# Thematic Roles: From Concepts to Utterances

John R. Kender

Claire Tsai, Michelle Alexander, Nnenna Okwara

Columbia University

and the IBM-Columbia team

Acknowledgement: Supported by the Intelligence Advanced Research Projects Activity (IAPRA) via Department of Interior National Business Center contract number D11PC20070. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of IARPA, DoI/NBC, or the U.S.Government.

# Outline

- Motivation and problem definition
- Summary of related work
- Proposed novel approach
- Results and waypoint experiments
- Summary of technical readiness
- Next steps

# Motivation: semantic classifiers as template fillers

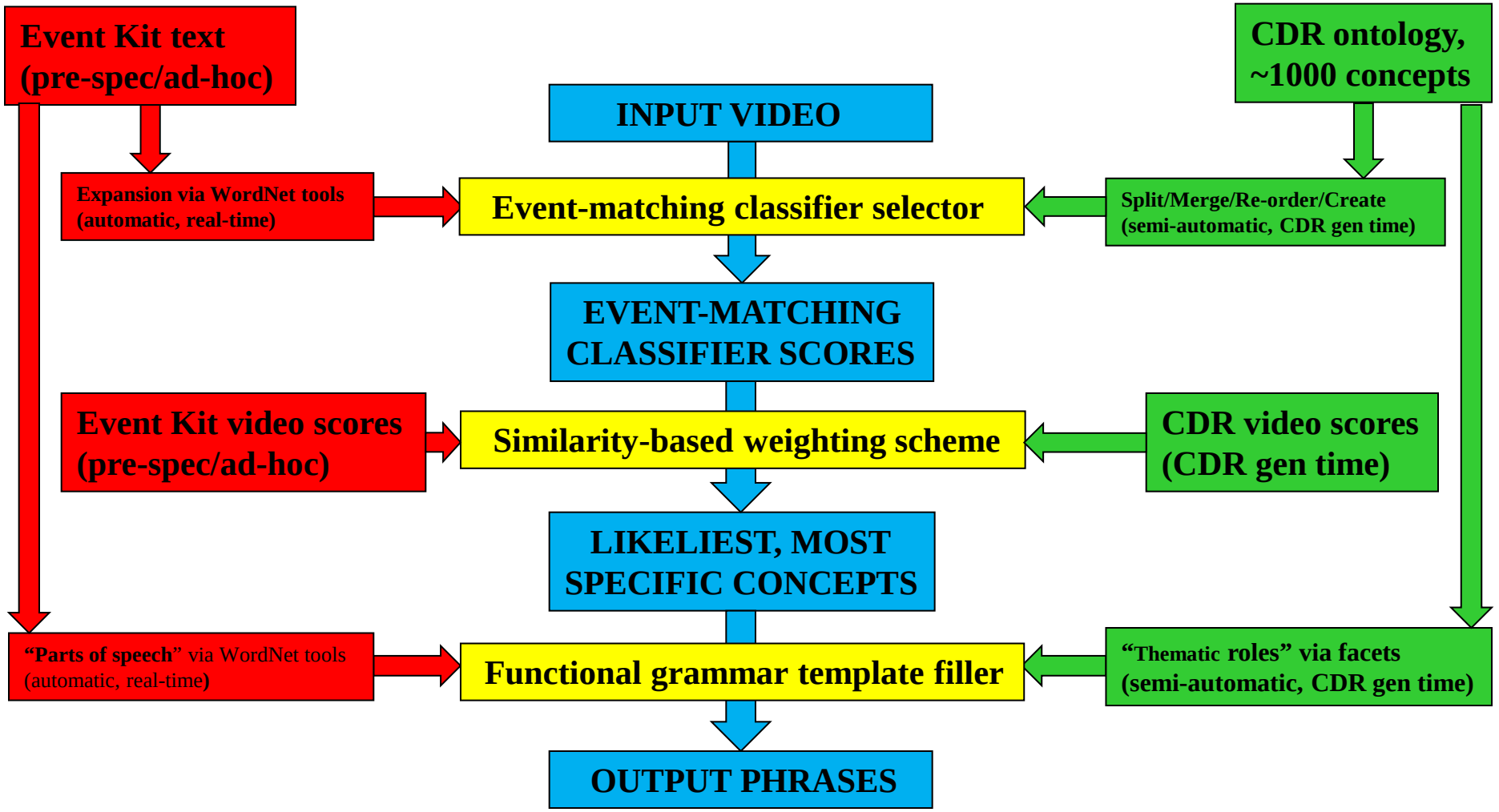
- Our semantic classifiers
  - About 1000 visual, 100 sound, 100 action; named
  - Designed midway between features and English text
  - Well suited for XML-based output: just use the name
- Existing “facet” structure reflects MER task well

