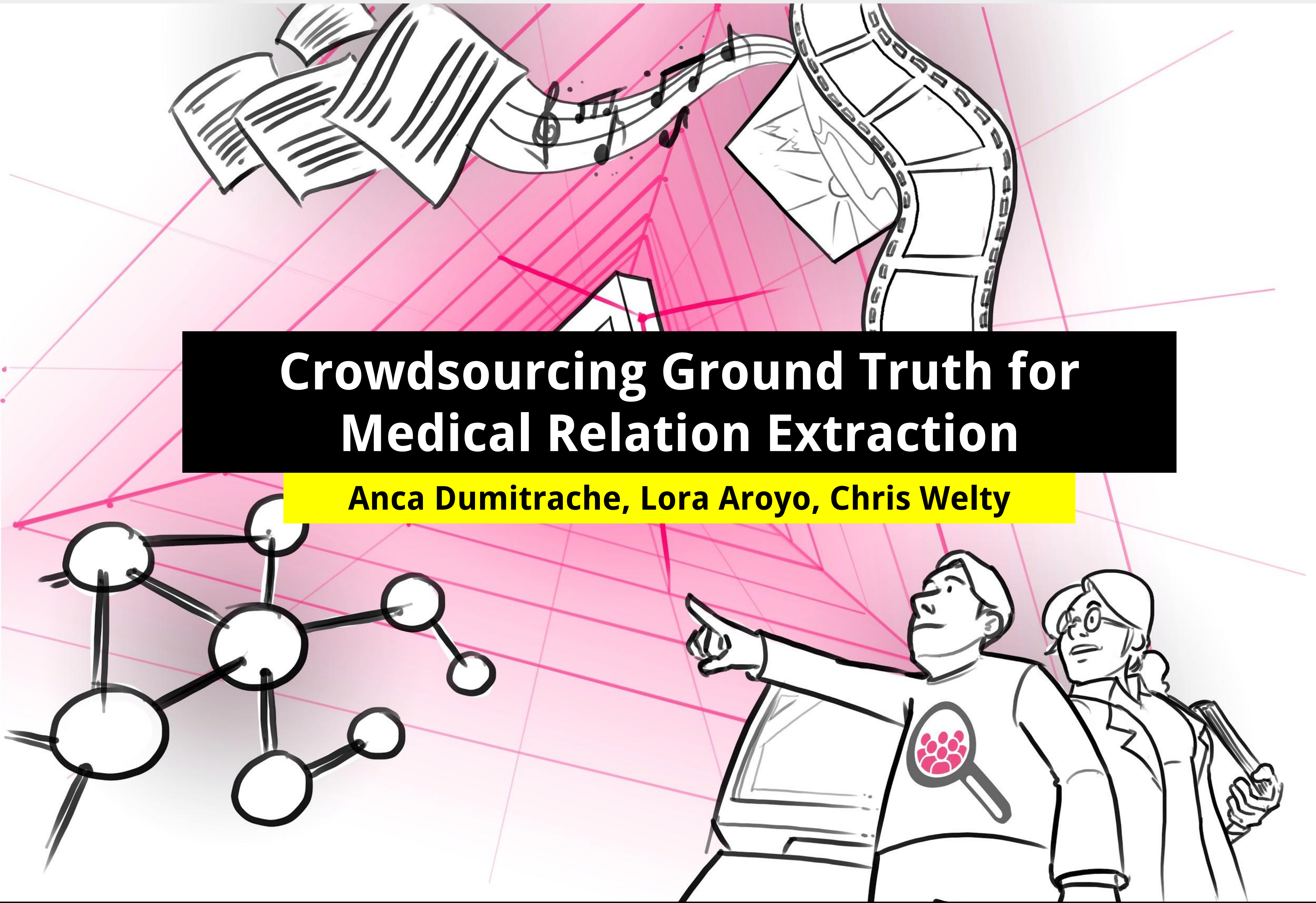
Crowdsourcing Ground Truth for Medical Relation Extraction

Anca Dumitrache, Lora Aroyo, Chris Welty



http://crowdtruth.org/

Published @TiiS Journal, Presented at IUI2019

CrowdTruth

The framework for crowdsourcing ground truth data

IS ONE ENOUGH?

MYTHS ABOUT HUMAN ANNOTATION

“Truth is a Lie: 7 Myths about Human Annotation”, *AI Magazine 2014*, L. Aroyo, C. Welty

One truth: knowledge acquisition for the semantic web assumes one correct interpretation for every example

All examples are created equal: triples are triples, one is not more important than another, they are all either true or false

Disagreement bad: when people disagree, they don’t understand the problem

Experts rule: knowledge is captured from domain experts

One is enough: knowledge by a single expert is sufficient

DOES THIS SENTENCE EXPRESS TREATS RELATION?

Treats: Chloroquine, Malaria

Rheumatoid arthritis and **MALARIA** have been treated with **CHLOROQUINE** for decades.

For prevention of malaria, use only in individuals traveling to malarious areas where **CHLOROQUINE** resistant P. falciparum **MALARIA** has not been reported.

Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.

WHAT DO EXPERTS SAY?

Treats: Chloroquine, Malaria

Rheumatoid arthritis and **MALARIA** have been treated with **CHLOROQUINE** for decades. ✓

For prevention of malaria, use only in individuals traveling to malarious areas where **CHLOROQUINE** resistant P. falciparum **MALARIA** has not been reported. ✓

Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood. ✗

WHAT DOES THE CROWD SAY?

Treats: Chloroquine, Malaria

Rheumatoid arthritis and **MALARIA** have been treated with **CHLOROQUINE** for decades.

95%

For prevention of malaria, use only in individuals traveling to malarious areas where **CHLOROQUINE** resistant P. falciparum **MALARIA** has not been reported.

75%

Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.

50%

Intuition: This is better

WHAT DOES THE CROWD SAY?

Treats: Chloroquine, Malaria

Rheumatoid arthritis and **MALARIA** have been treated with **CHLOROQUINE** for decades.

95%

BETTER

There's a difference between these two

For prevention of malaria, use only in individuals traveling to malarious areas where **CHLOROQUINE** resistant P. falciparum **MALARIA** has not been reported.

75%

WORSE

Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.

