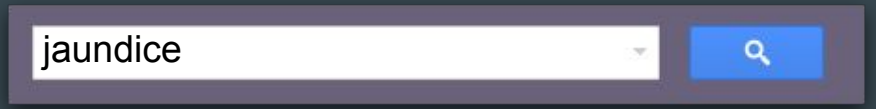
Why Is That Relevant?

Collecting Annotator Rationales for Relevance Judgments

Presenter: Tyler McDonnell
Department of Computer Science
The University of Texas at Austin

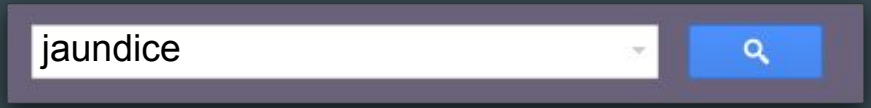
Tyler McDonnell, Matthew Lease, Mucahid Kutlu, Tamer Elsayed
2016 AAAI Conference on Human Computation & Crowdsourcing

Search Relevance



What are the symptoms of jaundice?

Search Relevance



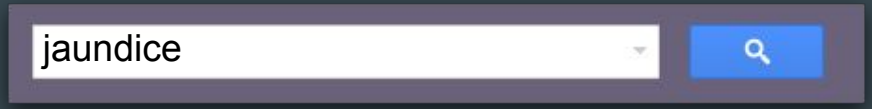
What are the symptoms of jaundice?

Search Relevance

25 Years of the National Institute of Standards & Technology Text REtrieval Conference (NIST TREC)

- Expert assessors provide relevance labels for web pages.
- Task is highly subjective: even expert assessors disagree often.*

Google: Quality Rater Guidelines (150 pages of instructions!)



What are the symptoms of jaundice?

* Voorhees 2000

A First Experiment

- Collected sample of relevance judgments on Mechanical Turk.
- Labeled some data myself.
- Checked agreement.
  - Between workers.
  - Between workers vs. myself.
  - Between workers vs. NIST gold.
  - Between myself vs. NIST gold.

- Why do I disagree with NIST? Who knows!

Search Relevance

Can we do better?

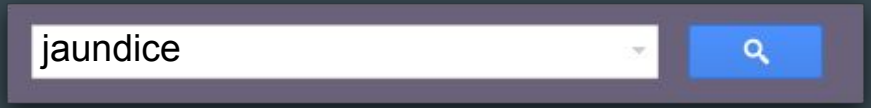
The Rationale

jaundice

What are the symptoms of jaundice?

Jaundice, also known as **icterus**, is a yellowish or greenish pigmentation of the skin and whites of the eyes due to high bilirubin levels.[1][2] It is commonly associated with itchiness.[3]

The Rationale

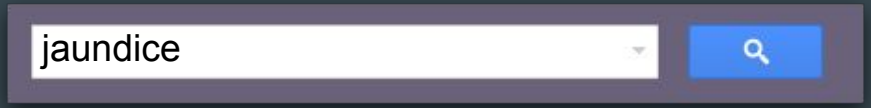


What are the symptoms of jaundice?

Why Rationales?

1. **Transparency**

- Focused context for interpreting objective *or* subjective answers.
- Workers can justify decisions and establish alternative truths.
- Useful for immediate verification and future users of collected data.

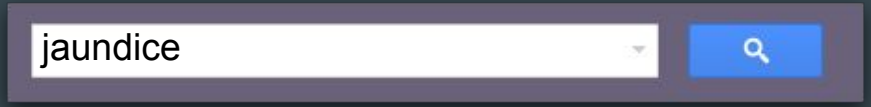


What are the symptoms of jaundice?

Why Rationales?

2. Reliability & Verifiability

- Logical insight into reasoning reduces temptation to cheat.
- Makes **explicit** the implicit reasoning underlying labeling tasks.
- Enables sequential task design.



