Is Human Label Variation Really so Bad for AI?

Barbara Plank
Center for Information and Language Processing,
MaiNLP lab, University of Munich (LMU)
(& ITU Copenhagen)

October 11, 2022
Prague
CLARIN 2022

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

A Typical “AI” Pipeline

X

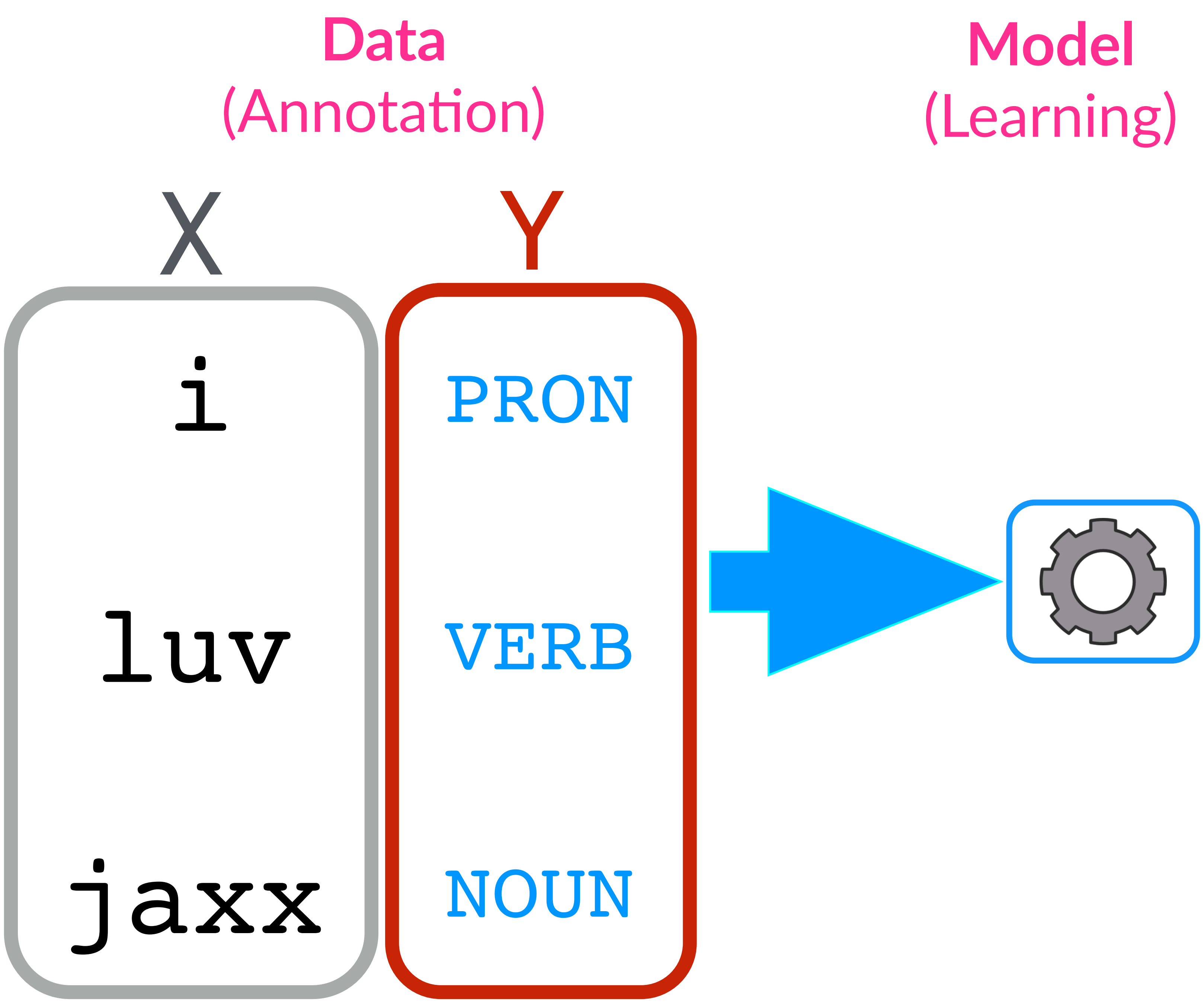
```
i

luv

jaxx
```

A Typical "AI" Pipeline

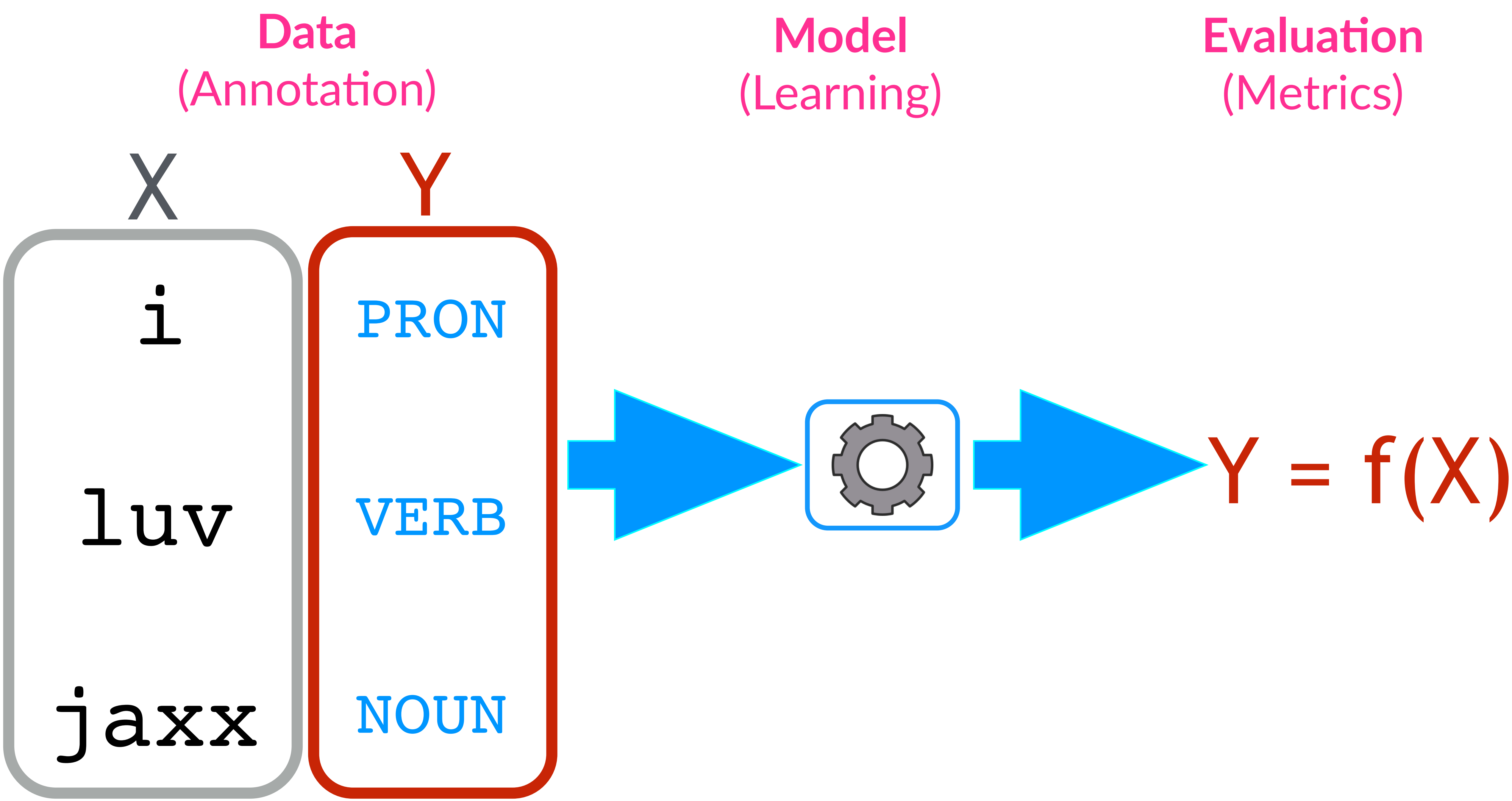
Data
(Annotation)

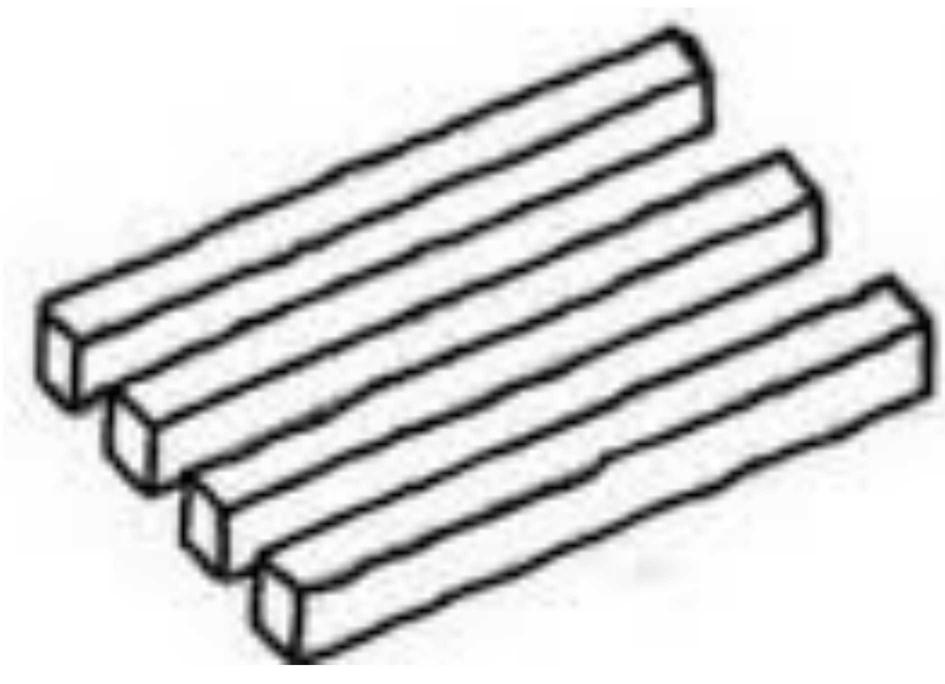
X	Y
i	PRON
luv	VERB
jaxx	NOUN

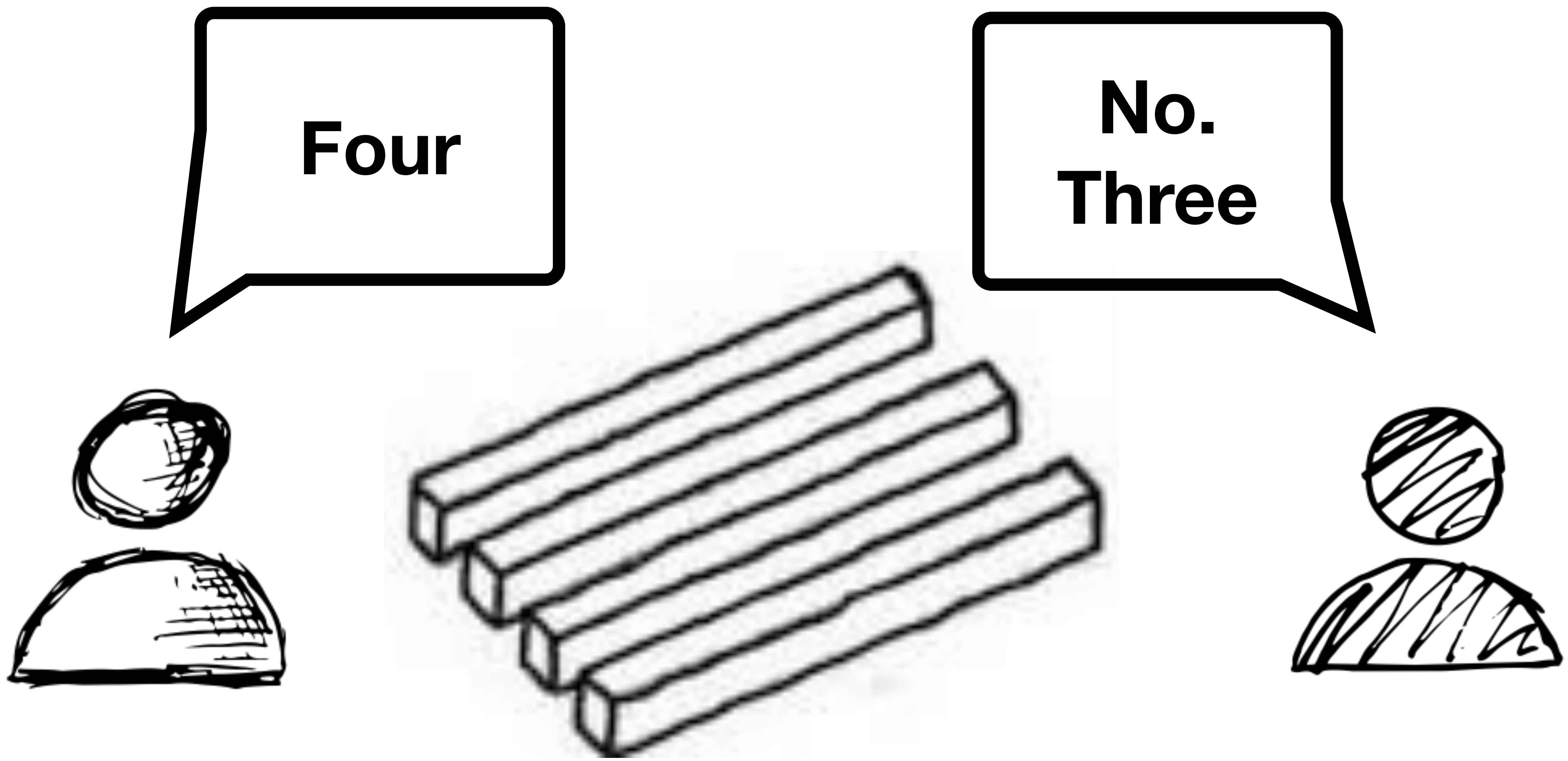
A Typical “AI” Pipeline



A Typical “AI” Pipeline





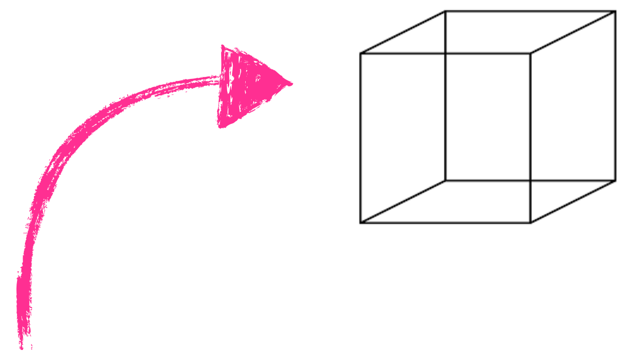

Four
No.
Three

Disagreement in human annotation is ubiquitous

Disagreement in human annotation is ubiquitous

— This impacts all 3 stages of the NLP pipeline.
— Human disagreement is one important form of uncertainty.

Can we turn disagreement into ***advantage***?

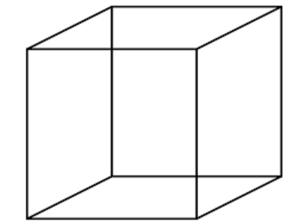


Fortuitous
data

Disagreement in human annotation is ubiquitous

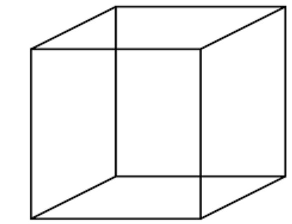
— This impacts all 3 stages of the NLP pipeline.
— Human disagreement is one important form of uncertainty.

Typology of fortuitous data



Type / Side benefit of	Examples	Availability	Readiness
meta-data	hyperlinks, HTML markup, genre labels, symbolic knowledge..	+	+
annotation	annotator disagreement	-	+
behavior	cognitive processing data	+	-

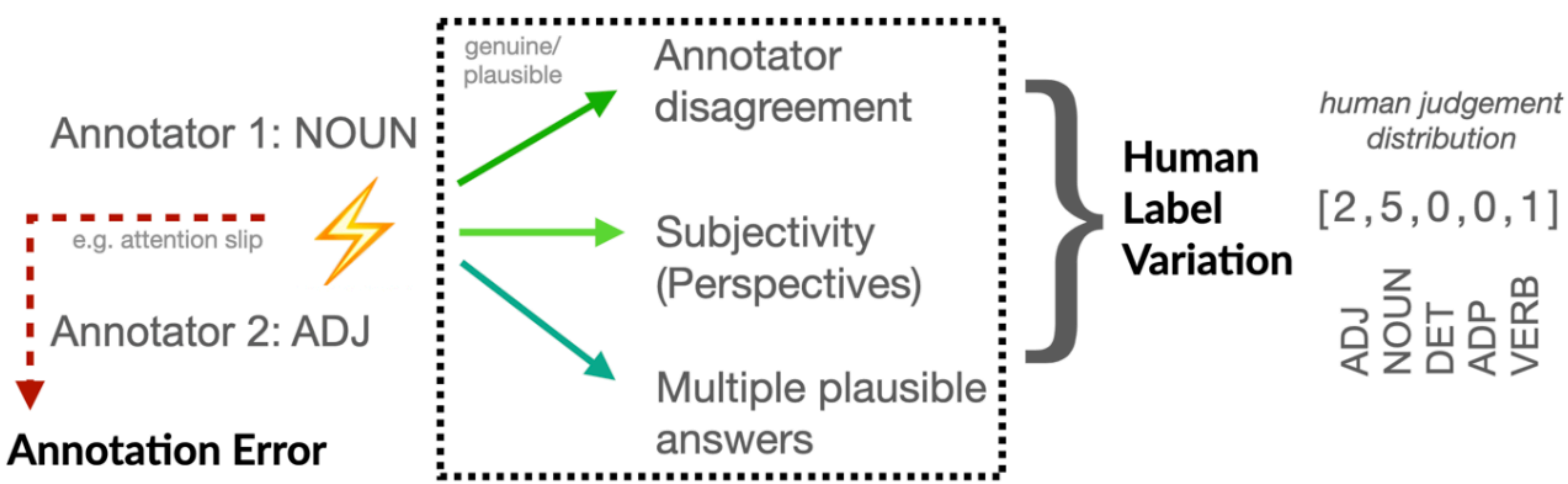
Typology of fortuitous data



Type / Side benefit of	Examples	Availability	Readiness
meta-data	hyperlinks, HTML markup, genre labels, symbolic knowledge..	+	+
annotation	annotator disagreement	-	+
behavior	cognitive processing data	+	-

Disagreement or variation?

- I propose to call it **Human label variation (HLV)** = plausible variation in annotation
  - Preferred over 'disagreement' as that implies two or more views cannot all hold
  - To reconcile different notions in the literature ('human uncertainty', 'perspectives', 'hard cases', 'disagreement' etc)
  - In contrast: annotation errors



(Plank, 2022 EMNLP)

Roadmap: Three perspectives

1. Data: Is human label variation (HLV) random noise or signal?
2. Modelling: How can we leverage human label variation?
3. Evaluation: How to evaluate in light of human label variation?

Selected examples

Act I: Data

there are linguistically hard cases, even for POS tagging

e.g. Manning (2011). *Part-of-Speech tagging from 97% to 100%. Is It Time for Some Linguistics?*

Part-of-Speech (POS)

NOUN	NOUN	VERB	ADJ
ADJ	NOUN	VERB	ADJ
social	media	are	massive

Medical Relations Extraction (MRE)

These data suggest that subclinical
RIBOFLAVIN DEFICIENCY may occur in adolescents and
that deficiency may be related to dietary intake of
RIBOFLAVIN

Data and example from Aroyo & Welty (2015) and Palomaki et al., (2018)

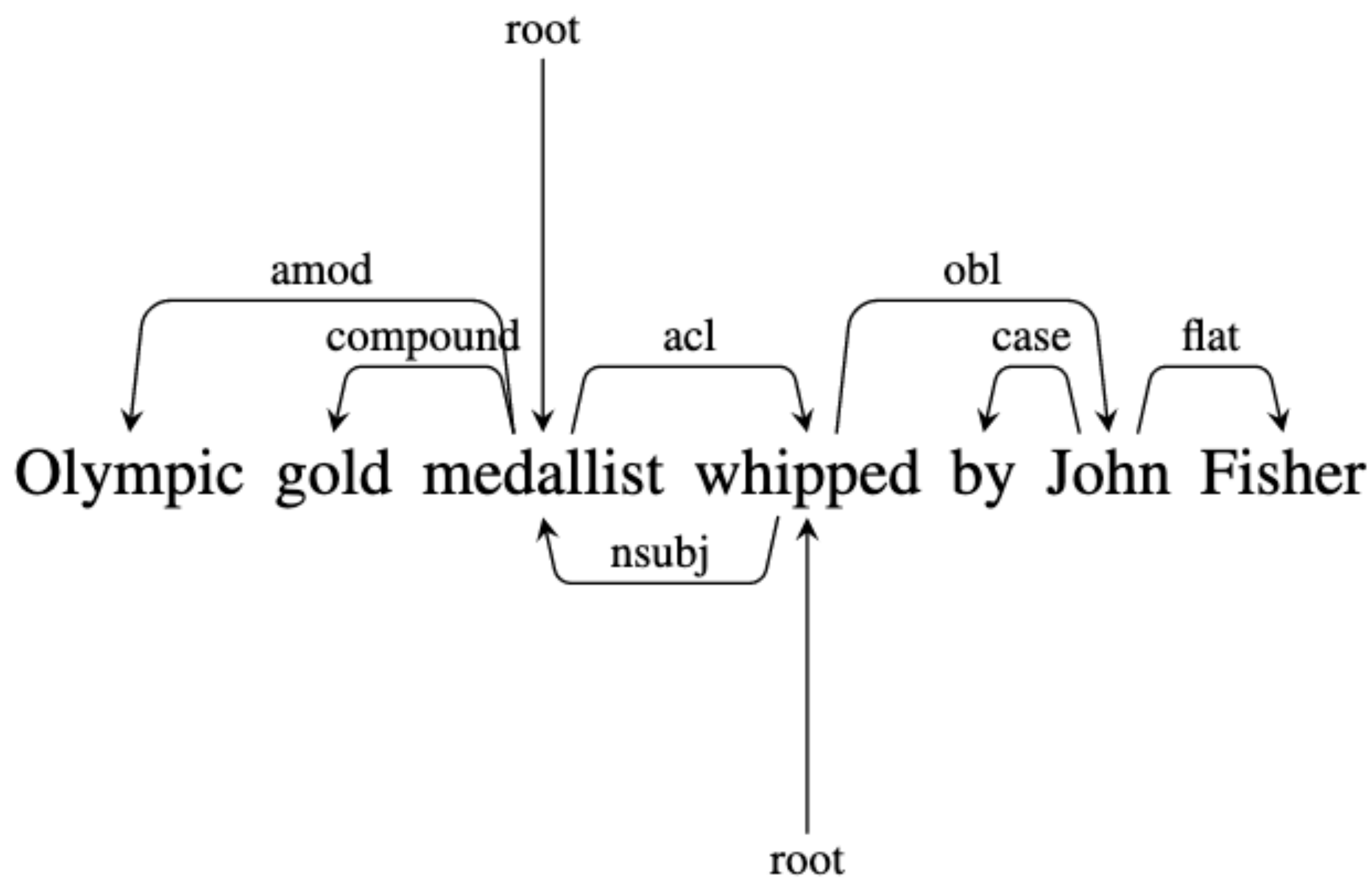
Medical Relations Extraction (MRE)

relation, count
ASSOCIATED_WITH 4
SYMPTOM 3
CAUSES 3
PREVENTS 1
SIDE_EFFECT 1
MANIFESTATION 1
PART_OF 1
DIAGNOSE_BY_TEST_OR_DRUG 1
OTHER 1

These data suggest that subclinical RIBOFLAVIN DEFICIENCY may occur in adolescents and that deficiency may be related to dietary intake of RIBOFLAVIN

Data and example from Aroyo & Welty (2015) and Palomaki et al., (2018)

Dependency Parsing



“Depending on whether this is an example of a zero copula construction, or a clause-modified noun, either annotation is plausible”

Example from Liu et al., 2018

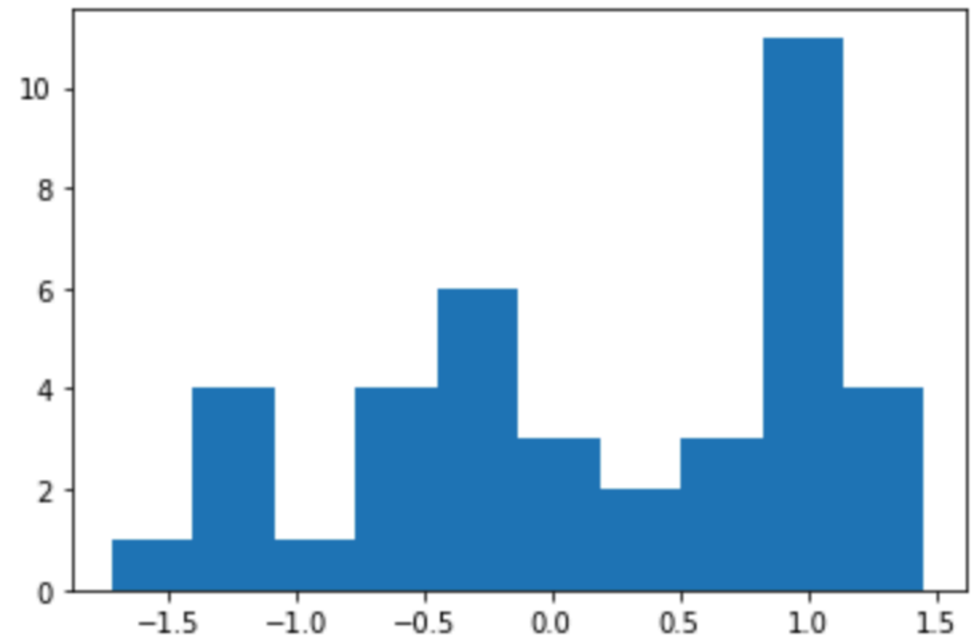
Natural Language Inference (NLI)

Premise *p*: Amanda carried the package from home .
Hypothesis *h*: Amanda moved .

