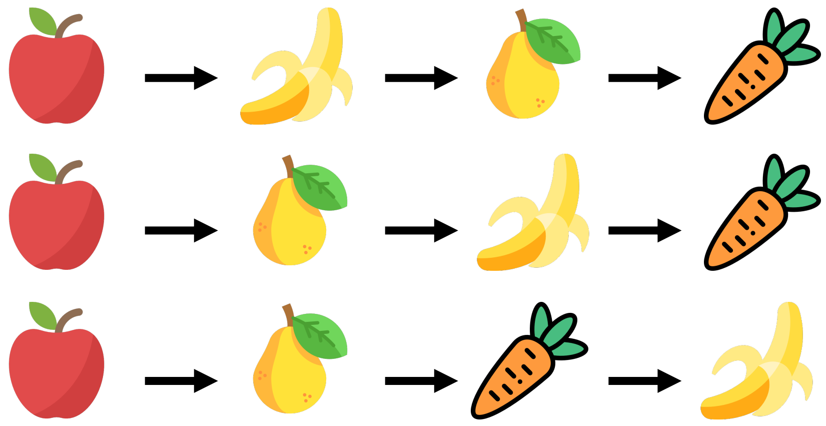


Introducing A Structure Among the Entities

Now with 3 known edges, we have a "partial order."

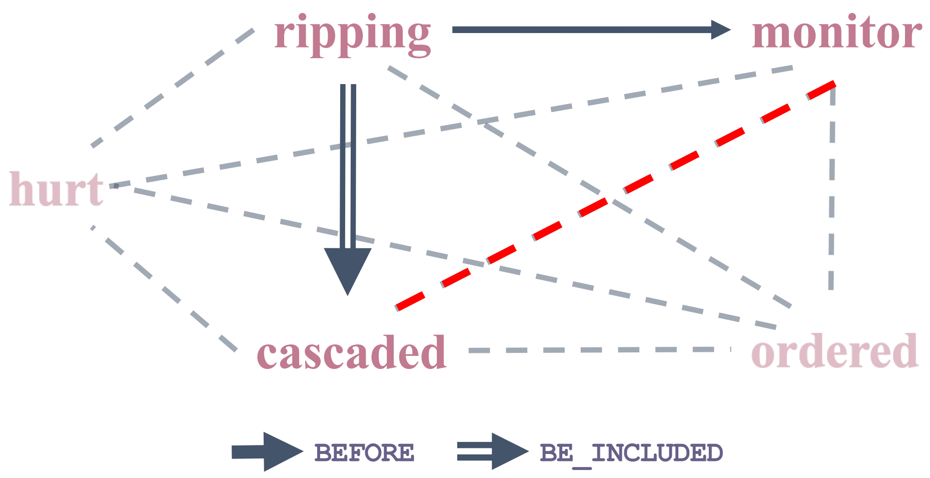


There are only 3 possibilities to describe the entities now (also known as the linear extensions of the partial order).



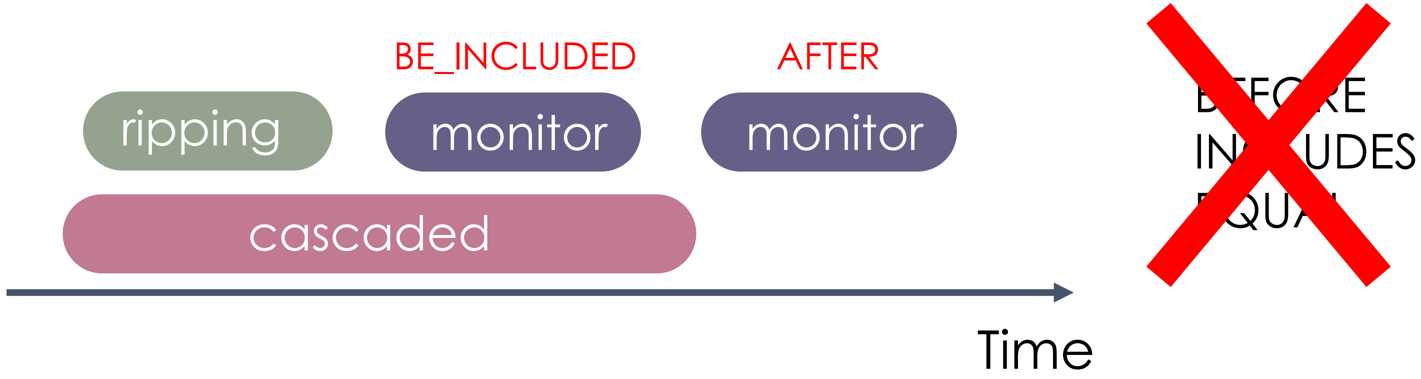
Remember the number of possibilities would have been $2^3=8$ if we hadn't known this structure.

A Relevant Example in NLP is Temporal Relationship Classification



Temporal relation graph: Nodes are events and edges are temporal relationships. It is more complex than a DAG because the edges can choose from more than two directions (depending on the setup, there can be as many as 13[1] labels representing the temporal relationship between two events).

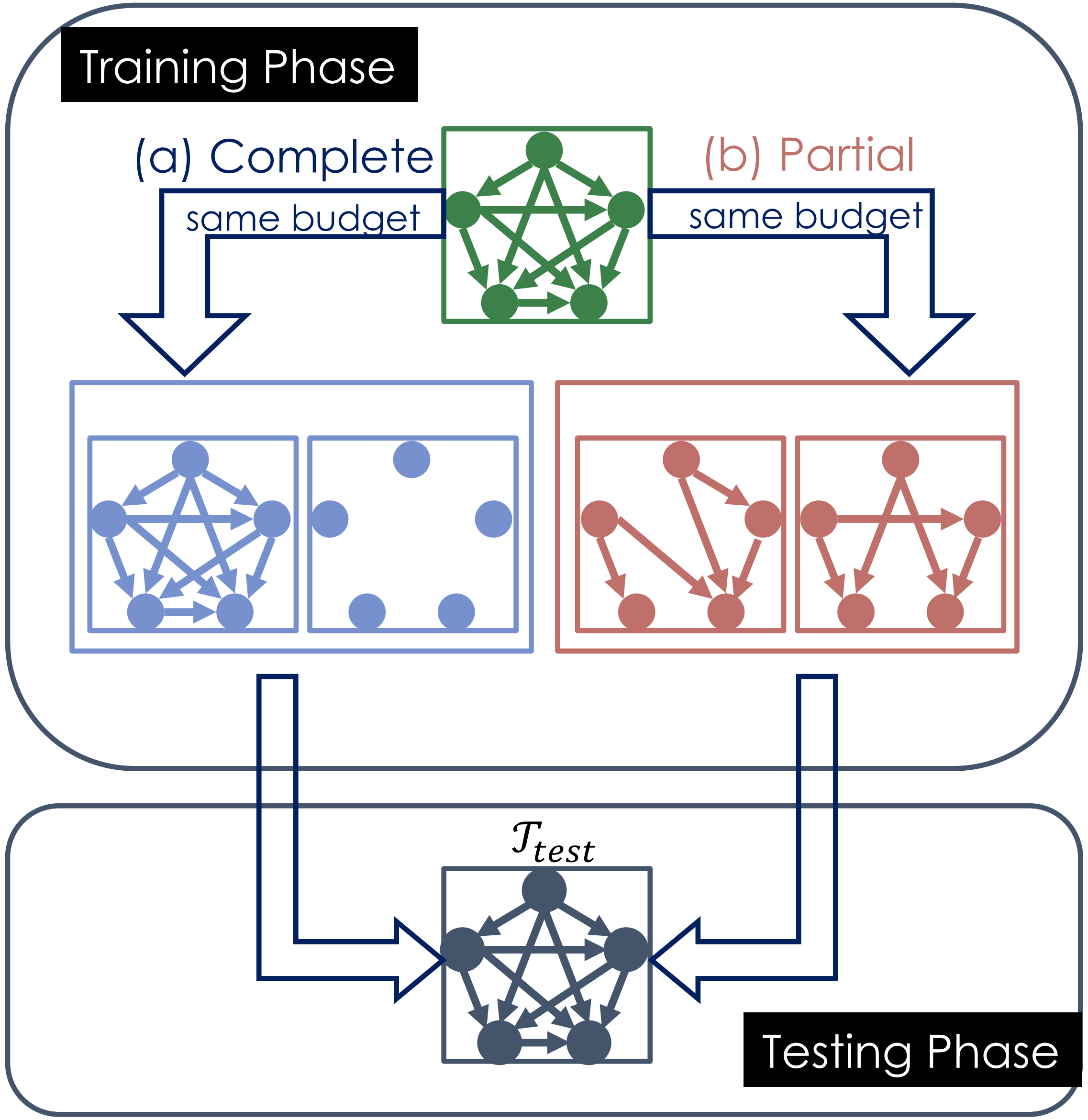
But the concept remains the same – the uncertainty is reduced because of the structure of the problem.



