Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge

Eddy Maddalena*, Marco Basaldella*, Dario De Nart*, Dante Degl’Innocenti*, Stefano Mizzaro*, **Gianluca Demartini**+

* University of Udine, Italy

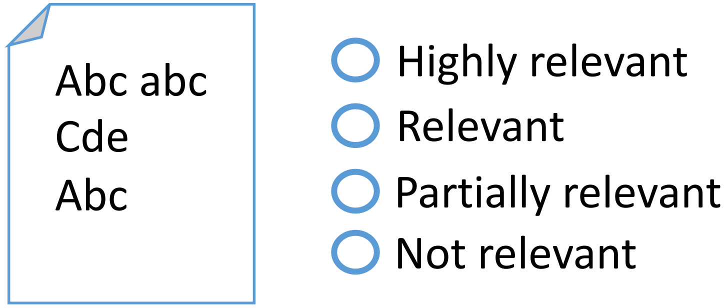
+ University of Sheffield, UK

EPSRC
Engineering and Physical Sciences
Research Council

Crowdsourcing Relevance Judgements

- Task: Given a Query, Document pair
  Is the doc
  highly relevant, relevant, partially relevant, not relevant?
- Ask multiple workers
- Aggregate answers to obtain a relevance label

Query: jaguar



Abc abc
Cde
Abc

○ Highly relevant
○ Relevant
○ Partially relevant
○ Not relevant

Our Research Question

Can we **limit the time to judge** to **reduce the cost ($$)** of creating IR test collections?

Hypothesis

Yes, but with quality loss

Our Experimental Setup

- **TREC8** Topics and documents (binary and 4-level expert judgements)
- **CrowdFlower**, repeated for USA and IND
- Majority vote aggregation
- Quality control: topic understanding question + high quality workers
- HIT Reward adapted based on the expected completion time

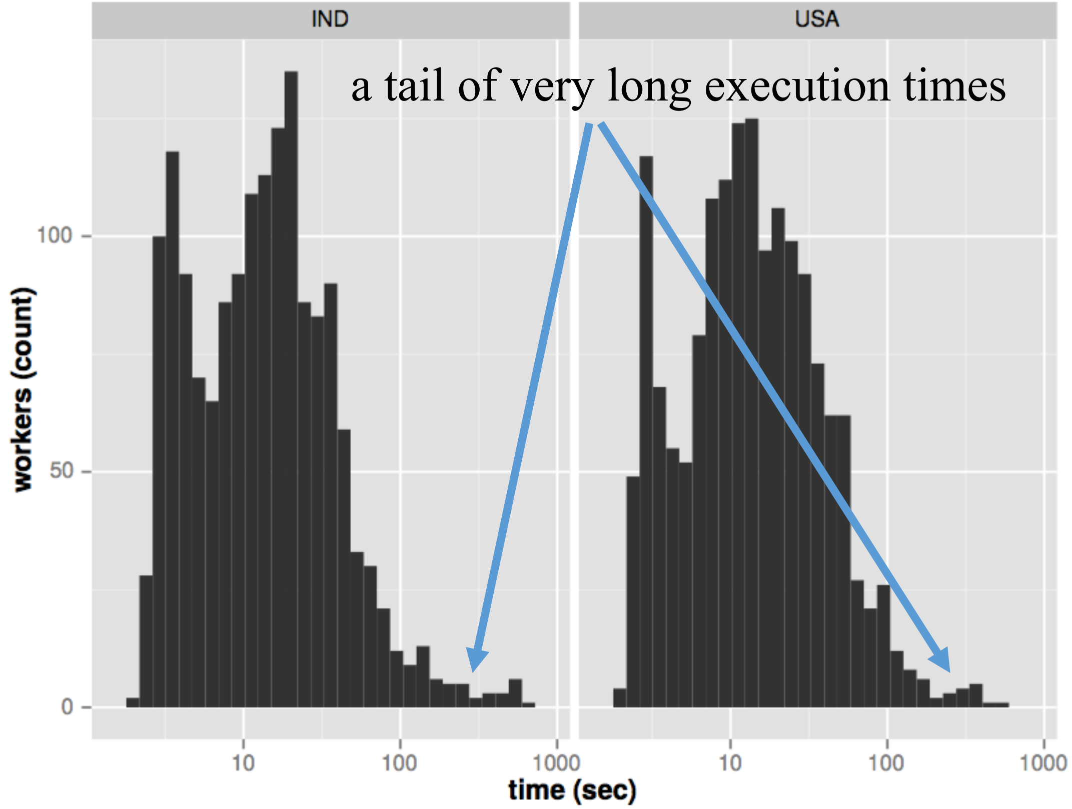
- Quality of a judgement: **Agreement** with editorial judgements
  - Cohen's Kappa and distance with 4-level labels

