Information Ordering

LING 573 — Systems & Applications

Ryan Georgi – 4/19/2018

Begin Recording!

Issues With Information Ordering

- Coherence & Cohesion
- Multiple document context:
  - Different documents contain superset/subset of information
  - Different documents express same topic with different wording
  - Different documents express same topics in different order

Information Ordering: Approaches

- “Dumb” Document Order
- Order based upon shared cluster information
  - Based on order presented within document cluster
  - Based on order of topic’s first mention
- Coherence

Information Ordering: Approaches

- "Dumb" Document Order
- **Order based upon shared cluster information**
  - **Based on order presented within document cluster**
  - **Based on order of topic's first mention**
- Coherence

Document Cluster Orderings

(Barzilay et al. 2002)

- Build on their existing system (MULTIGEN) (Barzilay et. al, 1999)
  - Ordering approach upon language generation systems at the time
    - Generation using knowledge representations
  - Information ordering as "Content Planning"
    - Sentences drawn from knowledge base propositions
    - Propositions ordered logically
    - Propositions realized as sentences

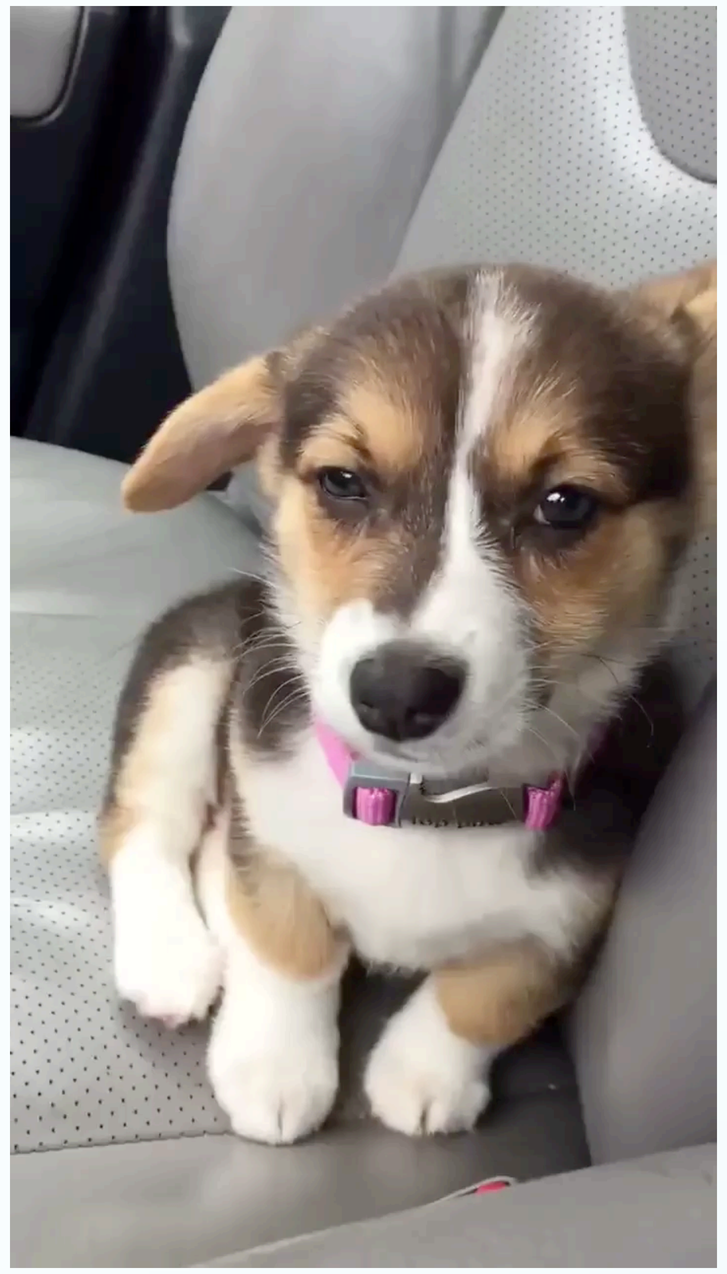
Document Cluster Orderings

(Barzilay et al. 2002)