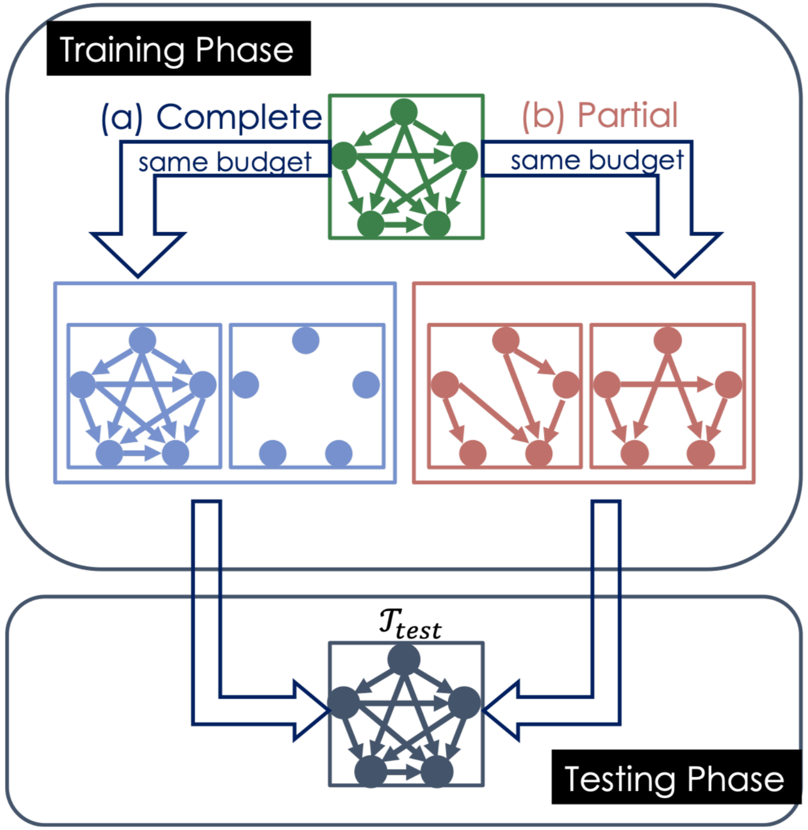
We "incidentally" learn something about the red edge from other edges.

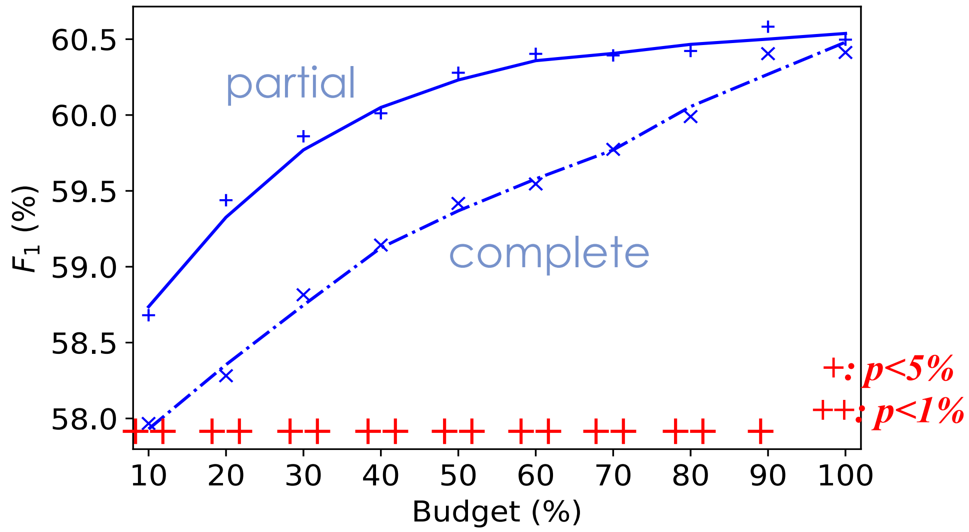
[1] Joint Reasoning for Temporal and Causal Relations. Ning et al., ACL'18.

Partial or complete, that's the question *[NAACL'19]*



Partial or complete, that's the question *[NAACL'19]*





- *Even if some annotations are partial, we "incidentally" learn information about the unannotated edges, so when we have a fixed budget, we can gain more "information" and achieve higher performance.*
- *How do we* ***quantify*** *the information brought by the structure?*

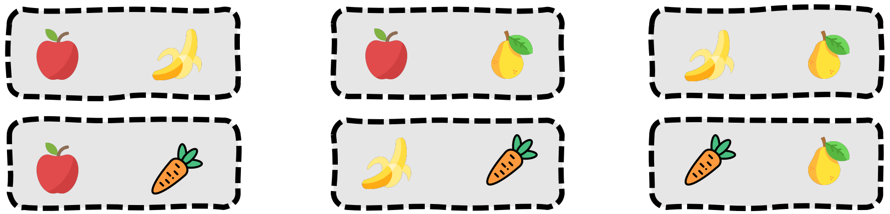
Quantifying Information: Problem Setup

Structure: a vector of random variables: $Y = [Y_1, Y_2, \dots, Y_d]$

Let $\mathcal{L}$ be the label set

$$Y \in C(\mathcal{L}^d) \subseteq \mathcal{L}^d$$

Not all assignments are valid (aka "constraints")



d=6 variables to be labeled

The relation network should be a DAG

Annotation:

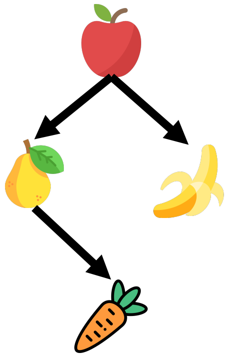
k out of d variables are labeled → Y is further limited to a subset of $C(\mathcal{L}^d)$

Let f_k **be the size of the feasible subset**

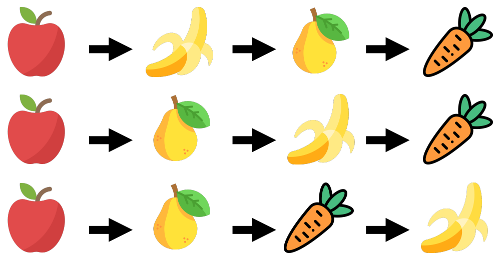
$$f_0 = |C(\mathcal{L}^d)| \geq f_1 \geq f_2 \geq \cdots \geq f_d = 1$$

No annotation

Complete annotation



k=3 out of d=6 variables are labeled



$f_3 = 3$

Quantifying Information: Problem Setup

Structure: a vector of random variables: $Y = [Y_1, Y_2, \ldots, Y_d]$

Let $\mathcal{L}$ be the label set

$$Y \in C(\mathcal{L}^d) \subseteq \mathcal{L}^d$$

Not all assignments are valid (aka "constraints")

