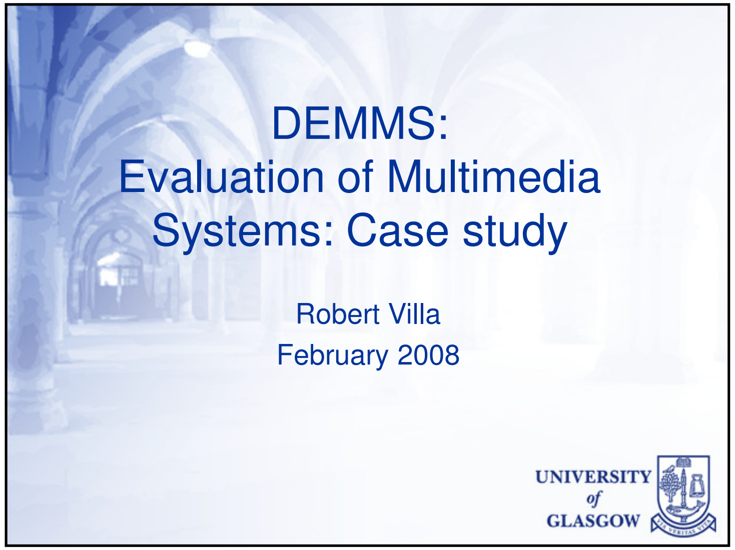
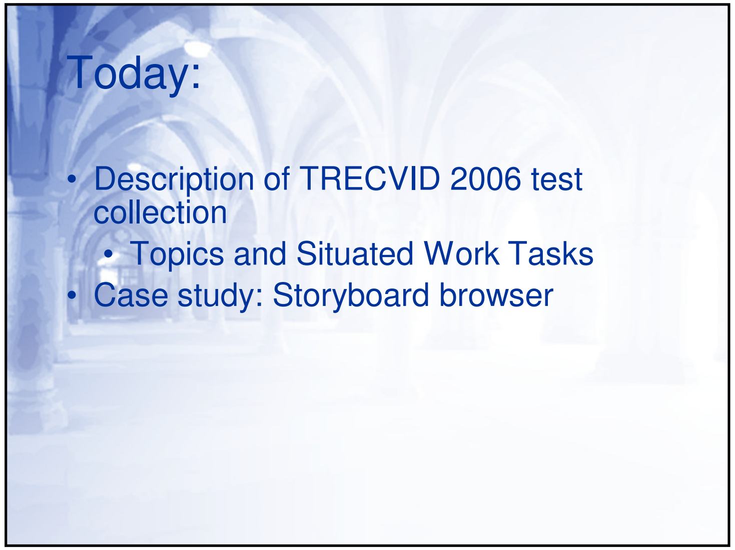