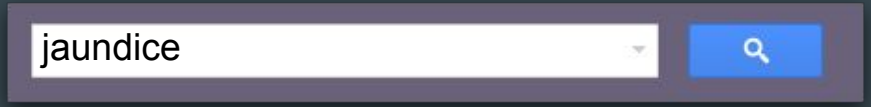
What are the symptoms of jaundice?

Why Rationales?

3. Increased Inclusivity

Hypothesis: With improved transparency and accountability, we can remove all traditional barriers to participation so **anyone** interested is allowed to work.

- Scalability
- Diversity
- Equal Opportunity



What are the symptoms of jaundice?

Experimental Setup

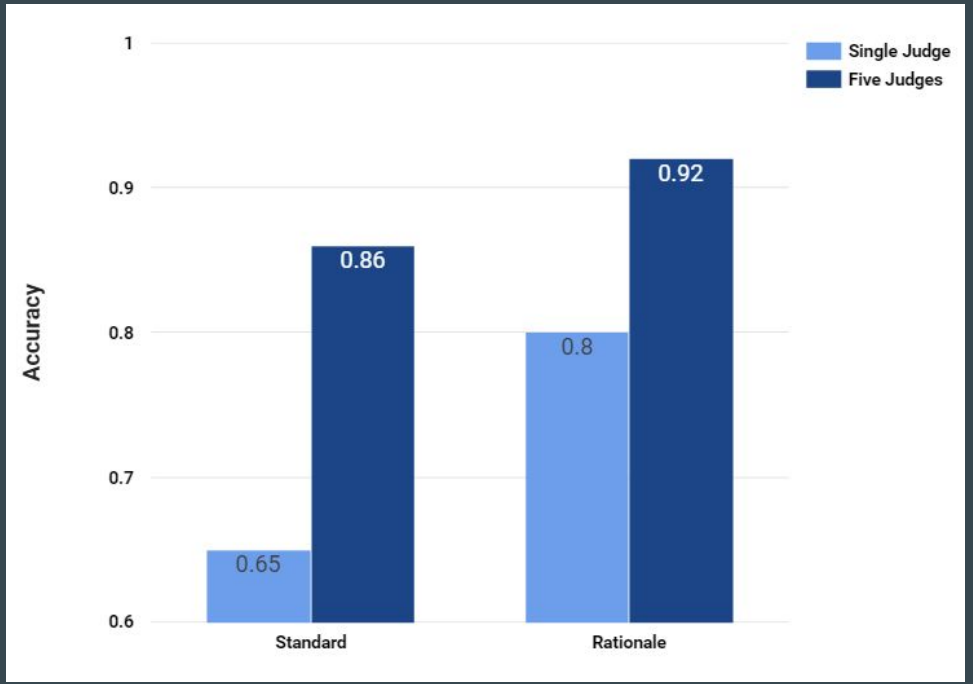
- Collected relevance judgments through Mechanical Turk.

- Evaluated two main task types.
  - Standard Task (Baseline): Assessors provide a relevance judgment for a given query, web page.
  - Rationale Task: Assessors provide a relevance judgment and rationale from the document.
  - (will mention two other variants later)

- No worker qualifications.
- No “honey-pot” or verification questions.
- Equal pay across all evaluated tasks.
- 10,000 judgments collected. (Available online*)

Results - Accuracy

- Workers who provide rationales produce higher quality work.
- Rationale tasks provided higher binary accuracy (92-96%) than comparable studies (80-82%).*
- Collecting one rationale provides only marginally lower accuracy than five standard judgments.