Annotation:

k out of d variables are labeled → a subset of $C(\mathcal{L}^d)$

Let $\boldsymbol{f_k}$ **be the size of the feasible subset**

$$f_0 = |C(\mathcal{L}^d)| \geq f_1 \geq f_2 \geq \cdots \geq f_d = 1$$

No annotation

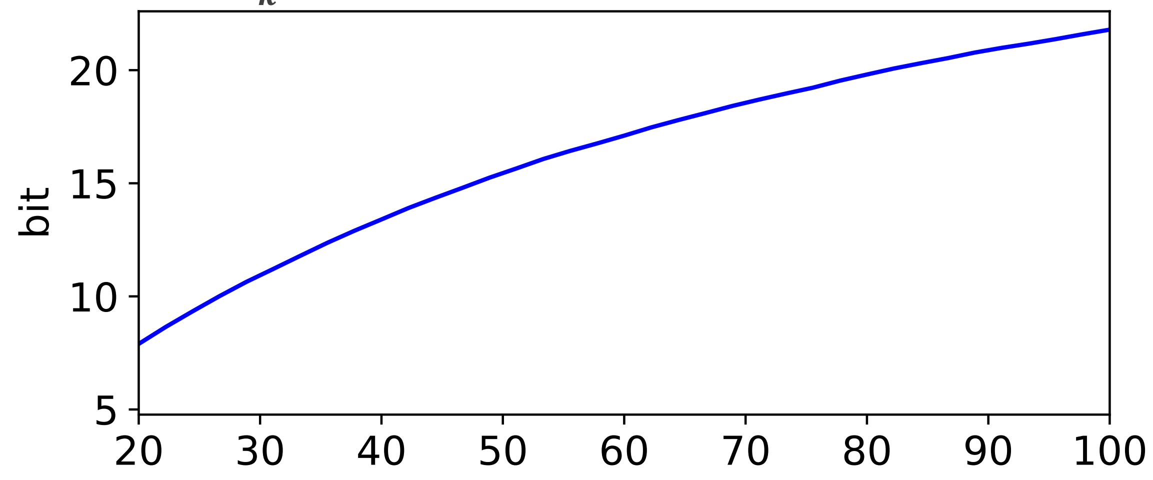
Complete annotation

Define the benefit of k labels: $\boldsymbol{I_k \triangleq \log |C(\mathcal{L}^d)| - E[\log f_k]}$

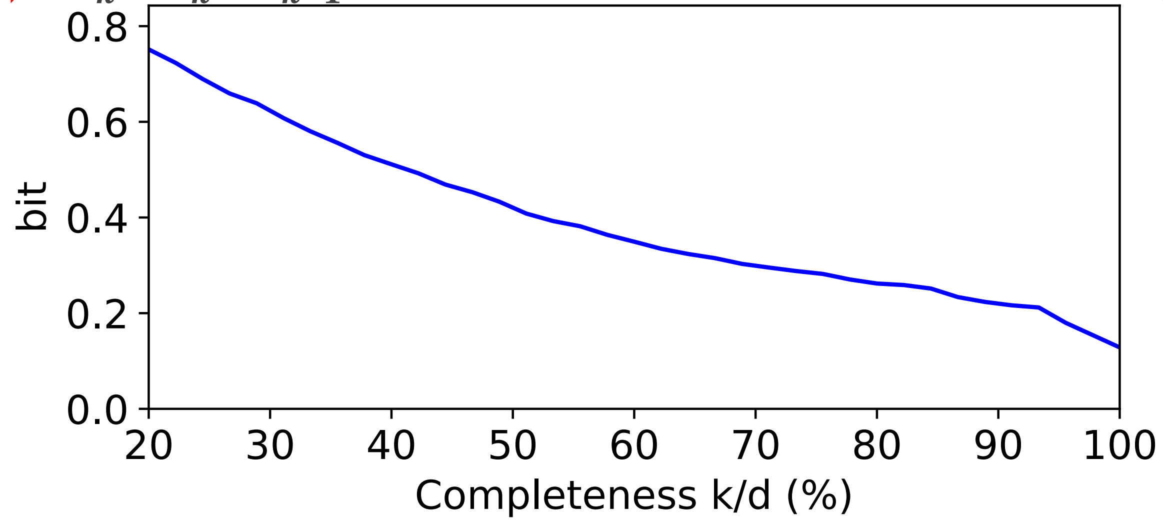
how much of $C(\mathcal{L}^d)$ has been **disqualified** by k labels

Quantifying Information: Diminishing Return

I_k: The benefit of k labels is concave

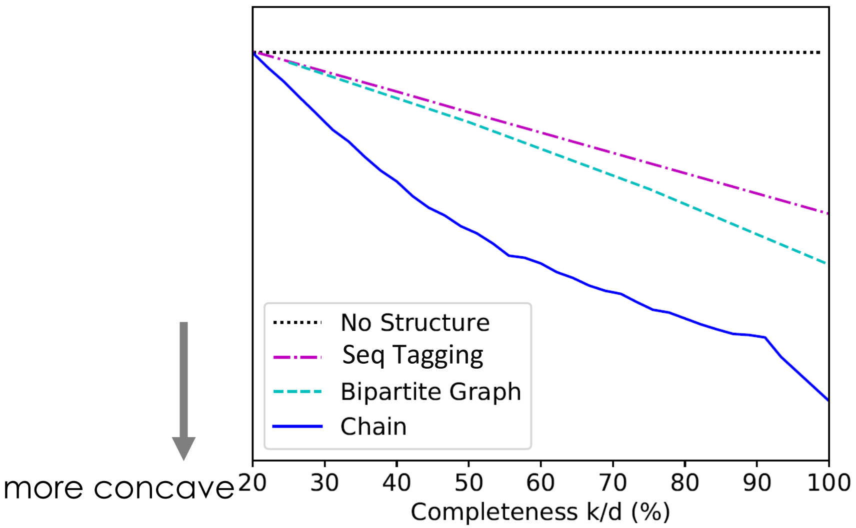


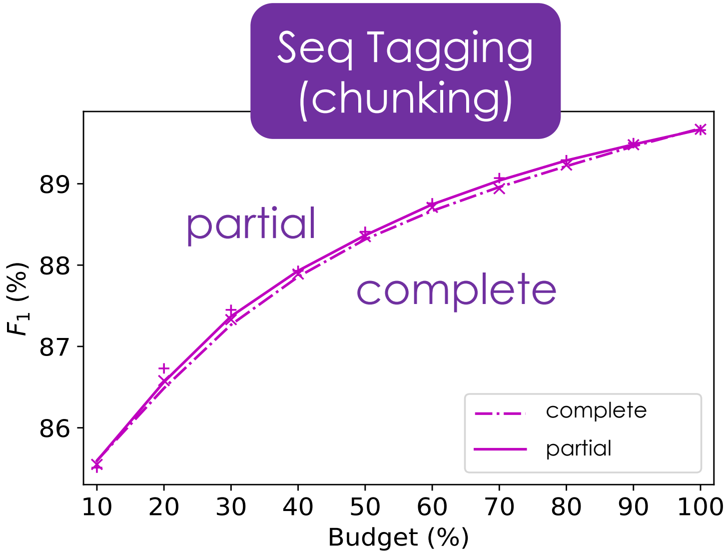
$\Delta_k = I_k - I_{k-1}$: The benefit of a new label is diminishing

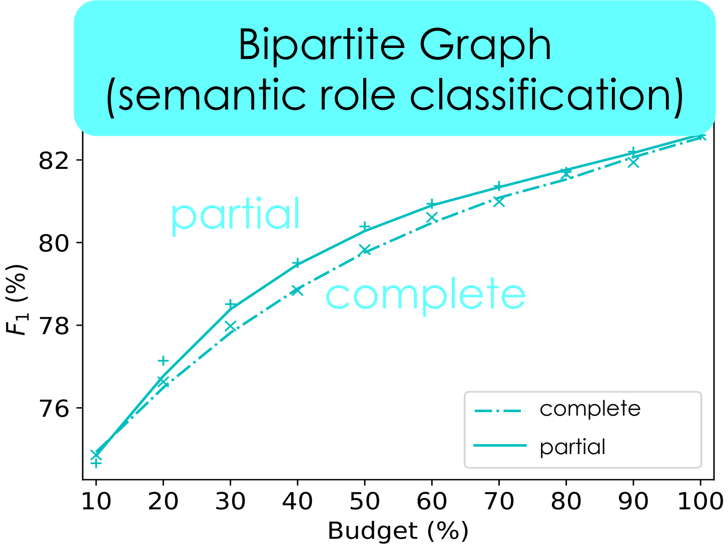


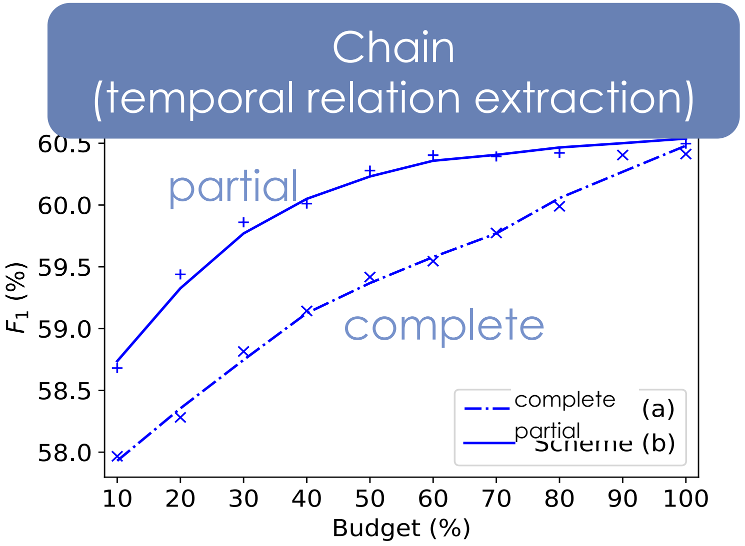
The more concave I_k is, more benefit partial brings!