50%

This one isn't utterly wrong

CROWDTRUTH

Annotator disagreement is **signal, not noise**

It is indicative of the **variation of human semantic interpretation**

It can indicate **ambiguity, vagueness, similarity, over-generality,** and most importantly **quality**

MEDICAL RELATION EXTRACTION

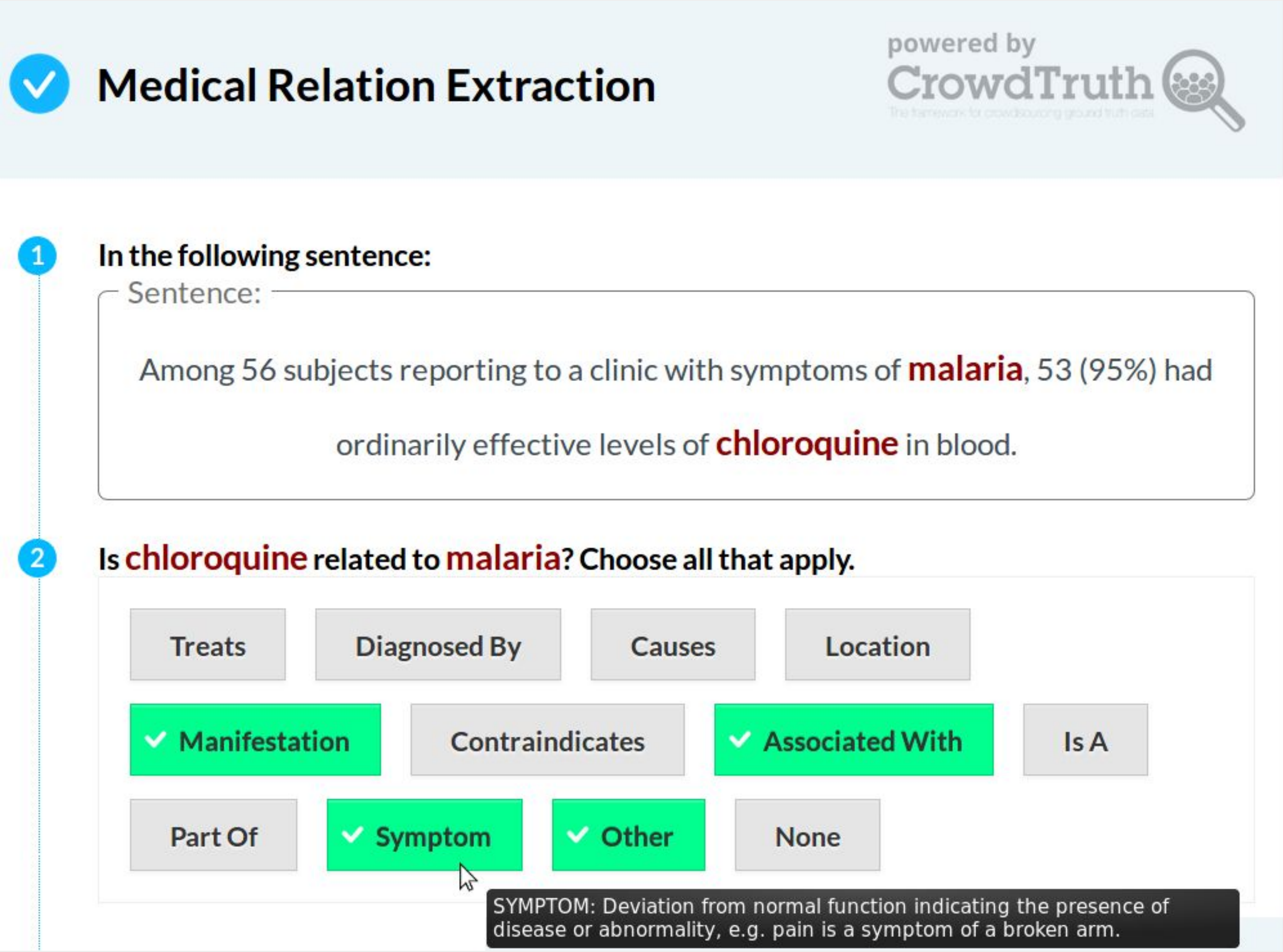
Goals:

- crowdsource a **gold standard** for *treat* & *cause* medical relation extraction
- improve performance of manifold model sentence-level classifier

Approach:

- compare crowd & medical expert on 900 sentences
- compare crowd & distant supervision on 3,900 sentences

CROWD TASK



Medical Relation Extraction

1 **In the following sentence:**

Sentence:

Among 56 subjects reporting to a clinic with symptoms of **malaria**, 53 (95%) had ordinarily effective levels of **chloroquine** in blood.

2 **Is chloroquine related to malaria? Choose all that apply.**

Treats · Diagnosed By · Causes · Location

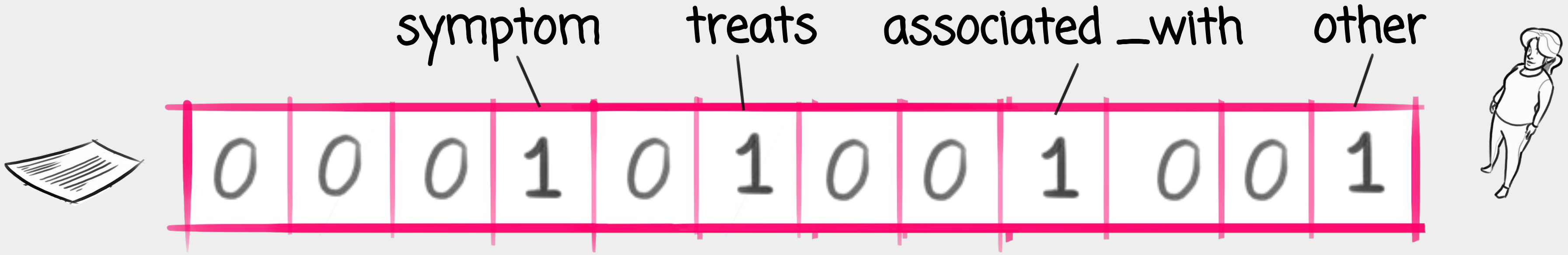
✓ Manifestation · Contraindicates · ✓ Associated With · Is A

Part Of · ✓ Symptom · ✓ Other · None

SYMPTOM: Deviation from normal function indicating the presence of disease or abnormality, e.g. pain is a symptom of a broken arm.