| 9 MER XML Element Types | 7 IMARS Facets |
|---|---|
| Scene | Setting |
| Persons | People |
| Objects | Objects |
| Action | Activity |
| Text | Type |
| Linguistic audio | [own representation] |
| Non-linguistic audio | [own representation] |
| Videography | Domain |
| Other | Color Space |

# Related work at IBM-Columbia

- Map of: Event Kit text → relevant classifiers
- Map of: classifier scores → confidences
  - Assumes synchronization to .5Hz sampling
  - Assumes standardization into probabilities
- Map of: semantic classifiers → thematic roles
  - Visual → Agent, Theme, Patient, Instrument, Location
  - Sound → prepositional phrase: “, with sounds of”
  - Action → Theme
- Map of: video semantic clusters → sentences
  - Heuristic video segmentation on a *semantic* level

# Proposed novel approach



# Results and waypoint experiments

- Processed 30 DryRun, 30 MERTest, and 14,000 MED-detected videos in 20 events
  - MERTest used Visual plus Sound, but MED only Visual
  - No learning of “differentials” across or within events
  - System not tightly tuned
    - Semantic segmentation is fooled by bad camera work
    - One single template for all NLG is redundant and repetitive
- Event identification high, clip identification low
  - Our user studies show:
    - Text (ASR, OCR) appears critical for clip identification
    - But it is culturally dependent

# Results and waypoint experiments

1) Get matrix of classifiers × time (at 2 second intervals)

| | :00 | :02 | :04 | :06 | :08 | :10 | :12 | :14 | :16 | :18 | :20 | :22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Adult | | | | | | | | | | | | |
| Food | | | | | | | | | | | | |
| White | | | | | | | | | | | | |
| Rug | | | | | | | | | | | | |
| Mountain | | | | | | | | | | | | |
| Knife | | | | | | | | | | | | |
| Outdoors | | | | | | | | | | | | |
| Hockey | | | | | | | | | | | | |
| Stick | | | | | | | | | | | | |
| Cheer | | | | | | | | | | | | |
| Traffic | | | | | | | | | | | | |

# Results and waypoint experiments

2) Select classifiers relevant to event

| | :00 | :02 | :04 | :06 | :08 | :10 | :12 | :14 | :16 | :18 | :20 | :22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Adult | | | | | | | | | | | | |
| Food | | | | | | | | | | | | |
| White | | | | | | | | | | | | |
| Rug | | | | | | | | | | | | |
| Mountain | | | | | | | | | | | | |
| Knife | | | | | | | | | | | | |
| Outdoors | | | | | | | | | | | | |
| Hockey | | | | | | | | | | | | |
| Stick | | | | | | | | | | | | |
| Cheer | | | | | | | | | | | | |
| Traffic | | | | | | | | | | | | |

# Results and waypoint experiments

3) Select significant scores (globally)

| | :00 | :02 | :04 | :06 | :08 | :10 | :12 | :14 | :16 | :18 | :20 | :22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Adult | | | | | | | | | | | | |
| White | | | | | | | | | | | | |
| Outdoors | | | | | | | | | | | | |
| Hockey | | | | | | | | | | | | |
| Stick | | | | | | | | | | | | |
| Cheer | | | | | | | | | | | | |