$\Delta_k = I_k - I_{k-1}$: The benefit of a new label









What is I_k actually?

Definition: A ***k*-partial annotation** A_k is a vector of random variables $A_k = [A_{k,1}, A_{k,2}, \dots, A_{k,d}] \in (\mathcal{L} \cup \sqcap)^d$, where ⊓ is a special character for no label yet, such that

$\sum_{i=1}^{d} \mathbb{I}(A_{k,i} \neq \sqcap) = k$

$P(Y|A_k = a_k) = P(Y|Y_j = a_{k,j}, j \in \mathcal{J})$, where $\mathcal{J} = \{j: a_{k,j} \neq \sqcap\}$

A_k means k variables in Y are labeled, and those k labels are correct

Theorem: I_k is the mutual information between Y and A_k when both Y and the k variables labeled in A_k follow uniform distributions.

What's annotation?

It is the reduction in the uncertainty of a target Y, by a random process A representing the annotation process

More generally, we argue: any signal that has non-zero mutual information with Y can be viewed as "annotation"

It points out a way to understand and quantify the value of indirect signals.

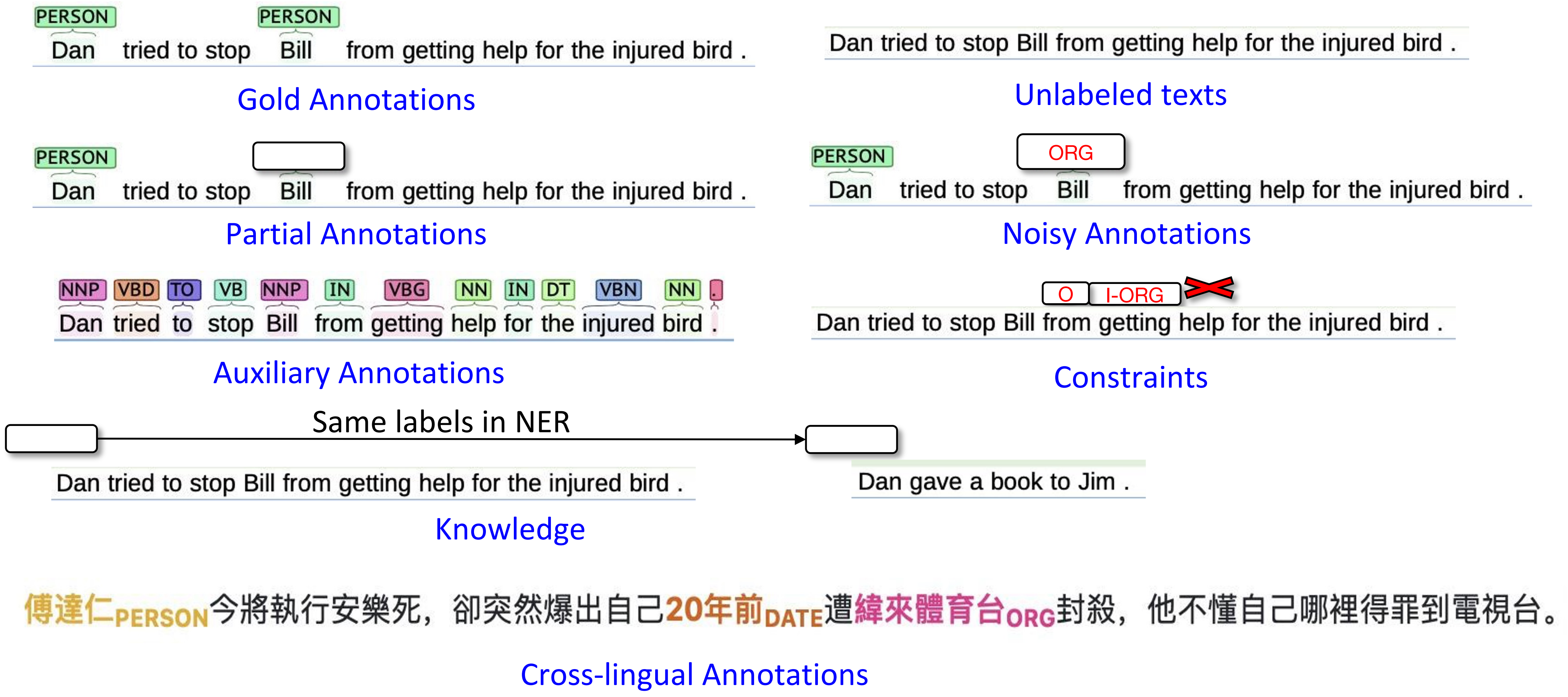
Incidental supervision

Measuring the Benefits of Incidental Signals

Foreseeing the Benefits of Incidental Supervision. He et al., EMNLP21.

Can we provide a unified framework for incidental signals, and quantify the extent to which various incidental signals can help the target task?

- Given the task of NER, what types of signals can we use?



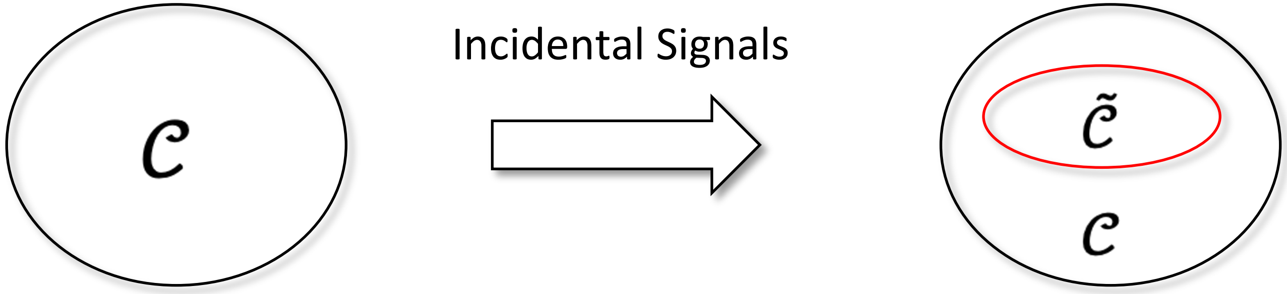
PABI: Impact of Incidental Signals

- $c: X \rightarrow Y, where\ c \in \mathcal{C}$
- Learning theory shows that the size of the concept class determines the "easiness" of the learning problem
  - E.g. the generalization bound $R(c) \leq \hat{R}(c) + \sqrt{\frac{\ln|\mathcal{C}| + \ln\frac{2}{\delta}}{2m}}$
- We will show that the use of incidental signals reduces the size of the concept class, and then will use the relative size of the reduction as a measure for the informativeness of the incidental signals