WORKER VECTOR FOR A SENTENCE

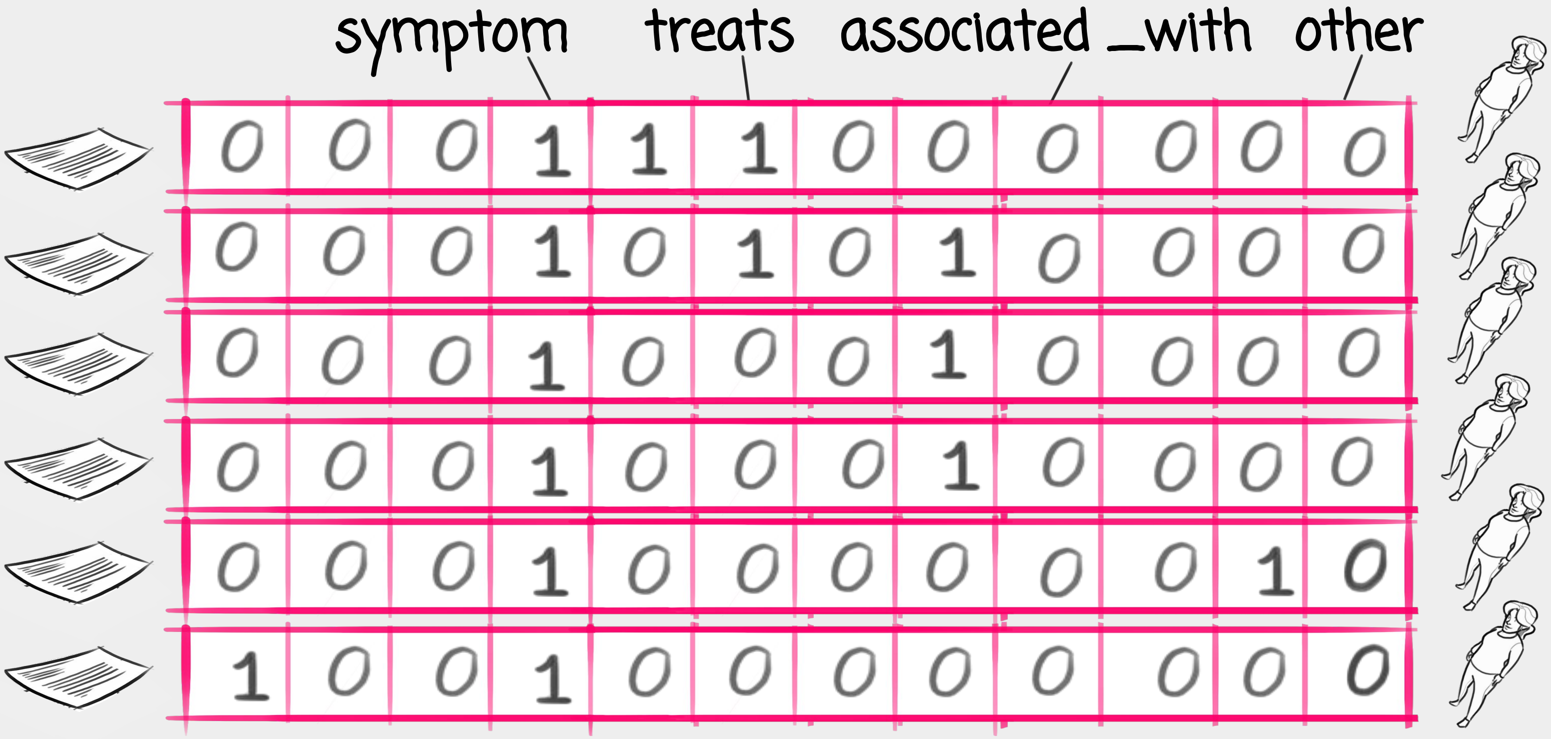
Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.



			symptom		treats			associated_with			other
0	0	0	1	0	1	0	0	1	0	0	1

MANY WORKERS FOR THE SAME SENTENCE

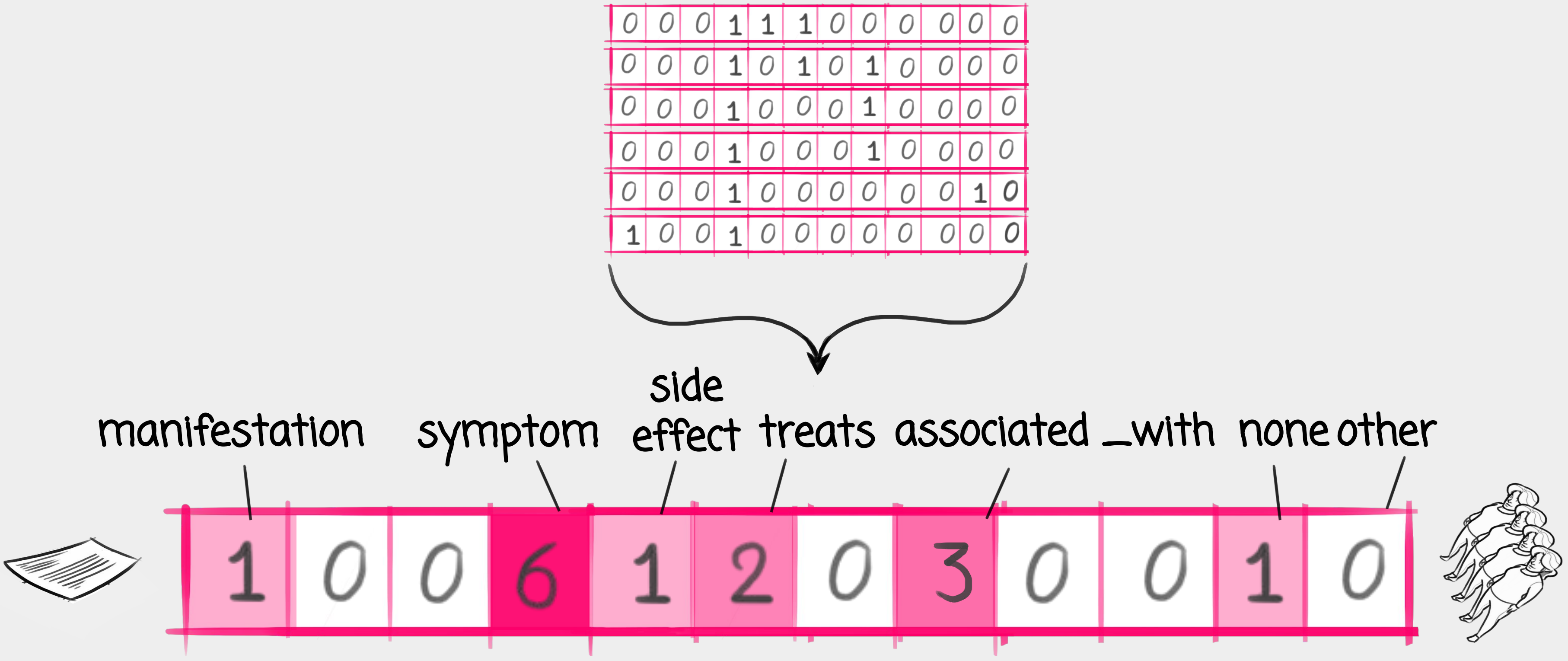
Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.