Does p->h?
RTE (Recognising Textual Entailment) original-dataset-label: entailed

Data with 50 annotators by Pavlick & Kwiatkowski (2019)
Newer ChaosNLI with 100 a. by Nie, Zhou, Bansal (2020)

Natural Language Inference (NLI)



contradiction ~neutral entailment

Premise *p*: Amanda carried the package from home .
Hypothesis *h*: Amanda moved .

Does p->h?
RTE (Recognising Textual Entailment) original-dataset-label: entailed

Data with 50 annotators by Pavlick & Kwiatkowski (2019)
Newer ChaosNLI with 100 a. by Nie, Zhou, Bansal (2020)

More examples (selected)

- Abusive & offensive language (Akhtar et al, 2021; Leonardelli et al., 2021; Ceras Curry et al., 2021)

- Visual Question Answering: Difficulty of VQA examples (Jolly et al., 2021)

Q: What is the pattern of the little girl's dress?
GT: plaid: 4, checks and flowers: 1, checkered with flowers: 1, polka dots, squares, plaid: 1, squares and flowers: 1, flowers: 1, plaid and floral: 1
EaSe: 1.0

Q: Where is this?
GT: road: 4, outside: 2, pakistan: 1, outdoors: 1, sidewalk: 1, sweden: 1
EaSe: 0.30

Figure 1: One image from VQA2.0 with two questions and the answers by 10 annotators. Frequency of each unique answer (e.g., *plaid* : 4) and EaSe values of the samples (the higher, the easier) are reported.

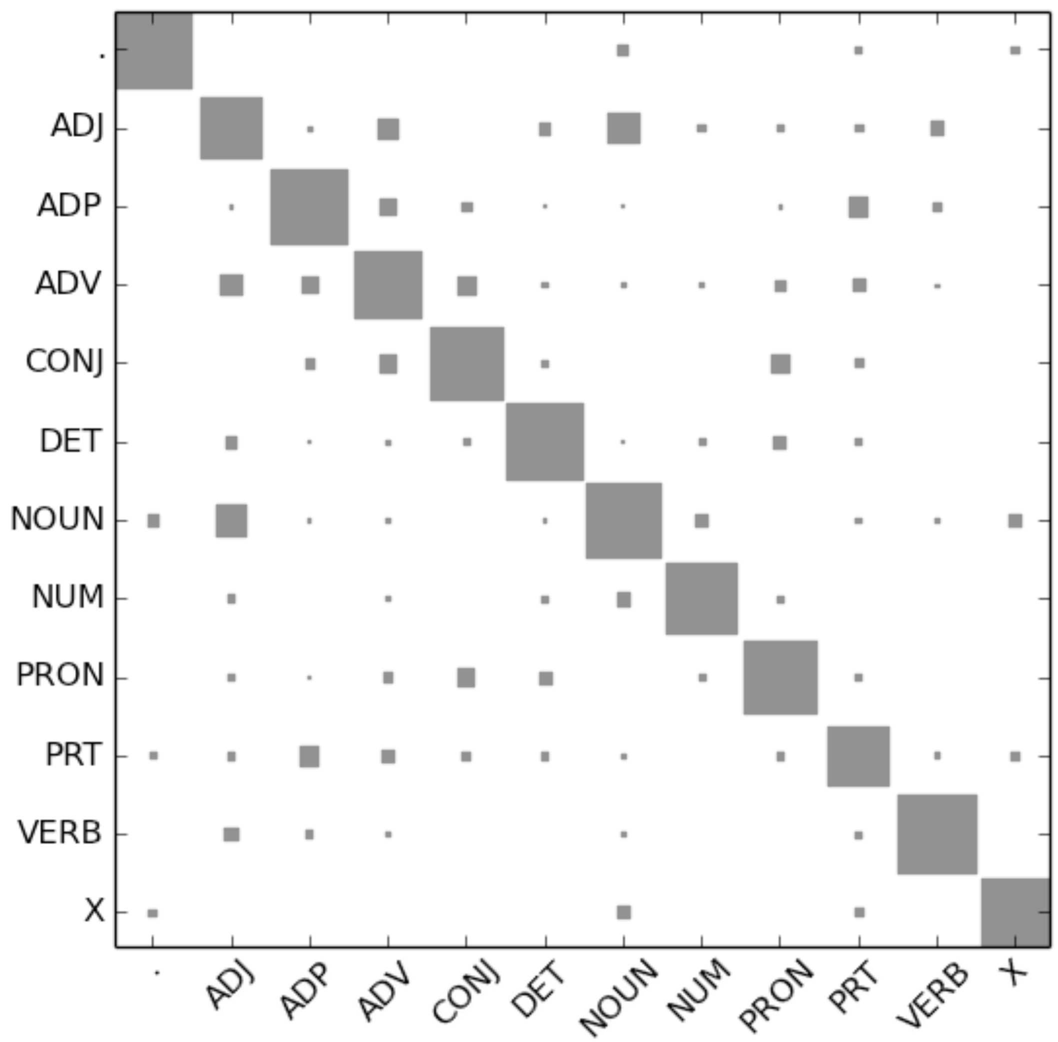
Is human label variation randomly distributed?

(Plank et al., 2014)

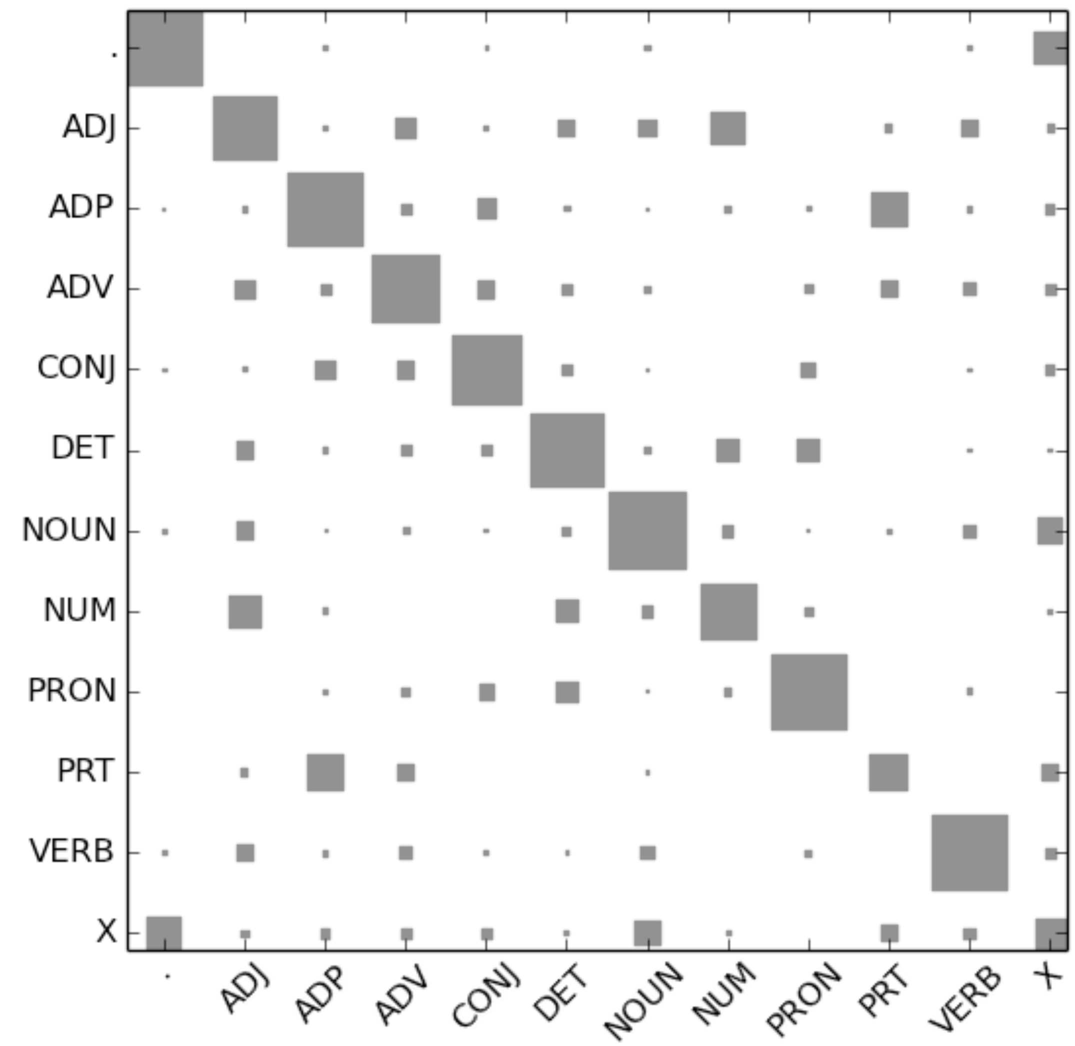
Is human label variation randomly distributed?

... and can we estimate disagreements from small samples?

(Plank et al., 2014)

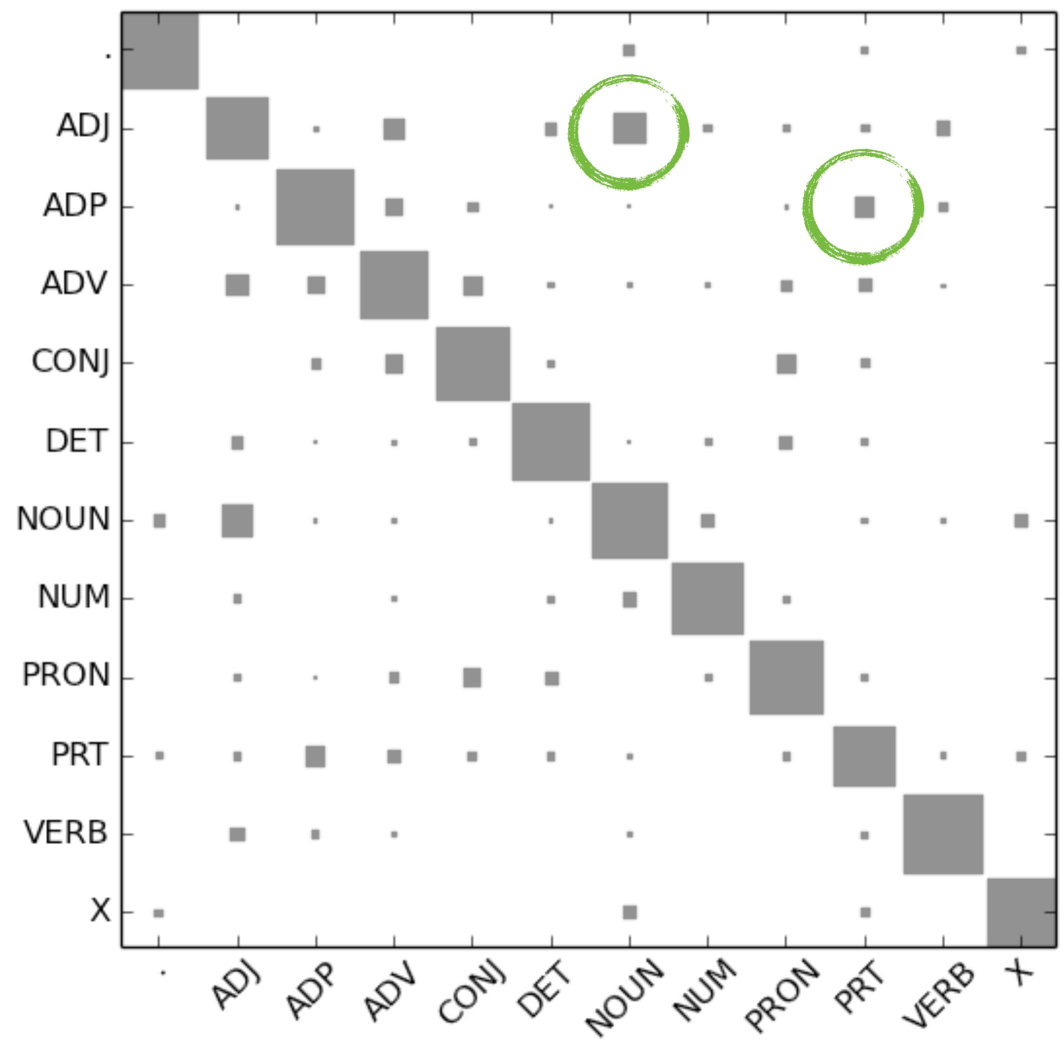


Wall Street Journal PTB-00

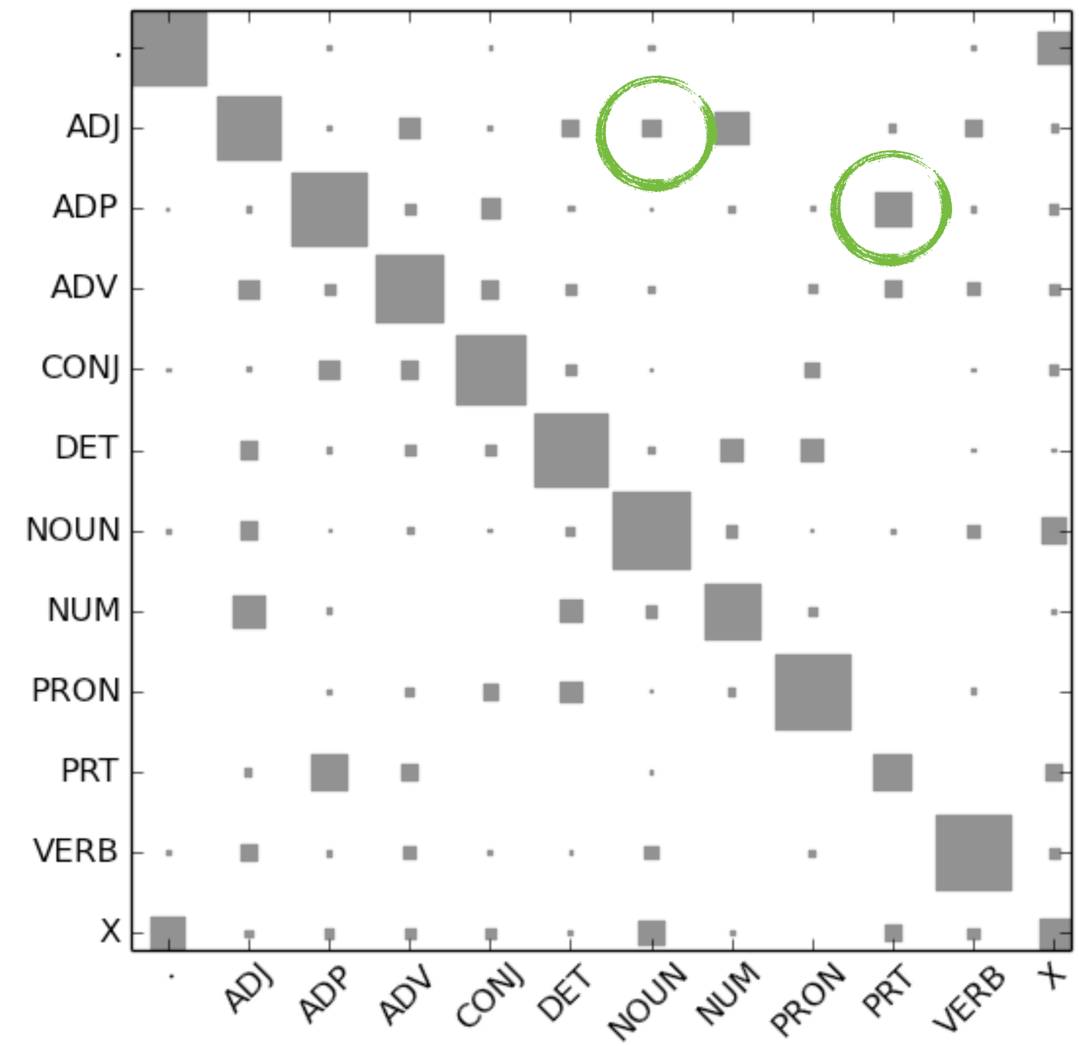


Twitter

(Plank et al., 2014)

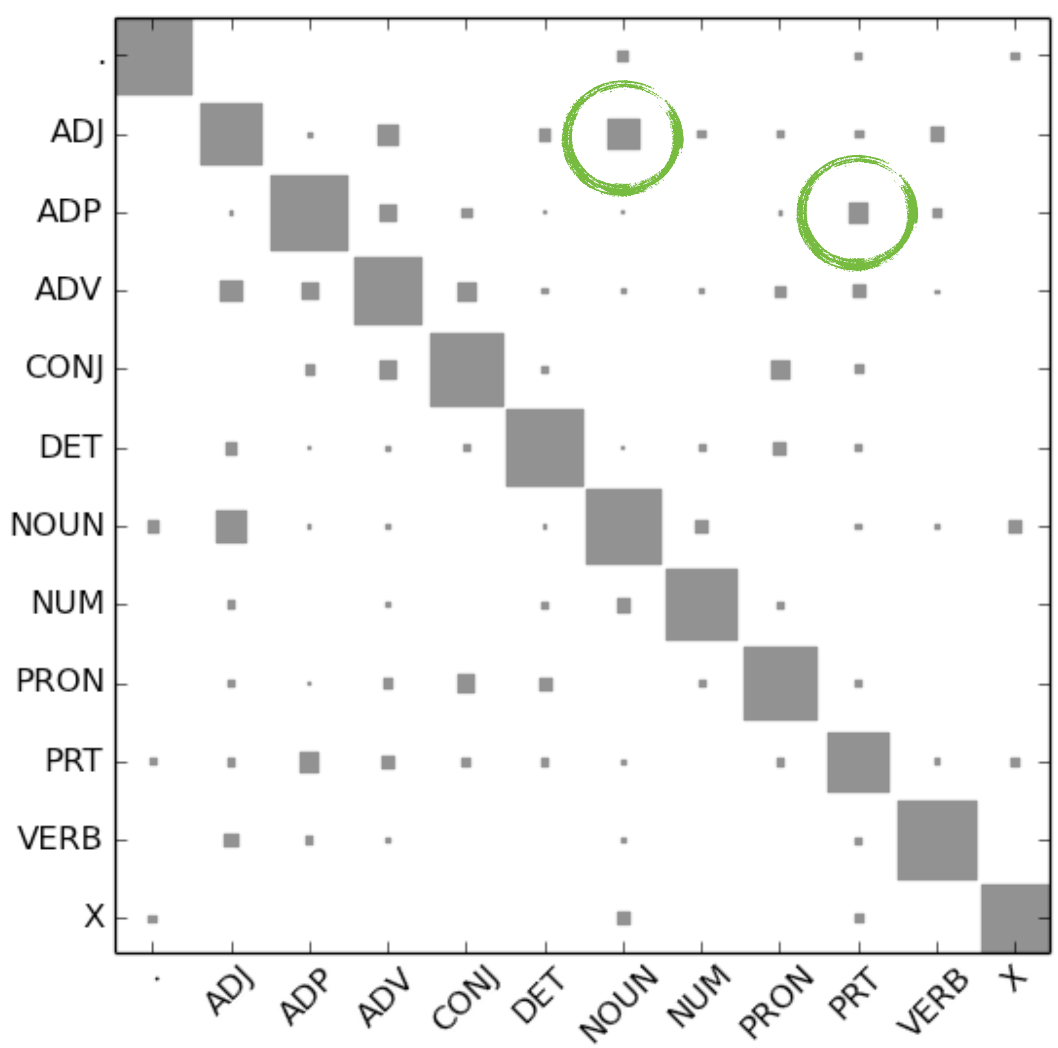


Wall Street Journal PTB-00

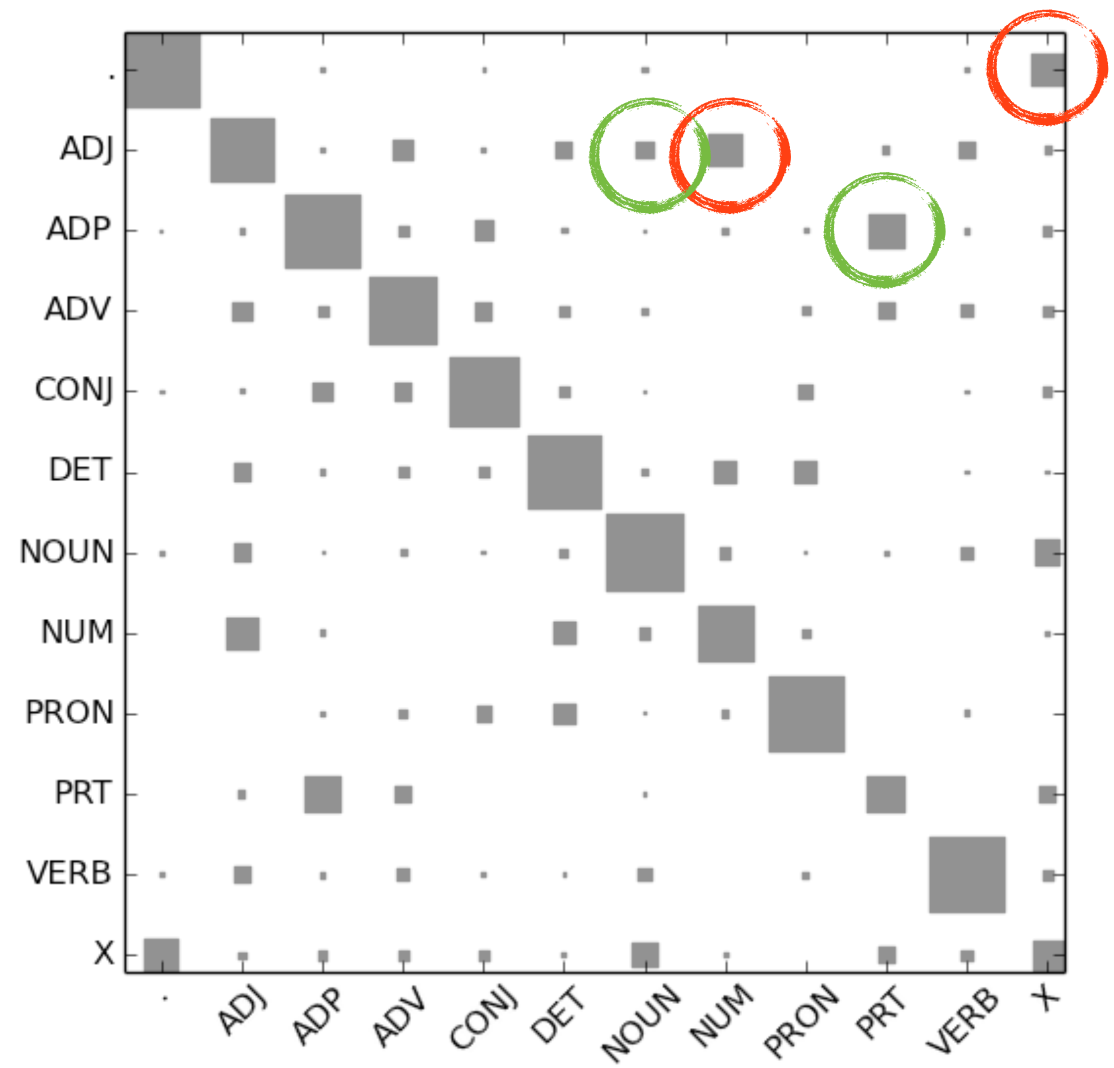


Twitter

(Plank et al., 2014)



Wall Street Journal PTB-00



Twitter

(Plank et al., 2014)

Is human label variation randomly distributed?

... and can we estimate it from small samples?

(Plank et al., 2014)

Is human label variation randomly distributed? No.

... and can we estimate it from small samples? Yes!

(Plank et al., 2014)

Are human label variation distributions unimodal?