- DUC/TAC Approach
  - Knowledge base is "implied" within the corpus
  - How to determine what is logical proposition order from this "knowledge base"?
- Analysis groups sentences into "themes"
  - Text units from different documents with repeated information
  - Roughly clusters groups of sentences with similar content
  - Intersection of their information is summarized
- Ordering is done on this selected content

Document Cluster Orderings

(Barzilay et al. 2002)



- “Theme” clusters
  - Sets of sentences from different documents containing repeated information
  - Do not necessarily contain sentences from all documents
  - Use SIMFINDER (Hatzivassiloglou et. al 2001)

Theme 1	Mr. Salvi, 24, apparently killed himself in his prison cell last November. The state wouldn’t execute him for killing two abortion clinic workers in 1994, so John C. Salvi III took his own life. John C. Salvi III, who was convicted of killing two people in a shooting spree on two abortion clinics in 1994, killed himself in prison.
Theme 2	His attorneys said he attempted suicide twice before in prison. His lawyers said that he twice had tried to commit suicide in jail, a charge authorities have denied.

Document Cluster Orderings

(Barzilay et al. 2002)

- Two basic strategies explored:
  - Chronological Ordering (**CO**)
  - Majority Ordering (**MO**)

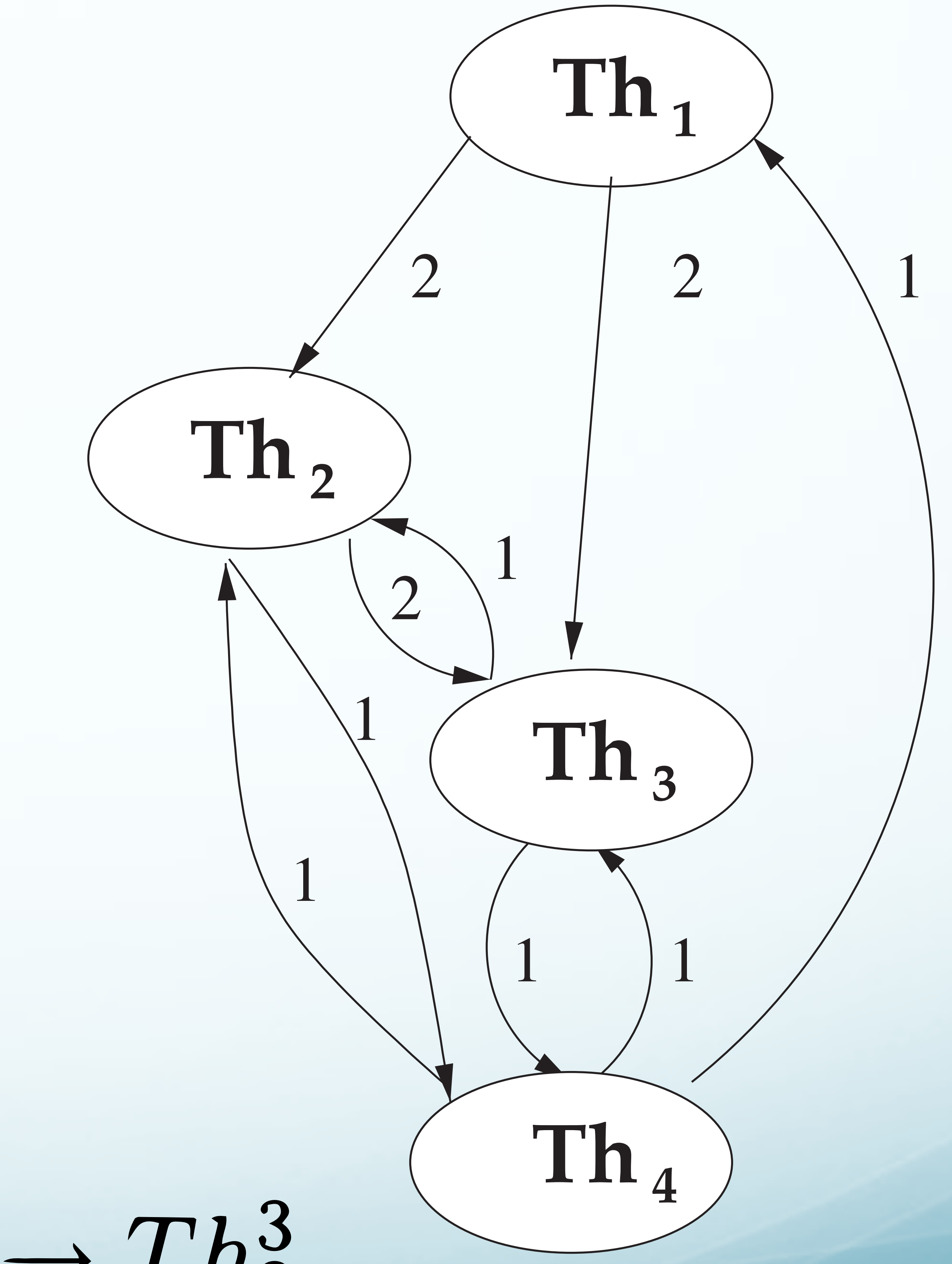
Majority Ordering

- Across all documents, find ordering for each pairwise set of themes
  - i.e. Th_1 precedes Th_2
- Use documents to "vote" on order of themes.
  - If Th_1 comes before Th_2 in 60% of documents
  - put Th_1 in summary first.
- Note! — Pairwise sorting not necessarily transitive
  - $Th_1 < Th_2 < Th_3$ in doc1
  - $Th_3 < Th_1 < Th_2$ in doc2