			symptom		treats			associated_with			other
0	0	0	1	1	1	0	0	0	0	0	0
0	0	0	1	0	1	0	1	0	0	0	0
0	0	0	1	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0	0	0	1	0
1	0	0	1	0	0	0	0	0	0	0	0

ALL WORKER VECTORS AGGREGATED IN A SENTENCE VECTOR

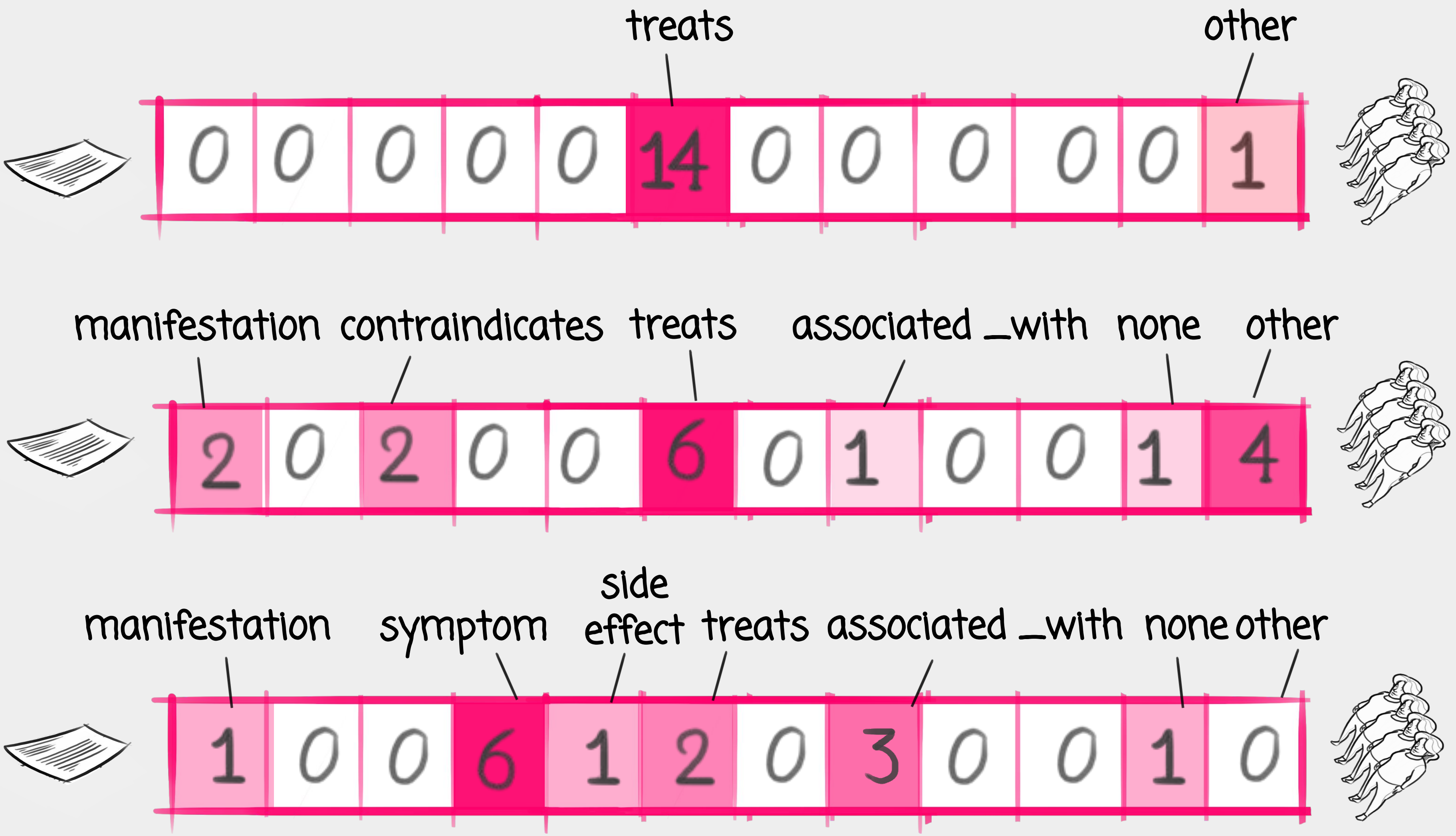
Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.



0	0	0	1	1	1	0	0	0	0	0	0
0	0	0	1	0	1	0	1	0	0	0	0
0	0	0	1	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0	0	0	1	0
1	0	0	1	0	0	0	0	0	0	0	0

manifestation			symptom	side effect	treats		associated_with			none	other
1	0	0	6	1	2	0	3	0	0	1	0