DEMMS: Evaluation of Multimedia Systems: Case study

Robert Villa

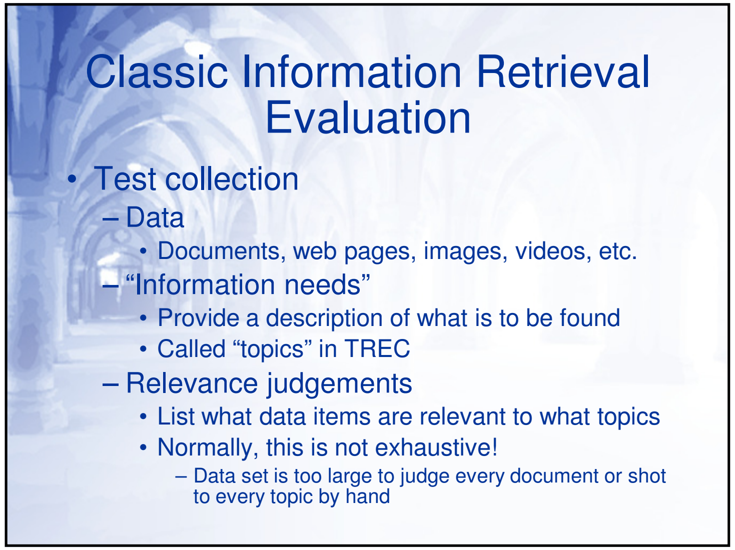
February 2008

UNIVERSITY *of* GLASGOW



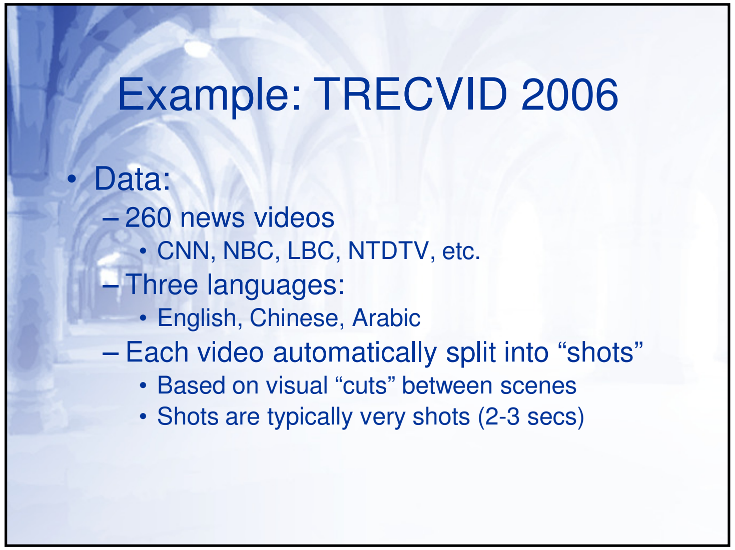
Today:

- Description of TRECVID 2006 test collection
  - Topics and Situated Work Tasks
- Case study: Storyboard browser



Classic Information Retrieval Evaluation

- Test collection
  - Data
    - Documents, web pages, images, videos, etc.
  - "Information needs"
    - Provide a description of what is to be found
    - Called "topics" in TREC
  - Relevance judgements
    - List what data items are relevant to what topics
    - Normally, this is not exhaustive!
      - Data set is too large to judge every document or shot to every topic by hand



Example: TRECVID 2006

- Data:
  - 260 news videos
    - CNN, NBC, LBC, NTDTV, etc.
  - Three languages:
    - English, Chinese, Arabic
  - Each video automatically split into "shots"
    - Based on visual "cuts" between scenes
    - Shots are typically very shots (2-3 secs)

Example of a shot

- Every shot has an associated text transcript:
  - E.g. “A dramatic arrival”
- Generated by Automatic Speech Recognition (ASR)

Information needs

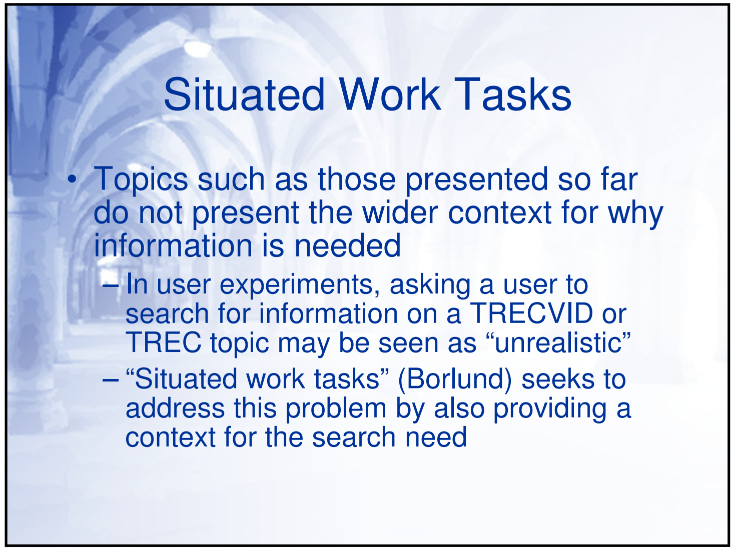
- Topic:
  - A statement of an information need
  - E.g. “Find out about George Bush’s youth”
- Query:
  - The statement sent to the IR system
  - E.g. “young George Bush”
- There are many different forms of “topic”

TRECVID 2006 Topic

- <videoTopic num="0173">
  - <textDescription text="***Finds shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.)***"/>
  - <videoExample src="20041124_110000_MSNBC_MSNBCNEWS11_ENG.mpg" start="04m04.177s" stop="04m06.179s" desc="tracks front of state police car up close"/>
  - <videoExample src="20041105_140000_LBC_LBCNAHAR_ARB.mpg" start="03m00.313s" stop="03m03.817s" desc="driving into gate"/>
- </videoTopic>