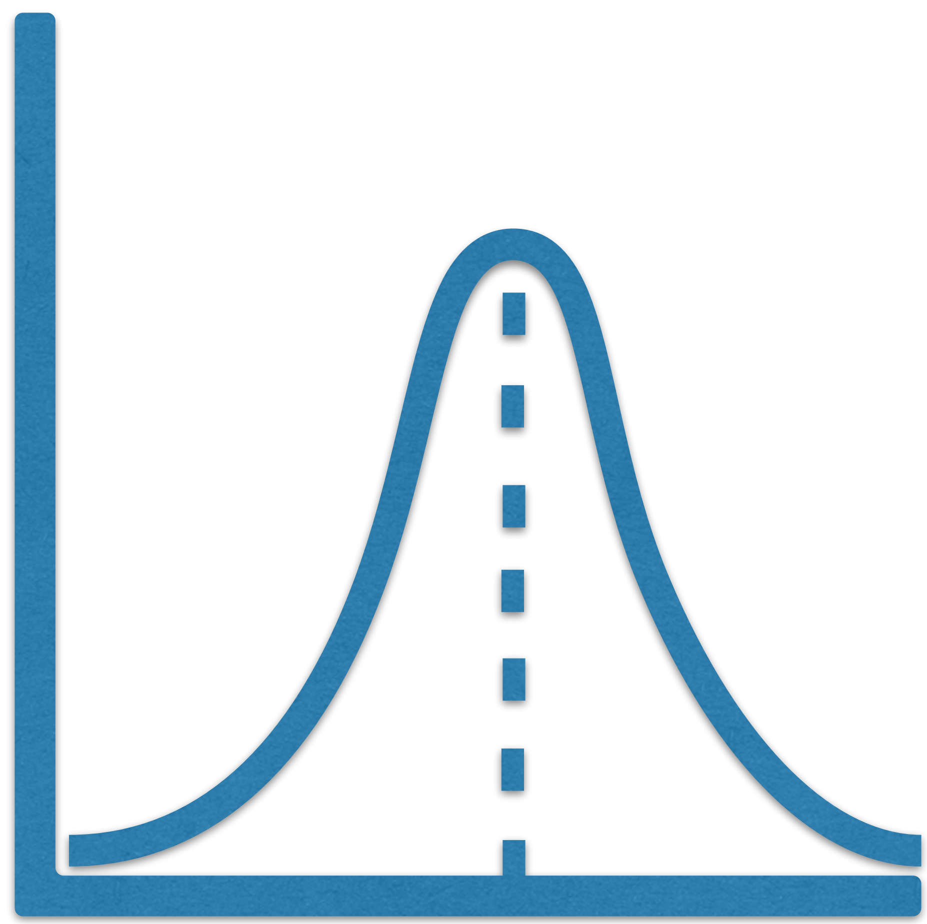
(Pavlick & Kwiatkovski, 2019)

Are human label variation distributions unimodal?

… do they contain inherent variation signal?

(Pavlick & Kwiatkovski, 2019)

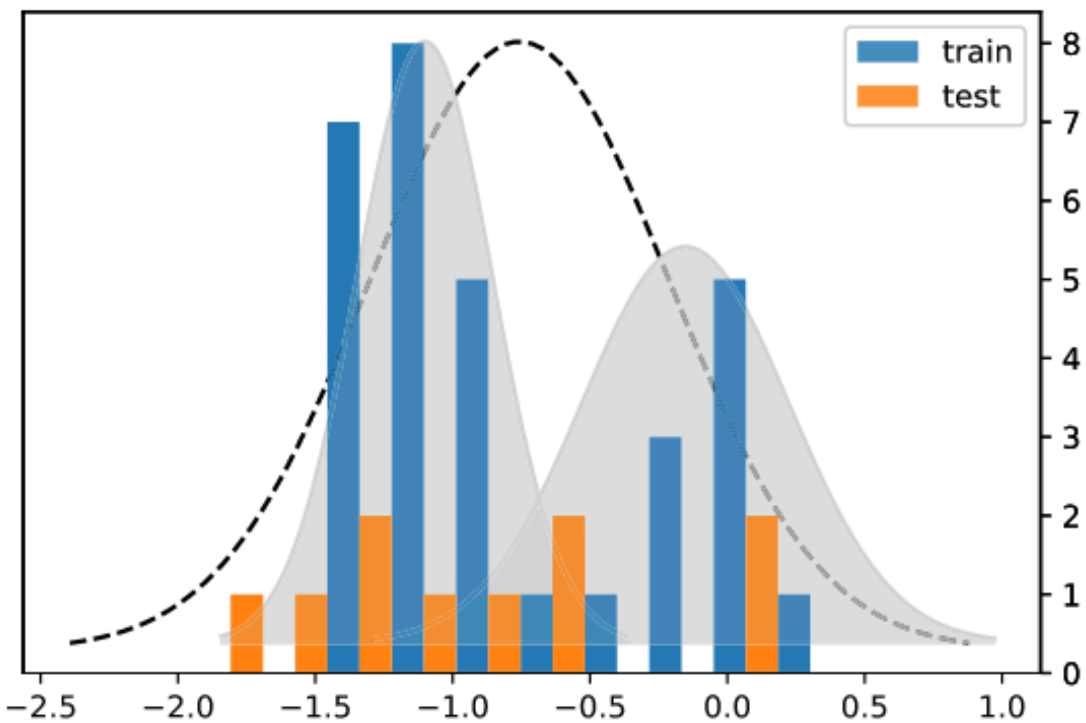
Unimodal (= Single Ground Truth)?



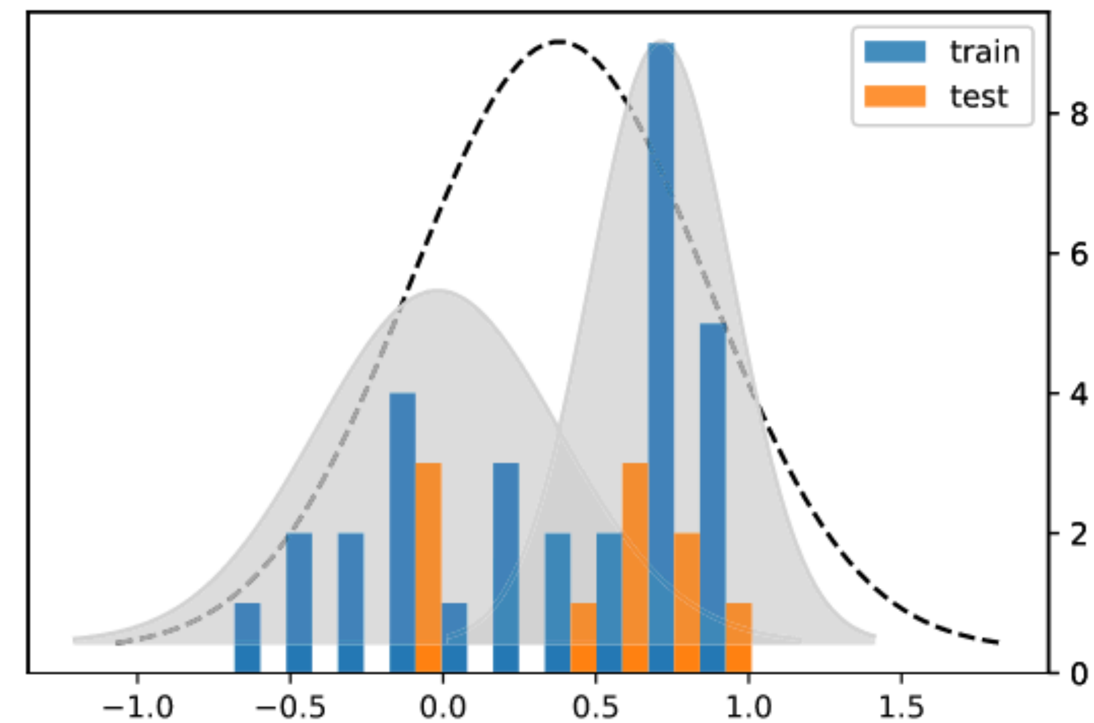
(Pavlick & Kwiatkovski, 2019)

Examples with bi-modal human judgement distributions

p: A homeless man being observed by a man in business attire.
h: Two men are sleeping in a hotel.



p: Paula swatted the fly.
h: The swatting happened in a forceful manner.

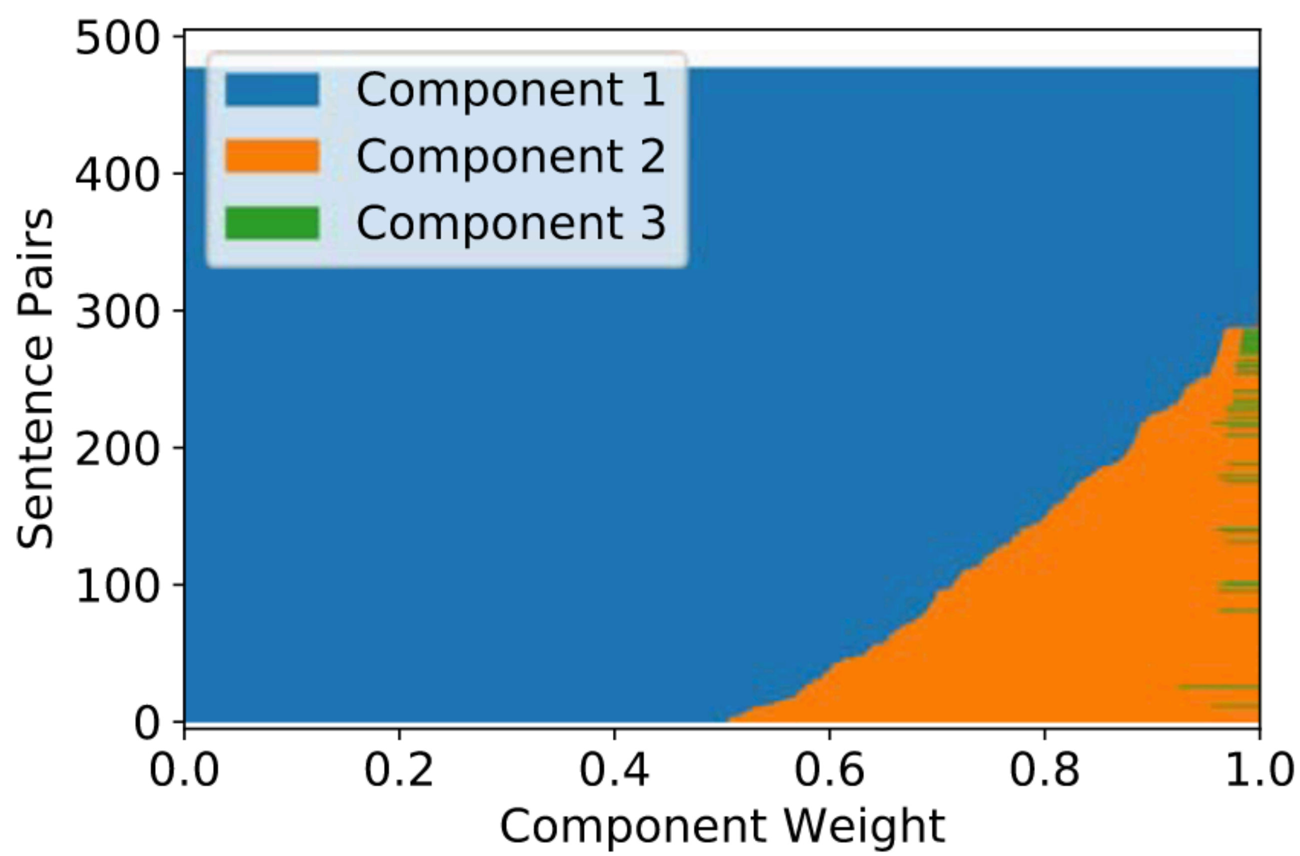


GMM with 1 *component* vs *k components*

(Pavlick & Kwiatkovski, 2019)

RTE Re-Annotation Analysis

"For 20% of the sentence pairs, there is a non-trivial second component"



(Pavlick & Kwiatkovski, 2019)

Are human label variation distributions unimodal?

… do they contain inherent variation signal?

(Pavlick & Kwiatkovski, 2019)

Are human label variation distributions unimodal?

No.

… do they contain inherent variation signal?

Yes!

(Pavlick & Kwiatkovski, 2019)

Human label variation is signal.

Sources of human label variation

(Basile et al., 2021)

- **Stimulus characteristics** (ambiguity, task difficulty)
- **Individual differences** (incl. cultural and socio-demographics): for example in hate speech or sentiment
- **Context and attention** (Intra-coder disagreement; attention slips play a non-negligible role as well; Beigman Klebanov et al., 2008)

- Very recent work: Taxonomy of disagreement reasons for NLI (Jiang & de Marneffe, TACL 2022 paper)

Roadmap: Three perspectives

1. Data: Is human label variation (HLV) random noise or signal?
2. Modelling: How can we leverage human label variation?
3. Evaluation: How to evaluate in light of human label variation?

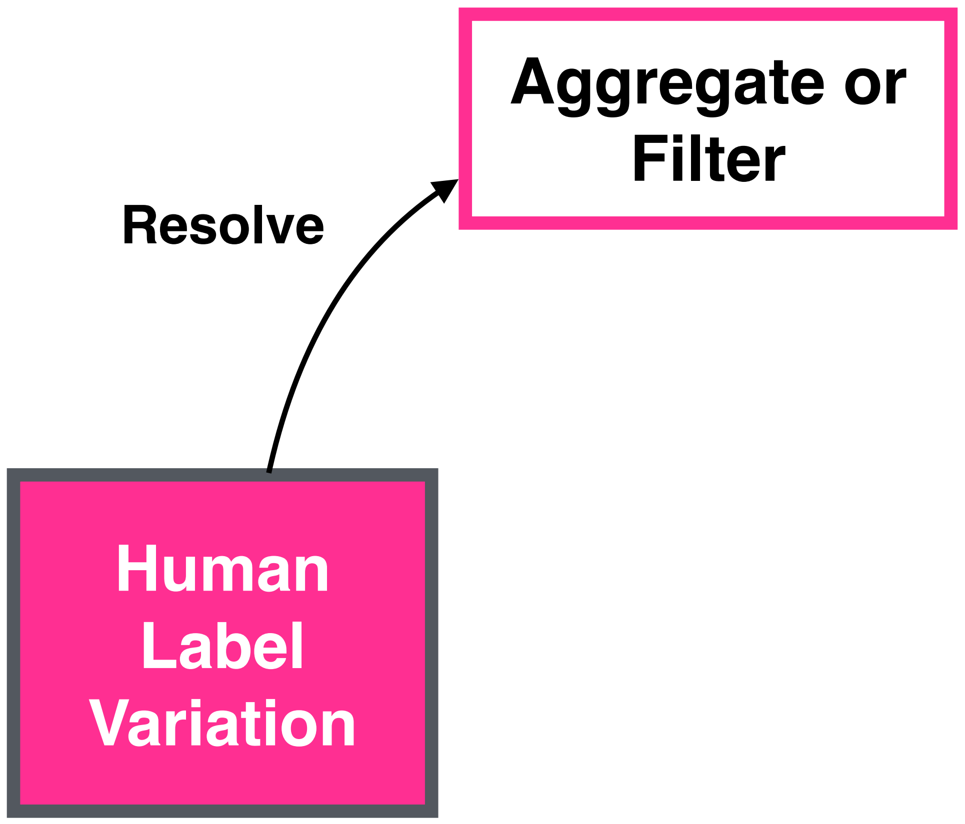
So what can we do?

Act II: Modelling

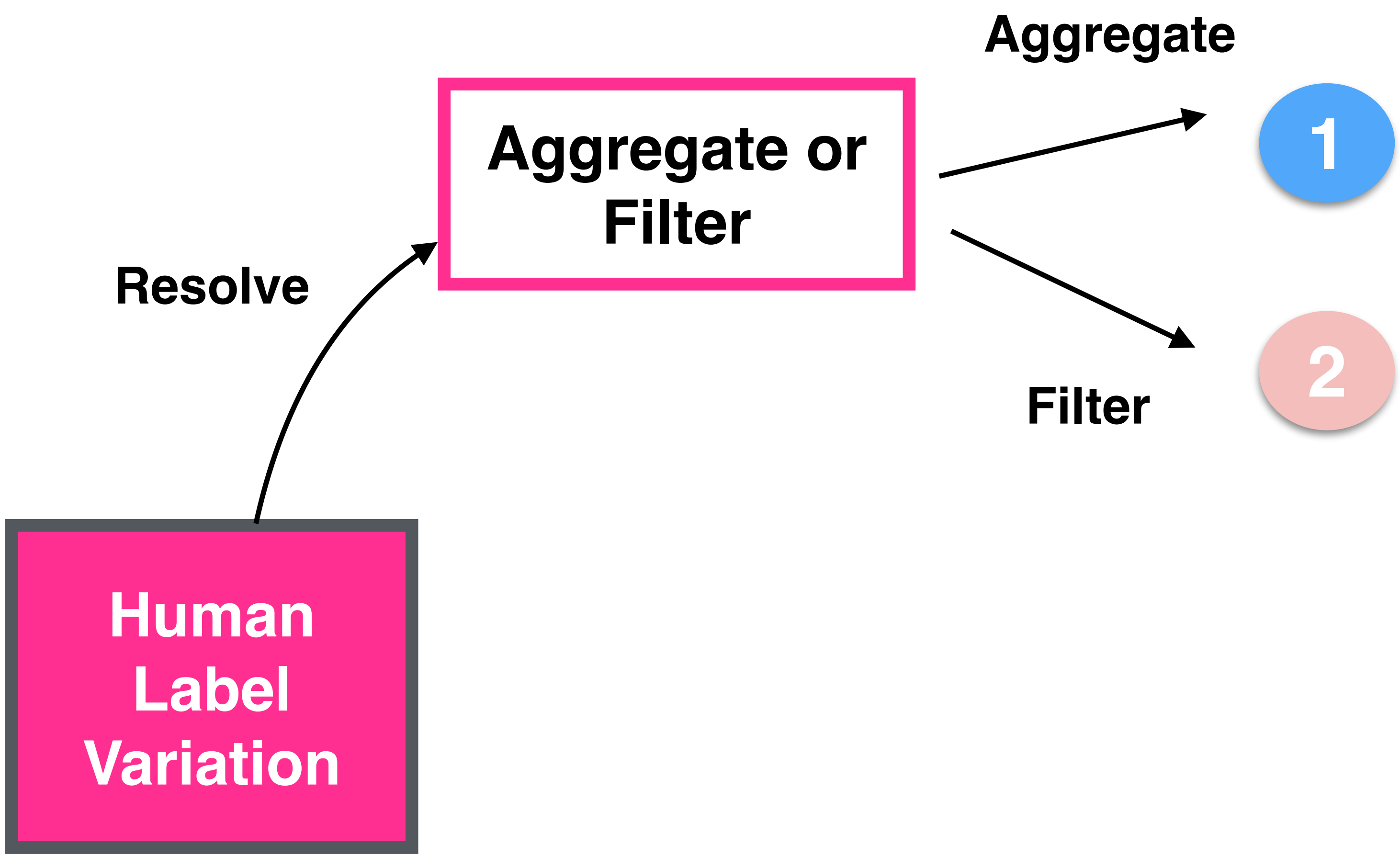
A Taxonomy for Learning with HLV

Human Label Variation

A Taxonomy for Learning with HLV



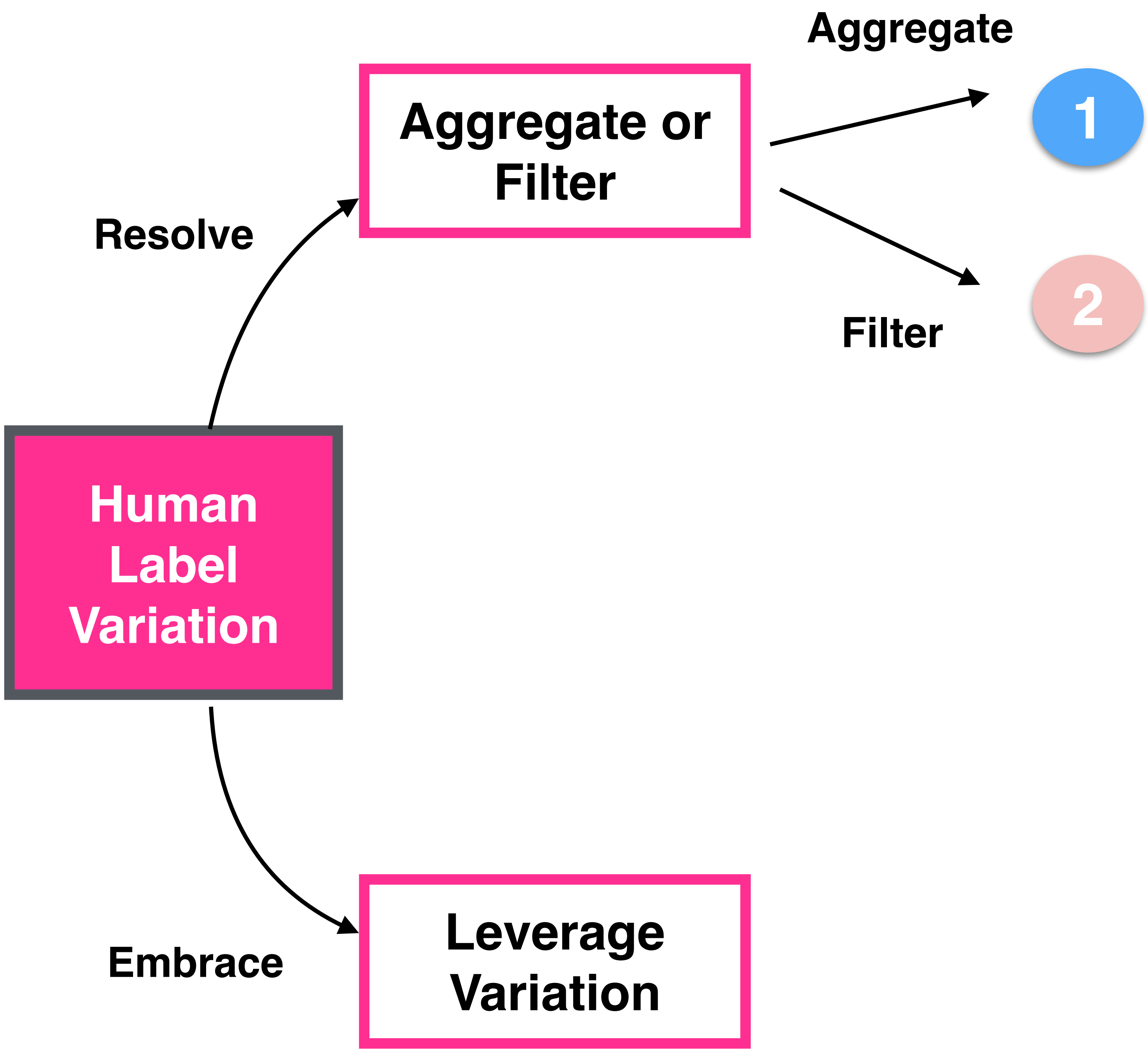
A Taxonomy for Learning with HLV



Aggregation methods
(e.g. Dawid & Skene, 1979; Hovy et al., 2013;
Paun, Artstein, Poesio, Morgan & Claypool book)

Filter data
(e.g. hard filtering by Reidsema &
van den Akker, 2008; Beigman-Klebanov et al.)

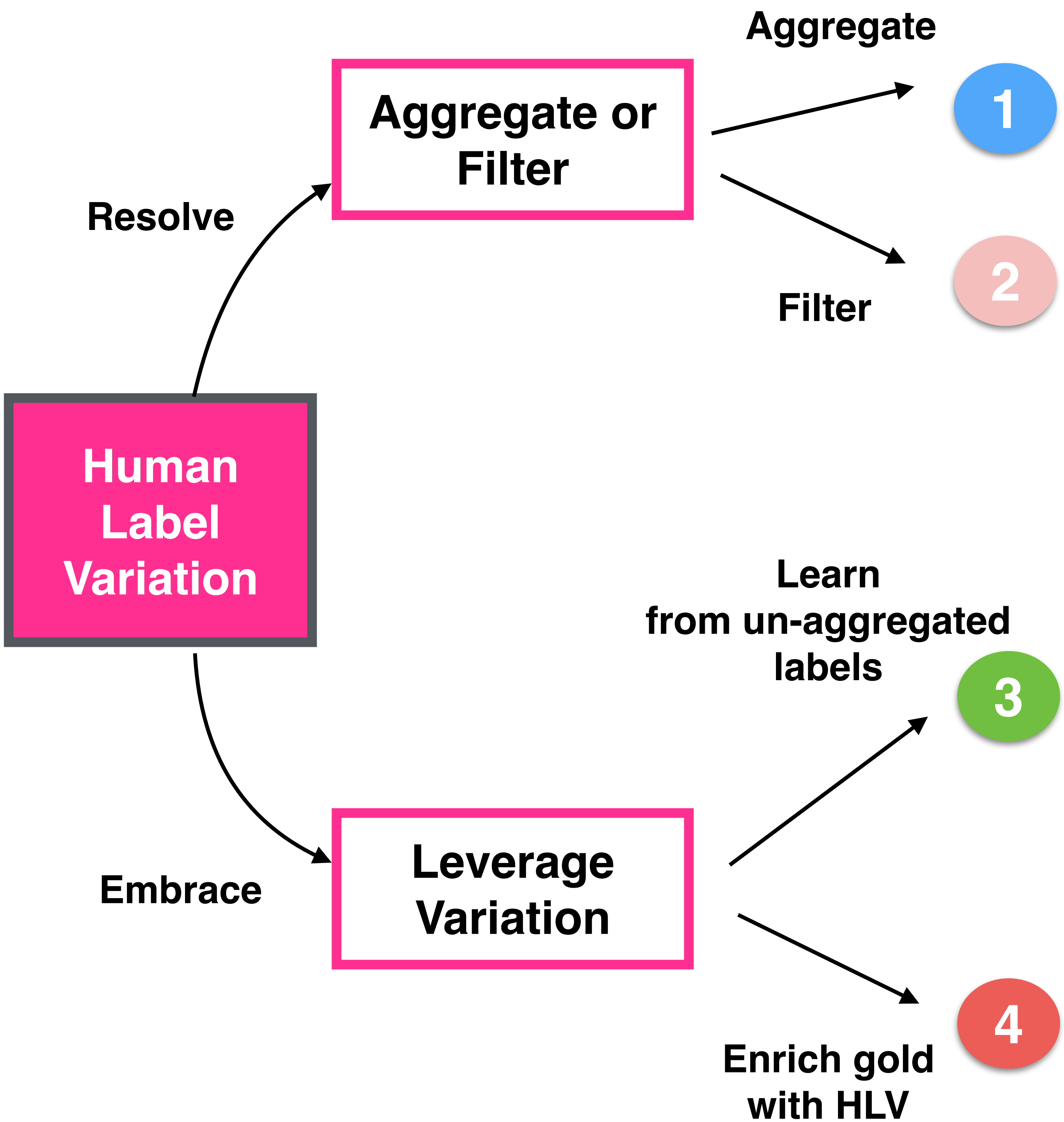
A Taxonomy for Learning with HLV



Aggregation methods
(e.g. Dawid & Skene, 1979; Hovy et al., 2013;
Paun, Artstein, Poesio, Morgan & Claypool book)

Filter data
(e.g. hard filtering by Reidsema &
van den Akker, 2008; Beigman-Klebanov et al.)