Majority Ordering

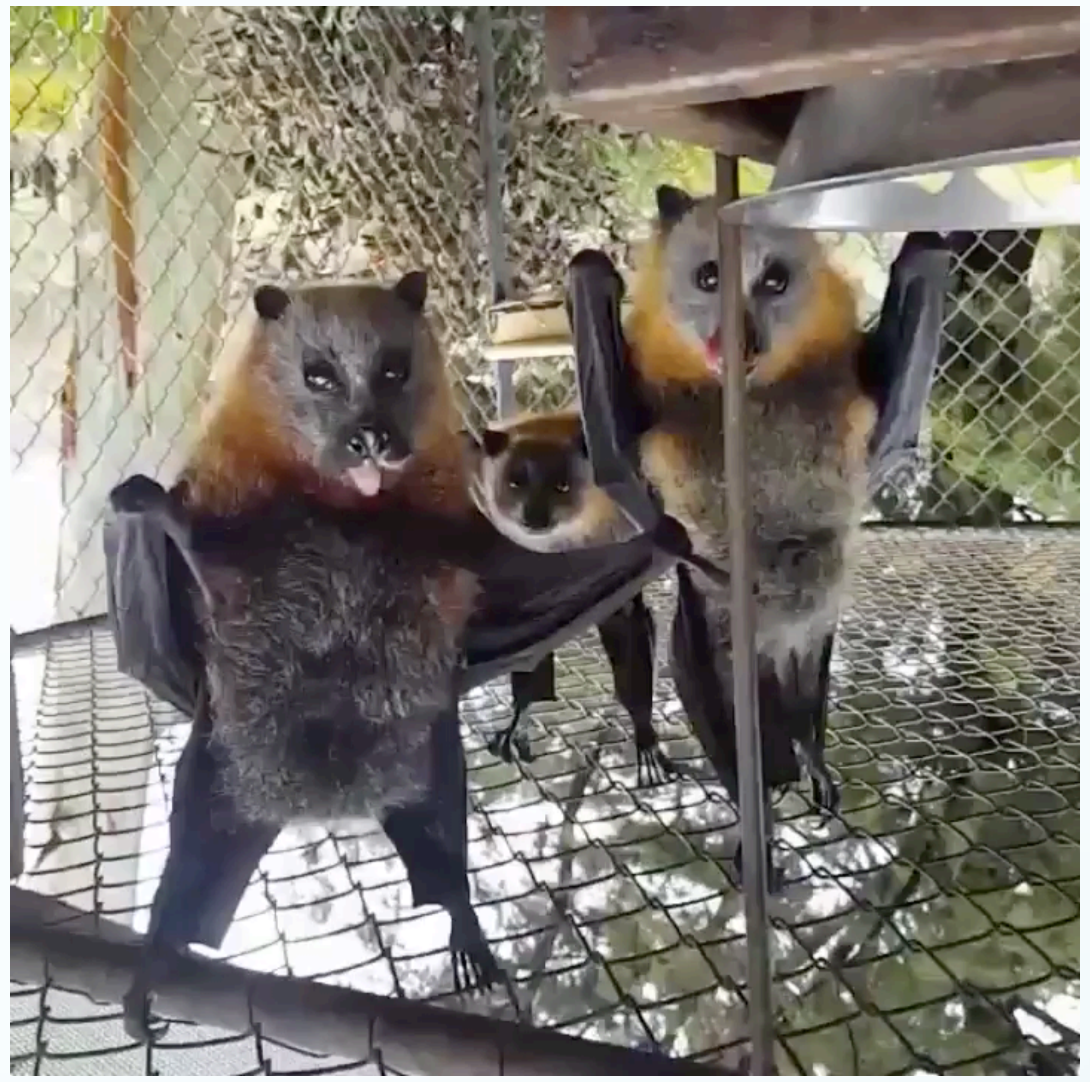
- Solution? — follow (Cohen et. al, 1999)
  - Represent themes as DAG:
    - Nodes are themes
    - $Weight_{Node}$ is (Σ $Weight_{Outgoing\ Edge}$) – (Σ $Weight_{Incoming\ Edge}$)
    - $Weight_{Edge}$ is # of documents with that ordering
  - Use topological sort to find approximate solution
  - Pop maximum weighted node
  - Iterate until graph is empty

$$\begin{array}{l} Th_1^1 \rightarrow Th_2^1 \rightarrow Th_3^1 \\ Th_3^2 \rightarrow Th_2^2 \rightarrow Th_4^2 \\ Th_4^3 \rightarrow Th_1^3 \rightarrow Th_2^3 \rightarrow Th_3^3 \end{array}$$



Chronological Ordering (CO)

- Look at document publication times
- Assign a date to each **theme** based on first publication date
  - If dame dates, use MO
- Intuition — early reports center on initial events
  - Less on reactions/results



Oct 5, 11:35am	Hours after the crash, U.S. officials said that the tragedy had been caused by an S-200 missile fired by Ukraine during military exercises on the Crimean Peninsula.
Oct 6, 6:13am	U.S. officials said immediately after the crash that they had evidence the passenger jet was hit by a Ukrainian missile.
Oct 5, 10:20am	But U.S. officials said that the crash had been caused by an S-200 missile fired mistakenly by Ukrainian forces during military exercises on the Crimean Peninsula.

Evaluation: CO vs. MO

- For 25 summaries, with human evaluations, neither system fares particularly well:

	Poor	Fair	Good
MO	3	14	8
CO	10	8	7

- MO works when presentation order consistent
  - When inconsistent, produces own brand new order
- CO problematic on:
  - Themes that aren't tied to document order
    - e.g. quotes about reactions to events
- Multiple topics not constrained by chronology

Evaluation: CO vs. MO

- Test data is still available!
  - http://www.cs.columbia.edu/~noemie/ordering/
  - Largely superseded by DUC/TAC MDS data
    - but may still be useful.