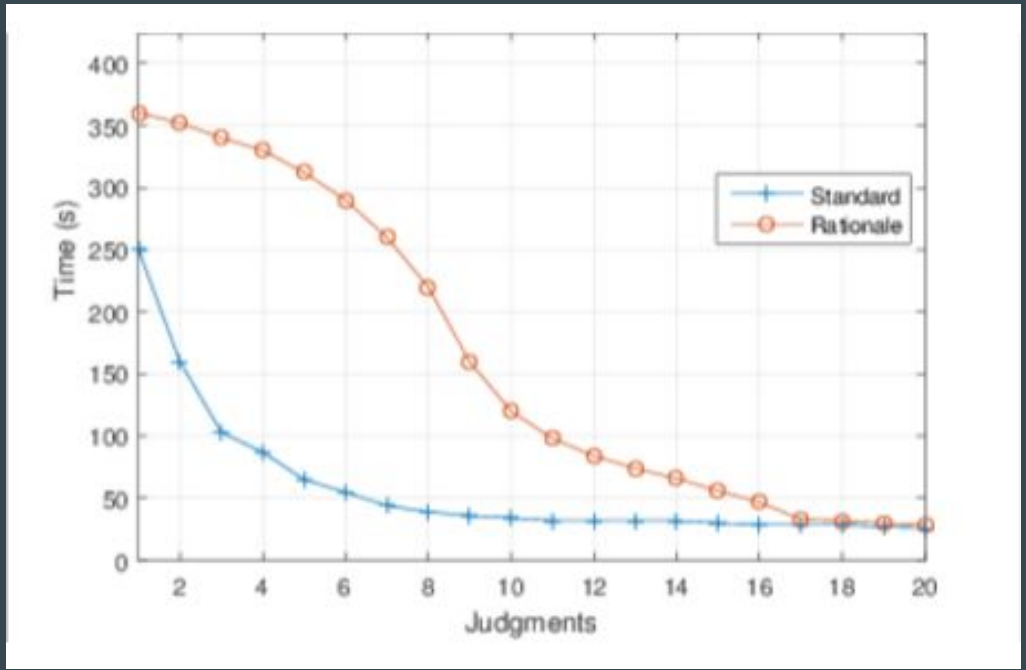


* Hosseini et al. 2012

Results - Cost-Efficiency

- Rationale tasks initially take longer to complete, but the difference becomes negligible with task familiarity.

- Rationales make **explicit** the implicit reasoning process underlying labeling.



But wait, there's more!

What about the rationale?

Using Rationales: Overlap

Assessor 1 Rationale

Assessor 2 Rationale

Using Rationales: Overlap

Assessor 1 Rationale

Assessor 2 Rationale

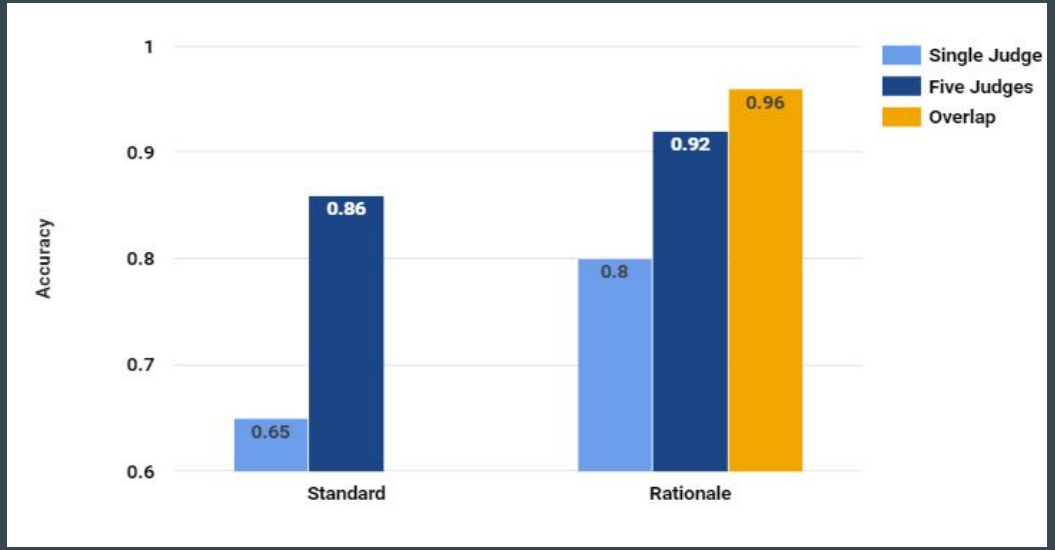
Overlap

Idea: Filter judgments based on pairwise rationale overlap among assessors.