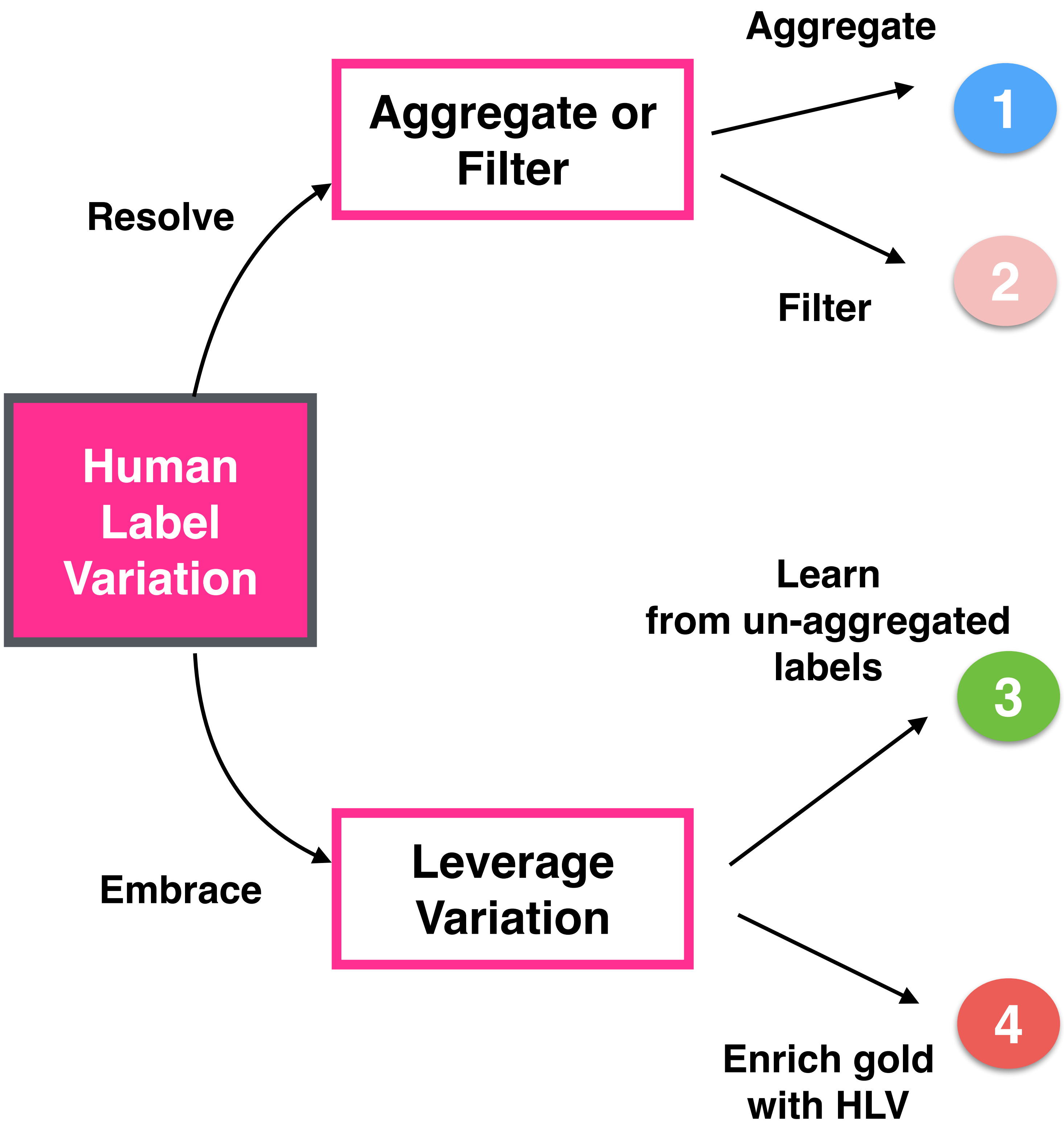
A Taxonomy for Learning with HLV



Aggregation methods
(e.g. Dawid & Skene, 1979; Hovy et al., 2013; Paun, Artstein, Poesio, Morgan & Claypool book)

Filter data
(e.g. hard filtering by Reidsema & van den Akker, 2008; Beigman-Klebanov et al.)

A Taxonomy for Learning with HLV



1 **Aggregation methods**
(e.g. Dawid & Skene, 1979; Hovy et al., 2013; Paun, Artstein, Poesio, Morgan & Claypool book)

2 **Filter data**
(e.g. hard filtering by Reidsema & van den Akker, 2008; Beigman-Klebanov et al.)

3 **Repeated Labeling; Soft-labels, CrowdTruth, CrowdLayer, MTL**
(e.g. Peterson et al., 2019, Uma et al. 2020; Aroyo & Welty, 2014; Rodrigues & Peireia, 2017; Sheng et al., 2008; Specia & Cohn, 2013; Davani et al., 2021 TACL)

4 **Weighting, Multi-task Soft Loss**
(e.g. Plank et al., 2014; Fornaciari et al., 2021)