Recall: $I_k \triangleq \log|C(\mathcal{L}^d)| - E[\log f_k]$

$$S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}}$$

Smaller $\tilde{\mathcal{C}}$ leads to higher Informativeness S

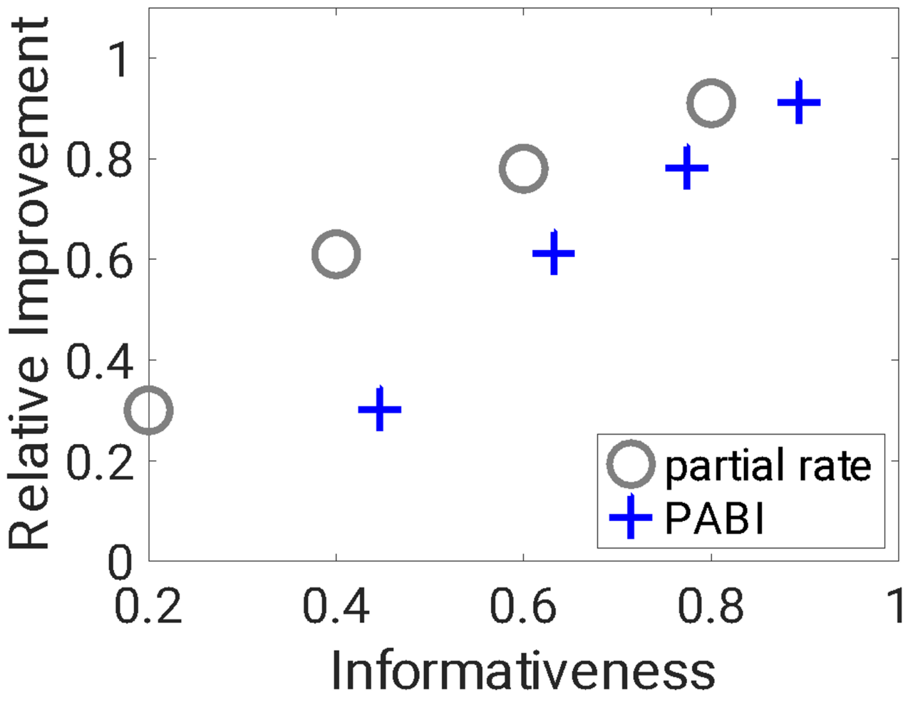


Original Concept Class

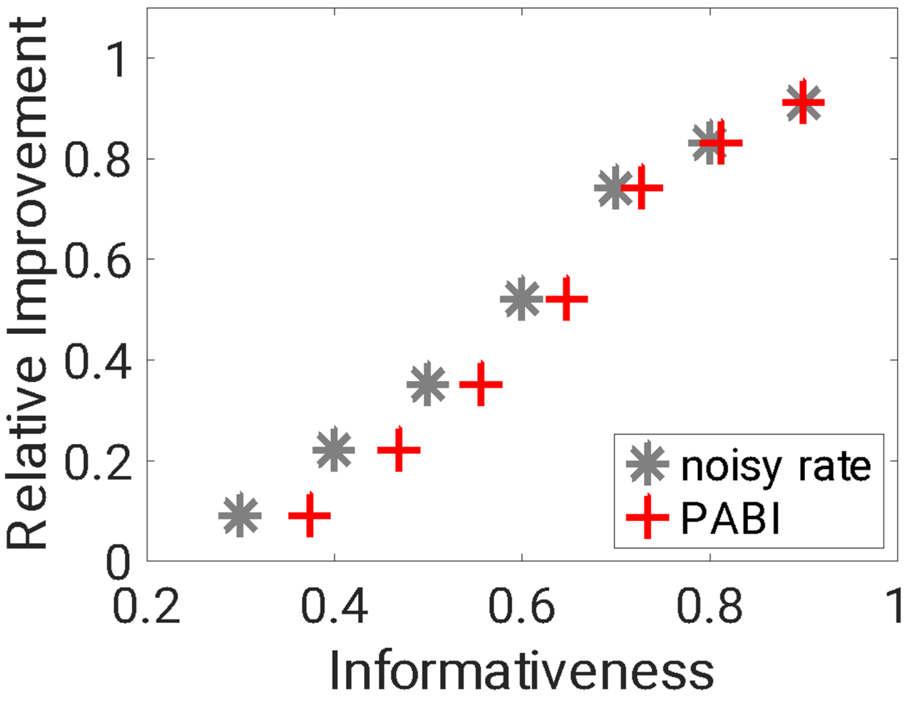
Reduced Concept Class

Reduce the concept class from $\mathcal{C}$ to $\tilde{\mathcal{C}}$

Results on NER (Ontonotes 5.0)



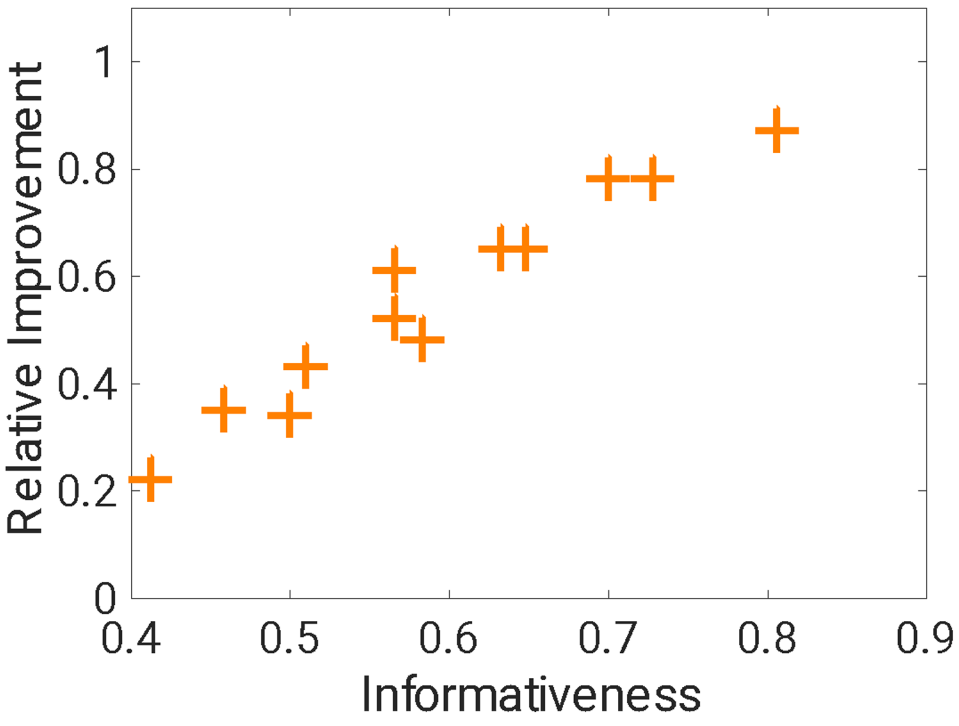
Partial supervision: relative improvement vs. the PABI score for partial signals with different partial rates



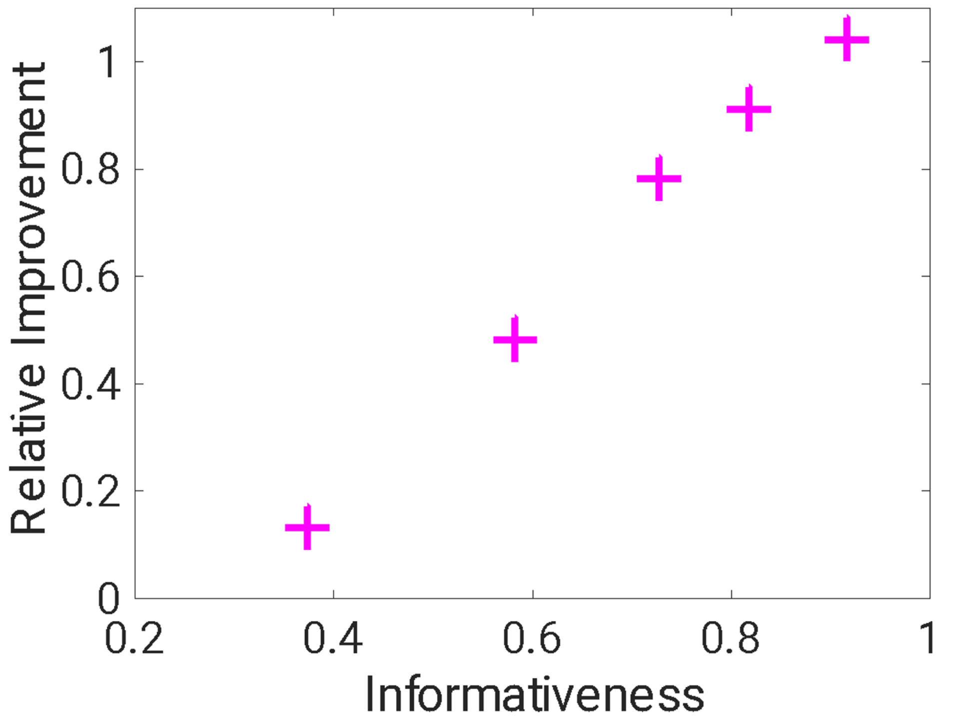
Noisy supervision: relative improvement vs. the PABI score for noisy signals with different noise rates

Before PABI, one might use partial annotation rate / noise rate as a proxy for the usefulness of an incidental dataset; it's indeed a good proxy.