Motivation: Workers who converge on similar rationales are likely to agree on labels as well.

Results - Accuracy (Overlap)

- Filtering collected judgments by rationale overlap prior to aggregation increases quality.



Using Rationales: Two-Stage Task Design

Assessor 1: Relevant

Assessor 1 Rationale

Assessor 2:

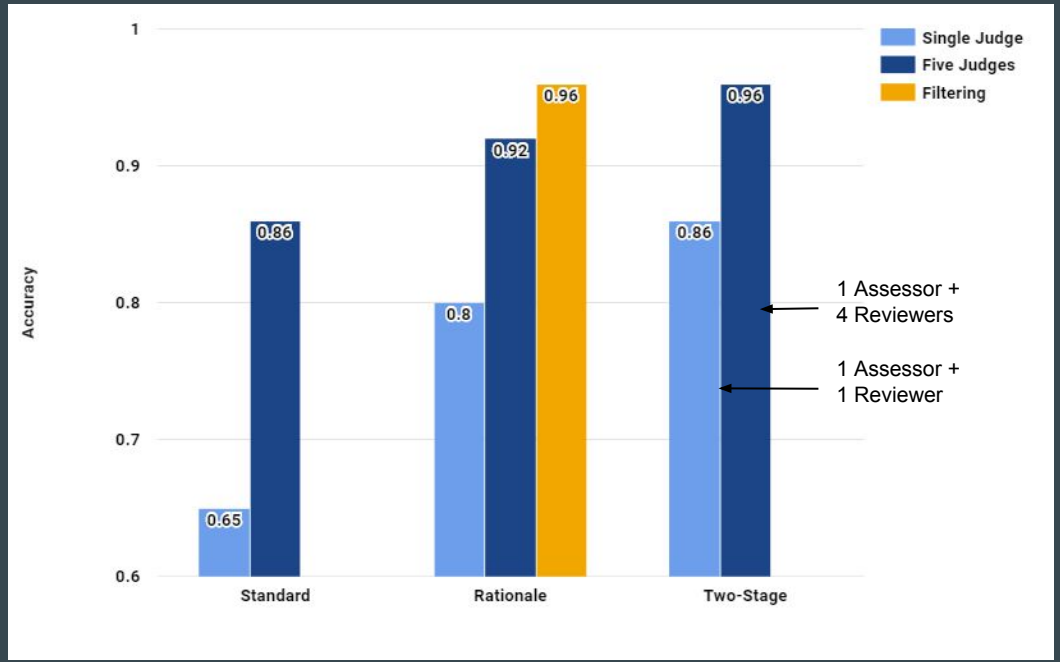
?

Idea: Reviewer must confirm or refute judgment of initial reviewer.

Motivation: Worker must consider their response in the context of peer's reasoning.

Results - Accuracy (Two-Stage)

- Single review offers same accuracy as five aggregated standard judgments.
- Aggregating reviewers reaches same accuracy as filtered approaches.



The Big Picture

- Transparency
  - Context for understanding and validating subjective answers.
  - Convergence on justification-based crowdsourcing. (e.g., Microtalk HCOMP 2016)
- Improved Accuracy
  - Rationales make the implicit reasoning for labeling *explicit* and hold workers accountable.
- Improved Cost-Efficiency
  - No additional cost for collection once workers are familiar with task.
- Improved Aggregation
  - Rationales are a *signal* that can be used for filtering or aggregating judgments.

Future Work

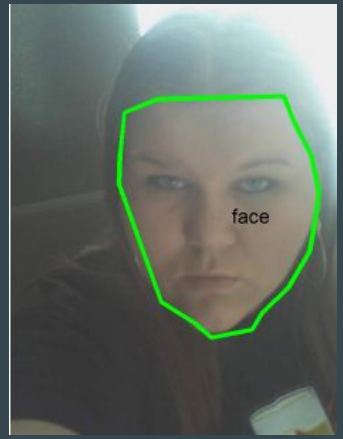
Dual Supervision: How can we further leverage rationales for aggregation?