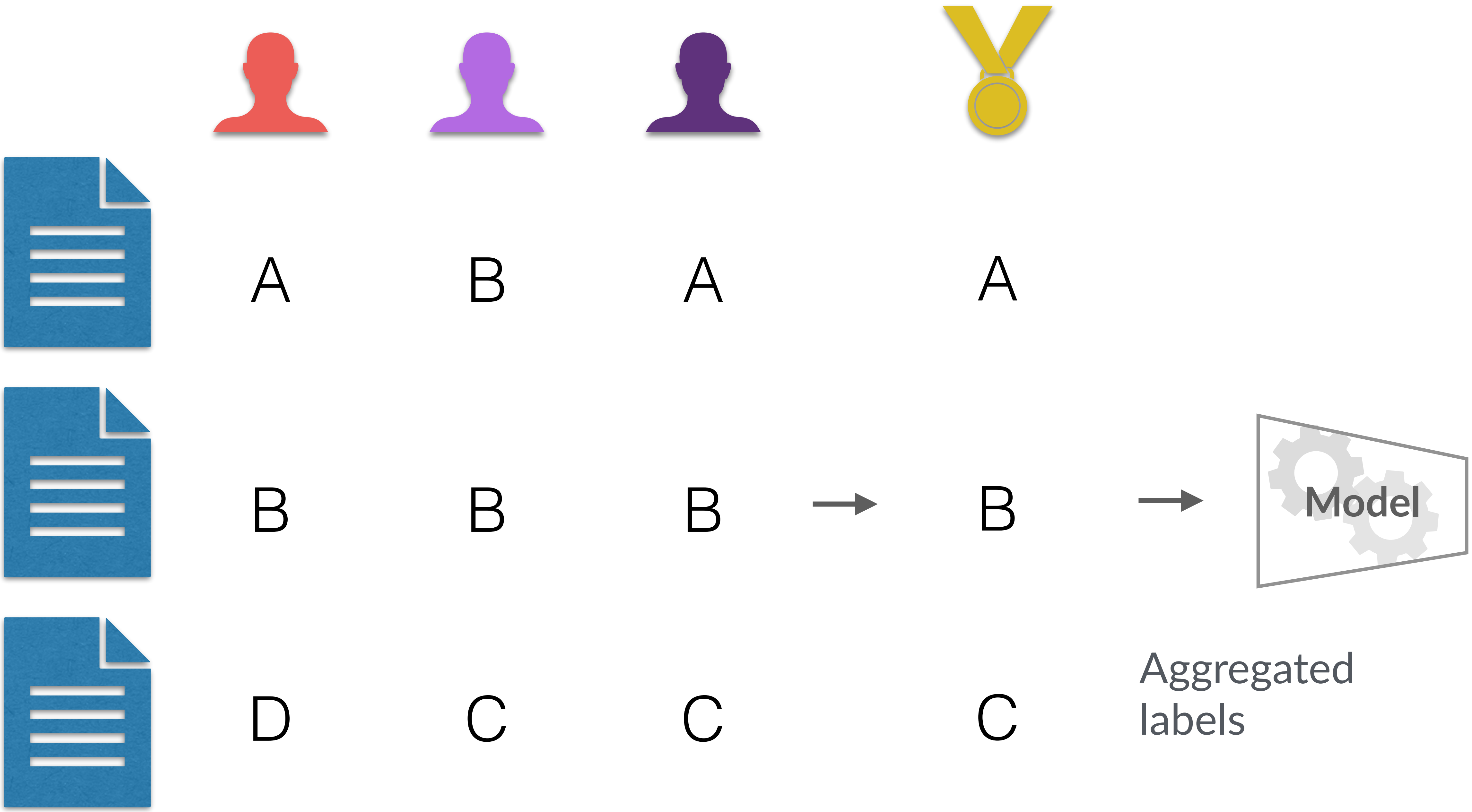
1 Aggregation

A	B	A
B	B	B
D	C	C

1 Aggregation

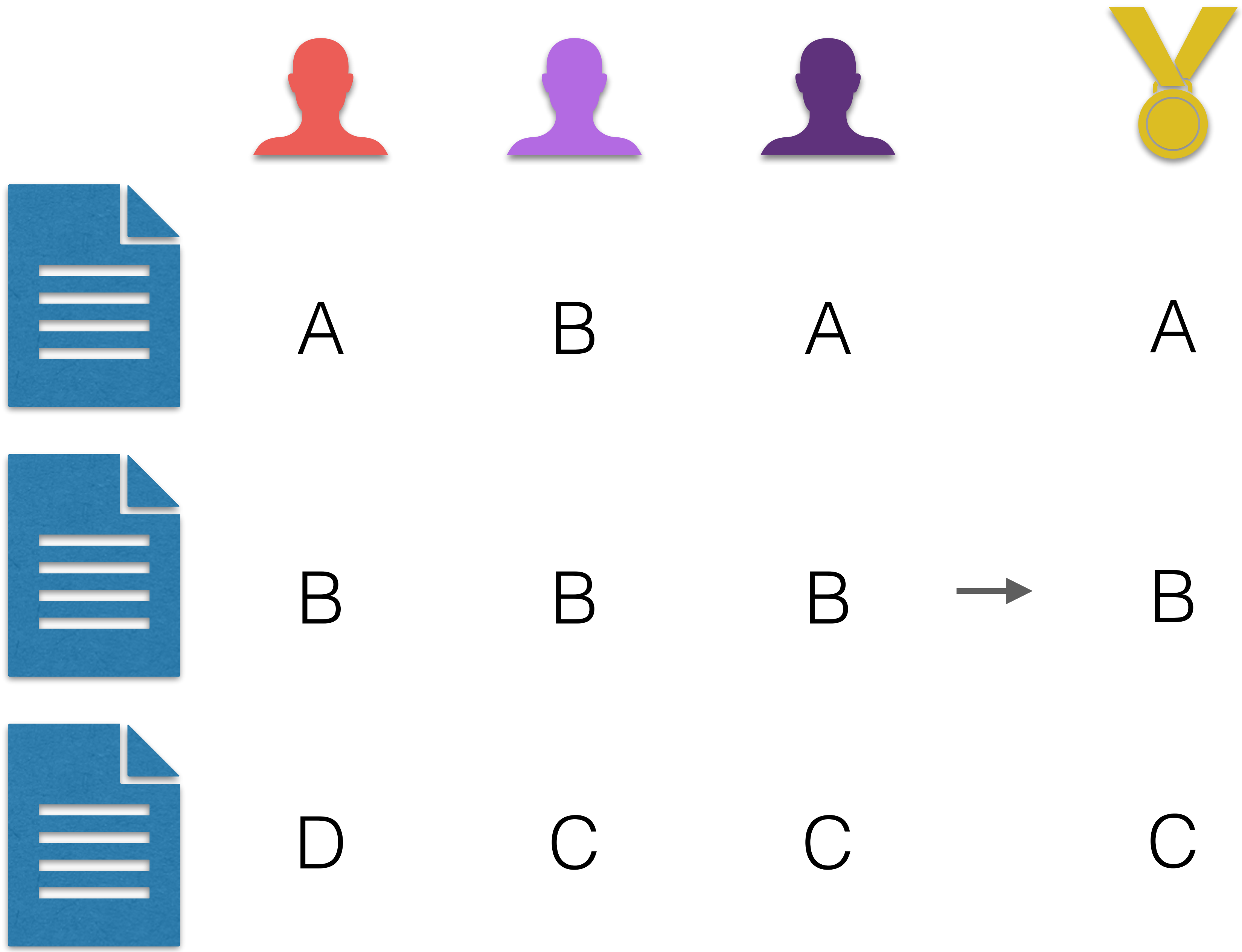
A	B	A		A
B	B	B	→	B
D	C	C		C

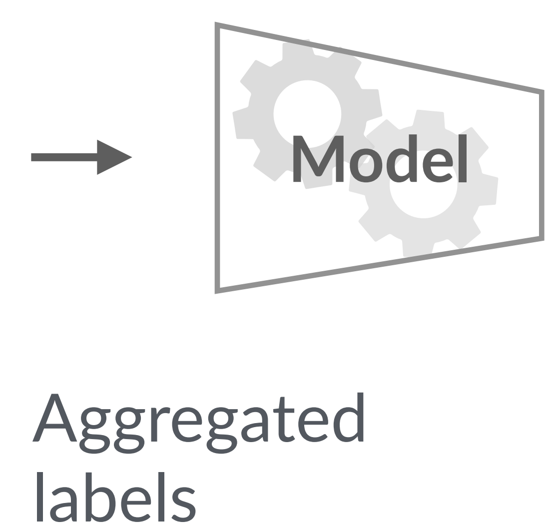
1 Aggregation



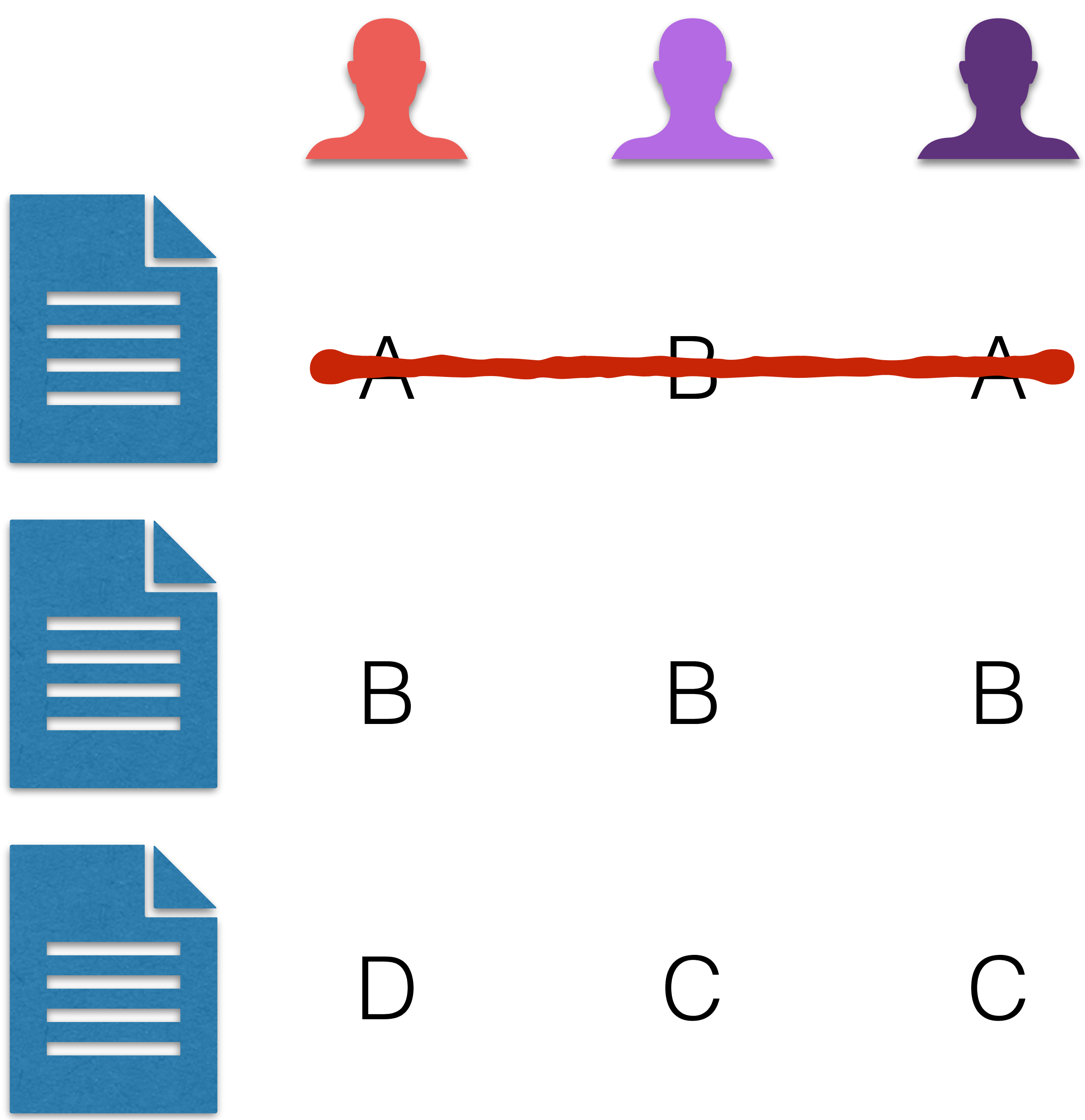
1 Aggregation

- **Widely-adopted**
- Allows only **one** label/belief/view

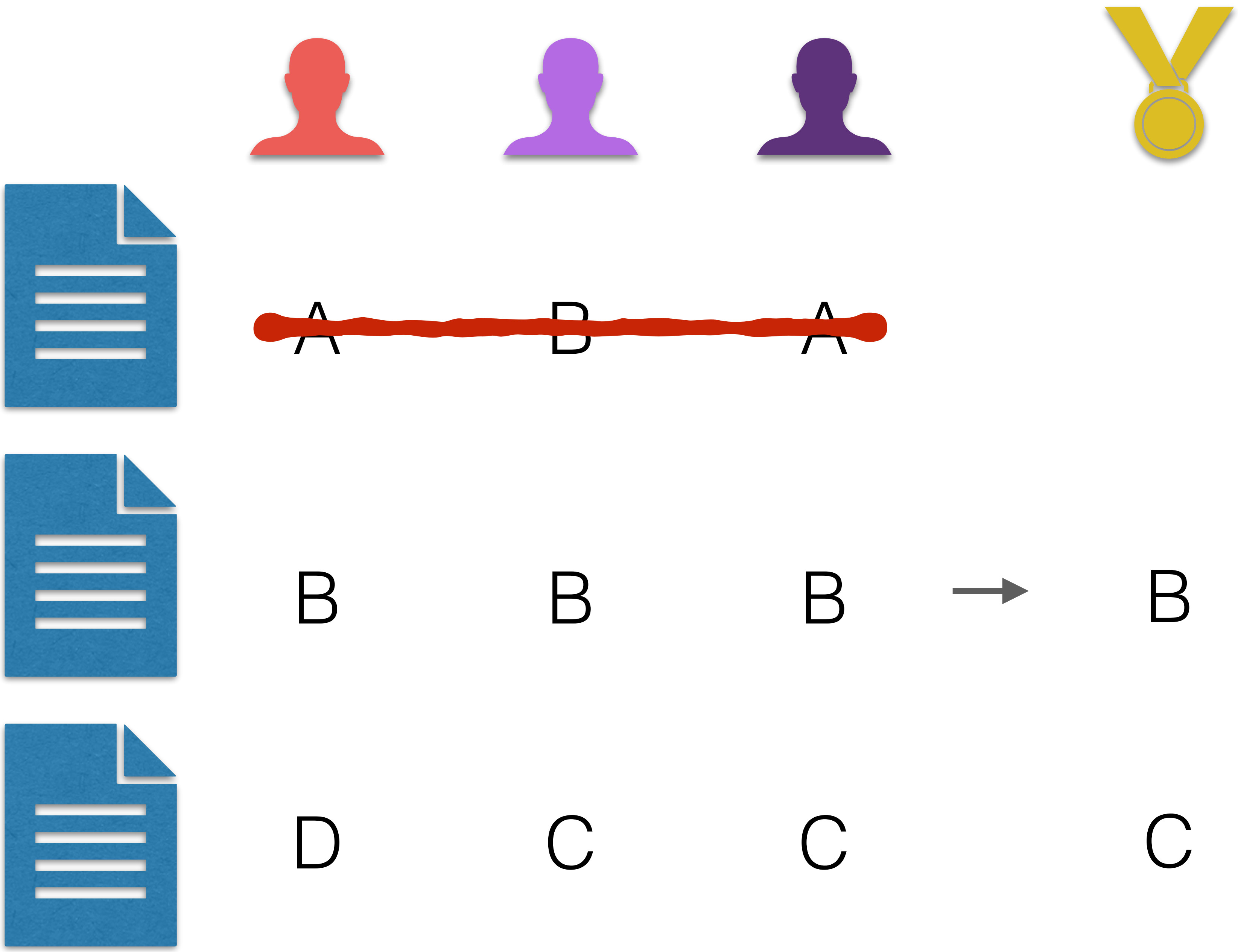




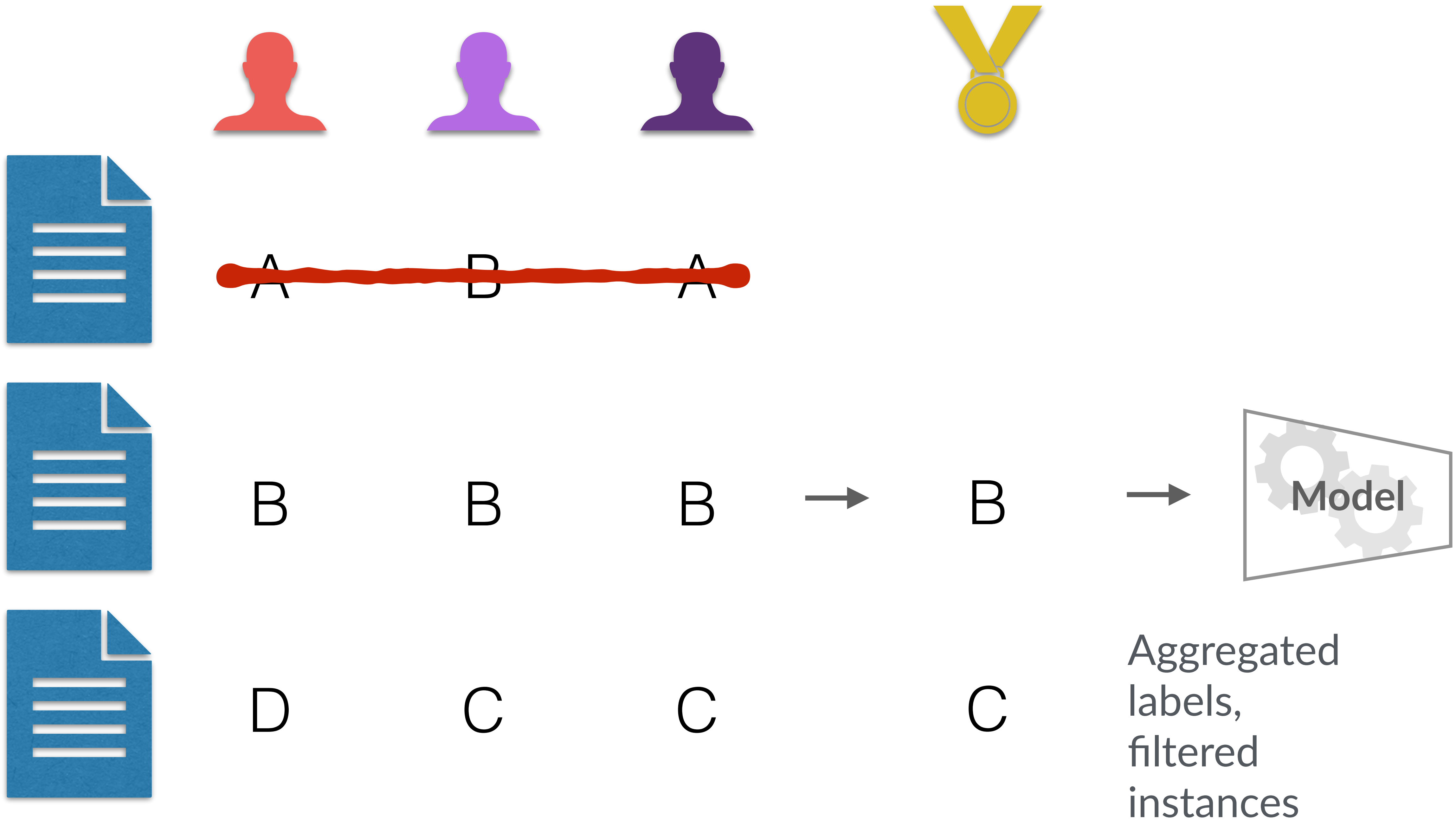
2 Filter



2 Filter

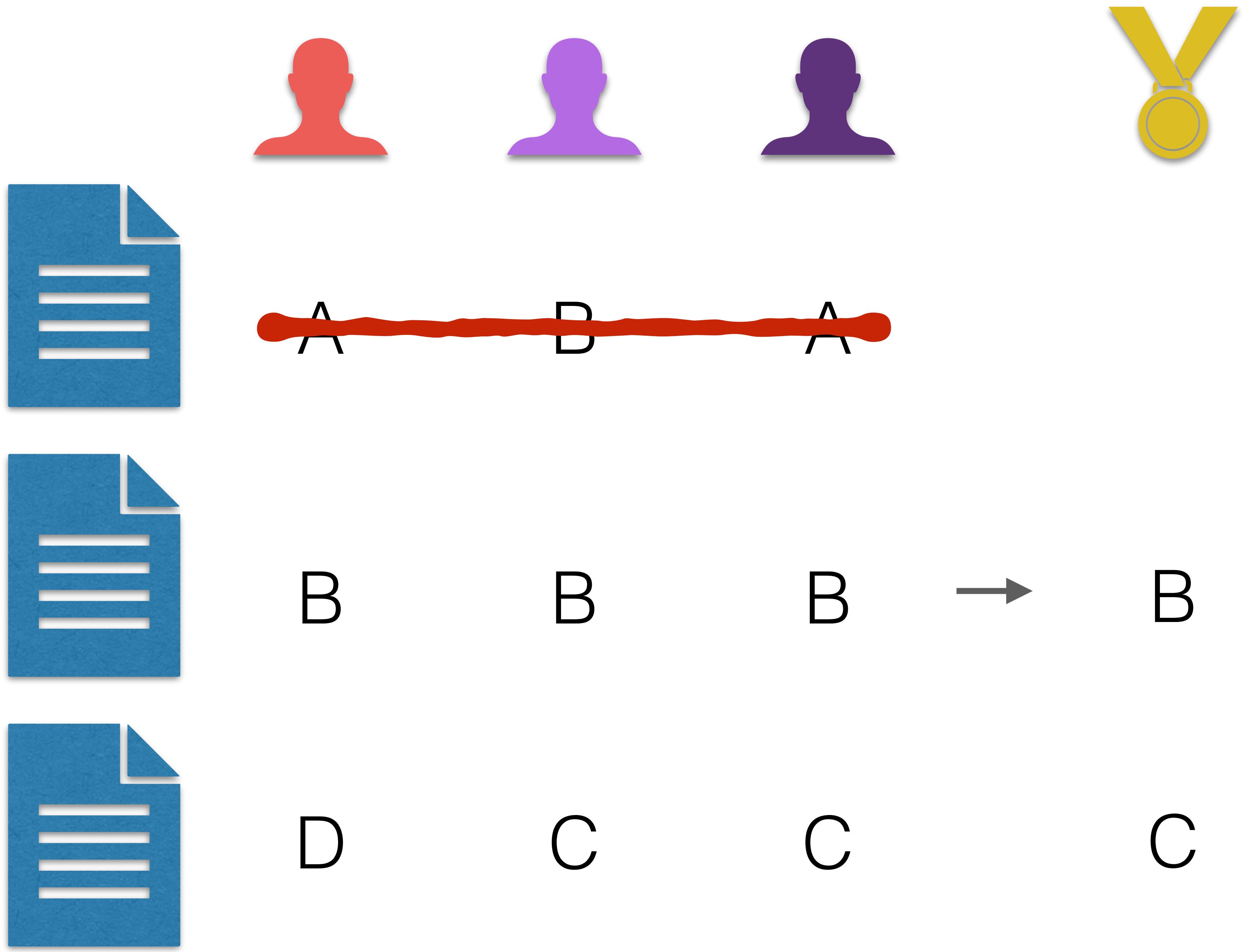


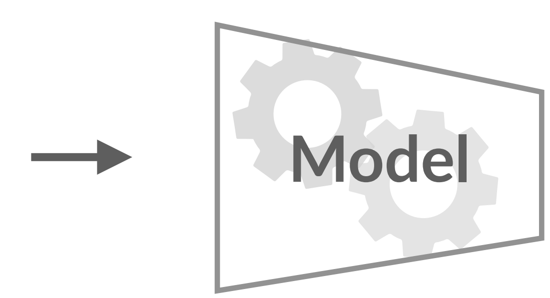
2 Filter



2 Filter

- **Neglects** genuine human nuances
- **Waste** of data



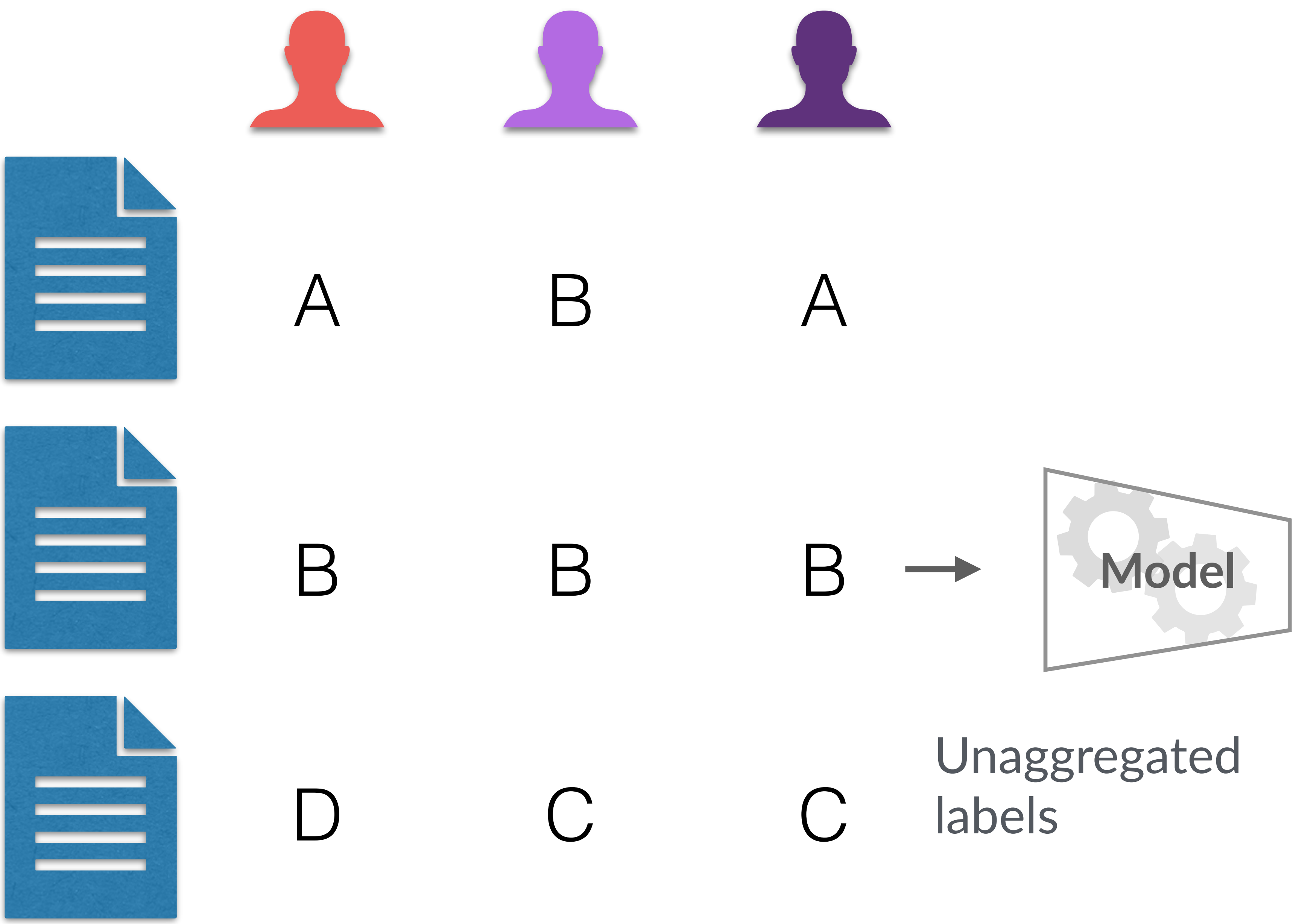


Aggregated labels, filtered instances

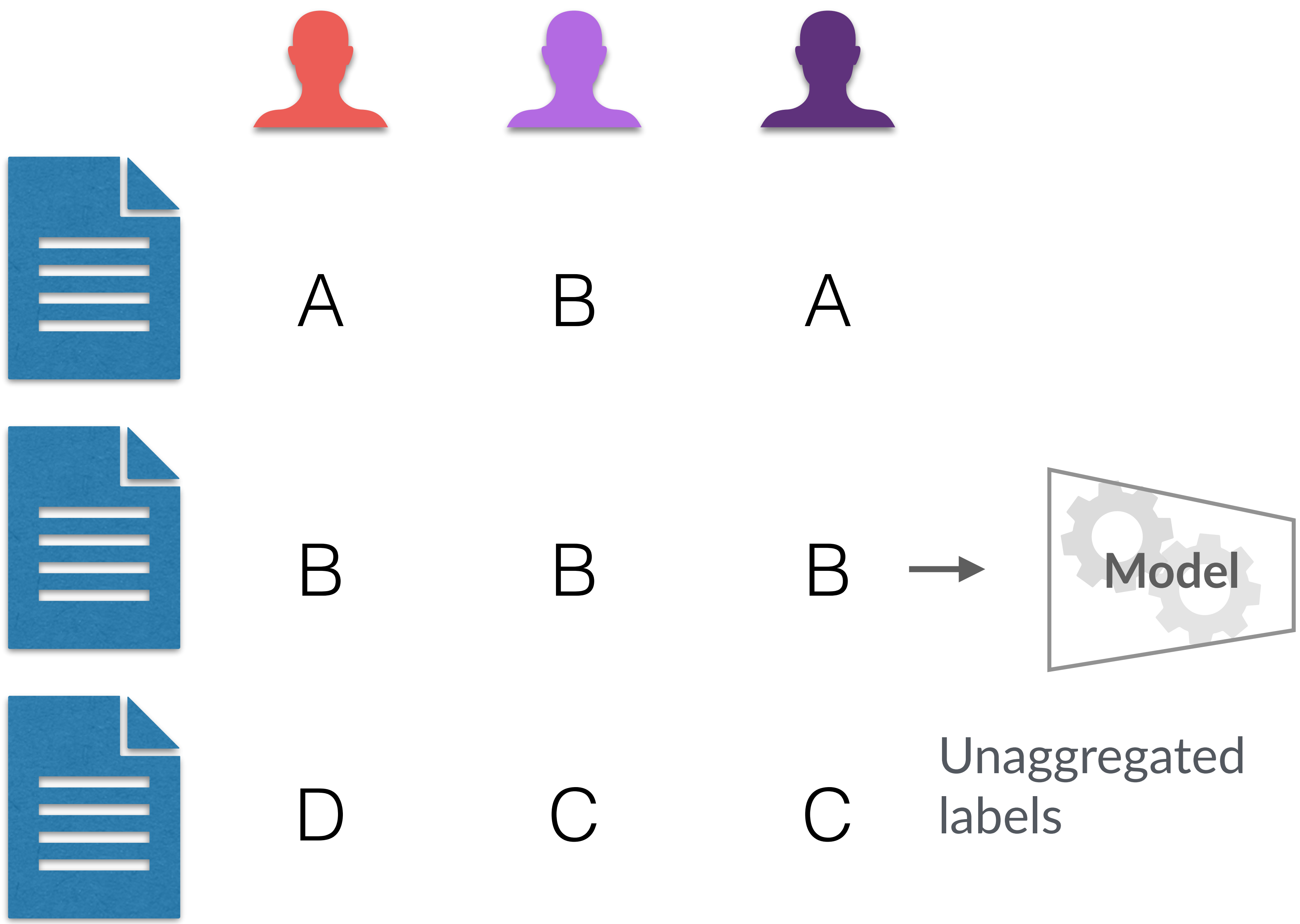
3 Learn from un-aggregated labels

A	B	A
B	B	B
D	C	C

3 Learn from un-aggregated labels



3 Learn from un-aggregated labels



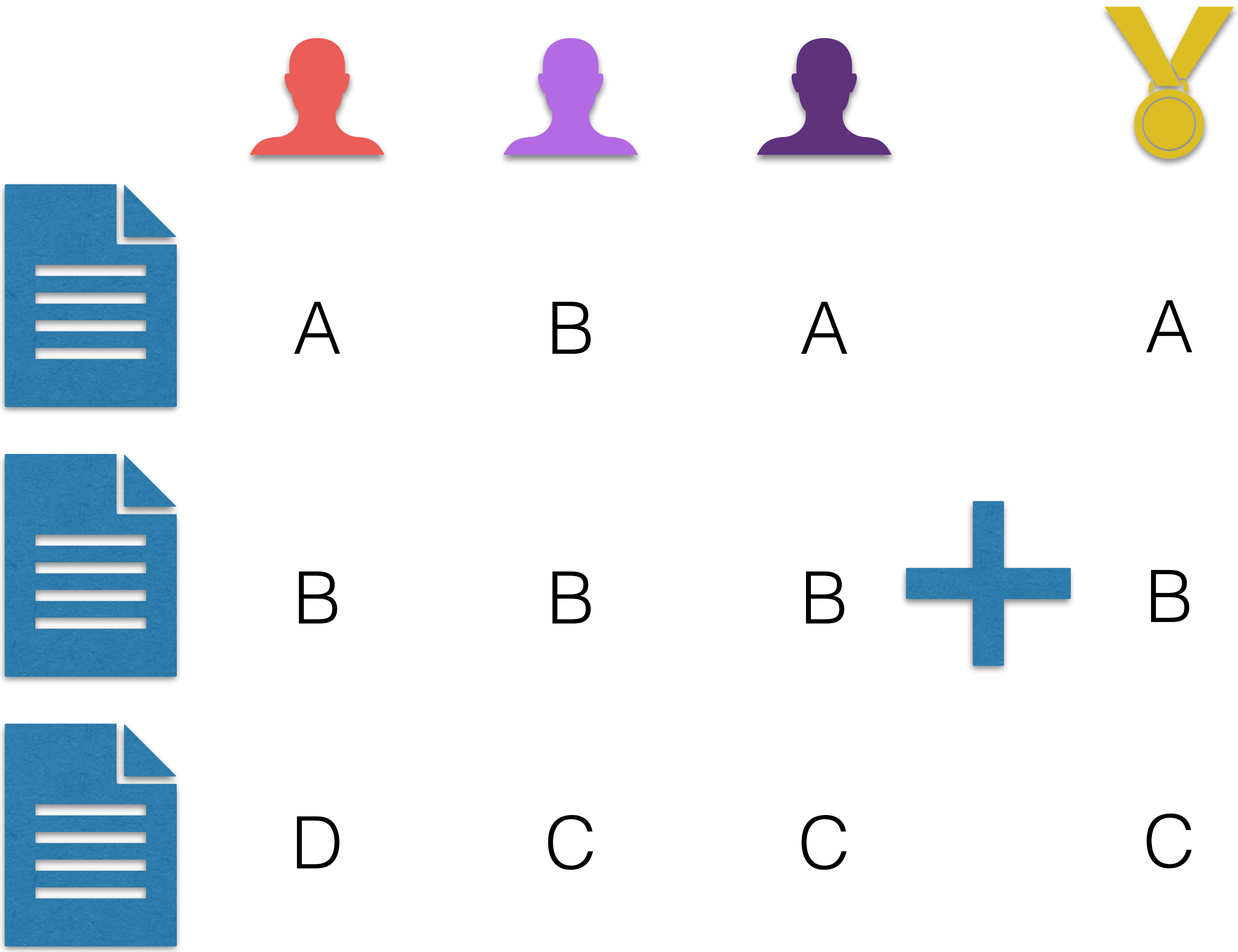
- **Embraces** nuances
- Methods of **varying** complexity: from general multi-label (Sheng et al., 2008) to architecture-specific
- So far **varying** success in NLP (see Uma et al., 2021)

4 Enrich gold with Distributions

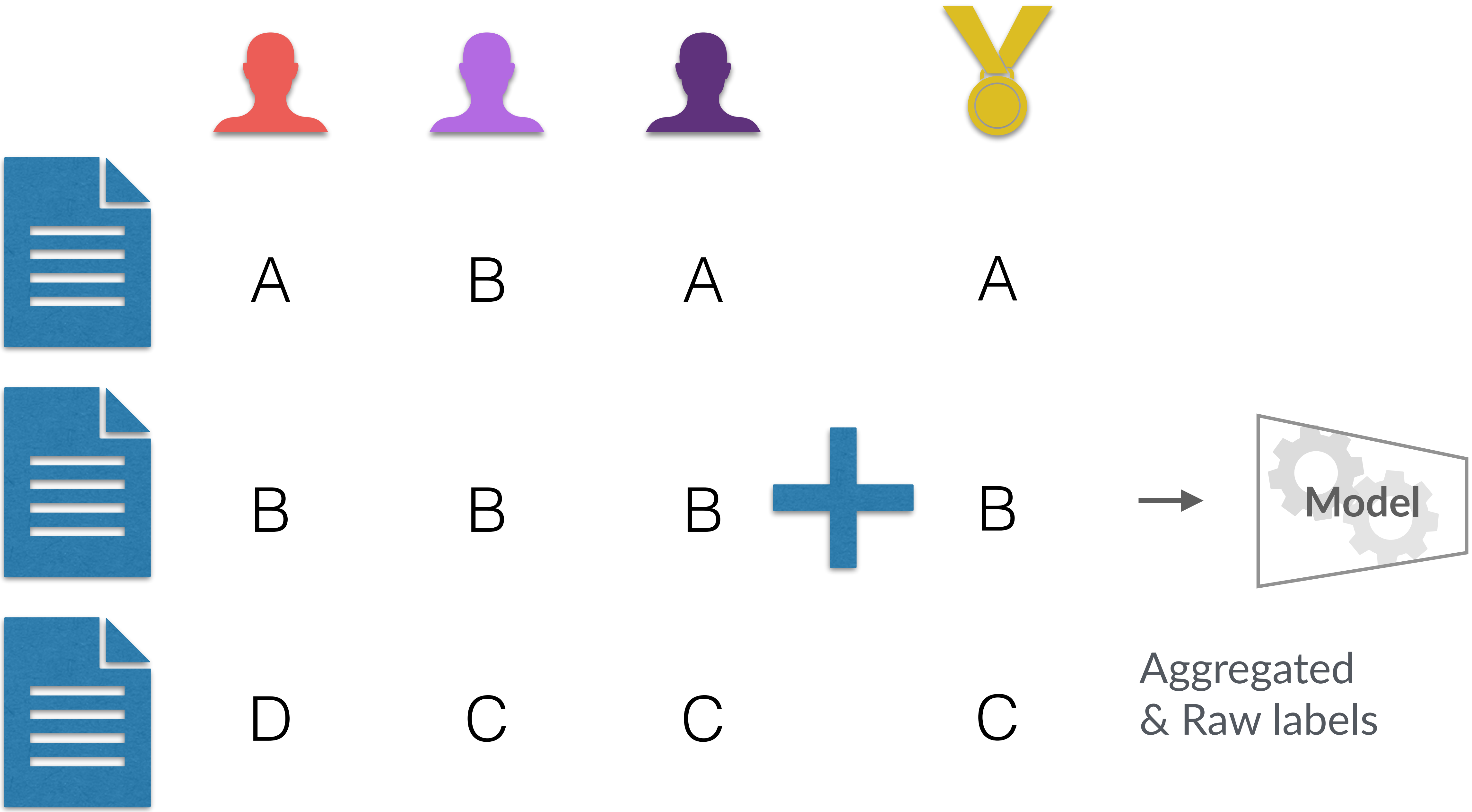
A	B	A
B	B	B
D	C	C

4

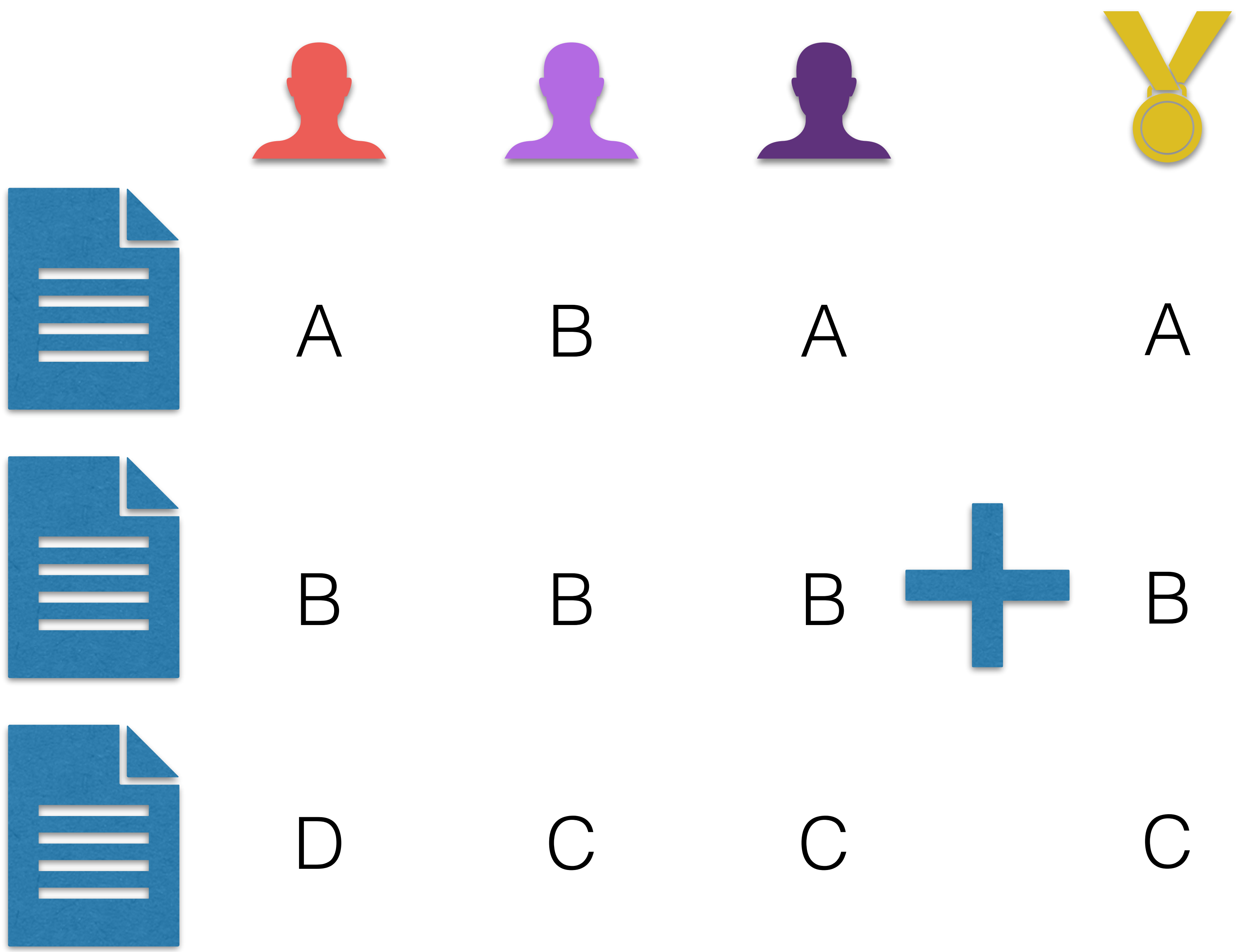
Enrich gold with Distributions



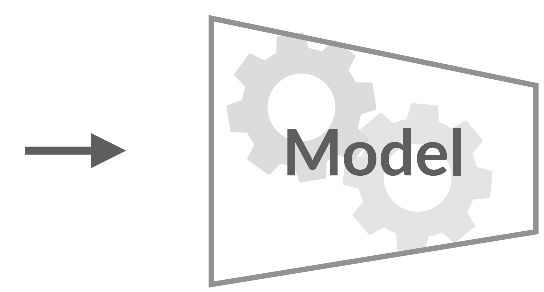
4 Enrich gold with Distributions



4 Enrich gold with Distributions



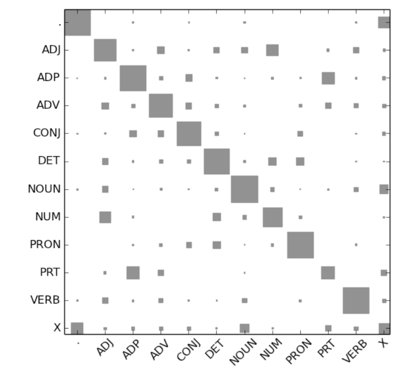
- **Embraces** nuances besides a "gold"
- **Regularization** effect