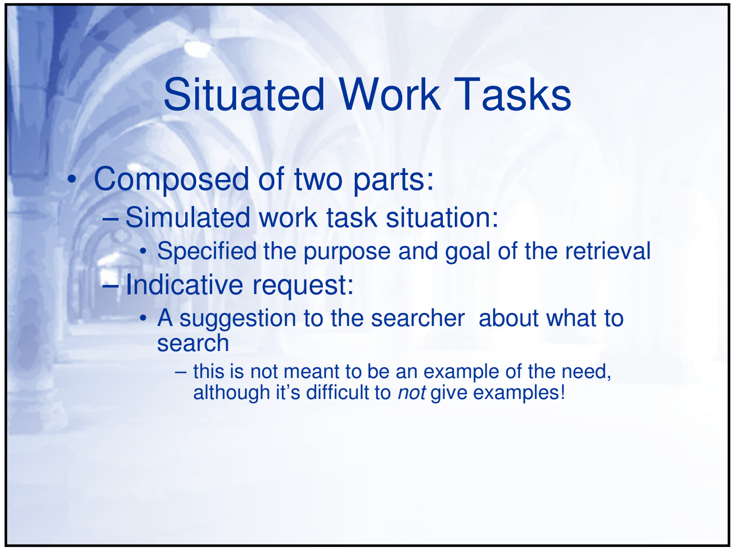
TREC Topic

- <num>
  - Number: 501
- <title>
  - deduction and induction in English?
- <desc>
  - Description: What is the difference between deduction and induction in the process of reasoning?
- <narr> Narrative:
  - A relevant document will contrast inductive and deductive reasoning. A document that discusses only one or the other is not relevant.



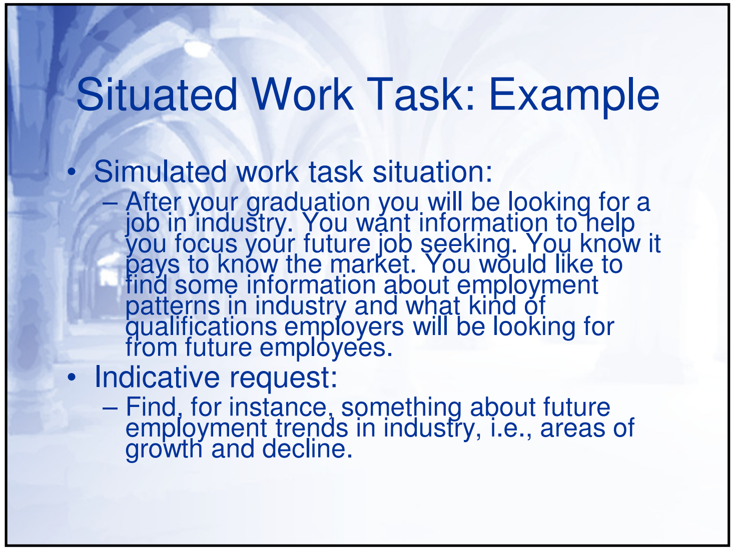
Situated Work Tasks

- Topics such as those presented so far do not present the wider context for why information is needed
  - In user experiments, asking a user to search for information on a TRECVID or TREC topic may be seen as “unrealistic”
  - “Situated work tasks” (Borlund) seeks to address this problem by also providing a context for the search need