Our Experimental Setup

- E1 **Unbound time** (i.e., the standard approach)
  - 5 judgements per doc, 8 documents, 5 topics, 2 crowds = 400 workers
- E2 Document shown for a **predefined amount of time**
  - 30, 15, 7, 3 seconds. Each worker to judge 8 docs
- E3 **Same timeout** for all 8 documents (15 or 30 sec)
- E4 **Fixed budget**: comparison between
  - more quick judgements
  - few slow judgements

E1: We Have All the Time in the World

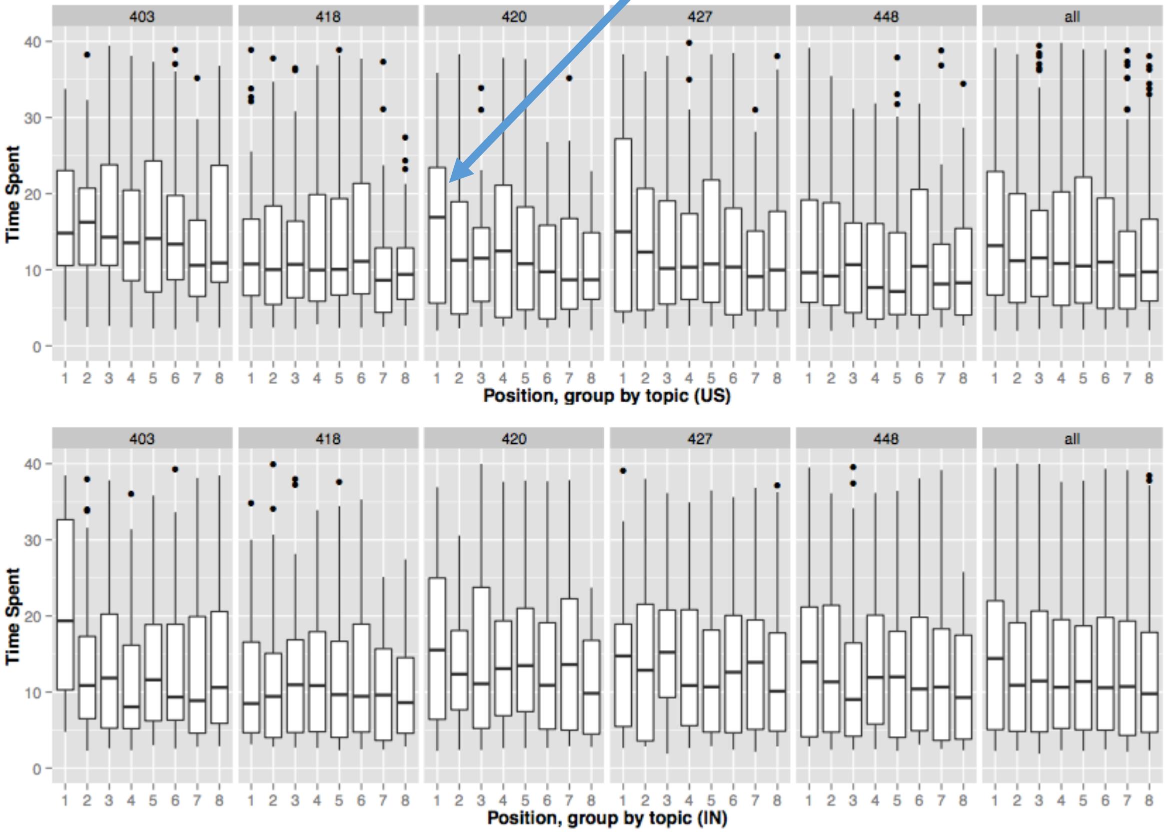
- RQ: **How much time** do crowd workers take to judge the relevance of a document **if no time constrain** is set?
    - 5 workers to judge a permutation of 8 docs



a tail of very long execution times

Median:
13 sec
Mean
24-25
sec

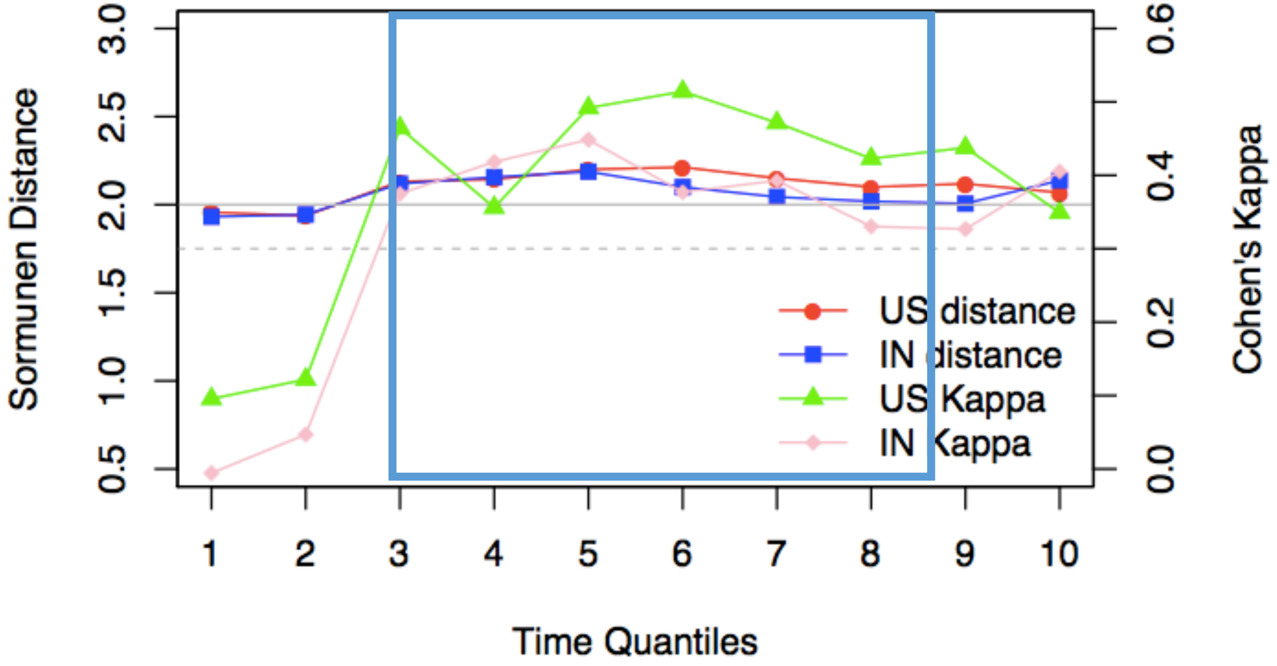
First doc takes longer (learning)

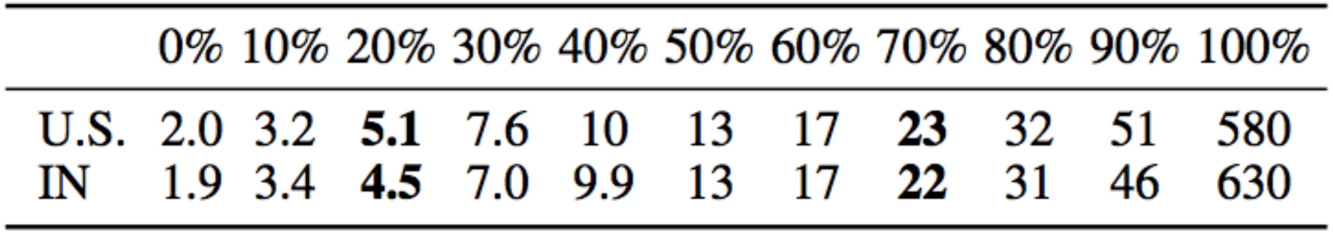


E1: We Have All the Time in the World

- No correlation of time with
  - Doc length
  - Doc readability
  - Topic
  - Relevance level
- Time vs Quality

5-30 seconds



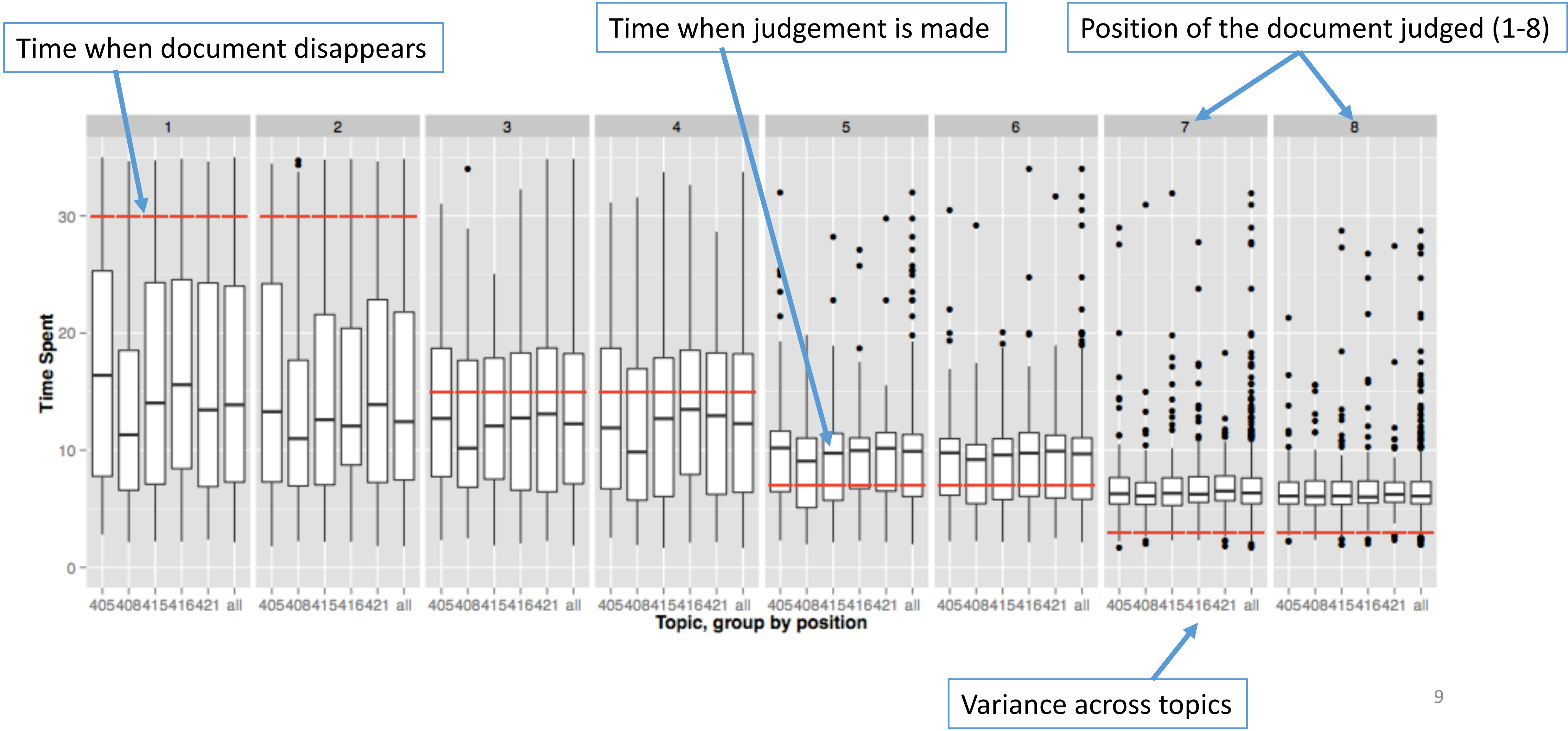


	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
U.S.	2.0	3.2	**5.1**	7.6	10	13	17	**23**	32	51	580
IN	1.9	3.4	**4.5**	7.0	9.9	13	17	**22**	31	46	630

E2: Faster! Faster! Sorry, Too Late

- Understand which is the **minimum amount of time required** to perform relevance judgments
- (max) timeouts: 30, 15, 7, 3 seconds
- Each worker to judge 8 docs, 2 for each timeout (one long, one short)
- Looking at Quality measures:
  - 3 and 7 secs are not enough
  - 15 slightly better than 30 (learning bias for position 1-2?)

E2: Faster! Faster! Sorry, Too Late



Time when document disappears

Time when judgement is made

Position of the document judged (1-8)

Variance across topics

E3: Selecting the Best Timeout

- We repeated E1 using 15 and 30 sec timeouts
- 15 seconds timeouts yield consistently better quality judgements
    - Than 30 seconds timeouts
    - Than no timeouts (E1 quality values)

Our Research Question

Can we **limit the time to judge** to **reduce the cost ($$)** of creating IR test collections?

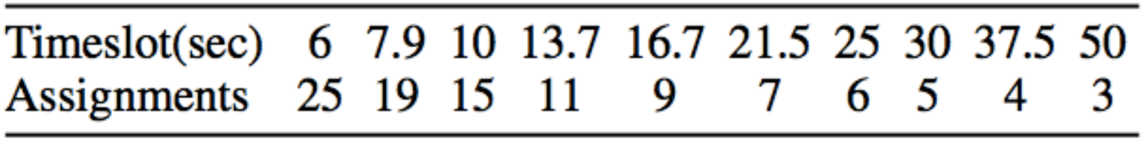
Hypothesis

~~Yes, but with quality loss~~

Yes, and it improves the quality!

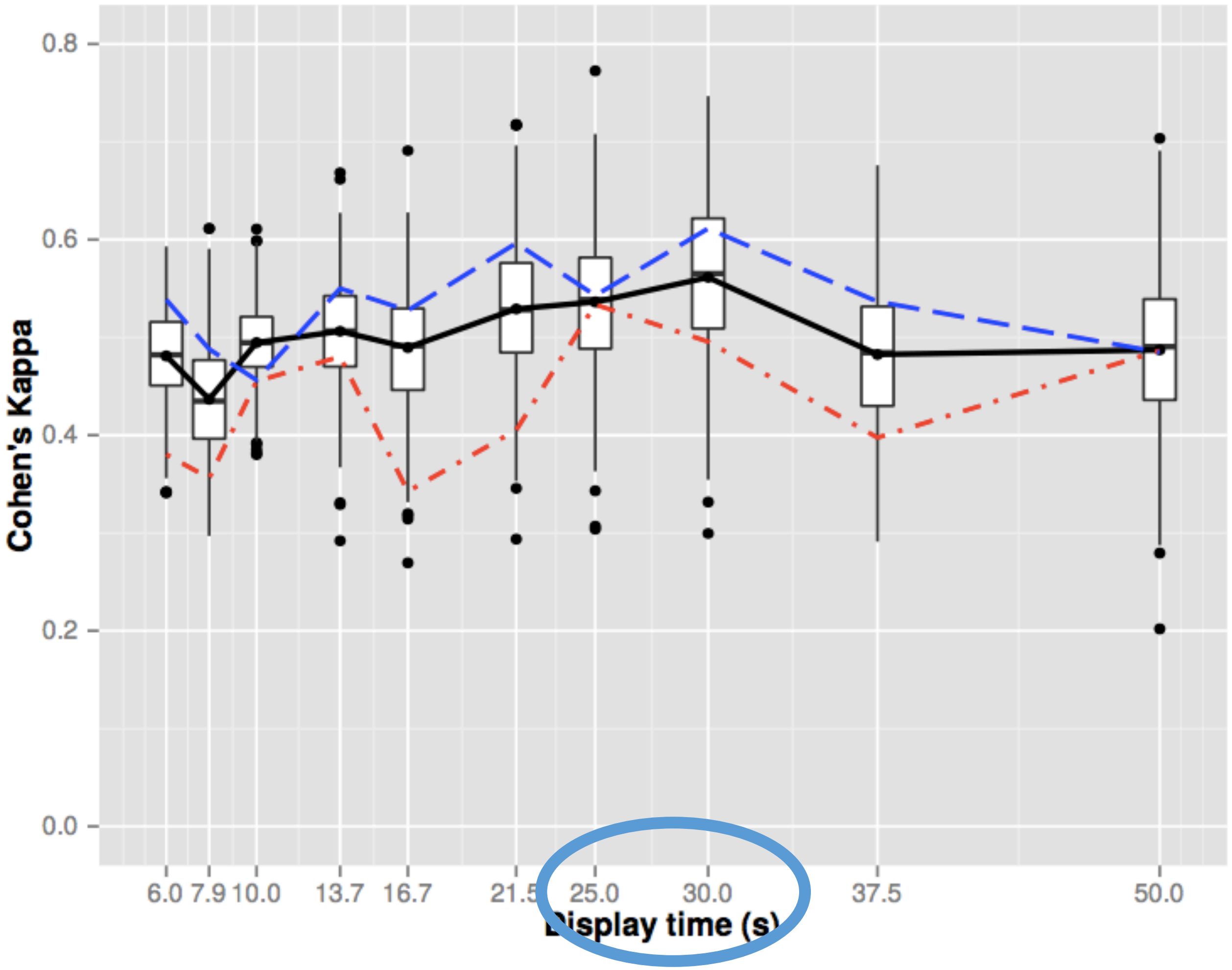
E4: Many Fast&Furious or a Few Laid-Back?

- **Fixed budget**:
  - small timeout, more workers
  - Long timeout, less workers
- We compared 10 combinations with the same budget



Timeslot(sec)	6	7.9	10	13.7	16.7	21.5	25	30	37.5	50
Assignments	25	19	15	11	9	7	6	5	4	3

- **Highest quality at 25-30 sec**



Findings

- The **first** couple of judgments done by a worker are of **lower quality**
- Judgements that take **more than 30** seconds are of **lower quality**
- **Time-outs** in relevance judgements HITs can **increase quality**
- The **best timeout** to be used lies in the interval of **25-30 seconds** and does not depend on topic, document, or crowd.

Conclusions

- Crowdsourcing Relevance Judgements for IR Evaluation can be **expensive to scale**
- **Limiting the time** to judge can **control the cost**
- But can also **increase the quality**!
  - By inducing workers to look at the document for a predefined amount of time
  - With a balance between boredom and stress -> “in the flow”

http://gianlucademartini.net