# Results and waypoint experiments

## 4) Aggregate activity at each time

| | :00 | :02 | :04 | :06 | :08 | :10 | :12 | :14 | :16 | :18 | :20 | :22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Adult | | | | | | | | | | | | |
| White | | | | | | | | | | | | |
| Outdoors | | | | | | | | | | | | |
| Hockey | | | | | | | | | | | | |
| Stick | | | | | | | | | | | | |
| Cheer | | | | | | | | | | | | |

| | :00 | :02 | :04 | :06 | :08 | :10 | :12 | :14 | :16 | :18 | :20 | :22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Σ activity | 1 | 2 | 2 | 2 | 0 | 1 | 0 | 3 | 3 | 3 | 1 | 0 |

# Results and waypoint experiments

## 5) Segment activity vector

| | :00 | :02 | :04 | :06 | :08 | :10 | :12 | :14 | :16 | :18 | :20 | :22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Adult | | | | | | | | | | | | |
| White | | | | | | | | | | | | |
| Outdoors | | | | | | | | | | | | |
| Hockey | | | | | | | | | | | | |
| Stick | | | | | | | | | | | | |
| Cheer | | | | | | | | | | | | |

| | :00 | :02 | :04 | :06 | :08 | :10 | :12 | :14 | :16 | :18 | :20 | :22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Σ activity | 1 | 2 | 2 | 2 | 0 | 1 | 0 | 3 | 3 | 3 | 1 | 0 |

# Results and waypoint experiments

6) Map concepts to thematic role and fill slots in grammar

| | :02 | :04 | :06 | Role |
|---|---|---|---|---|
| Adult | | | | Agent |
| White | | | | Patient |
| Outdoors | | | | Location |
| Hockey | | | | Theme |
| Stick | | | | Instrumnt |
| Cheer | | | | Sound |

"Agent does Theme to a Patient using a Instrument at a Location [, with sounds of Sound]" →
"Somebody does Hockey to a White using something at a Outdoors"

# Results and waypoint experiments

6) Map concepts to thematic role and fill slots in grammar

| | :14 | :16 | :18 | Role |
|---|---|---|---|---|
| Adult | | | | Agent |
| White | | | | Patient |
| Outdoors | | | | Location |
| Hockey | | | | Theme |
| Stick | | | | Instrumnt |
| Cheer | | | | Sound |

"Agent does Theme to a Patient using a Instrument at a Location [, with sounds of Sound]" →

"A Adult does Hockey to something using a Stick at someplace, with sounds of Cheer"

# Results and waypoint experiments

- Sample outputs:
  - "A Demonstration_Crowd or a Group_of_People or a Crowd does Sitting_Down or does Talking or does Cheering or does Speaking_To_Camera or does Demonstration or does Press_Conference or does Politics to a Government-Leader or to a Man_Wearing_A_Suit or to a Politicians using a Demonstration_Banners at a Flags, with sounds of Cheer or Graduation or One_Person or Cheer or Speech or Crowds.“
  - Somebody does Rock_Clibming or does Cliff_Diving to some object using a Knife at a Canyons_and_Rock_Formations or at a Outdoors or at a Mountains or at a Rocky_Ground, with sounds of Clap or Group_of_Three_or_More or Vocals."

# Results and waypoint experiments

| 1 | Sequence of Activities | |
|---|---|---|
| Somebody does some action to a White using some object at a Unknown_Fire. | C = 0.90 | I = 0.50 |
| *None additional* | | |
| Somebody does some action to some object using a Power_Drill at a Unknown_Fire. | C = 0.90 | I = 0.67 |
| *None additional* | | |
| Somebody does some action to a White using some object at a Unknown_Fire. | C = 0.88 | I = 0.50 |
| *None additional* | | |
| Somebody does some action to a White using some object at a Unknown_Fire, with sounds of Vocals. | C = 0.87 | I = 0.50 |
| *None additional* | | |
| Somebody does some action to a White using some object at a Unknown_Fire. | C = 0.87 | I = 0.50 |
| *None additional* | | |
| Somebody does some action to a White using some object at a Unknown_Fire. | C = 0.89 | I = 0.50 |
| *None additional* | | |
| Somebody does some action to a White using some object at a Unknown_Fire. | C = 0.89 | I = 0.50 |
| *None additional* | | |
| Somebody does some action to a White using some object at a Unknown_Fire. | C = 0.89 | I = 0.50 |
| *None additional* | | |
| Somebody does some action to a White using a Knife at a Unknown_Fire, with sounds of Vocals. | C = 0.89 | I = 0.56 |
| *None additional* | | |