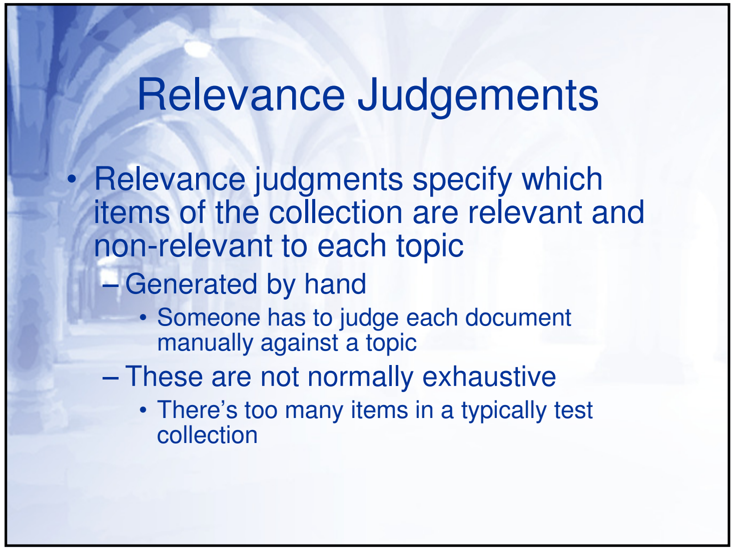
Situated Work Tasks

- Composed of two parts:
  - Simulated work task situation:
    - Specified the purpose and goal of the retrieval
  - Indicative request:
    - A suggestion to the searcher about what to search
      - this is not meant to be an example of the need, although it’s difficult to *not* give examples!



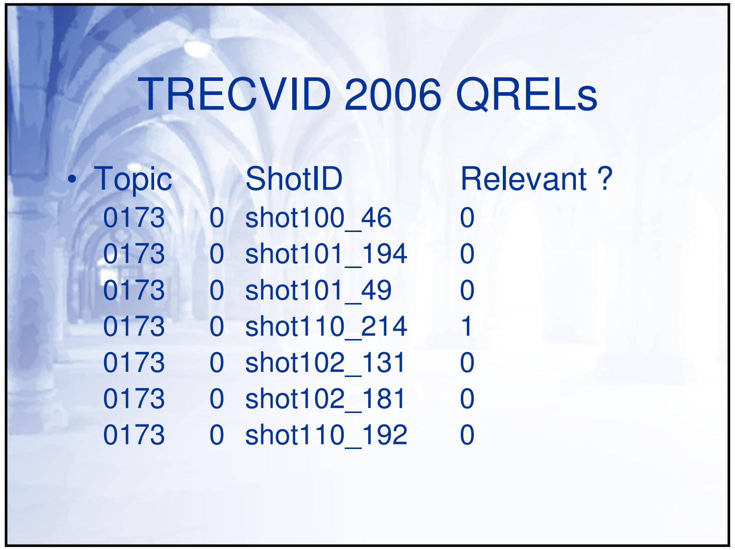
Situated Work Task: Example

- Simulated work task situation:
  - After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.
- Indicative request:
  - Find, for instance, something about future employment trends in industry, i.e., areas of growth and decline.



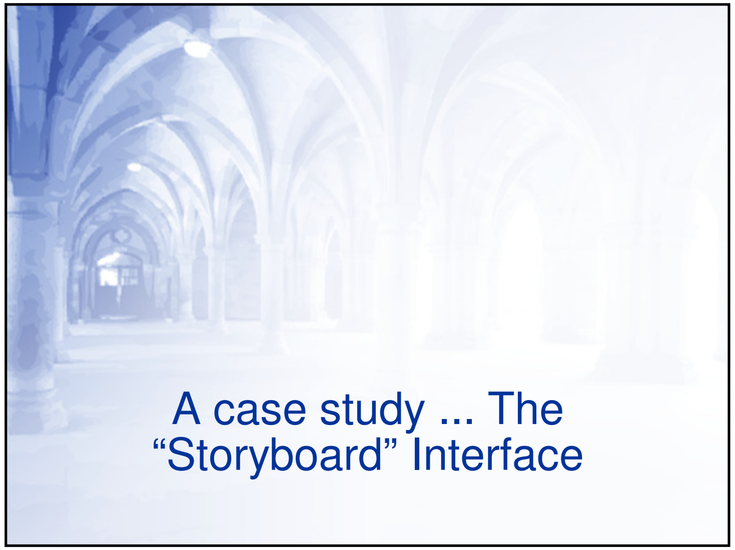
Relevance Judgements

- Relevance judgments specify which items of the collection are relevant and non-relevant to each topic
  - Generated by hand
    - Someone has to judge each document manually against a topic
  - These are not normally exhaustive
    - There’s too many items in a typically test collection