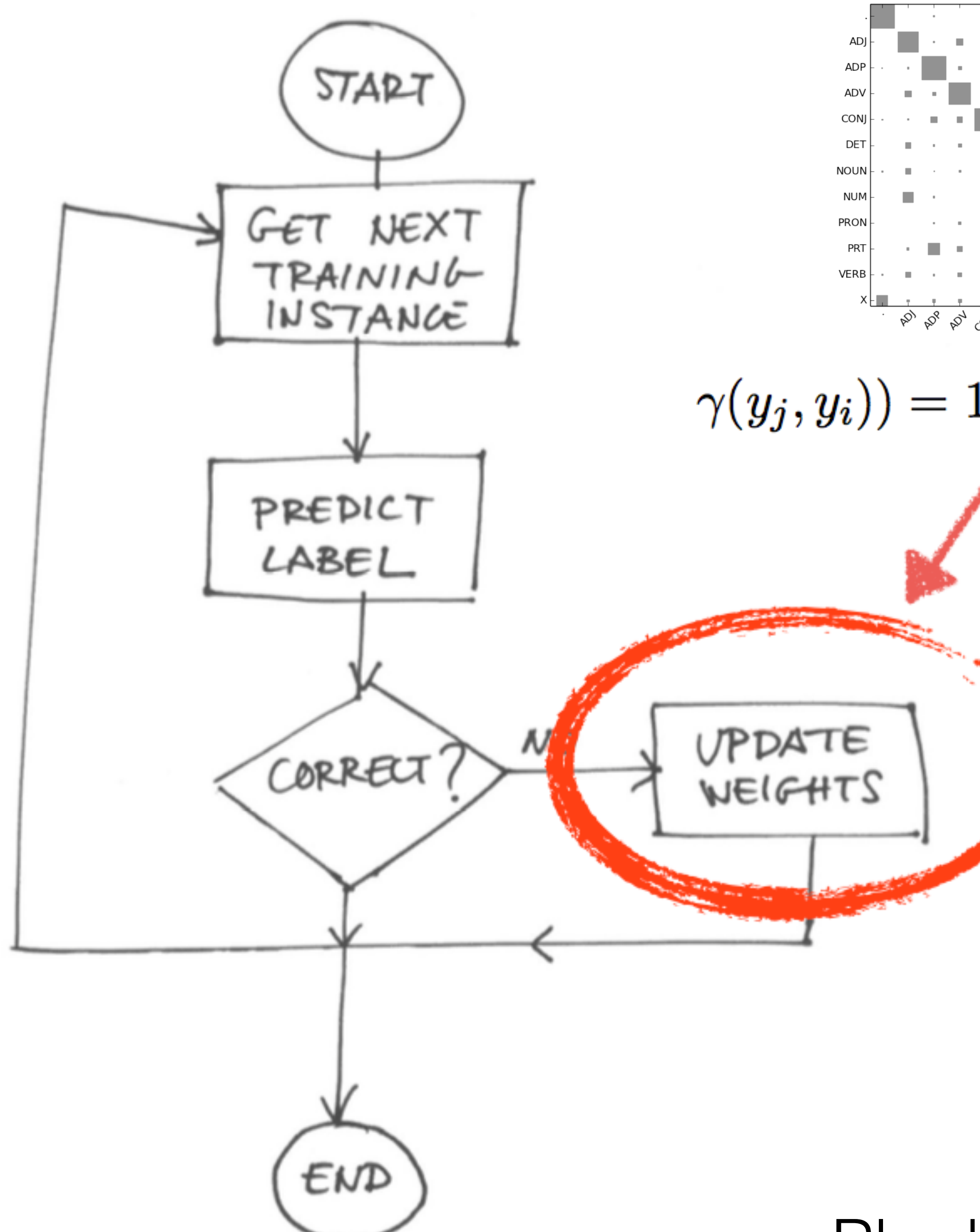
Aggregated & Raw labels

Example of 4: Weighting by Disagreement

CM (confusion matrix)

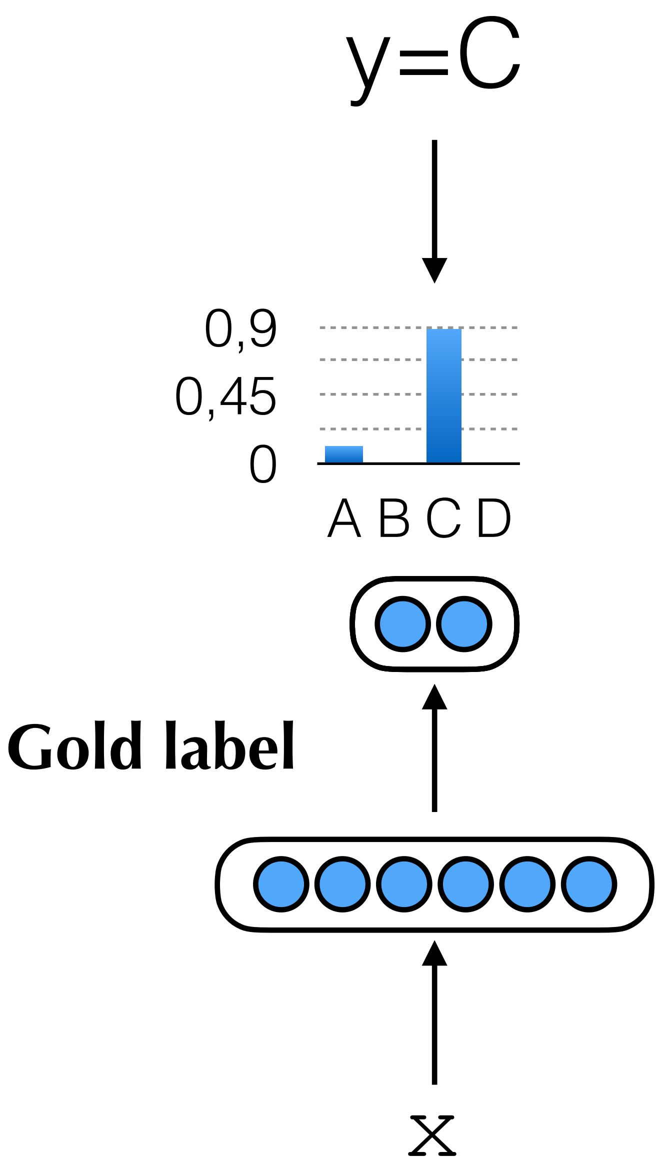


$$\gamma(y_j, y_i)) = 1 - P(\{A_1(X), A_2(X)\} = \{y_j, y_i\})$$



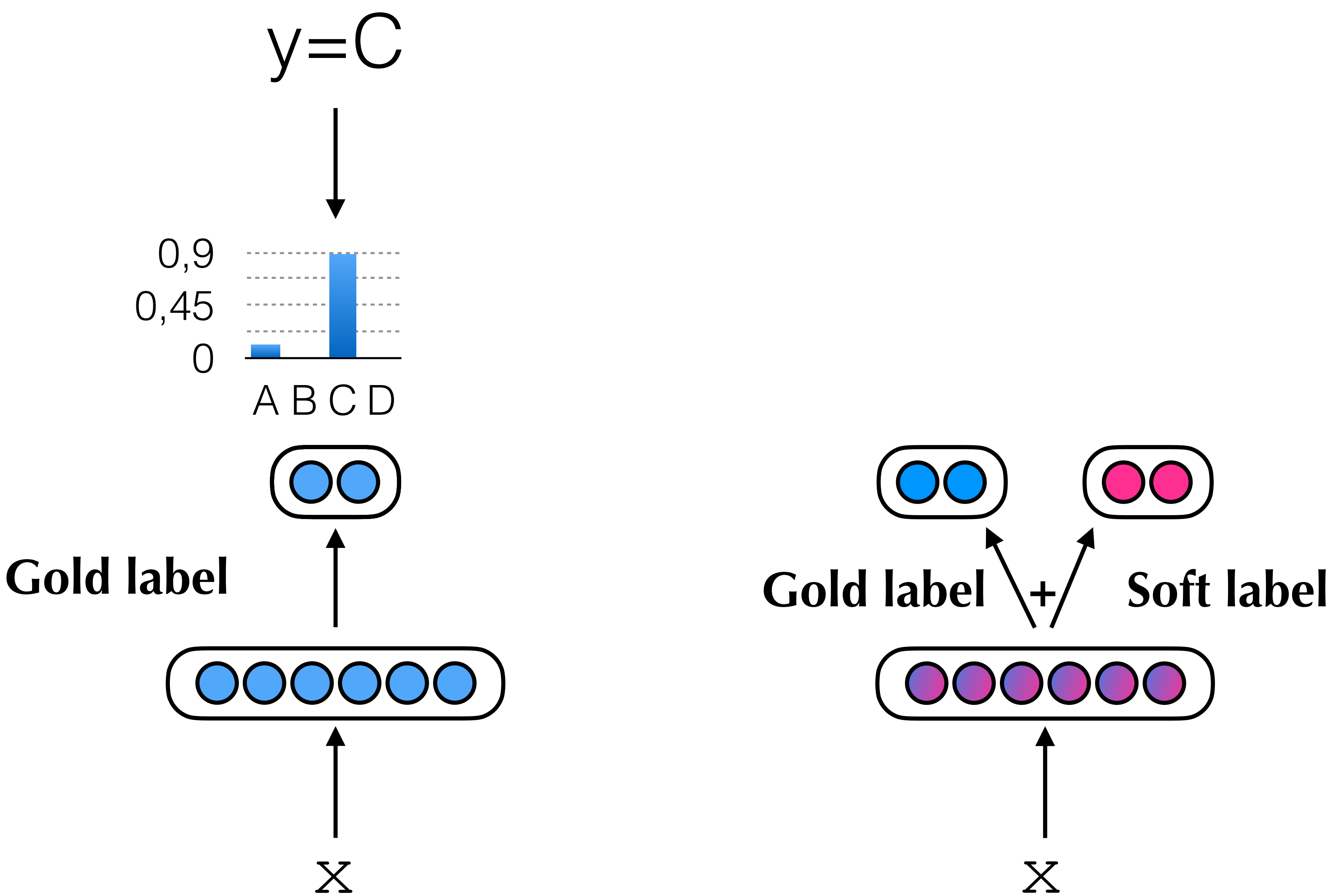
Plank, Hovy, Søgaard (2014)

Example of 4: Soft-label MTL



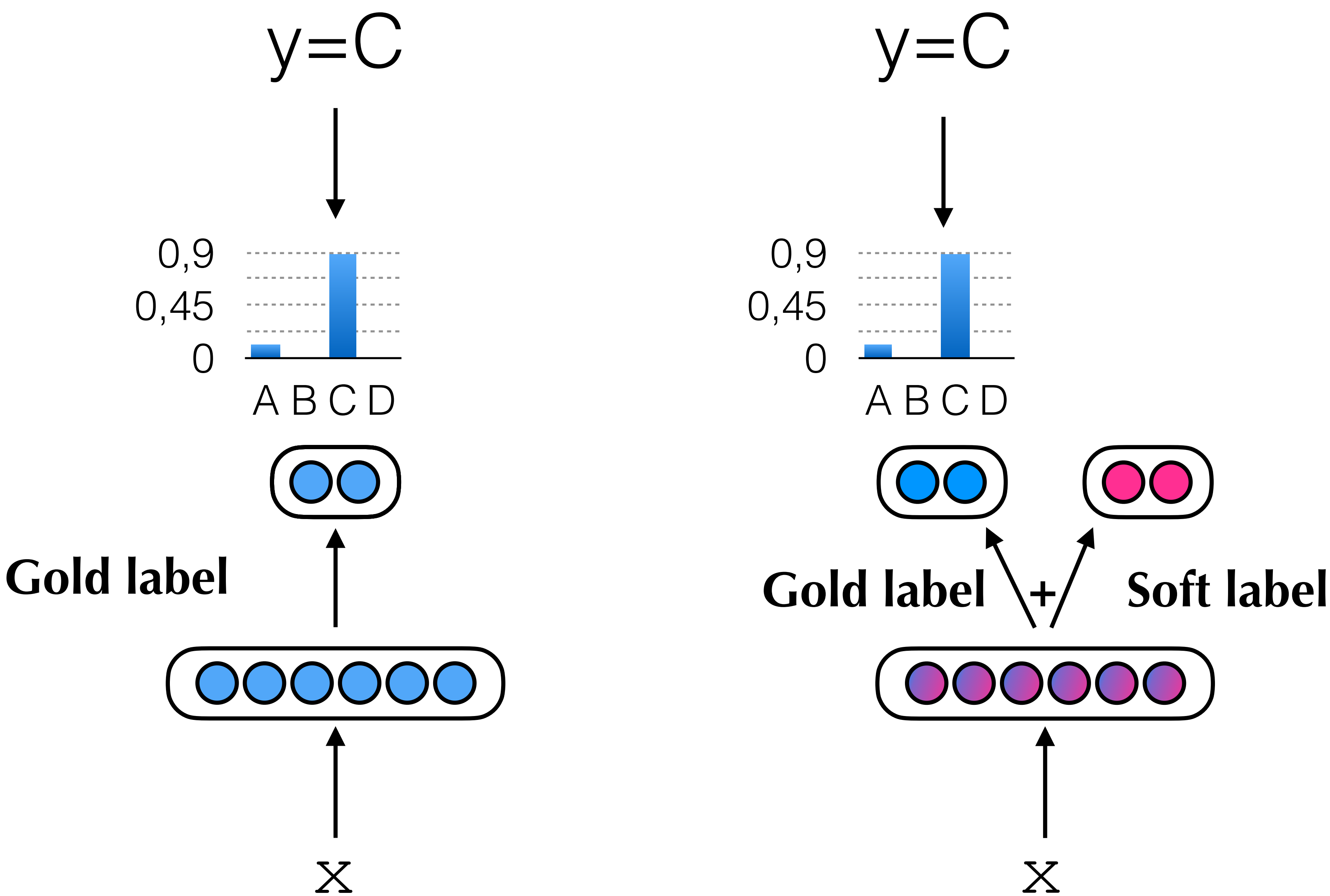
(Fornaciari, Uma, Paun, Plank, Hovy, Poesio 2021 NAACL)

Example of 4: Soft-label MTL



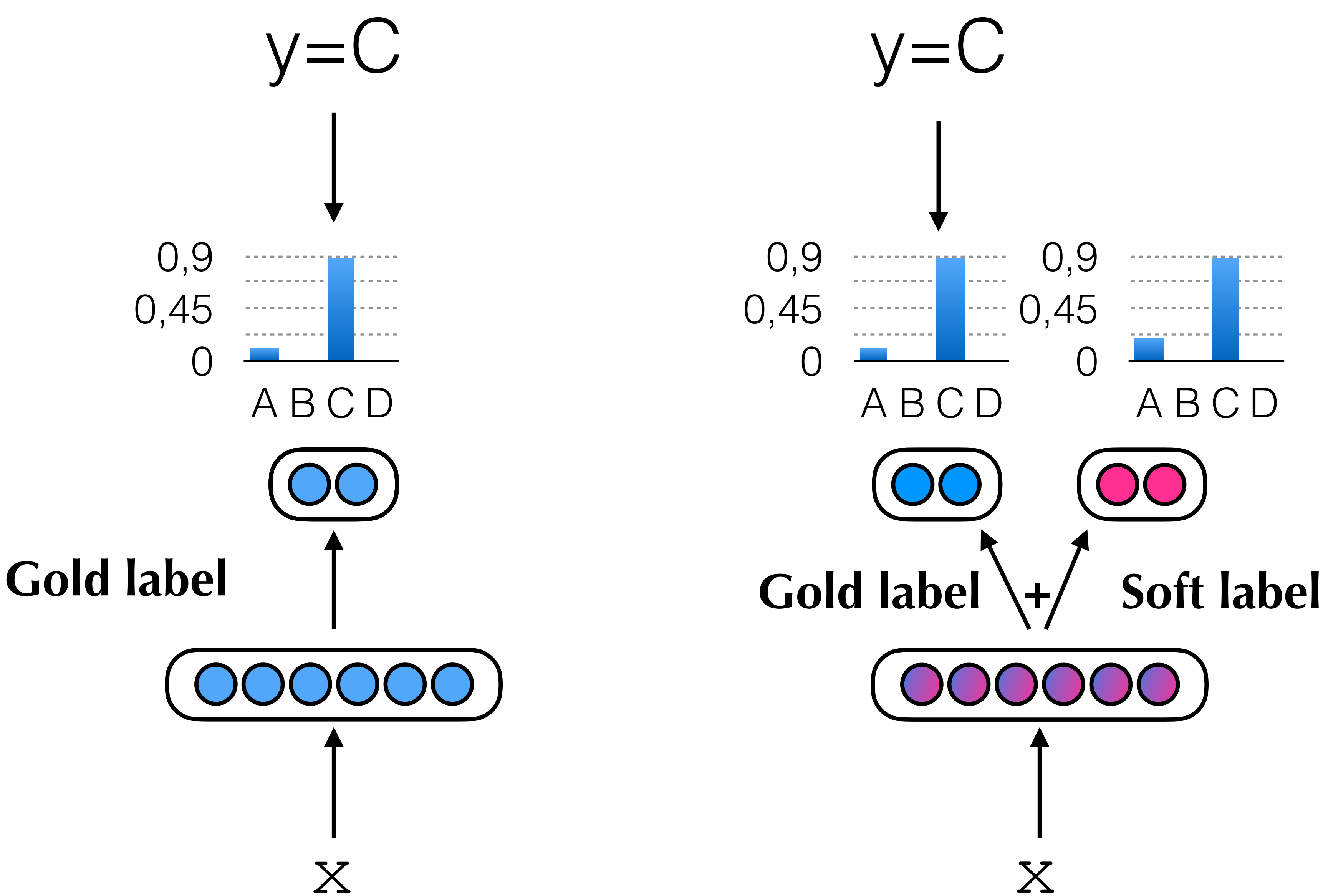
(Fornaciari, Uma, Paun, Plank, Hovy, Poesio 2021 NAACL)

Example of 4: Soft-label MTL



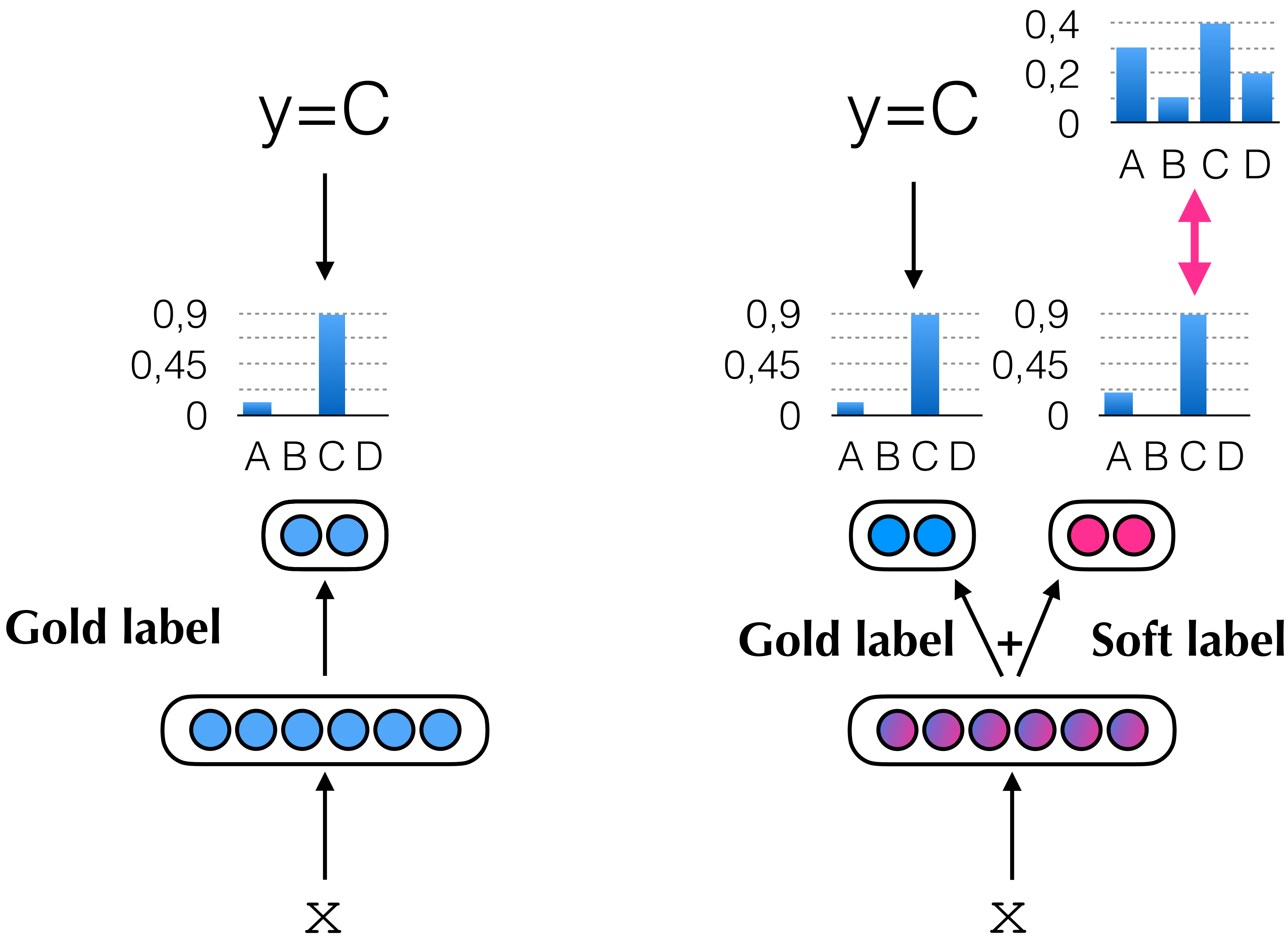
(Fornaciari, Uma, Paun,
Plank, Hovy, Poesio 2021 NAACL)

Example of 4: Soft-label MTL



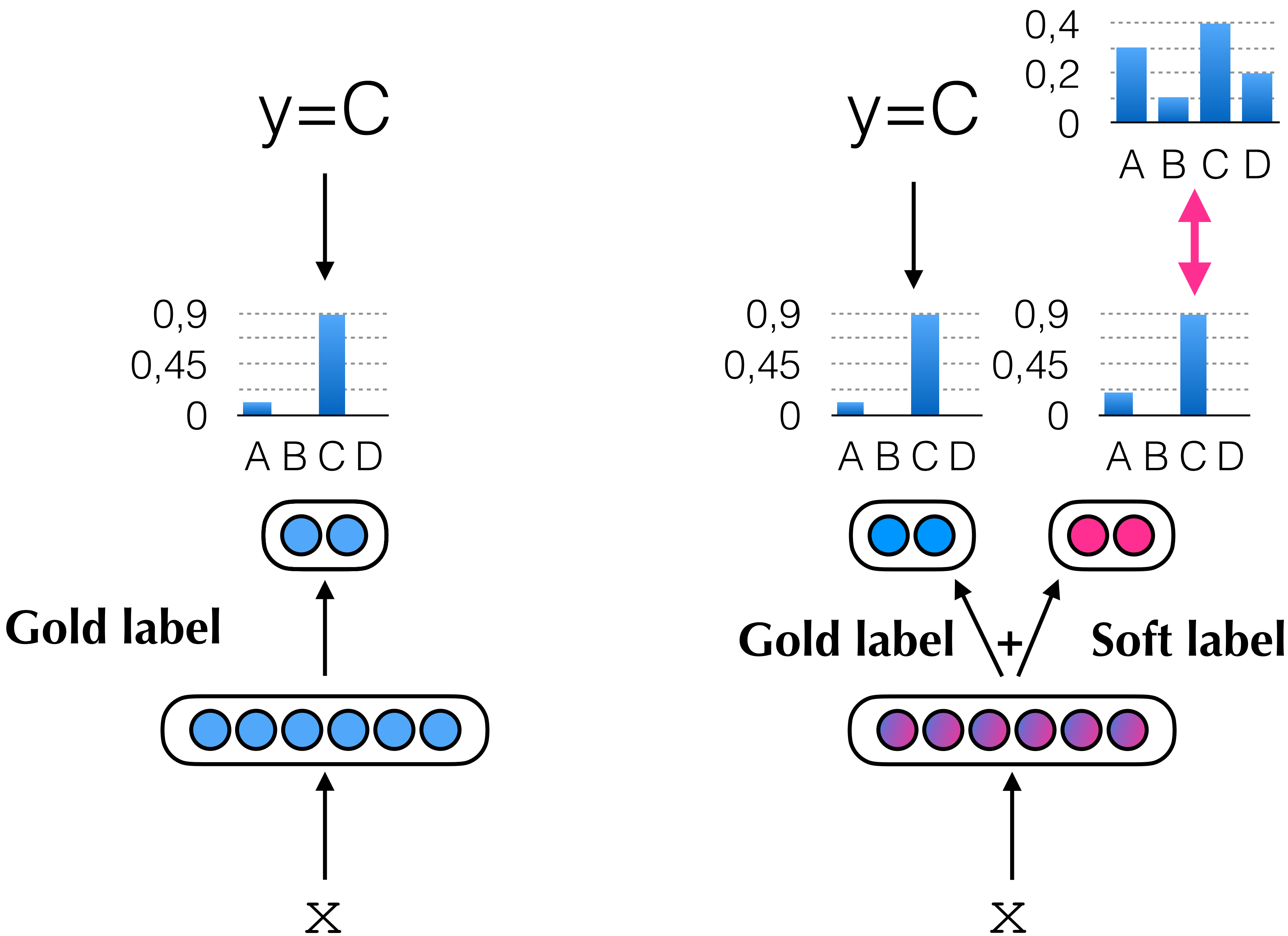
(Fornaciari, Uma, Paun, Plank, Hovy, Poesio 2021 NAACL)