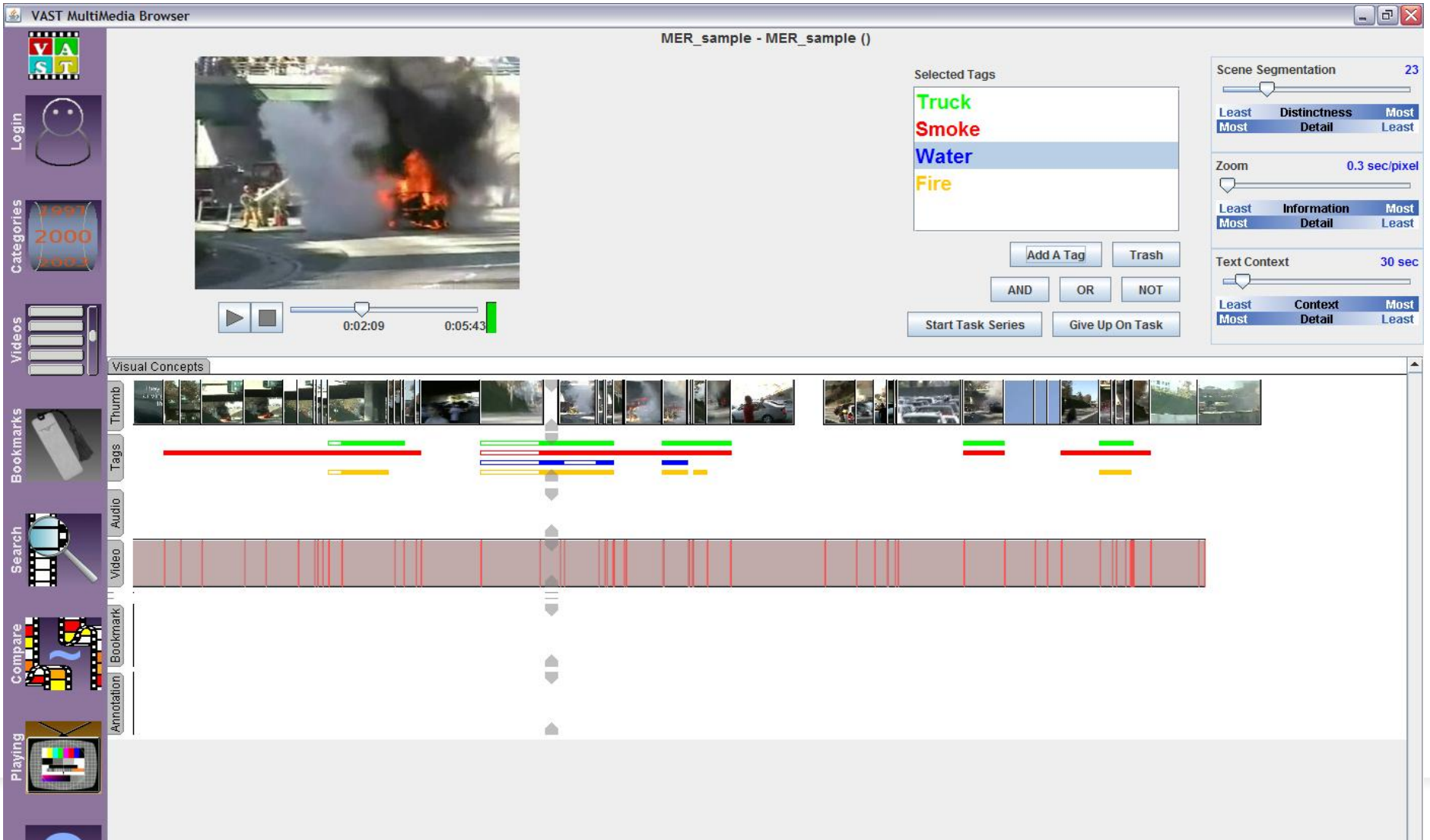
# Results and waypoint experiments

- Map of: event kits → relevant classifiers
  - Assisted by Lucene tool and WordNet; difficult
  - But: event kit text is free-form, ambiguous, incomplete, arbitrary in inclusions/exclusions, atemporal, causal, assumptive of extensive real-world knowledge
  - Would be clearer if event kit were in XML or a checklist
- Matlab is a good vehicle
  - Supports XML and matrix operations of (scores × time)
  - Quick intermediate visualizations for debugging
  - Lightweight and fast: (1100 × 60) input in .2 seconds