Improving on MO/CO:

(Barzilay et. al, 2008)

- Error analysis — experiments on sentence ordering by subjects
  - Many possible orderings but far from random
    - Blocks of sentences group together (cohere) ← *Aha!*
- *Add cohesion!*
- Main insight:
  - Maintain continuity — maintain references to same entity in adjacent sentences.
- Themes "related" if, when two themes appear in same text, frequently appear in same segment (threshold)

Improving on MO/CO:

(Barzilay et. al, 2008)

- Order over ***groups of themes*** by CO
  - Then order within groups by CO
- Significantly better!

Before and After

1. Thousands of people have attended a ceremony in Nairobi commemorating the fir bombing attacks against u.s. Embassies in Kenya and Tanzania.
2. Saudi dissident Osama bin Laden, accused of masterminding the attacks, and nine o
3. President Clinton said, "The intended victims of this vicious crime stood for everyt country and the world."
4. U.S. federal prosecutors have charged 17 people in the bombings.
5. Albright said that the mourning continues.
6. Kenyans are observing a national day of mourning in honor of the 215 people who

1. Thousands of people have attended a ceremony in Nairobi commemorating the fir bombing attacks against u.s. Embassies in Kenya and Tanzania. Kenyans are observin mourning in honor of the 215 people who died there.
2. Saudi dissident Osama bin Laden, accused of masterminding the attacks, and nine o federal prosecutors have charged 17 people in the bombing.
3. President Clinton said, "The intended victims of this vicious crime stood for everyt country and the world." Albright said that the mourning continues.

Annotating Data!

Annotating Data

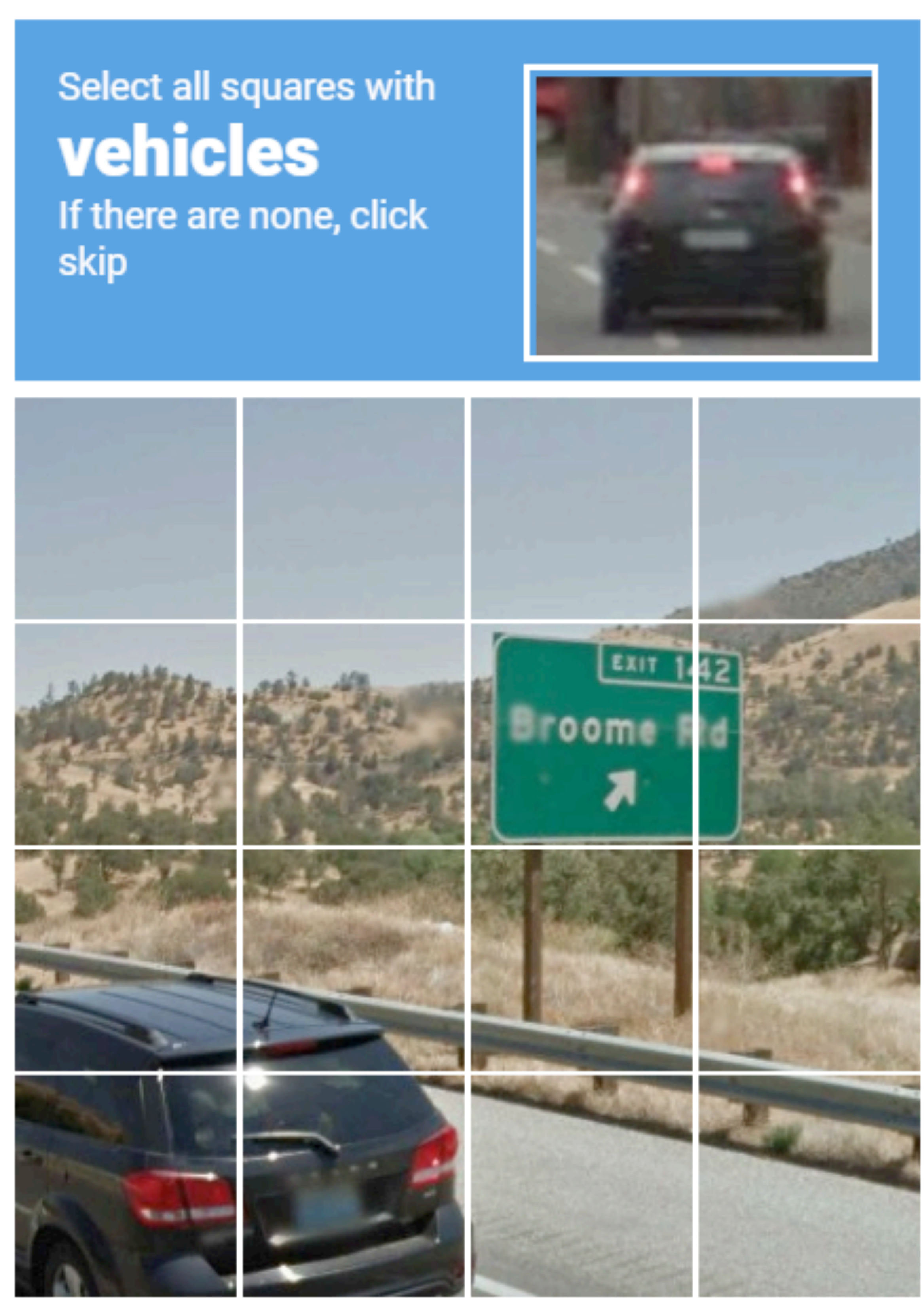
- Recurring problem in ML
  - What data is available?
  - If no data is available, how can it be created?
  - Humans are expensive
  - If we MUST use humans, is there any way to make the task as efficient as possible?