SENTENCE VECTORS FOR THE 3 SENTENCES

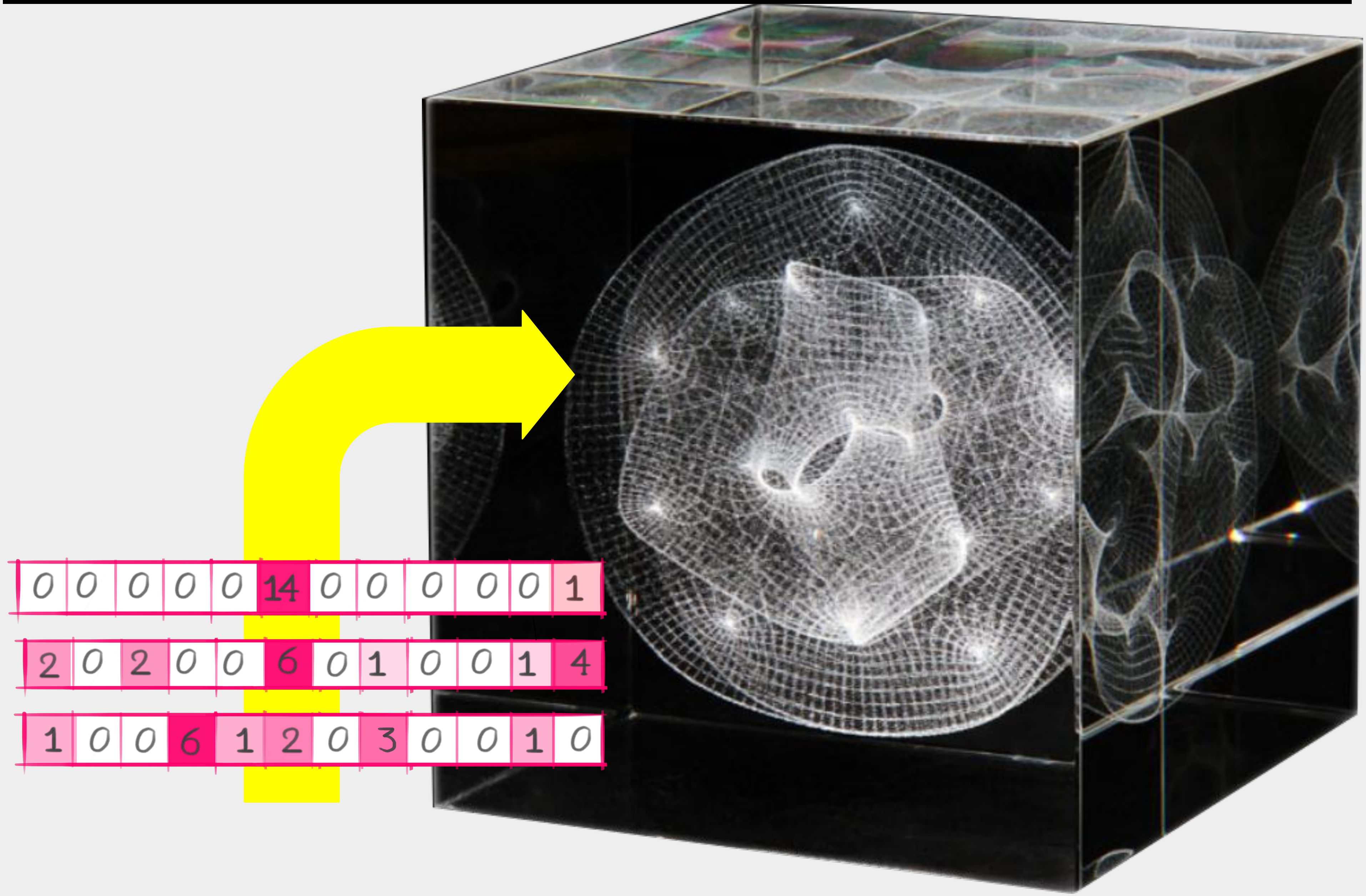


					treats						other
0	0	0	0	0	14	0	0	0	0	0	1

manifestation		contraindicates			treats		associated _with			none	other
2	0	2	0	0	6	0	1	0	0	1	4

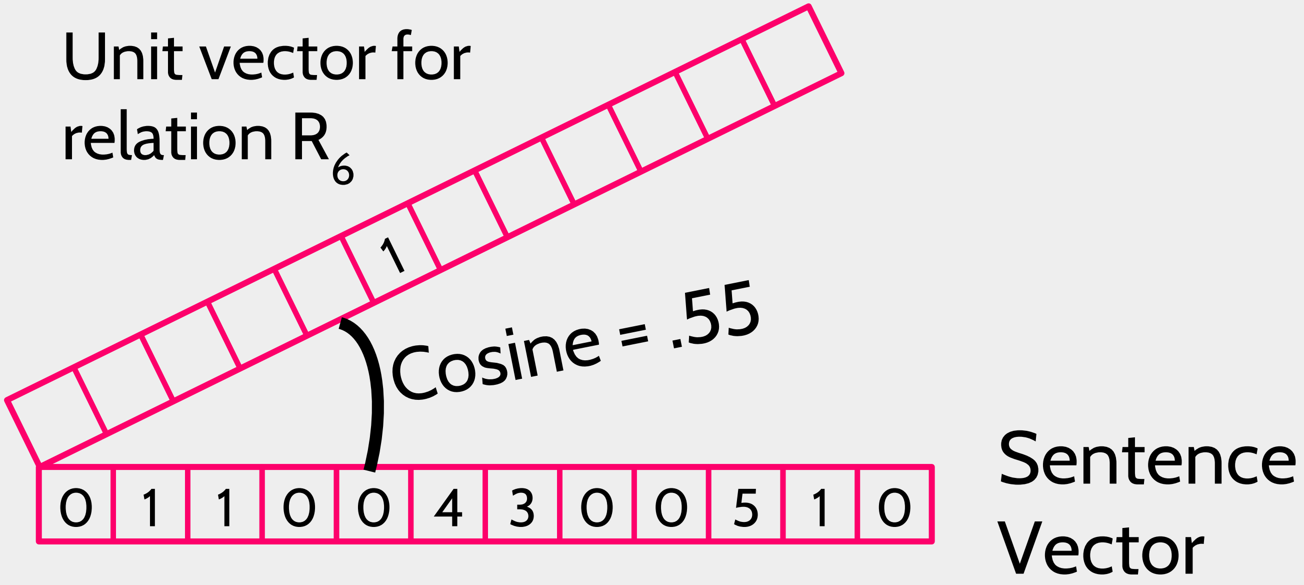
manifestation			symptom	side effect	treats		associated _with			none	other
1	0	0	6	1	2	0	3	0	0	1	0

SEMBEDDINGS : embeddings with semantic dimensions



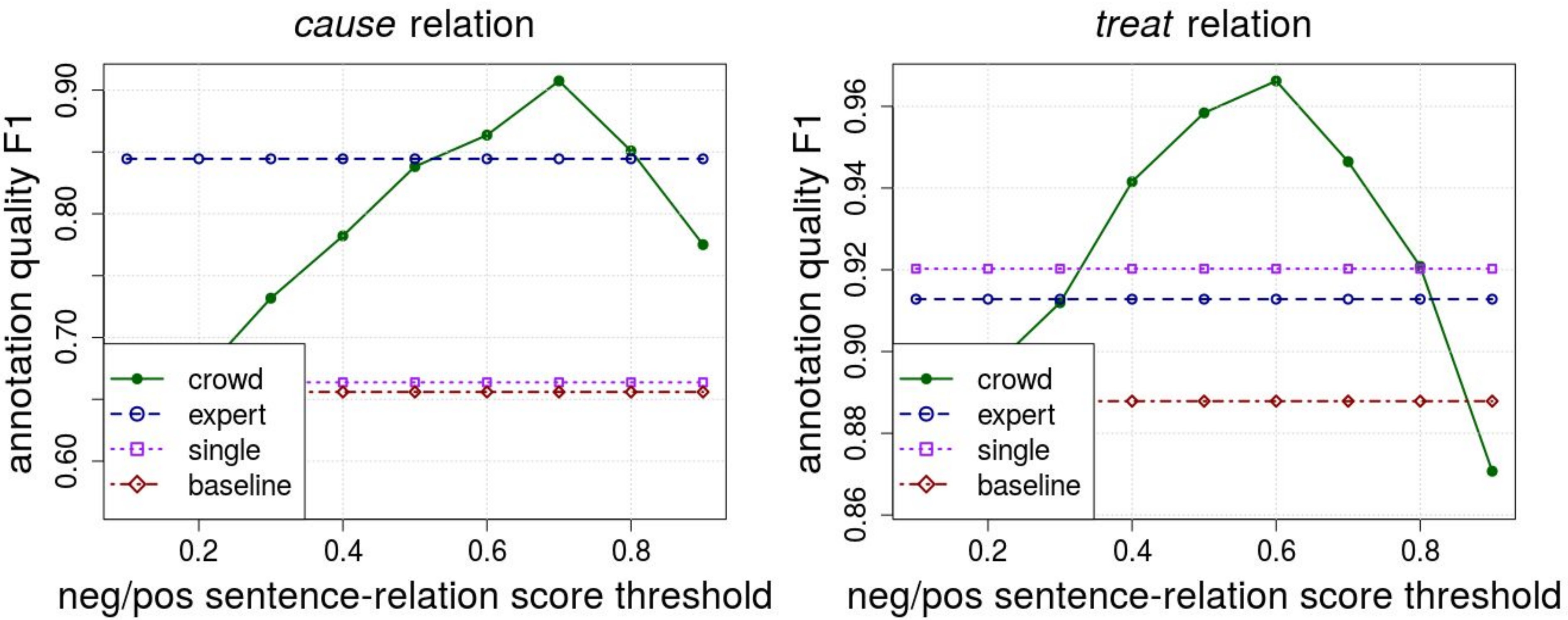
0	0	0	0	0	14	0	0	0	0	0	1
2	0	2	0	0	6	0	1	0	0	1	4
1	0	0	6	1	2	0	3	0	0	1	0