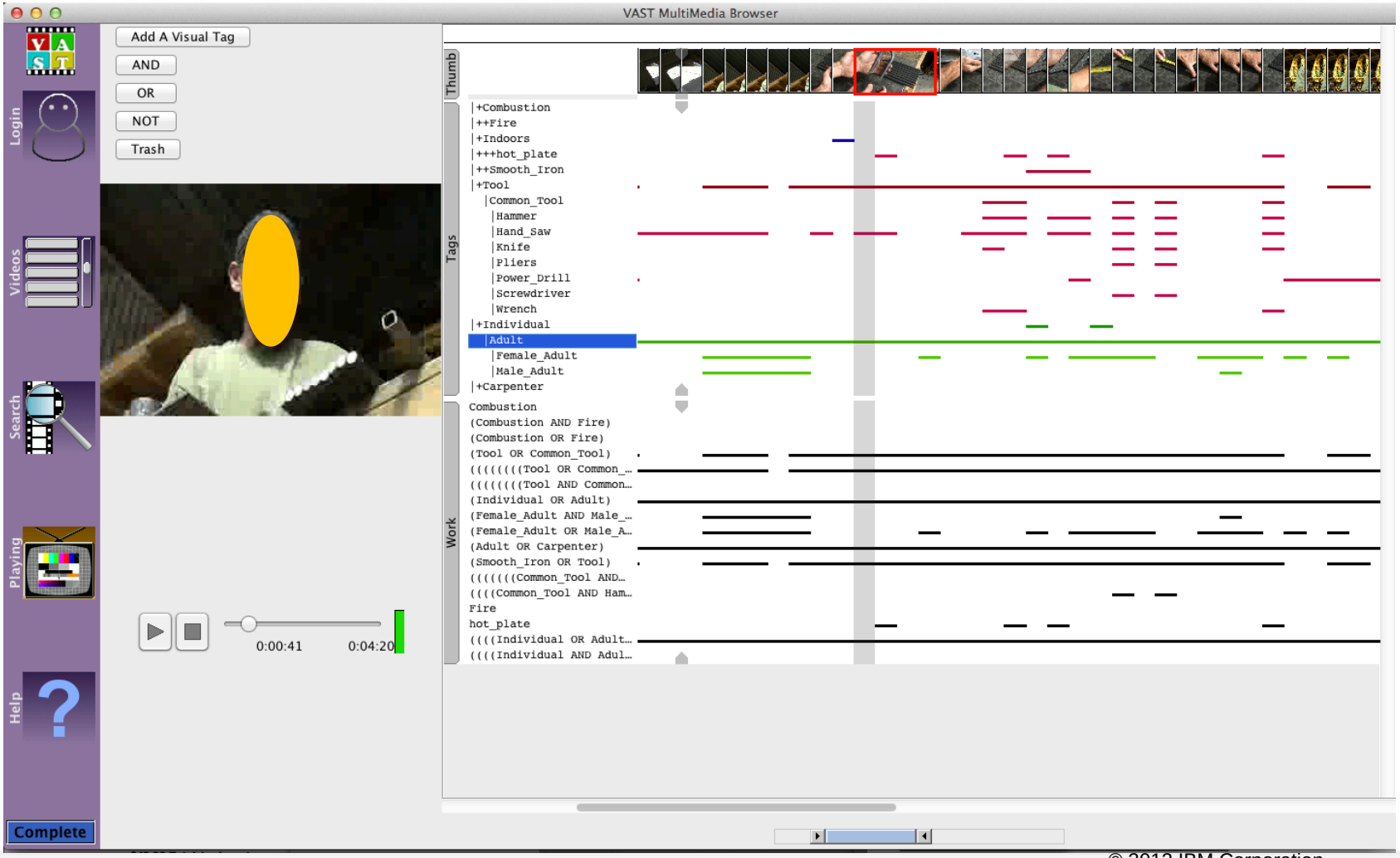
# Results and waypoint experiments

- New PhD, MS, and BS student user studies find:
  - Need better actions
  - Need “attributes”, especially for Agents: clothing → role
  - Text (ASR, OCR) is critical for clip identification
  - System output of 1 sentence/segment improvable by:
    - Narrative (Introduction, Development, Conclusion)
    - Anaphora (pronouns) and ellipsis (deletion of repetition)
    - Compaction of “or” conjunctions
  - But, much is lost when output is in natural language
    - Better: Visualization via thumbnails plus semantic timelines
    - Better: Output checklist, to compare to Event Kit input checklist

# Results and waypoint experiments



# Results and waypoint experiments



# Summary of technical readiness

- Generally, at about DoD TRL 4
  - "Component and/or breadboard validation in lab"
- Needs better "glue"
  - Between classifier score output and Matlab input
  - Between event kit text and importance
  - Among Visual, Sound, and Action concepts
- Many parameters can be more carefully tuned
  - Classifier score thresholds; may be concept-specific
  - Thresholds for video semantic segmentation
  - Semantic generalization (solve child/parent inversions)

# Next steps

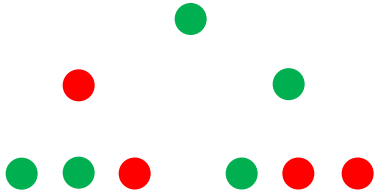
- User studies in progress on impact and output of:
  - Camera motion
    - PTZ from optic flow; zoom-in as a cue to significance
    - How to visualize videography?
  - Named entities
    - Particularly for ASR/OCR words, but more for “how-to” events
  - Attributes
    - Particularly for departures from a “standard model” of humans