Example of 4: Soft-label MTL



(Fornaciari, Uma, Paun, Plank, Hovy, Poesio 2021 NAACL)

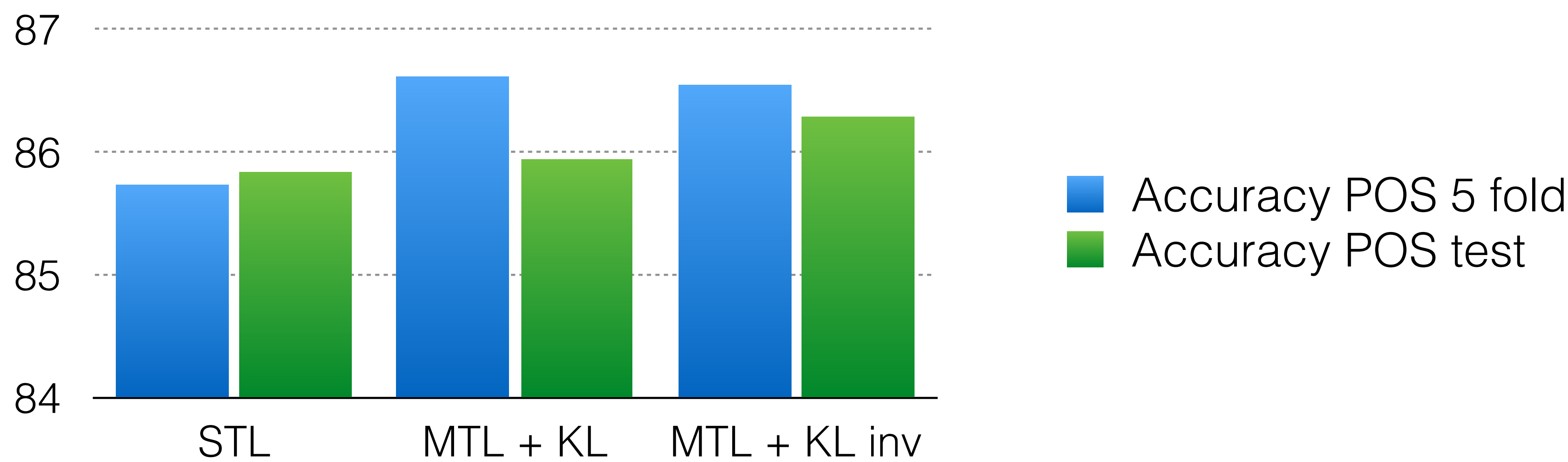
Example of 4: Soft-label MTL

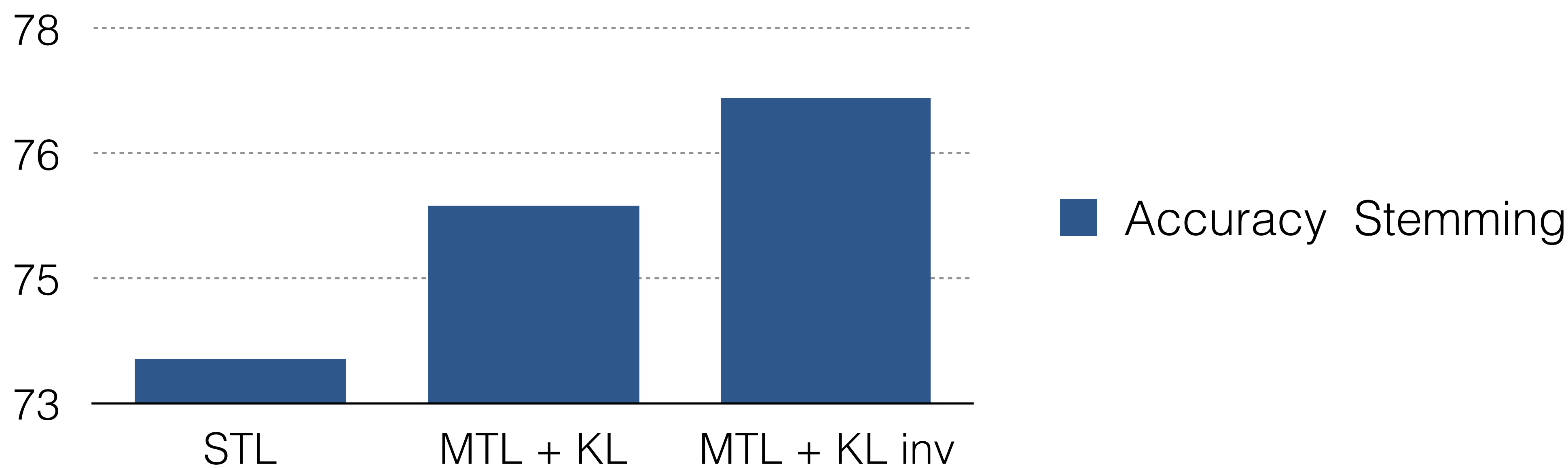


- Needs one auxiliary head (instead of one per annotator as proposed by Specia & Cohn, 2013 and Davani et al., 2021)
- Good results across tasks (Uma et al., 2021)

(Fornaciari, Uma, Paun, Plank, Hovy, Poesio 2021 NAACL)

Results: POS and Stemming





$D_{KL}(P||Q)$ $D_{KL}(Q||P)$

3 Learn from un-aggregated labels: Deep Learning from Crowds

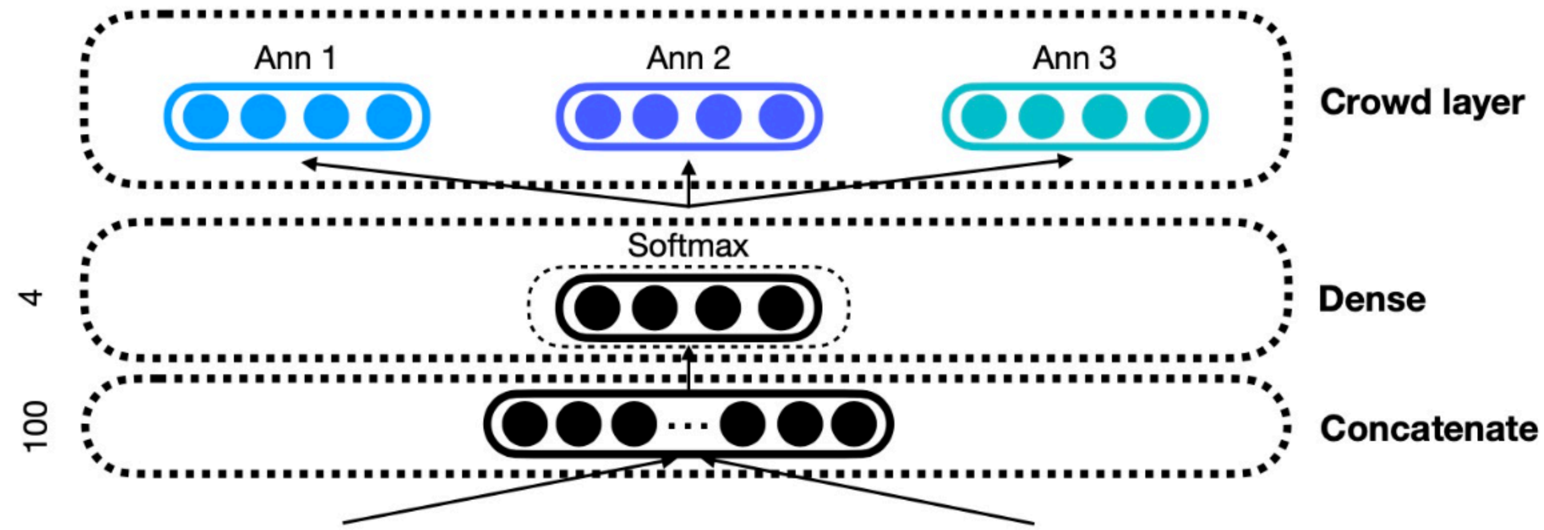


Figure 3: Illustration of deep learning from crowds proposed by Rodrigues and Pereira (2017).

Example: Understanding Indirect Answers to Polar Questions

- **Task:**

Q: Hey. Everything ok?
A: I'm just mad at my agent

☒ Yes
☒ No
☐ Yes, subject to some condition
☐ Neither Yes nor no

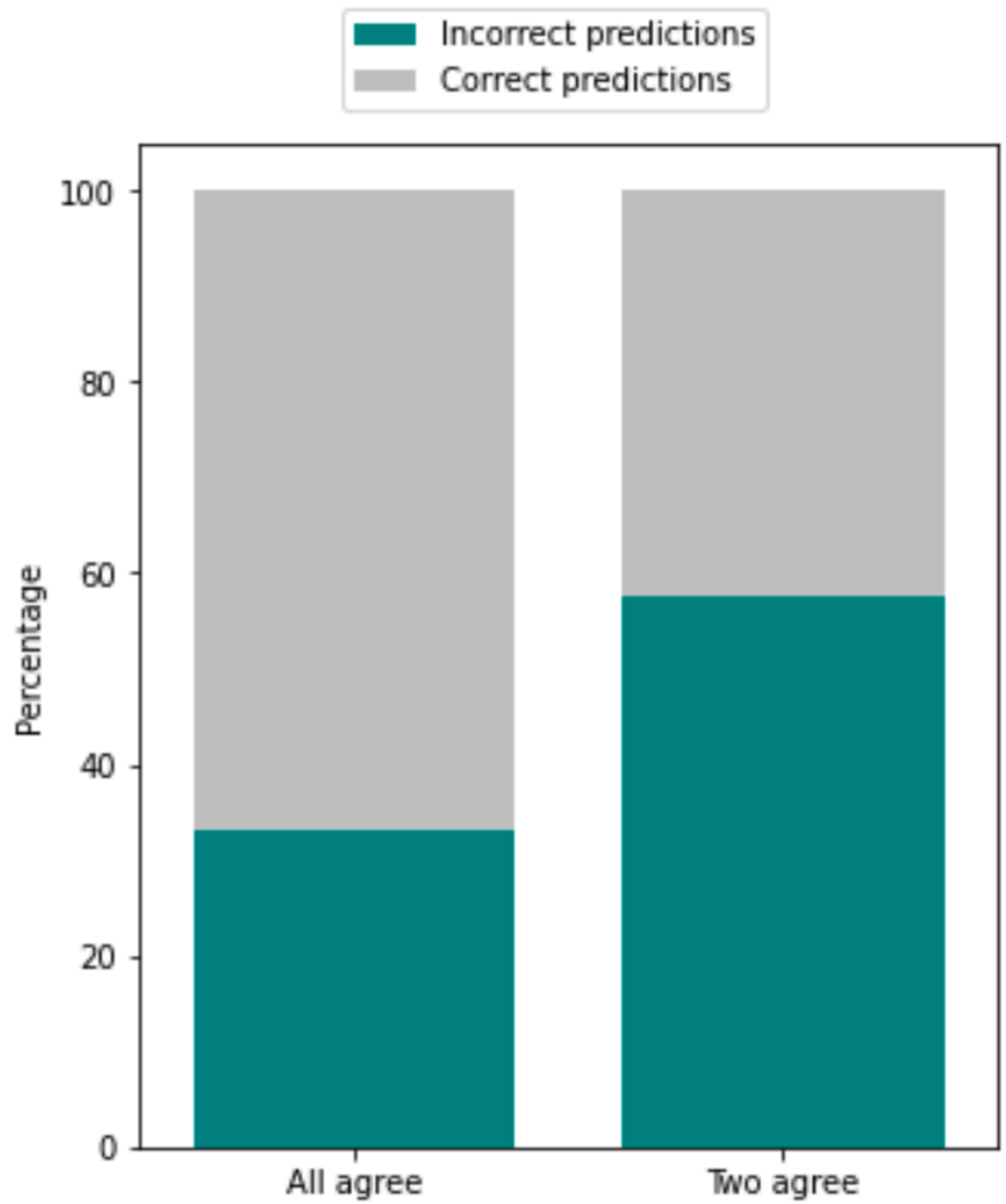
- **Dataset**: Friends-QIA dataset: 5.9k QA pairs
(Damgaard, Toborek, Eriksen & Plank, 2021)
https://aclanthology.org/2021.codi-main.1/

All agree	75.02%
Two agree	23.39%
All disagree	1.59%

Table 3: Annotator agreement.



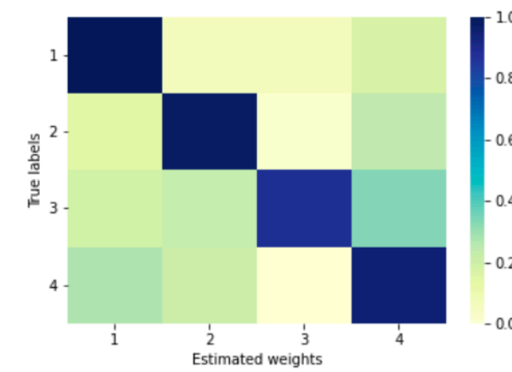
Most “incorrect” predictions on instances humans did not agree on



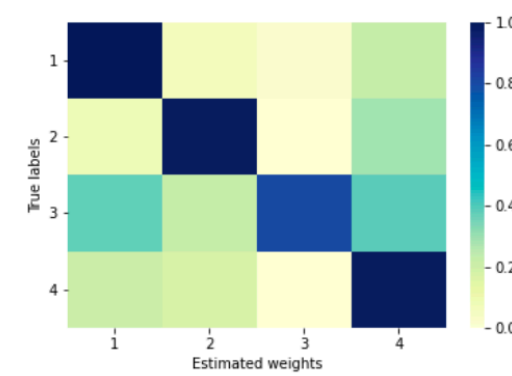
Correct and incorrect predictions of CNN with BERT vs. annotator agreement.

Understanding Indirect Answers

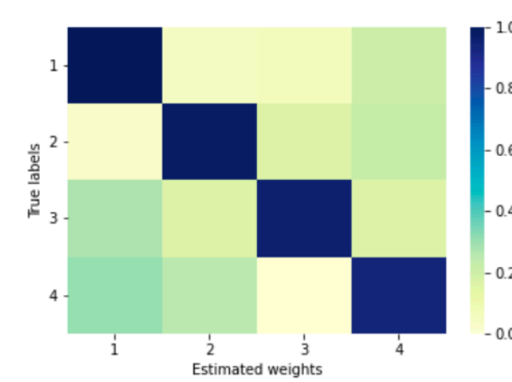
	Accuracy	F1-score
Majority baseline	49.07	16.46
Train on FRIENDS-QIA:		
CNN with BERT	**61.33**	45.65
CNN with BERT, multi-input	61.10	45.53
CNN with BERT + crowd layer	60.32	**47.89**



(a) Annotator 1



(b) Annotator 2

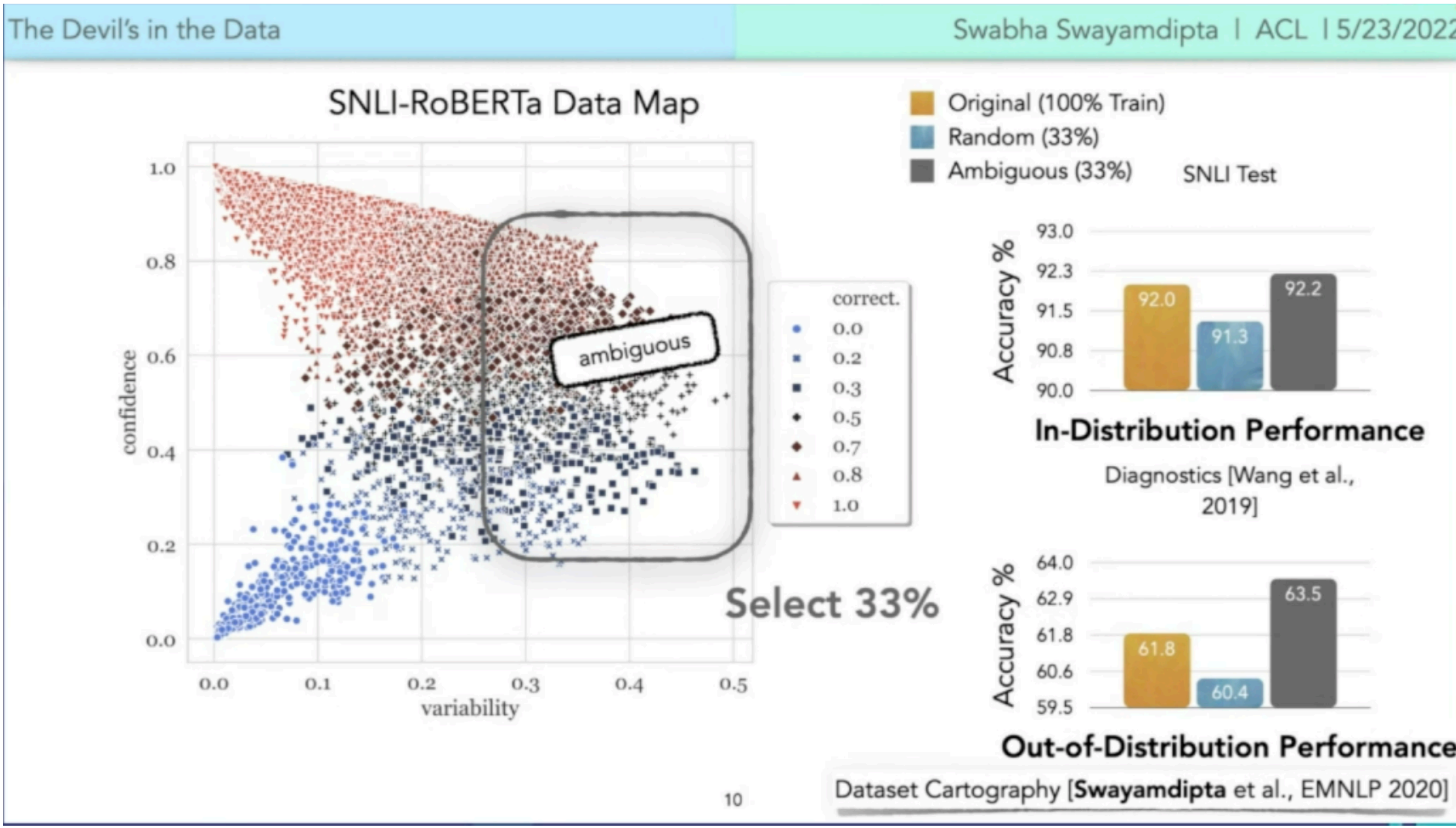


(c) Annotator 3

Supporting Evidence: Learning with humans-in-the-loop & insights from data difficulty

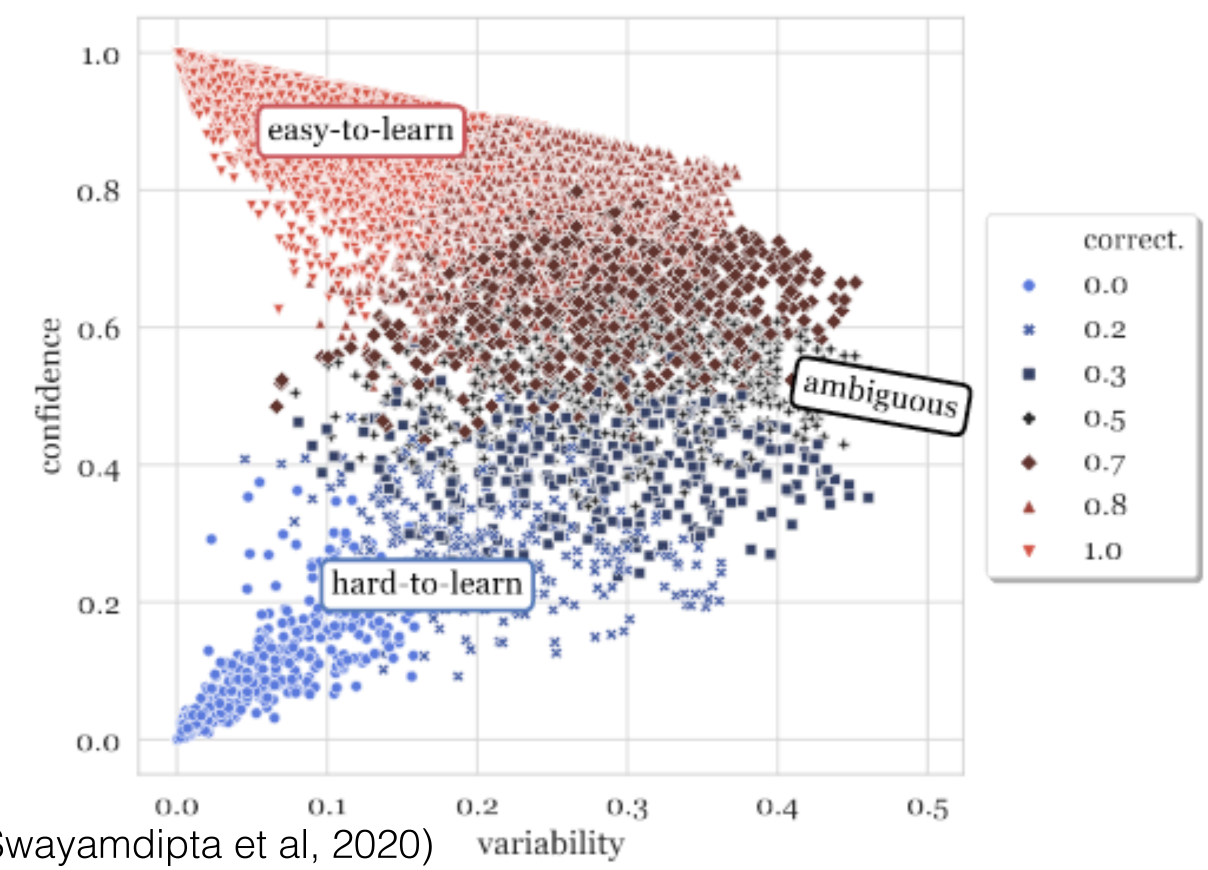
Further evidence: Ambiguous Instances help OOD generalisation

(Swabha Swayamdipta's ACL 2022 talk)

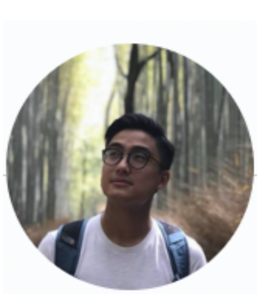


Further evidence: Ambiguous Instances help active learning

- **Key idea:** Data maps provide insights into training dynamics. We propose data maps for more effective active learning.