Humans Are Expensive

- **Solution #1 — Pay them less.**
  - This is the *raison d'être* for Mechanical Turk
  - Williamson, V. (2016) "Can crowdsourcing be ethical?" The Brookings Institution.
    - "In the course of my graduate work at Harvard University, I paid hundreds of Americans living in poverty the equivalent of about $2 an hour. It was perfectly legal for me to do so, and my research had the approval of my university's ethics board."

Humans Are Expensive

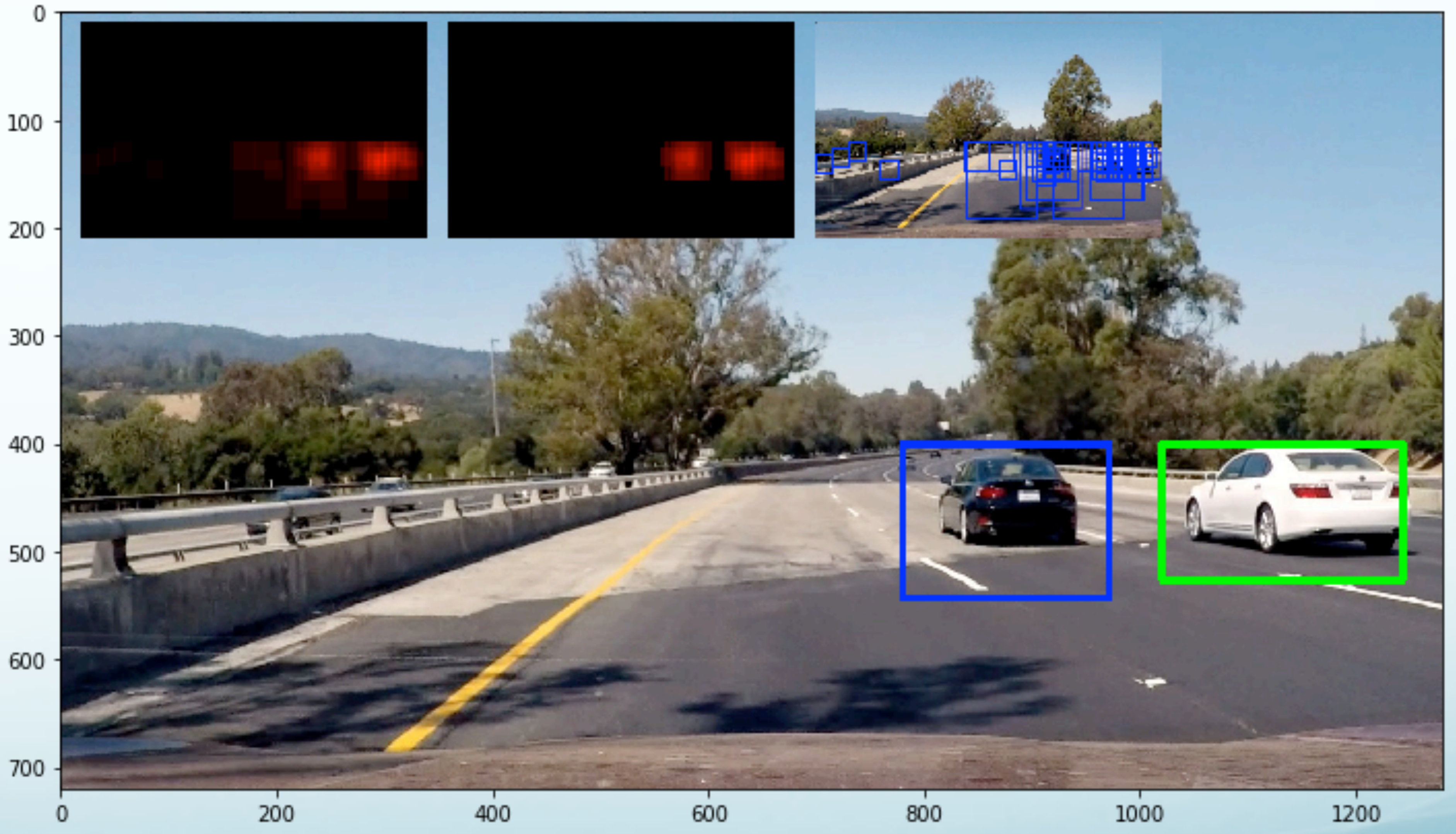
- **Solution #1a — Pay them nothing.**
  - Recaptcha — crowdsources HITs for bot reduction



Select all squares with

vehicles

If there are none, click skip



Humans Are Expensive

- **Solution #2 — Make them more Efficient**
  - Come up with simple, easy-to-interpret guidelines
    - Constrain the problem space as much as possible
    - e.g. define "Noun" very clearly
  - Increase ease of use
    - PLEASE, GOD, NO RAW XML EDITING!
  - Make task easily repeatable
    - Put it inside some kind of UI or other harness

Annotating Sentences for Extraction

- CNN/dm corpora — `/dropbox/17-18/573/other_resources/cnn-dm`
  - These have “highlights” — abstractive summary sentences
  - Cheng & Lapata mention an ngram-overlap rule-based system, 85% accurate
    - Code is no longer available.
    - 85% isn’t particularly great for this task
  - Can we do better?

Annotating Sentences for Extraction

- I have written an interactive script:
  - `/dropbox/17-18/573/code/cnn-dm-tools/highlight-selector.py`
  - Pulls 10 random articles from the CNN portion of CNN-DM
    - (Seeded by your username)
  - Splits document sentences and highlights
  - Calculates simple cosine overlap between sentences, ranks sentences
    - Presents highlight and n-best list of ranked sentences for interactive prompt
  - Allows user to choose one, multiple, or no sentences that "match" the given highlight
  - Writes users choices out to sidecar annotation file

Annotating Sentences for Extraction

- If you are all willing… we have nearly 40 students enrolled…
  - If we all take ~5-10 minutes a piece to do this task…
  - We should be able to come up with ~400 unordered extractive summaries!
  - Can be additional training data for any group wanting to do supervised classification approach

Running the Script

- /dropbox/17-18/573/code/cnn-dm-tools/highlight-selector.py

- May need to run:

- python
- >>> import nltk
- >>> nltk.download([‘punkt','brown'])

- (will download to your user directory)