SENTENCE - RELATION SCORE (SRS)



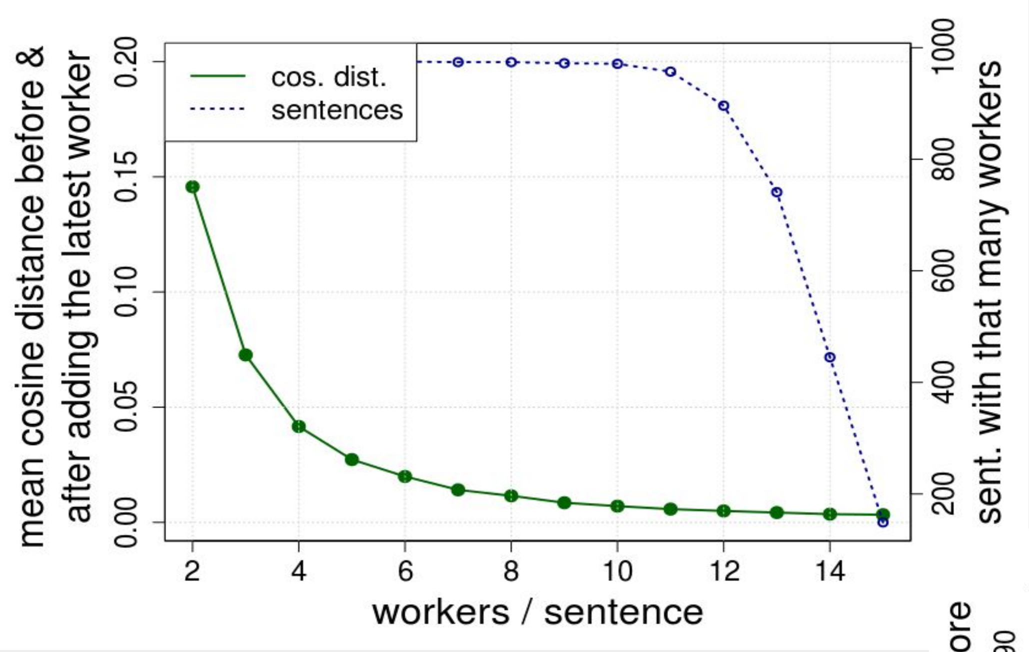
Measures how clearly a sentence expresses a relation

CROWD vs. EXPERT ANNOTATION QUALITY



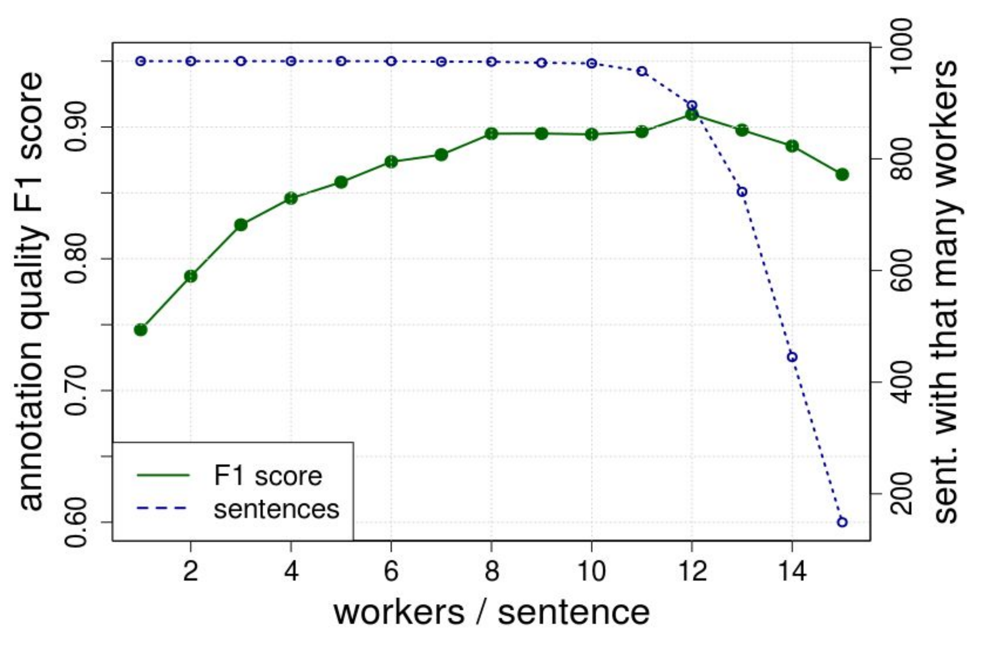
[0.6 - 0.8] crowd significantly out-performs expert

HOW MANY WORKERS / SENTENCE?