Results on NER (Ontonotes 5.0)



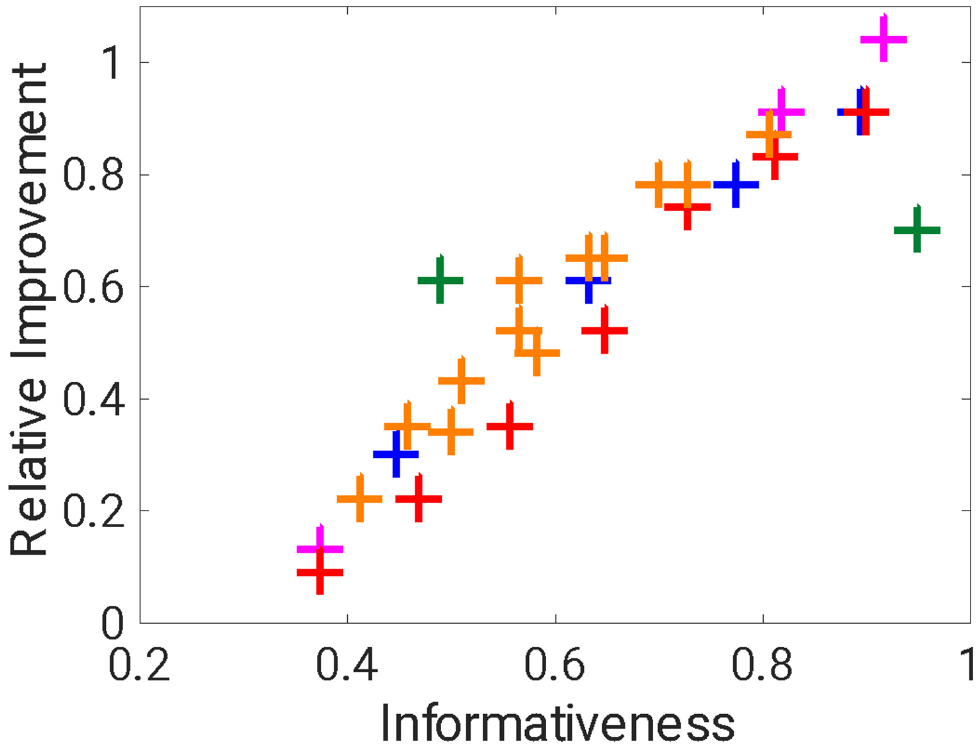
Partial + noisy supervision: relative improvement vs. the PABI score for data with both partial and noisy annotations



Partial + constraints supervision: relative improvement vs the PABI score for data with both partial labels and constraints

The (relative) benefits from the mixed signals (e.g., a dataset is both partial and noisy) cannot be determined in existing frameworks, but our PABI framework can handle it.

Results on NER (Ontonotes 5.0): Overlay



The relation between the relative improvement and PABI for various incidental signals: partial labels, noisy labels, auxiliary labels, partial + noisy, and partial + constraints.

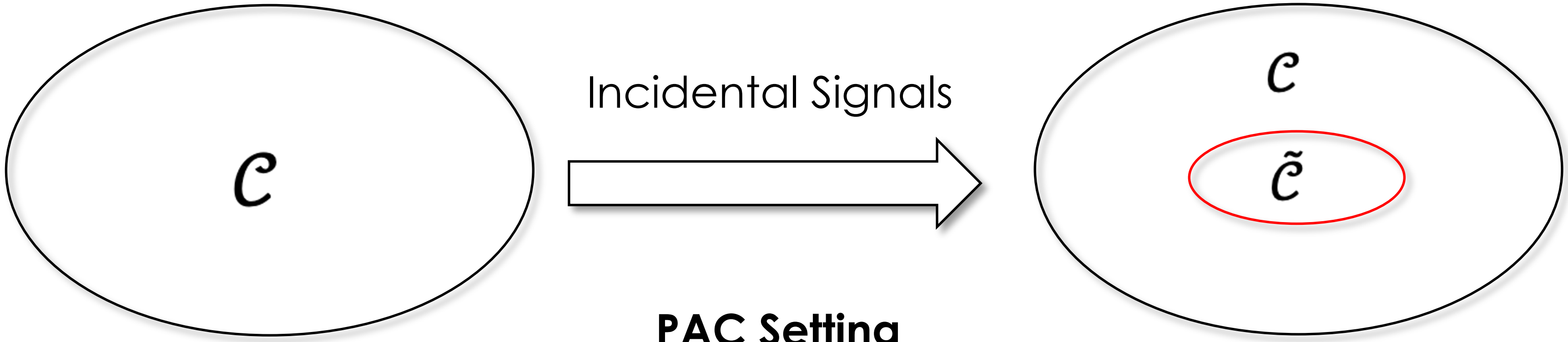
The Pearson's correlation coefficient is: 0.92

The Spearman's rank correlation coefficient is: 0.93

Take away:
The informativeness of a signal predicts the improvement provided by the signal.

Key Insight:
PABI is useful in comparison between the contribution of different types of incidental supervision signals.

PABI: A Unified PAC-Bayesian Informativeness Measure

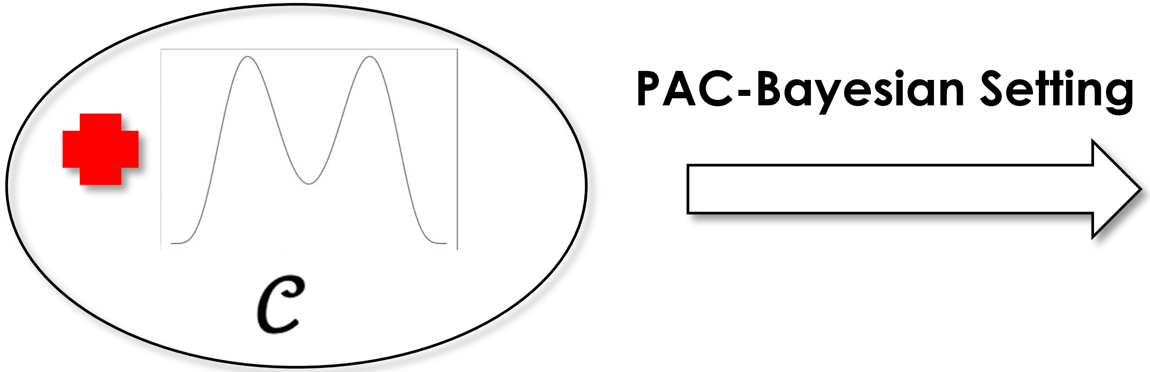


PAC Setting

Original Concept Class

$$S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln |\tilde{\mathcal{C}}|}{\ln |\mathcal{C}|}}$$

Reduce the concept class from $\mathcal{C}$ to $\tilde{\mathcal{C}}$



PAC-Bayesian Setting [1]

$$S'(\pi_0, \tilde{\pi}_0) = \sqrt{1 - \frac{D_{KL}(\pi^* || \tilde{\pi}_0)}{D_{KL}(\pi^* || \pi_0)}} \approx \hat{S}'(\pi_0, \tilde{\pi}_0) = \sqrt{1 - \frac{H(\tilde{\pi}_0)}{H(\pi_0)}}$$

Concept Class with Probability Measure

Make the prior π_0 closer to the gold posterior π^*

Can handle the infinite concept class case

For non-probabilistic cases, $S = S' = \hat{S}'$

[1] PAC-Bayesian supervised classification: the thermodynamics of statistical learning. Catoni, 2007.

Study of Learnability

Learnability with Indirect Supervision Signals. Wang et al., NeurIPS20.

Problem Setup

- To move one-step further in theoretical analysis, we consider a classification task where we predict the target label Y of an instance variable X.
- An *indirect supervision signal* is any random variable (denoted by O) that is correlated to the target label Y.
- We assume the learner only receives samples of (X, O) but does not observe Y directly.