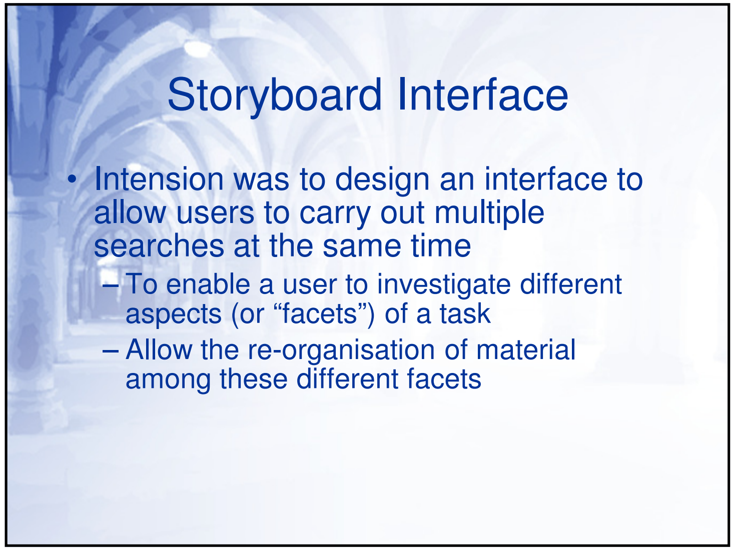


TRECVID 2006 QRELs

• Topic		ShotID	Relevant ?
0173	0	shot100_46	0
0173	0	shot101_194	0
0173	0	shot101_49	0
0173	0	shot110_214	1
0173	0	shot102_131	0
0173	0	shot102_181	0
0173	0	shot110_192	0



A case study ... The “Storyboard” Interface

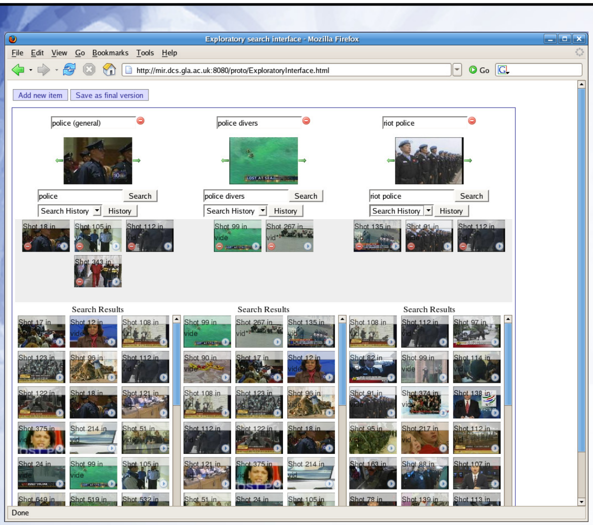


Storyboard Interface

- Intension was to design an interface to allow users to carry out multiple searches at the same time
  - To enable a user to investigate different aspects (or “facets”) of a task
  - Allow the re-organisation of material among these different facets

The Storyboard Interface

- Designed using a storyboard metaphor
  - User can have multiple searches on the go at the same time
  - User can drag and drop between the different searches
- Available at:
  - http://mir.dcs.gla.ac.uk:8080/demo/ExploratoryInterface.html



Exploratory search interface - Mozilla Firefox

http://mir.dcs.gla.ac.uk:8080/proto/ExploratoryInterface.html

police (general)

police divers

riot police

police

police divers

riot police

Done

Initial research question?

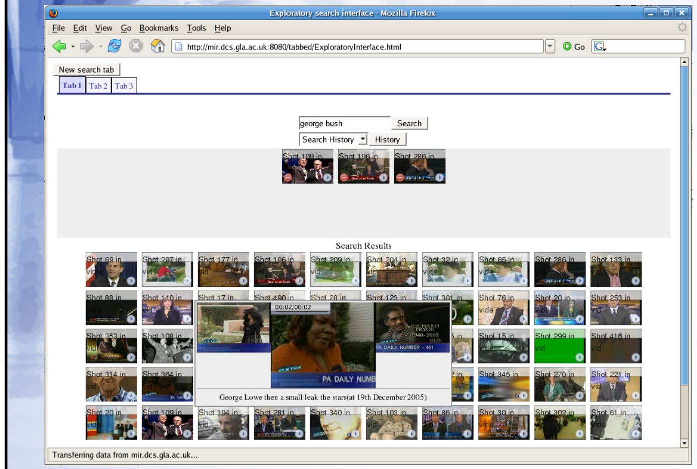
- does the storyboard interface enable the user to better explore a complex search task?