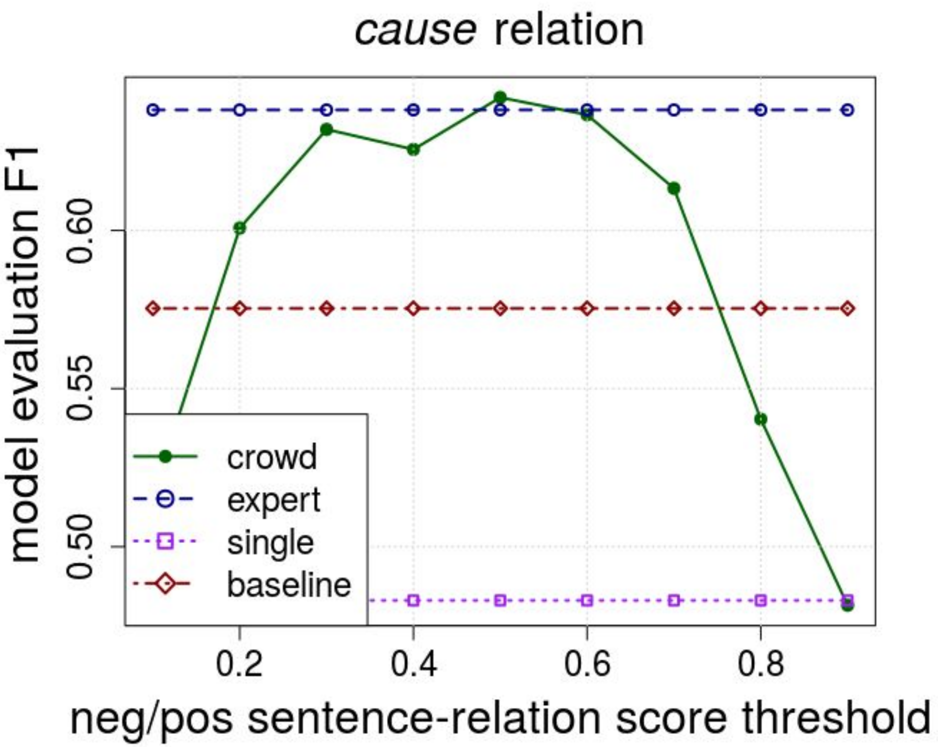
cosine & F1 scores are stable at 15 workers / sentence

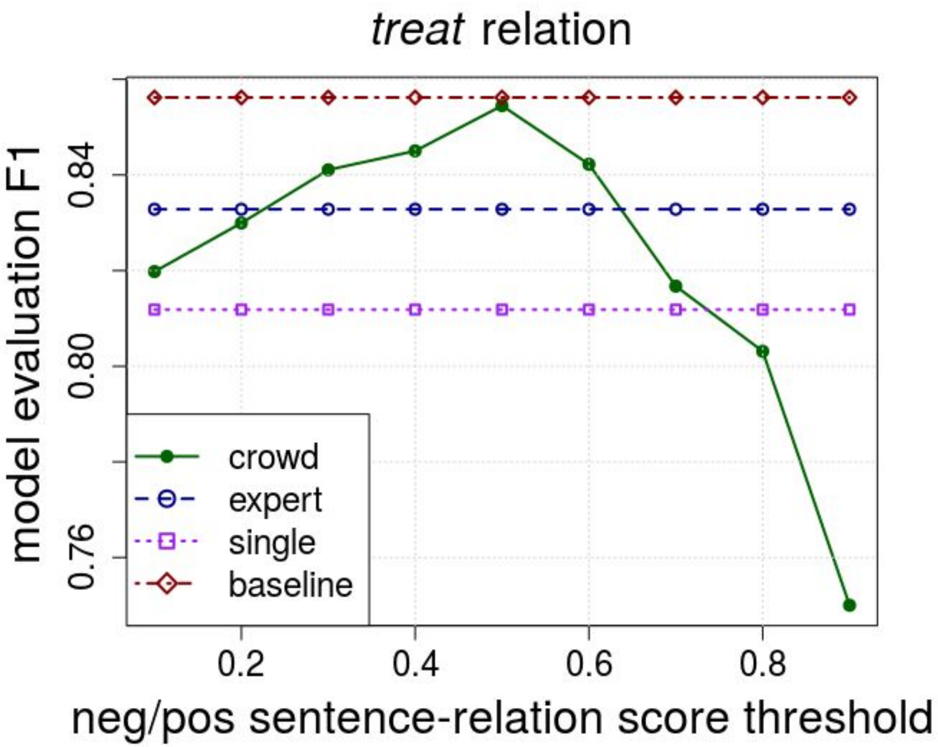
15 workers / sentence still costs less than 1 expert / sentence



CROWD vs. EXPERT MODEL QUALITY

RelEx model: Wang & Fan. *Medical relation extraction with manifold models*. ACL 2014





crowd provides training data that is at least as good, if not better than experts

EVALUATING WITH SRS-WEIGHTED METRICS

Weighted Precision:

$$P' = \frac{\sum_s srs(s) \cdot tp(s)}{\sum_s srs(s) \cdot tp(s) + (1 - srs(s)) \cdot fp(s)}$$

Weighted Recall:

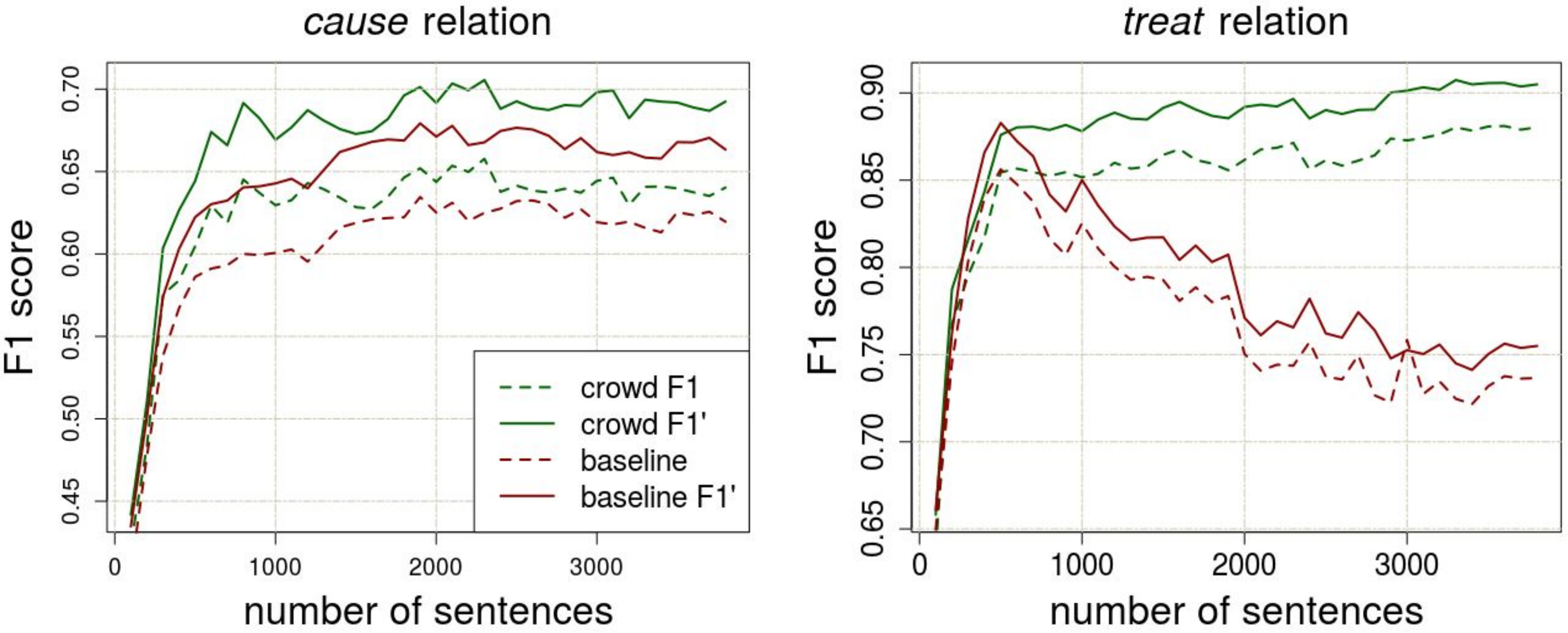
$$R' = \frac{\sum_s srs(s) \cdot tp(s)}{\sum_s srs(s) \cdot tp(s) + srs(s) \cdot fn(s)}$$