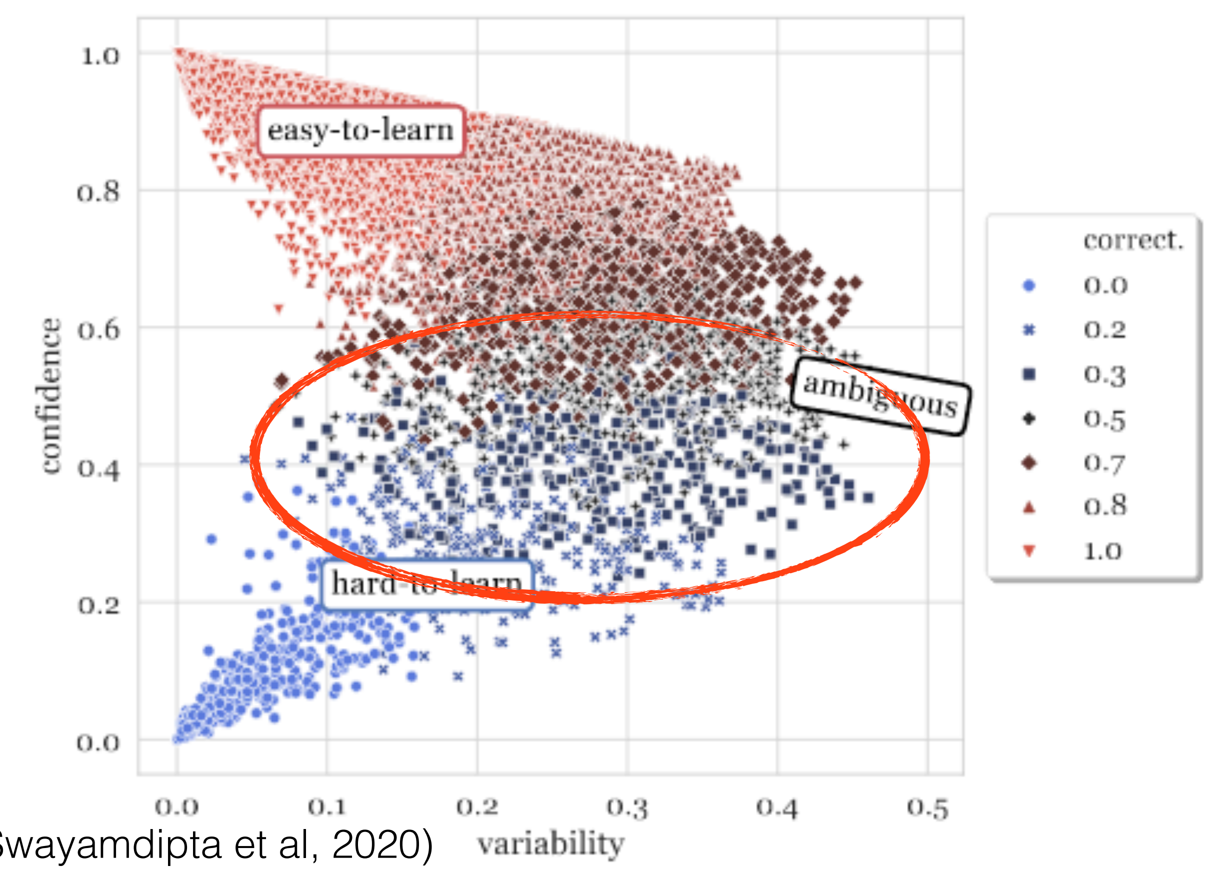
(Swayamdipta et al, 2020)



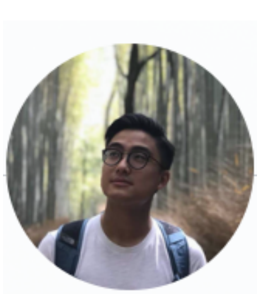
Zhang & Plank (EMNLP 2021 Findings). *Cartography Active Learning*

Further evidence: Ambiguous Instances help active learning

- **Key idea:** Data maps provide insights into training dynamics. We propose data maps for more effective active learning.



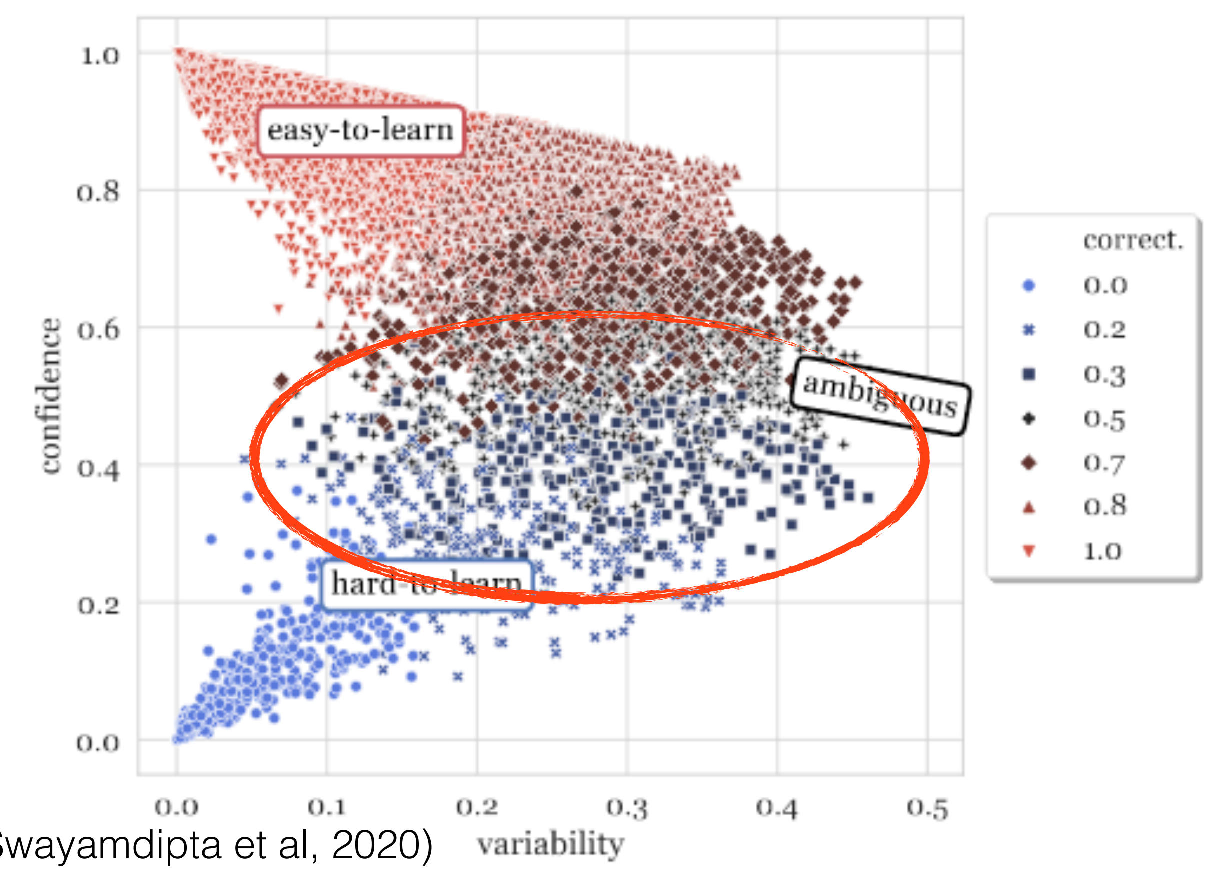
(Swayamdipta et al, 2020)



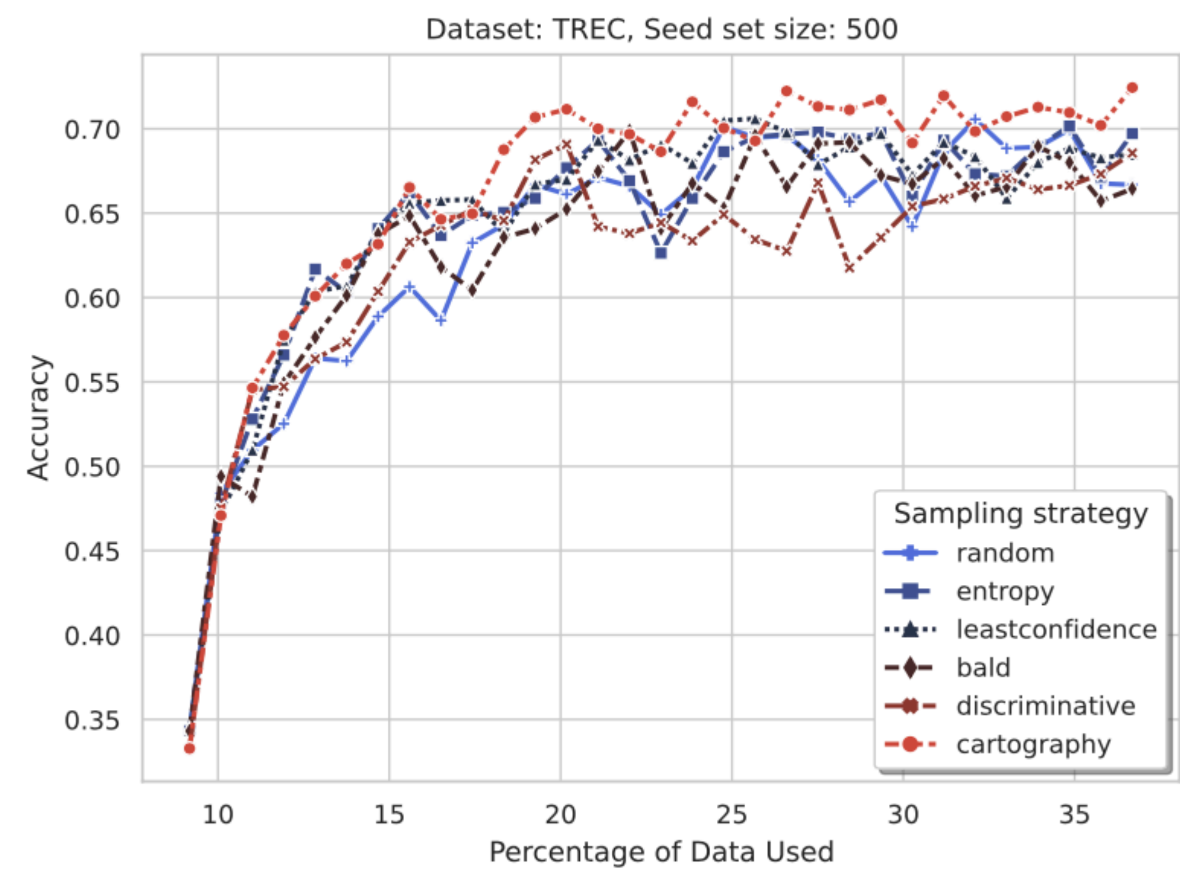
Zhang & Plank (EMNLP 2021 Findings). *Cartography Active Learning*

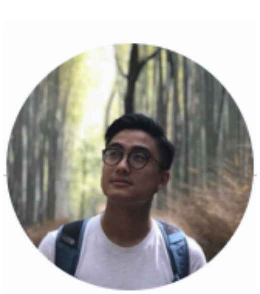
Further evidence: Ambiguous Instances help active learning

- **Key idea:** Data maps provide insights into training dynamics. We propose data maps for more effective active learning.



(Swayamdipta et al, 2020)





Zhang & Plank (EMNLP 2021 Findings). *Cartography Active Learning*

Learning with HLV: Open Challenges

- Increasing interest, yet existing research is fragmented (even within NLP), so are methods proposed so far

- Lack of diverse datasets (and challenge of balance between multiple annotations vs more data) - yet even small samples can be useful (e.g. Plank et al., 2014), learning from different amounts of labeled data is emerging (Zhang et al., 2021)

- Little explored connections to other related disciplines

More methods, overview and empirical evaluations:

JAIR survey by Uma et al., 2021:

Learning from Disagreement: A Survey

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio (2021 JAIR)

Roadmap: Three perspectives

1. Data: Is human label variation (HLV) random noise or signal?
2. Modelling: How can we leverage human label variation?
3. Evaluation: How to evaluate in light of human label variation?

We Need to Talk about Disagreement in Evaluation

Work in collaboration with Alexandra Uma, Dirk Hovy, Massimo Poesio, Michael Fell, Silviu Paun, Tommaso Fornaciari, Valerio Basile (BPPF workshop@ACL 2021)

https://aclanthology.org/2021.bppf-1.3.pdf

Evaluation in Interpretation Tasks

- Many works on learning from disagreement compare against an evaluation set assumed to encode a **single ground truth**
  - A single correct answers ignores the **subjectivity** and **complexity** of many tasks
    - ➡ Focus on “easy”, low-risk **evaluation**
    - ➡ Metrics not aligned with reality (Gordon et al., 2021)
- Research has started to evaluate with **hard** and **soft labels**

Examples

- F1 **against individual** annotator labels used in Davani et al. (2021) for hate-speech and emotion prediction, besides "gold standard" evaluation
- Evaluation against **cluster of users** (e.g. Akhtar et al., 2019; see Basile et al., 2021)
- **Disagreement Deconvolution** (Gordon et al., 2021) propose to compare predictions to **each annotator's belief**.
  - Across users: Stratified evaluation over user groups
  - Within a user: Primary label estimation

➡ Soft evaluation sheds light in **uncertainty** in models, important for more **trustworthy** AI

To sum up

Is Human Label Variation So Bad? No.

It provides opportunities for more trustworthy, human-facing AI.

Ways Forward (in light of the 3 "steps")

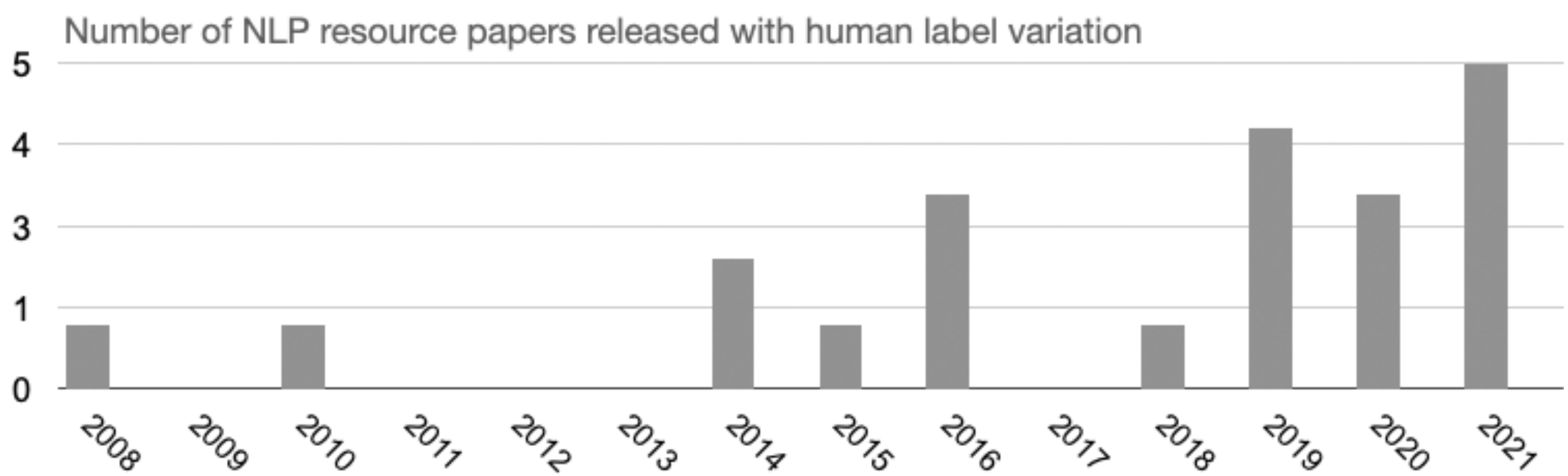
Ways Forward (1/3): Data

- Collect & release more **annotator-level (un-aggregated) labels** (Basile et al., 2021; Prabhakaran et al., 2021)

- In general, value in releasing meta-data at the (instance) level

Gleam of hope: Growth in resources

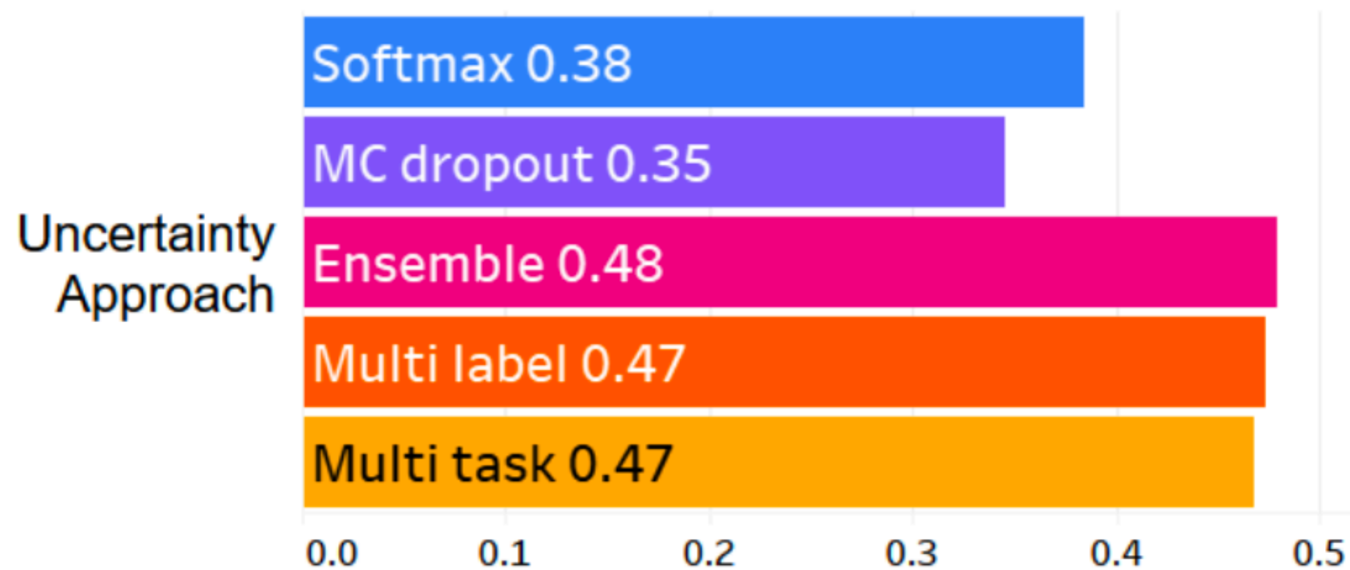
- Analysis of dataset papers with multiple-annotator release



(Plank, 2022 EMNLP)

Ways Forward (2/3): Evaluation

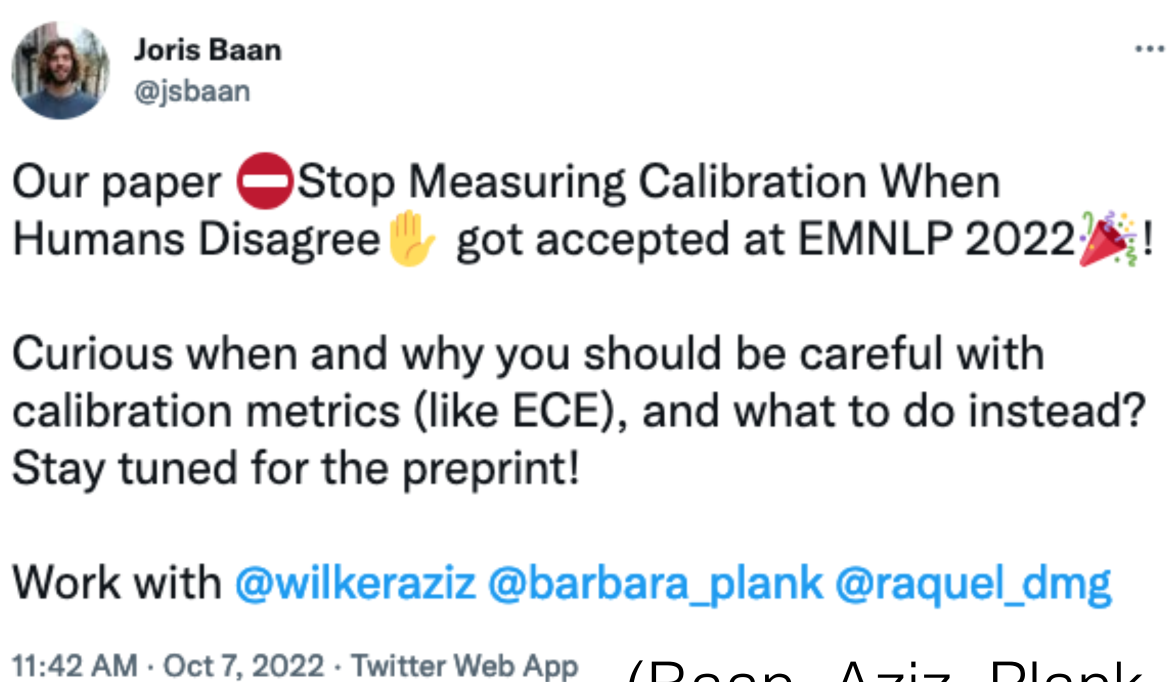
- Beyond accuracy (and single “mode” evaluation)
- Human label variation and **model uncertainty**



(Davani et al., 2021)

! Calibration to majority is flawed !

- Calibration is a popular framework to evaluate whether a classifier knows when it does not know
  - In an upcoming paper, we provide theoretical and empirical evidence that calibration to human majority is problematic
  - To address this, we devise instance-level measures of calibration to capture human label variation


Joris Baan
@jsbaan

Our paper ⛔Stop Measuring Calibration When Humans Disagree✋ got accepted at EMNLP 2022🎉!

Curious when and why you should be careful with calibration metrics (like ECE), and what to do instead? Stay tuned for the preprint!

Work with @wilkeraziz @barbara_plank @raquel_dmg

11:42 AM · Oct 7, 2022 · Twitter Web App

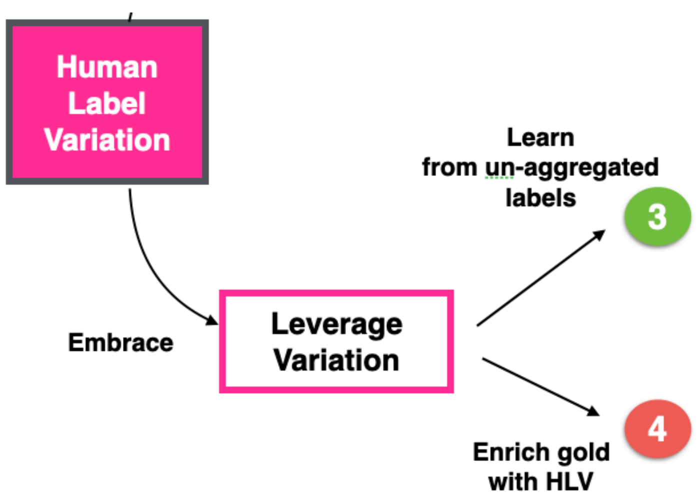
(Baan, Aziz, Plank, Fernandez, 2022 EMNLP)

Ways Forward (3/3): Learning

- Categories exist, but they are fluid; Let's not throw away signal!

Ways Forward (3/3): Learning

- Categories exist, but they are fluid; Let's not throw away signal!



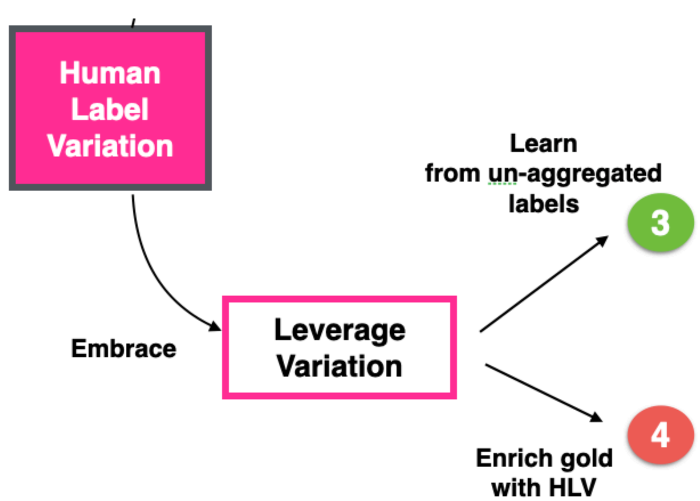
Ways Forward (3/3): Learning

- Categories exist, but they are fluid; Let's not throw away signal!

a continuum of plausible variation

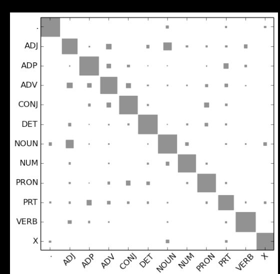
Noise vs. A range of

Human label variation



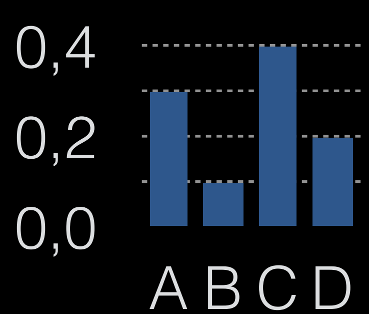
– To model Human Perspectives
– Provide highly-informative examples (less but more informative data)

Take-home message



✓ not all **human label variation** is noise

✓ embrace it during **learning** / Let's not continue to model only the "mode", but the collective human label variation!



✓ embrace it during **evaluation**

✦ Research opportunities in this space

✦ Plug: SemEval 2023 shared task

Key selected references

- JAIR Survey

Learning from Disagreement: A Survey

Alexandra N. Uma A.N.UMA@QMUL.AC.UK
Queen Mary University of London

Tommaso Fornaciari FORNACIARI.TOMMASO@UNIBOCCONI.IT
Dirk Hovy DIRK.HOVY@UNIBOCCONI.IT
Università Bocconi, Milano

Silviu Paun S.PAUN@QMUL.AC.UK
Queen Mary University of London

Barbara Plank BAPL@ITU.DK
IT University of Copenhagen

Massimo Poesio M.POESIO@QMUL.AC.UK
Queen Mary University of London

- EMNLP 2022 theme paper

The "Problem" of Human Label Variation:
On Ground Truth in Data, Modeling and Evaluation

Barbara Plank
Center for Information and Language Processing (CIS), LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany

Questions? Thanks!

IT UNIVERSITY OF COPENHAGEN

Is Human Label Variation Really so Bad for AI?

Interested?
I'm hiring!

@barbara_plank

http://mainlp.github.io

Thanks to all students, lab members and collaborators. Research in parts support by: