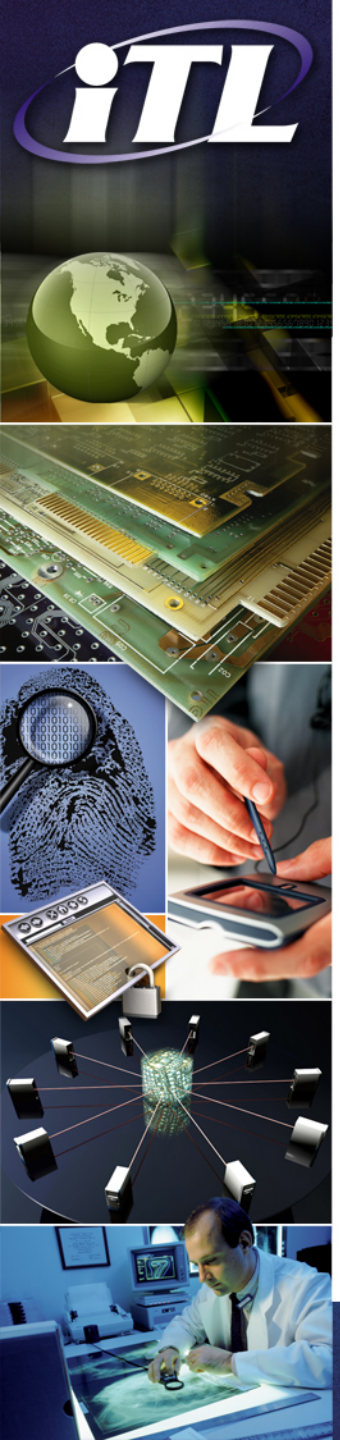


Multimedia Event Recounting (MER)

TRECVID 2014

Greg Sanders, David Joy, Jon Fiscus

NIST Information Technology Laboratory

Multimodal Information Group

Talk Outline

- MER Evaluation Overview
  - Tasks, data, evaluation, and caveats
- Results
  - Highlights of findings
- Panel Discussion Charge

The MER Task

- Execute a 10Ex MED Query generating a recounting for each video ranked above the R_0 rank threshold
  - "Recounting" is the annotation of the Event Query with scores and with key metadata evidence that was used to compute the score for the event.
    - *In effect, the recounting instantiates the query.*

Teams interpreted "key metadata evidence" differently

1. All evidence
2. All recountable evidence
3. Evidence optimizing MER

- For each piece of evidence
  - Localize the evidence
    - Temporally within the clip
    - Spatially within the video frame (optional)
  - Label as Key/Non-Key
    - Key evidence is "the minimal evidence that is needed to show that the video contains the event"

Some teams did not make this Key/Non-Key distinction

- Provide a textual description of the piece of evidence – we call this a "tag"

What Was Judged for Query/Recounting

- Judge whether or not the query was concise and logical
  - We later computed various objective measures of the length and structural complexity of the queries
- Judge each piece of key evidence by doing the following:
  - Read the tag's text and judge if the text accurately describes the snippet
  - Judge how well the evidence is temporally localized (for non-keyframe evidence)
  - Judge how well the evidence is spatially localized (for provided bounding box(es))
- After the judge has viewed all pieces of key evidence, the judge states whether the evidence convinced him/her that the clip contains an instance of the event

- All judgments made with Likert-style questions and a 5-point scale
  - Example: <tag name> correctly captures the contents of the snippet.
    - Strongly Disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly Agree

The judges weight concise vs. logical differently

When teams have differing Key/Non-Key distinctions, cross-team comparisons are not valid

Recountings Selected for Judgment

- Recountings were selected for:
  - 10 events
    - 6 Pre-specified events
    - 4 Ad-hoc
  - 15 highly ranked videos per event

- ≈ 5 independent judgments per recounting

Event Query Comparisons

The Event Queries were used by the MED systems

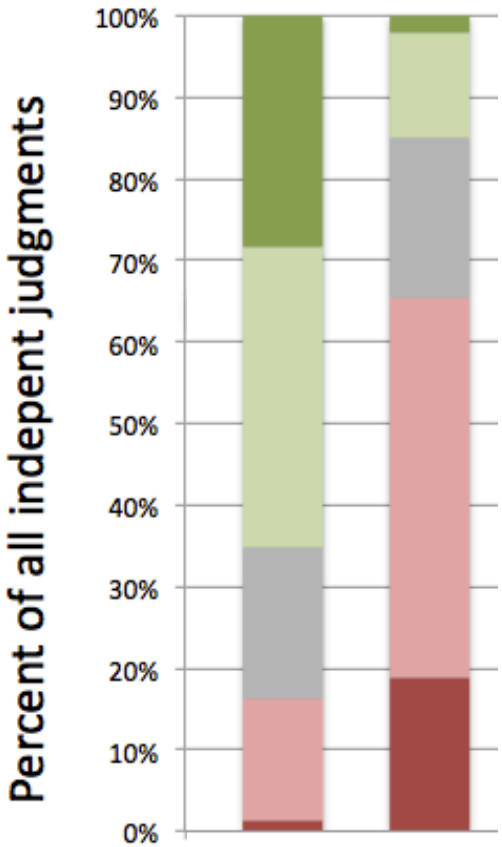
In general, each Event Query was judged by at least 10 different judges

Large differences in Query Size

Here is a short, concise query (5 nodes and 11 tags)

```
<query eventID="E043">
  <node id="E043" name="Busking" eq='SUM("D"=>0.66,"S"=>0.34)'>
<detector  id='D'  name='Detected Busking'>  <!
[CDATA[<parameters><classifier>svm</classifier><local_model_path>/svm/
ADEK10/E043.mat</local_model_path></parameters>]]>  </detector>
    <node id="S" name="Semantic busking" eq="SUM">
      <node id="S1" name="Objects" eq="WEIGHTED_SUM">
        <tag id="S1.1" name="musical instrument" weight="1.000" />
        <tag id="S1.2" name="street sign" weight="0.899" />
        <tag id="S1.3" name="instrument" weight="0.484" />
        <tag id="S1.4" name="dancer" weight="0.362" />
      </node>
      <node id="S2" name="Actions" eq="WEIGHTED_SUM">
        <tag id="S2.1" name="dancing" weight="0.735" />
        <tag id="S2.2" name="singing" weight="0.413" />
        <tag id="S2.3" name="performing" weight="0.390" />
      </node>
      <node id="S3" name="Scenes" eq="WEIGHTED_SUM">
        <tag id="S3.1" name="city street" weight="0.899" />
        <tag id="S3.2" name="street" weight="0.899" />
        <tag id="S3.3" name="parking lot" weight="0.574" />
        <tag id="S3.4" name="sidewalk" weight="0.502" />
      </node>
    </node>
  </node>
</query>
```

Human judgments also differed



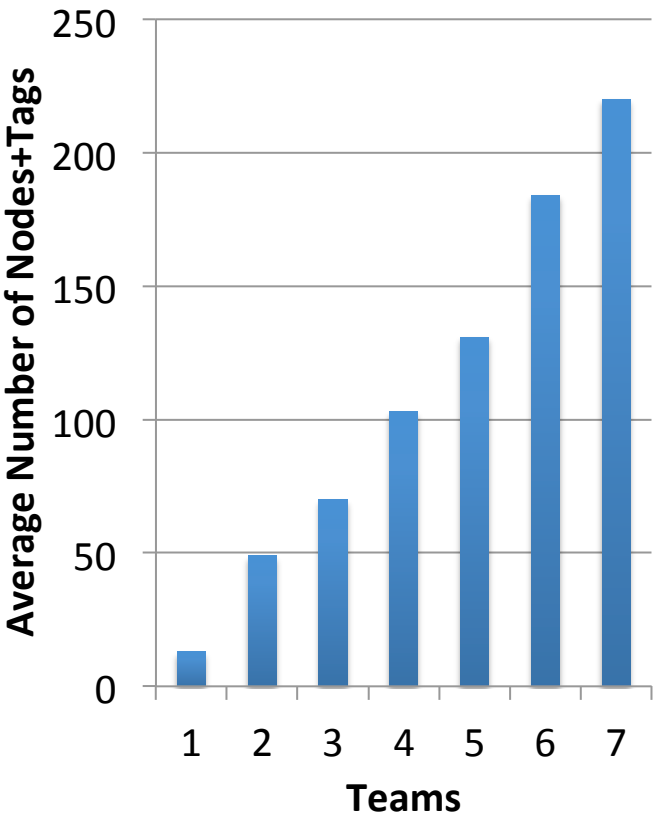
?

Query Size (number of nodes + number of tags)

Query Structural Metrics

- A Query is a tree structure of:
  - Nodes: contain nodes and tags
  - Tags: populated with evidence in the recounting
- Counts of Nodes and Tags are an objective measure of conciseness

Query size differed widely across teams



Summary Comments on Query Quality

- Event Query Quality judgments suggest the judges didn't pay attention to "concise"
  - We think the judges probably paid attention to whether the query seemed to make sense
  - My guess: judges liked queries containing plausibly relevant names of things and actions.

- Maybe we did not ask the judges the right question(s) about the queries
  - For example: we did not ask about "coverage"
    - Actually reading a number of the queries and comparing to the "Concise And Logical" scores from the judges suggests to me that judges did not pay attention to how thoroughly those queries ***covered*** the evidence that ought to have existed in recountings (the judges had not yet seen the recountings when they scored the queries).
    - I'll note that the judges were seeing only the one-sentence version of the event definitions.

  - It is my impression that because of inadequate **coverage,** I would have judged many queries more harshly (as not so logical) than our judges did.

  - How can we best judge Event Query Quality (or qualities)?

Recounting Comparisons

Evidence Quality:

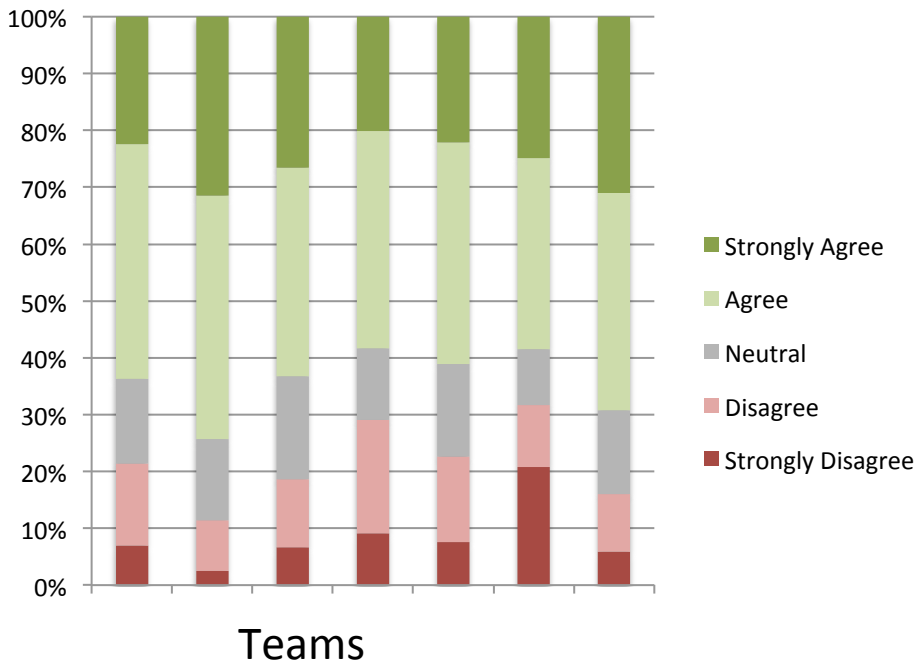
Question: How convincing was the evidence?

Answer: For all teams, it was more convincing for the positive clips (which is good).

Positive clips

red indicates judges were confused

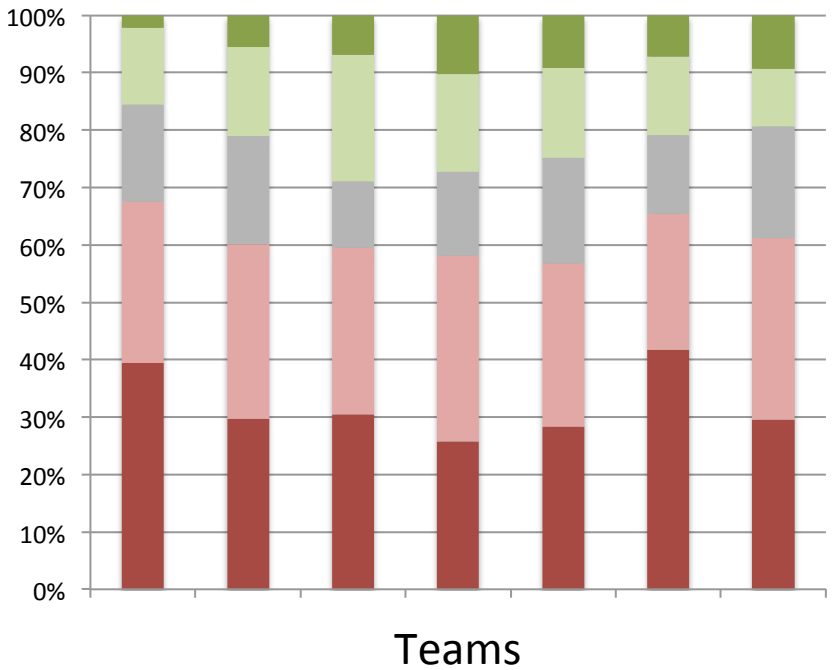
Key evidence (alone) was convincing: Targets only



Negative clips

green indicates judges were confused

Key evidence (alone) was convincing: Non-targets only

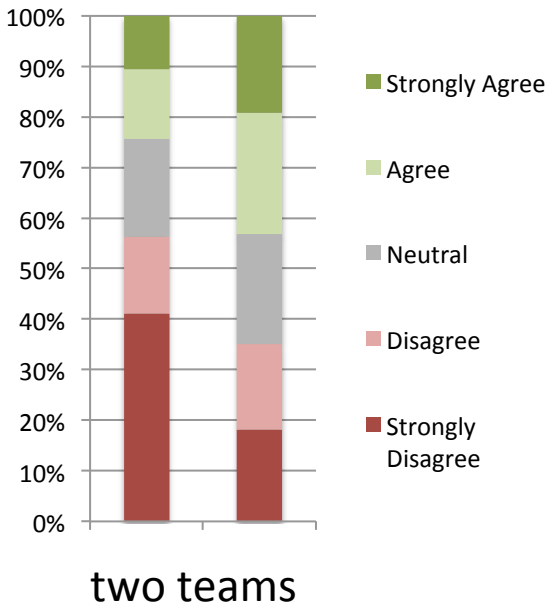


Temporal Localization of Evidence

There was a wide range of scores

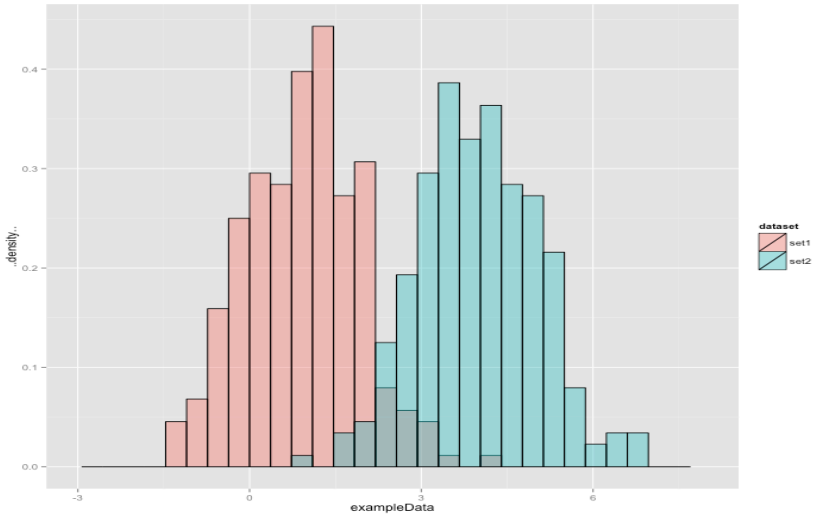
For each piece of evidence:
after the judge had viewed the snippet,
we asked the judge whether:
"The system chose the right window of time to present the evidence."

This question was not asked for pieces of evidence of type ***keyframe.***

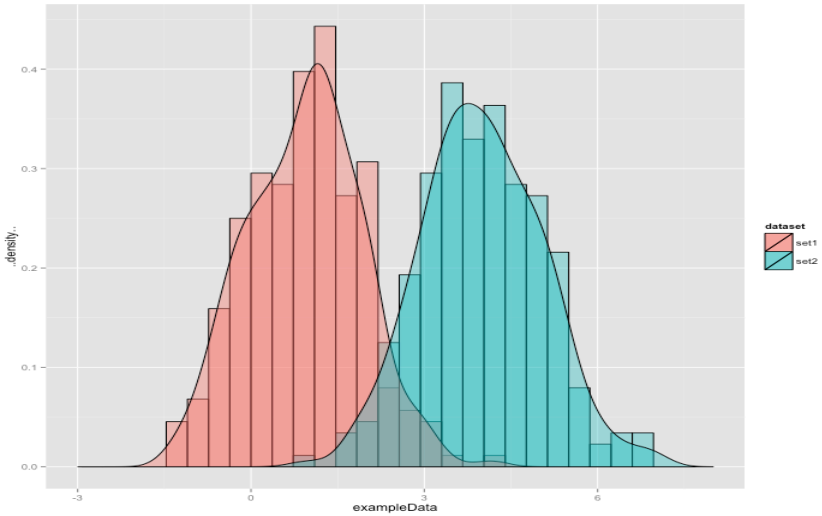


Digression: What's a "violin plot"
First, here are two distributions: pink and green

plain histograms



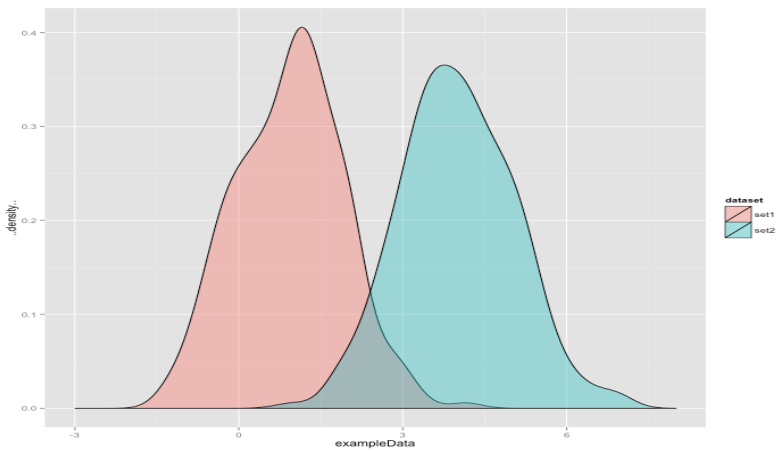
then add kernel density plots
(the smooth curves)



Digression: What's a "violin plot"

Second, From kernel density plots to violin plots

keeping just the kernel density plot, eliminating the histogram

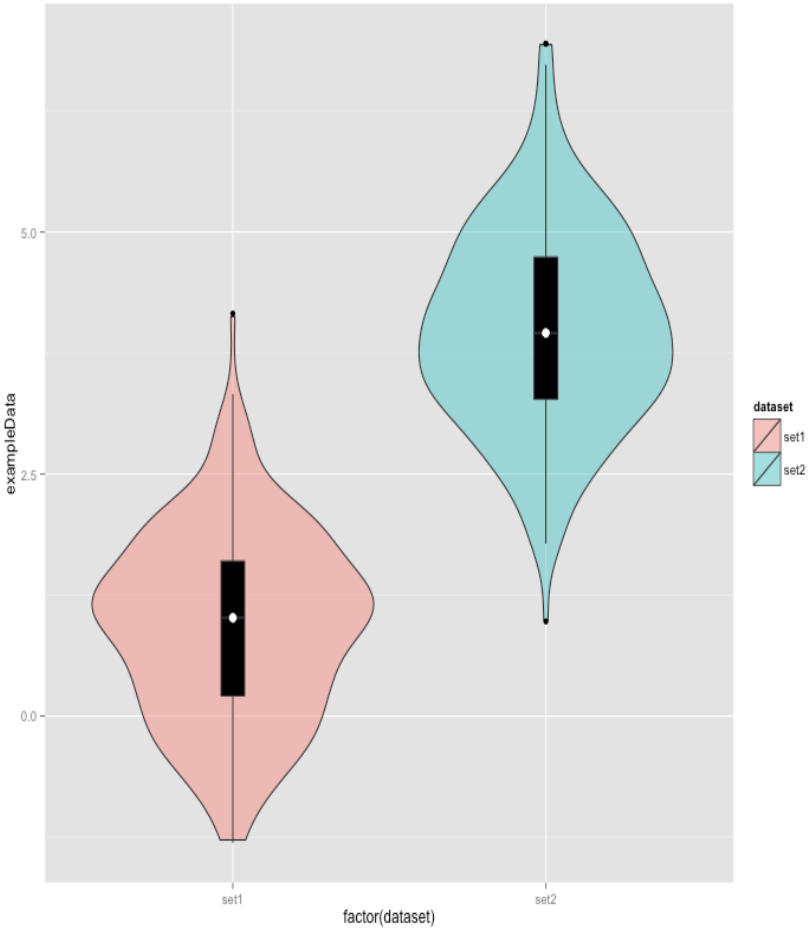


Rotate each kernel density plot counter-clockwise by 90° and "mirror" it. The result is two violin plots.

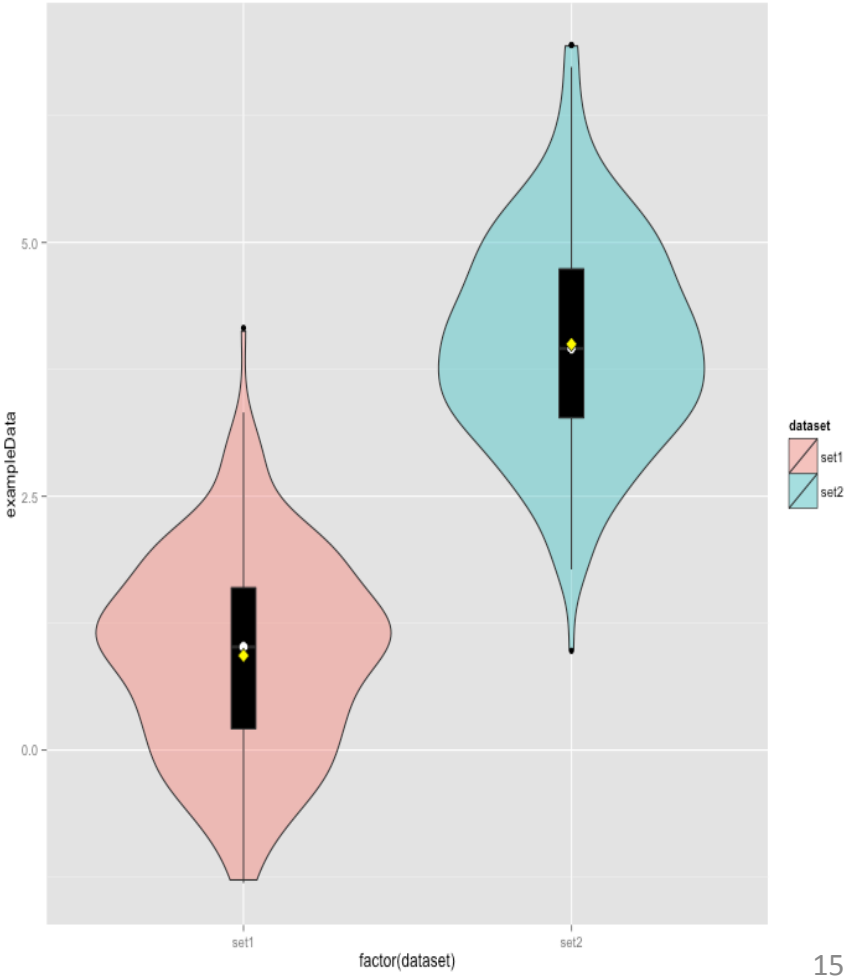
Digression: What's a "violin plot"

Third, One can add additional information to the violin plots

We can overlay a Tukey boxplot on top of each violin plot – here the white dot shows the median.



In addition, one could overlay a marker to show the mean (a yellow diamond here)



Violin plots of
Tag Quality vs. Confidence Score
by evidence type

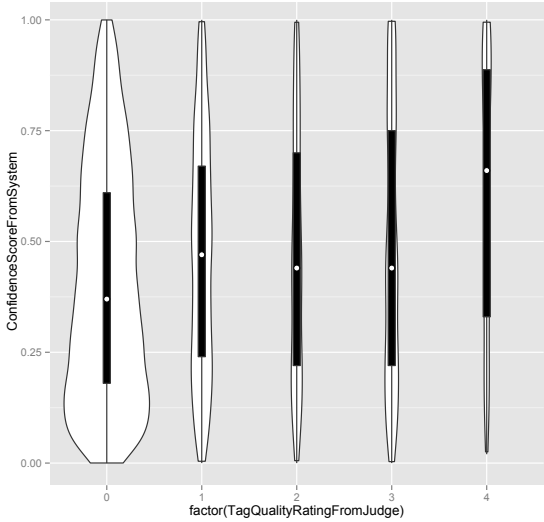
Tag Quality:

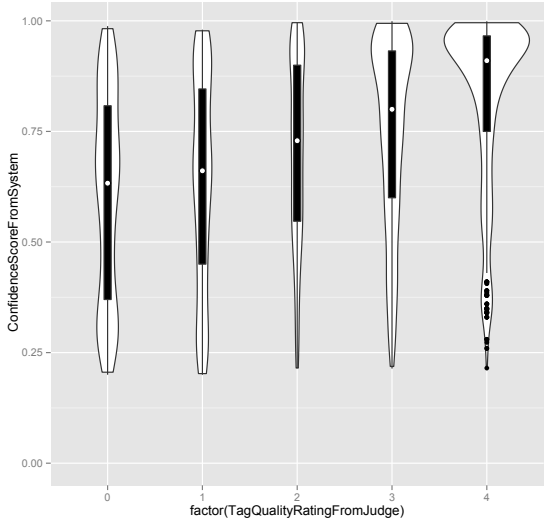
<tag name> correctly captures the contents of the snippet.

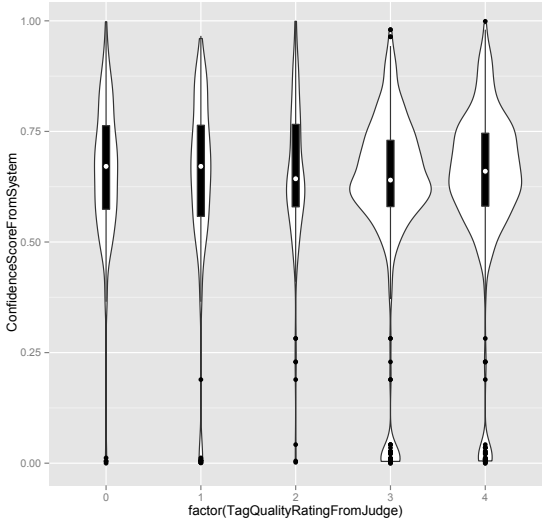
0 == Strongly Disagree
1 == Disagree
2 == Neutral
3 == Agree
4 == Strongly Agree

There was not a consistent correlation between

- the Tag Quality ratings from the judges and
- the Confidence Scores from the systems





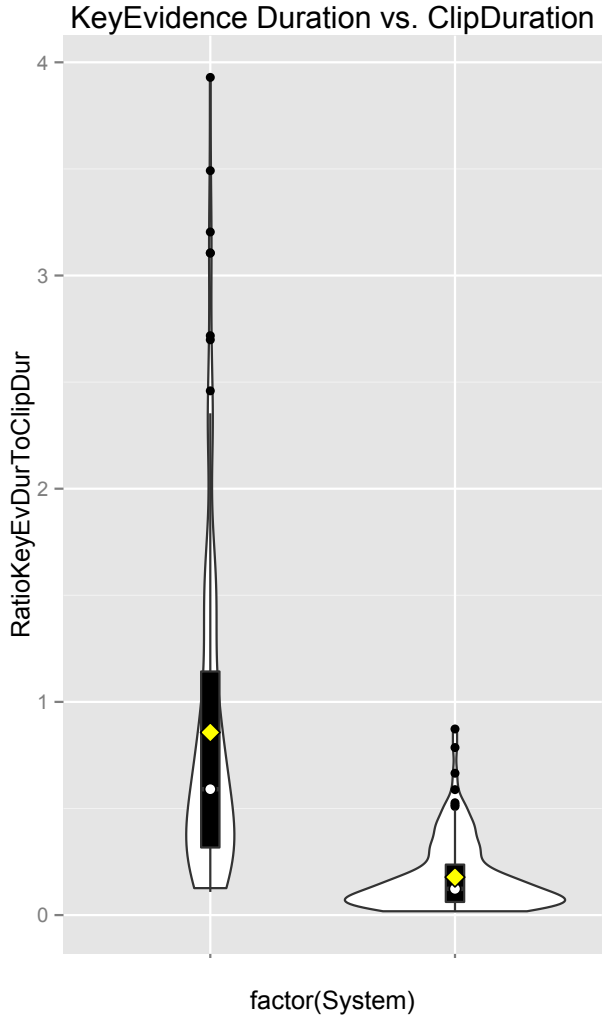


Recounted Percent

(in effect: Clip “Compression”)

Recounted Percent

How much of the clip time is in the snippets?
Distribution, over all clips, of KeyEvidenceDuration vs. ClipDuration



0.43 0.08 ← Sum(KeyEvidenceDurations) / sum(ClipDurations)

The white dot shows the median, and the yellow diamond shows the mean.

For some teams, the key evidence (the snippets) was only a small part of the overall clip durations.
So, it appears that it is possible to accomplish that (see plot on right).

HOWEVER, cross-team comparisons are NOT valid: The teams did not all make a key vs. non-key distinction, and the teams differed about what they considered to be key evidence.

We hope for interesting discussion during the upcoming panel.

Possible questions for discussion:

- What properties of the queries should we look at?
- What should be in the recountings – what should they consist of?
- What should we be measuring about the recountings, and how?
- What do the confidence factors from the systems actually mean?
-
-
-

Thank You!