Taking the named entity recognition (NER) tagging as an example:

	Instance X:	Warren	lives	in	New	York
	Gold label Y:	B-PER	O	O	B-LOC	I-LOC
Possible Indirect Signals	O_1:	B-PER	O	?	?	I
	O_2:	NNP	VBZ	IN	NNP	NNP
	O_3:	Two of the five labels are "O"				

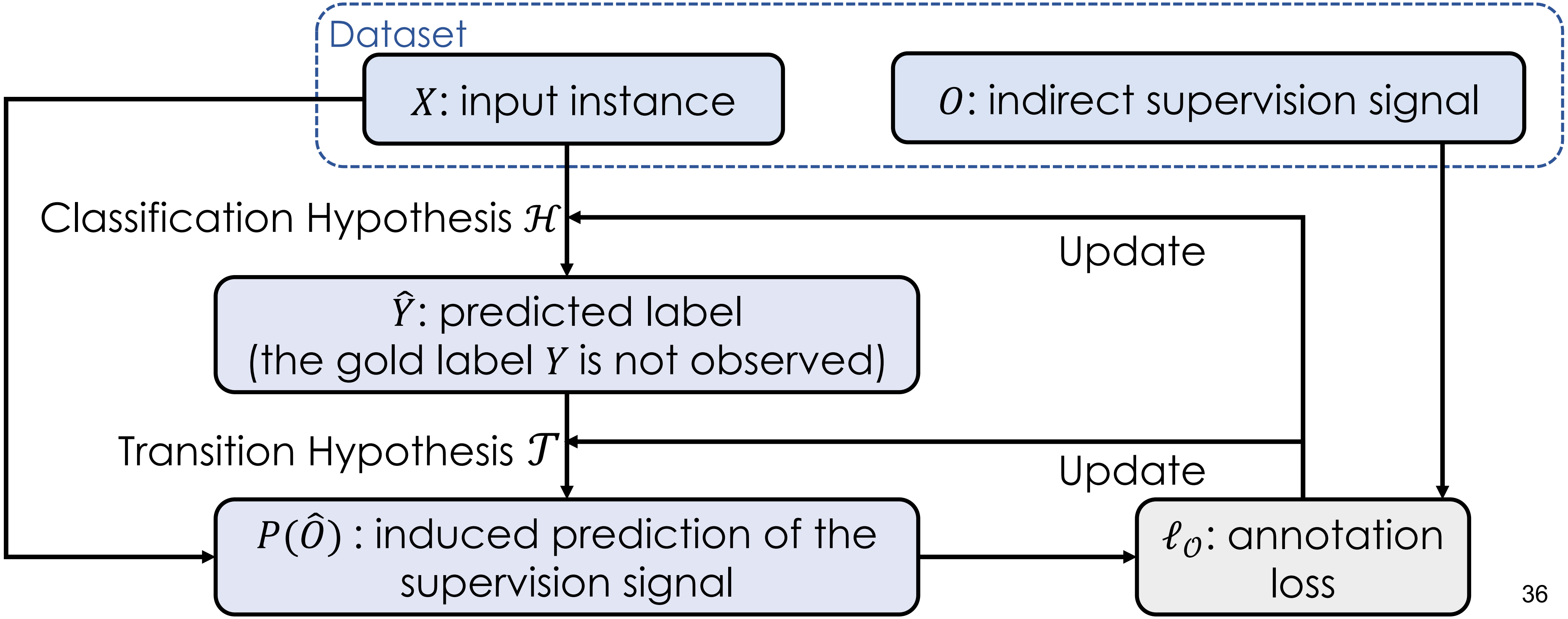
Problem Setup

The learnability problem concerns whether we can learn the optimal classifier in our model given sufficiently many incidental supervision samples (just like using gold labels).

- Intuitively, some incidental signals cannot guarantee learnability since they are *weak*.
  - For example, O_3 (i.e., 2 out of 5 labels are "Outside" in th B-I-O labeling task above) only tells a statistics of the label but there can be a lot of wrong predictions that satisfy this constraint.
  - In contrast, O_1 seems to be a promising choice if the missing rate is low.
- How do we formalize our intuition here?

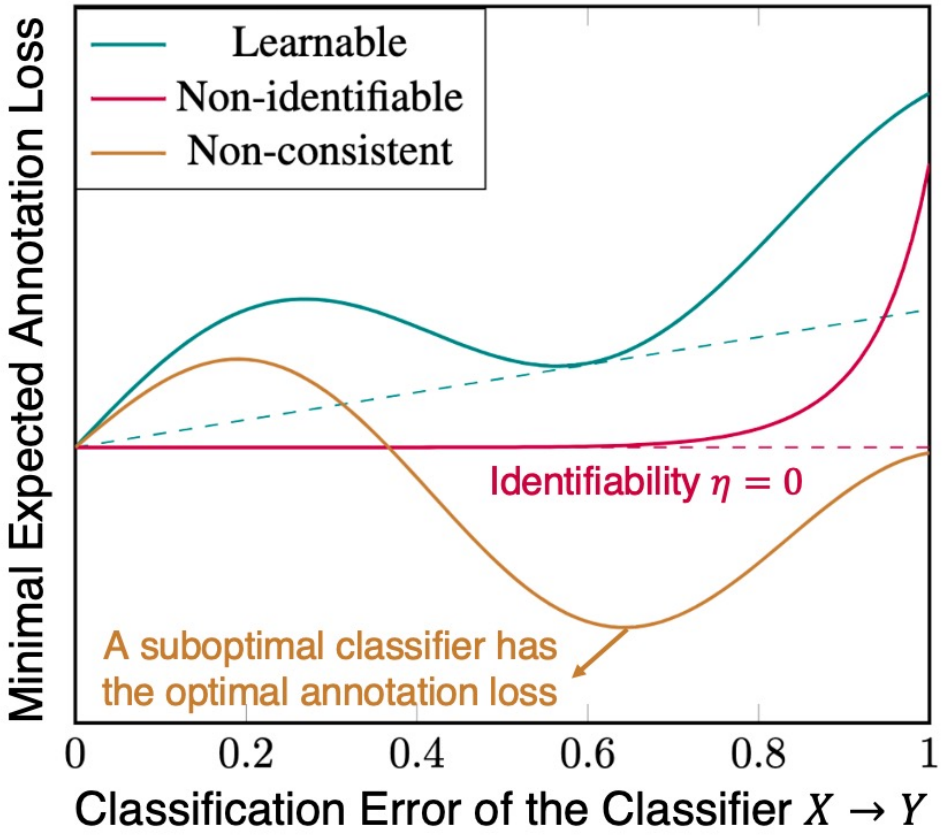
Problem Setup

Our learning framework is shown in the figure. The learner uses the prediction of Y to induce predictions about O. This prediction is then evaluated by the observed dataset. The annotation loss is used to update the classifier and the transition hypothesis.



Learnability Condition: Illustration

To illustrate the learnability condition, we plot the the relationship between the classification error of a hypothesis h and the minimum annotation loss (risk) it can have (over choices of transition hypotheses).



Learnability Condition 1: Consistency

The optimal classifier should be able to induce an optimal prediction of the incidental signal. Formally, we require:

Condition 1

The optimal classifier $h_0 \in \underset{h \in \mathcal{H}, T \in \mathcal{T}}{\operatorname{argmin}} \operatorname{Risk}_{\mathcal{O}}(T \circ h)$.

Remark: When the consistency condition does not hold (this can happen when our signal is very noisy), maximizing the likelihood of the observable will contradict our goal of maximizing the likelihood of the true label.

Learnability Condition 2: Identifiability

A suboptimal classifier should induce higher annotation loss than the lowest annotation loss on average. Formally, we require

Condition 2

Define and let

$$\eta := \inf_{h \in \mathcal{H}, T \in \mathcal{T} : R(h) > 0} \frac{\mathrm{Risk}_{\mathcal{O}}(T \circ h) - \inf_{T \in \mathcal{T}} \mathrm{Risk}_{\mathcal{O}}(T \circ h_0)}{\mathrm{Risk}(h)} > 0$$

Learnability Condition 3: Complexity

Our model should not be too complex. Complexity of a model can be described by (a generalized) VC-dimension. Formally, we require:

Condition 3

We assume $\ell_{\mathcal{O}} \circ \mathcal{T} \circ \mathcal{H}$ is weak VC-major with dim $d < \infty$.