Weighted F1:

$$F1' = \frac{2P'R'}{P' + R'}$$

CROWD vs. DISTANT SUPERVISION MODEL QUALITY

Distant Supervision: Mintz et al. *Distant supervision for relation extraction without labeled data*. ACL 2009



- **crowd is better training data than distant supervision**
- **weighing the eval metrics with SRS results in increase**

RESULTS SUMMARY

CrowdTruth performs **just as well as medical experts** at training a relation extraction classifier, while being **cheaper** and **always available**.

CrowdTruth performs **better than distant supervision** at training the classifier.

Metrics weighted with SRS evaluate **truth on a continuous scale**, as opposed to using binary ground truth labels.

https://sadworkshop.wordpress.com/
https://humlworkshop.github.io/HumBL-WWW2019/
http://crowdtruth.org/tutorial/

Subjectivity, Ambiguity and Disagreement (SAD) in Crowdsourcing workshop @ WebConf 2019

Augmenting Intelligence with Bias-Aware Humans--in--the--Loop (HumBL) workshop @ WebConf 2019

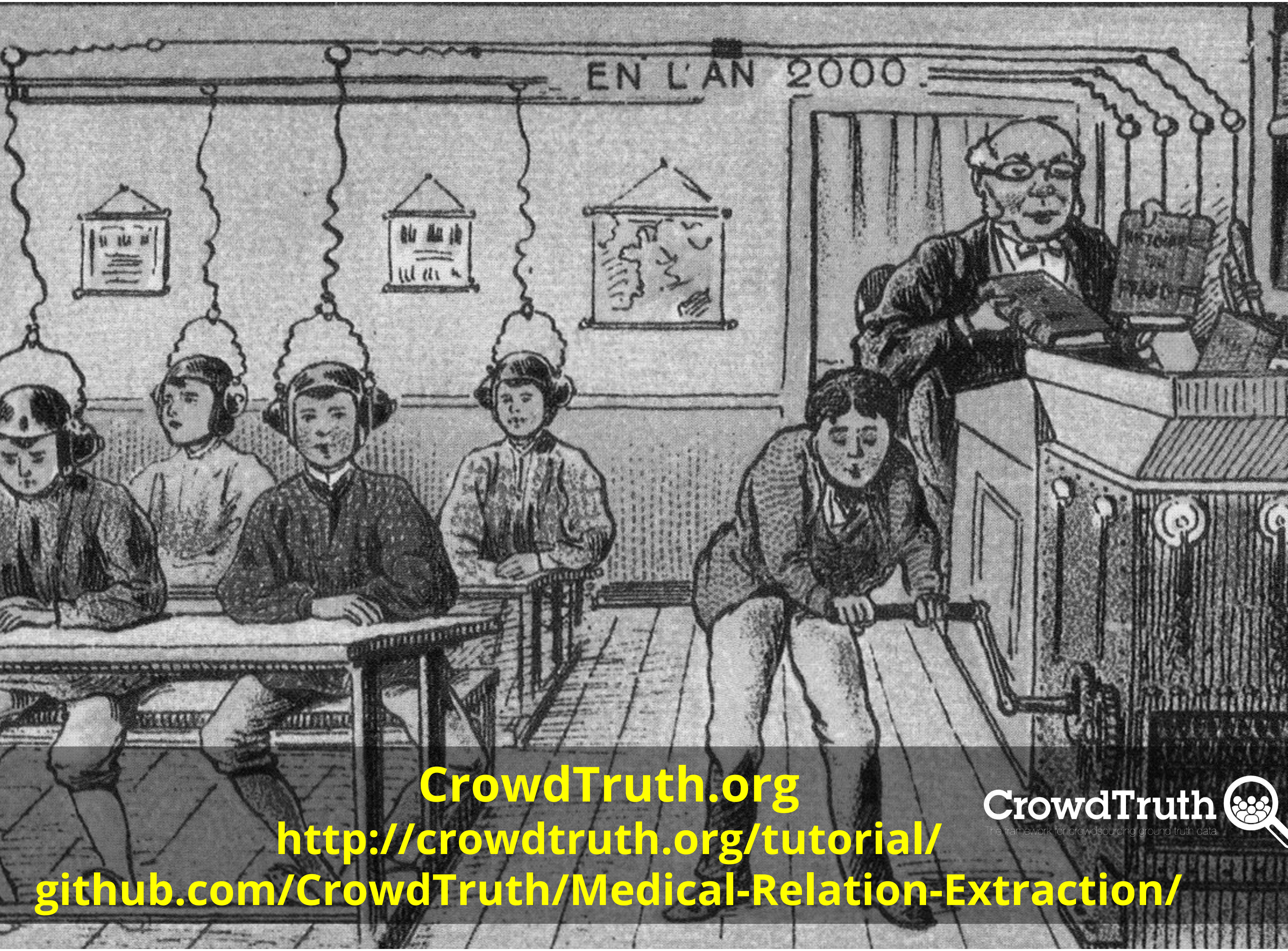
CrowdTruth Tutorial @ WebConf 2019
How to Build Ambiguity-Aware Ground Truth

THE WEB CONFERENCE
San Francisco
May 13-17, 2019

CrowdTruth.org

CrowdTruth
The framework for crowdsourcing ground truth data

github.com/CrowdTruth/Medical-Relation-Extraction/



CrowdTruth.org

http://crowdtruth.org/tutorial/

github.com/CrowdTruth/Medical-Relation-Extraction/

CrowdTruth

The framework for crowdsourcing ground truth data