- Supervised learning over labels/rationales.
  Zaidan, Eisner, Piatko 2007. NAACL 2007

Task Design: What about other sequential task designs? (e.g., multi-stage)

Generalizability: How far can we generalize rationales to other tasks? (e.g., images)

- Donahue, Grauman. *Annotator Rationales for Visual Recognition*. ICCV 2011.



Acknowledgements

We would like to thank our many talented crowd contributors.

This work was made possible by the Qatar National Research Fund, a member of Qatar Foundation.

Questions

				Binary		Ternary	
Row	Task	Filter	Judgments	Accuracy	Cohen's κ_C	Accuracy	Cohen's κ_C
1	Standard	-	Single Judge	0.65	0.36	0.47	0.34
2	Rationale	-	Single Judge	0.80	0.51	0.64	0.50
3	Two-Stage	-	Judge + Reviewer	0.85	0.58	0.75	0.60
4	Standard	-	5 Judges (EM)	0.86	0.59	0.75	0.46
5	Rationale	-	5 Judges (EM)	0.92	0.80	0.84	0.80
6	Rationale	TOP-3	5 Judges (EM)	0.93	0.81	0.91	0.82
7	Rationale	THRESHOLD	5 Judges (EM)	0.96	0.85	0.91	0.84
8	Two-Stage	-	Judge + 4 Reviewers (EM)	0.96	0.85	0.91	0.85

Query (Alice)	dogs for adoption
Narrative (Alice)	I want to find information on adopting a dog. This includes names and locations of rescue organizations or vehicles (e.g. classifieds) as well as documents with info on qualifications, fees (if any), what to expect, resources, etc. Organizations may be rescue organizations, pounds, shelters, etc. but not breeders or pet shops, unless the pet shop runs adoption fairs. A site providing general information on dog adoption is also relevant.

	Document 1	**Document 2**	**Document 3**
Worker 1 Judgment (Tom)	Probably Not Relevant	Definitely Relevant	Definitely Not Relevant
Worker 1 Rationale (Tom)	*Rooterville Sanctuary. For adoption: pets, pig, pigs, piggy, piggies, pork.*	*View our rescue dogs - visit our organization or contact us directly to see what is available.*	*The dogs listed here all require a new home. These dogs all deserve that second chance and you may be that special person to give it to them. View Rescue Dogs adoption fees. Contact us for more info.*
Worker 2 Judgment	Probably Not Relevant	Probably Relevant	Definitely Relevant
Worker 2 Reasoning	I agree that this organization is probably not likely to be one where Alice will find the animal she is looking for, since they seem to focus on pigs, although they mention dogs	It is a site that lists dog rescue organizations, which is what Alice is searching for. But it is an Australian website. I suspect Alice was looking for an organization in the US.	Tom provided a lot of information that shows why this page should be useful for Alice.
Gold Standard	Probably Not Relevant	Probably Relevant	Definitely Relevant

Algorithm 1 Threshold Filtering

```
1: procedure FILTER-BY-THRESHOLD(J_d)
2:     T ← SELECT-THRESHOLD(J_d)
3:     selected ← ∅
4:     for each (j_1, j_2) ∈ COMBINATIONS(J_d, 2) do
5:         if SIMILARITY(j_1, j_2) ≥ T then
6:             selected ← selected ∪ j_1 ∪ j_2
7:     return selected
8: procedure SELECT-THRESHOLD(J_d)
9:     T ← 0
10:    for each (j_1, j_2) ∈ COMBINATIONS(J_d, 2) do
11:        T ← max(T, SIMILARITY(j_1, j_2))
12:    return ROUND-DOWN(T, 10)
```

Algorithm 2 Top-N Filtering

```
1: procedure FILTER-BY-TOP-N(J_d, N)
2:     pairs = COMBINATIONS(J_d, 2)
3:     for each pair ∈ pairs do
4:         pair.sim ← SIMILARITY(pair.j_1, pair.j_2)
5:     Sort(pairs) by descending similarity
6:     return GETTOPJUDGMENTS(pairs, N)
```