Learning Bound

Now we are able to state the main result:

Theorem (Learnability)

If the above three conditions are satisfied, then for any $\delta < 1$, with probability of at least $1-\delta$, we have:

$$\text{Risk}(\text{ERM}(S^{(m)})) \leq \frac{2b}{\eta}\left(\sqrt{\frac{2\overline{\Gamma}_m(d)}{m}} + \frac{4\overline{\Gamma}_m(d)}{m} + \sqrt{\frac{2\log(4/\delta)}{m}}\right)$$

where $\overline{\Gamma}_m(d)$ is defined as

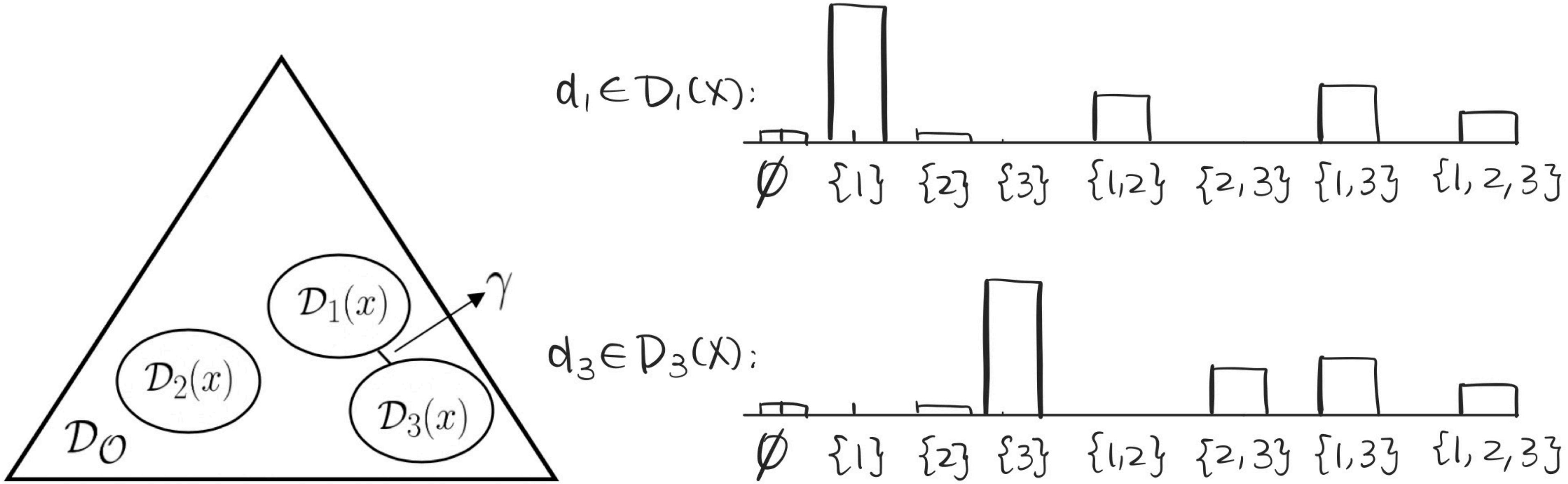
$$\overline{\Gamma}_m(d) := \log\left[2\sum_{j=0}^{\min\{d,m\}}\binom{m}{j}\right] = d\log m(1+o(1)) \text{ as } m \to \infty$$

This implies $R(\text{ERM}(S^{(m)})) \to 0$ in probability as $m \to \infty$.

In other words, we can find the optimal classifier as we have a large training set.

Separation

To check the first two conditions with the learner's prior knowledge, we further propose the **separation** condition. We illustrate the definition using the example of partial label O for a 3-class classification problem where O is identified as a subset of the label space {1,2,3}.



A (predicted) label y_i will induce a distribution family on the annotation space $\mathcal{O}$, denoted as $\mathcal{D}_i(x)$. **Separation** condition requires different families are separated by a minimal *distance* $\gamma > 0$.

Separation: Formal Definition

Theorem (Separation)

For all $x \in \mathcal{X}$*, we denote the induced distribution families by label* y_i *as* $\mathcal{D}_i(x) = \{(T(x))_i : T \in \mathcal{T}\} \subseteq \mathcal{D}_{\mathcal{O}}$*, and the set of all possible predictions of* y *as* $\mathcal{H}(x) = \{h(x) : h \in \mathcal{H}\} \subseteq \mathcal{Y}$*. If*

$$\gamma = \inf_{(x,i,j):p(x,y_i)>0, j\neq i, y_j \in \mathcal{H}(x)} \mathrm{KL}(\mathcal{D}_i(x) \,\|\, \mathcal{D}_j(x)) > 0 \qquad (1)$$

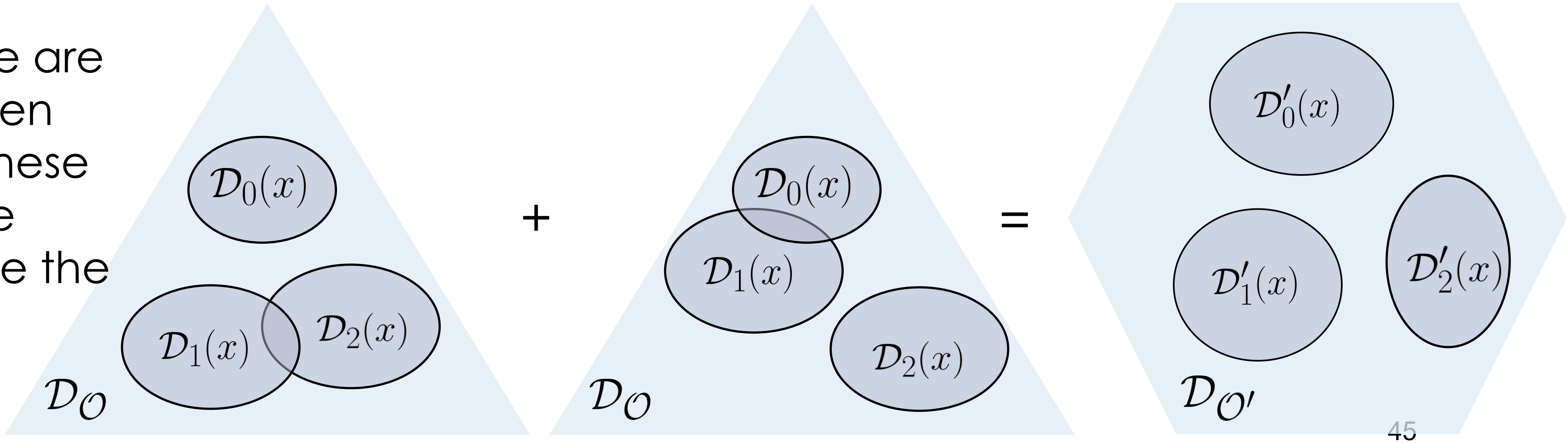
Then the first two conditions are satisfied with $\eta \geq \gamma > 0$ *via the ERM of the cross-entropy loss for the observables.*

Moreover, if (1) is not satisfied, then it can be shown the learning problem can be arbitrarily difficult since different labels can induce arbitrarily similar distributions over annotation space $\mathcal{O}$. In other words, the observation of $\mathcal{O}$ cannot help us to distinguish different labels.

Application of Separation: Joint Supervision

If a single source of supervision signal cannot ensure learnability, it should be used jointly with other signals. We show that a joint supervision can:

- Harm the separation if supervision signals are simply mixed. This is due to the convexity of the KL-divergence.
- Preserve the pairwise separation if modelled *properly*. This effect is visualized in the following figure, where each signal cannot separate one pair of labels, but can be combined to ensure global separation.
- Create new separation: If there are constraints between different signals, these constraints can be utilized to supervise the learning.



Summary

Summary

- We started with a toy example of DAG
  - Knowing part of a graph gives us information about the remaining of the graph
  - We used mutual information as a measure and demonstrated that partially annotating structured prediction problems led to better learning performance, because the uncertainty reduction was higher.
- We continued to argue that incidental signals are those that have non-zero mutual information with the label of the target task.
  - This is supported in PAC and PAC-Bayesian theory because the reduction of uncertainty is actually a term in generalization bounds.
  - We defined PABI as a measure of usefulness of an incidental supervision dataset, and demonstrated its prediction power for actual performance gain on various NLP tasks.
- We formally introduced the learnability conditions from incidental signals, and described a more convenient notion called "separation."

Thank You