  - Sounds (in isolation)
    - How accurate is “radio understanding”?
    - ASR/OCR fails on non-English

# Next steps

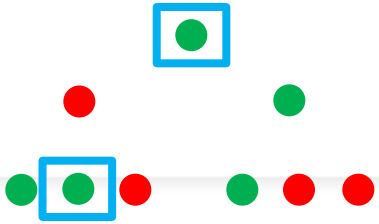
- More realistic Natural Language Generation
  - Extension to other thematic roles:
    - Direction *in the scene*
    - Time *in the scene*
  - Facets for “material roles
    - *Military* plane
    - *Football* field
  - Facets for attributes (adjectives and adverbs):
    - Color, number (absolute)
    - Size, manner (relative to a standard)

# Next steps

- More focused use of ontology tree
  - Resolution of subtree specificity, even if no NLG



    - "Tool", not "hammer or saw"; but "two", not "one or two"
  - Algorithm in development and in user studies:
    - Convert SVM scores to probabilities
    - Form local "dominations" between parents and children
    - Filter this "domination graph" for "sources":



# Next steps

- Visualizations of semantics in the video
  - As research tool:
    - Semantic timelines > media player
    - Gives feedback on classifier creation, training, reorganization
  - And possibly better than NLG, especially if interactive
  - Still, questions of scale:
    - Selecting a video from a collection
    - Contrasting event-specific interpretations of a single video
    - Controlling ratio of explication to video length or complexity
    - Selecting/controlling sampling rate of thumbnails
    - Selecting/controlling semantic resolution of classifiers

# Outline

- Motivation and problem definition
- Summary of related work
- Proposed novel approach
- Results and waypoint experiments
- Summary of technical readiness
- Next steps