Initial research question?

- does the storyboard interface enable the user to better explore a complex search task?
  - Question: Explore a task “better” relative to what?
  - We need a baseline system to compare to

The baseline

- Baseline system
  - Intended to reflect current practice
  - Based on a “tabbed” model as used in many current web browsers (e.g. Firefox)



Exploratory search interface - Mozilla Firefox

http://mir.dcs.gla.ac.uk:8080/tabbed/ExploratoryInterface.html

george bush

George Lowe then a small leak the stars(at 19th December 2005)

Transferring data from mir.dcs.gla.ac.uk...

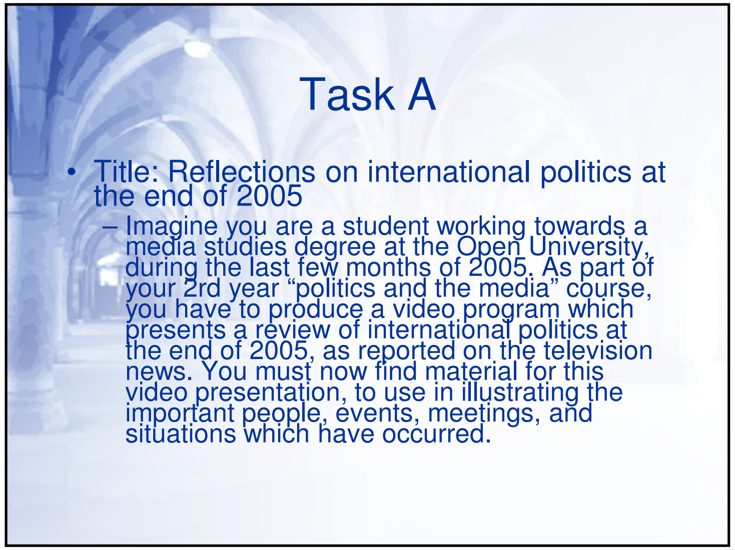
Next question: what are the "complex" search tasks?

Tasks

- Two search tasks were defined, both situated work tasks
  - Why only 2 tasks?
    - We wanted give users time to explore the tasks
    - This meant more time per task (30 minutes) ...
    - ... which means less tasks
  - 2 search tasks at 30 minutes each equated to a study lasting 2 hours
    - This is a long time for an experiment
    - Much of the extra time spent on training

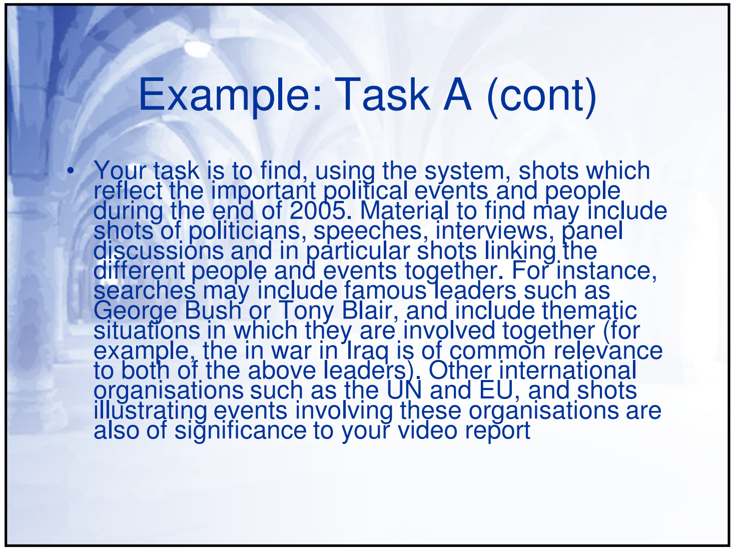
The Tasks

- Task A:
  - Reflections on international politics at the end of 2005
- Task B:
  - A Summary of the trial of Saddam Hussein



Task A

- Title: Reflections on international politics at the end of 2005
  - Imagine you are a student working towards a media studies degree at the Open University, during the last few months of 2005. As part of your 2rd year "politics and the media" course, you have to produce a video program which presents a review of international politics at the end of 2005, as reported on the television news. You must now find material for this video presentation, to use in illustrating the important people, events, meetings, and situations which have occurred.



Example: Task A (cont)

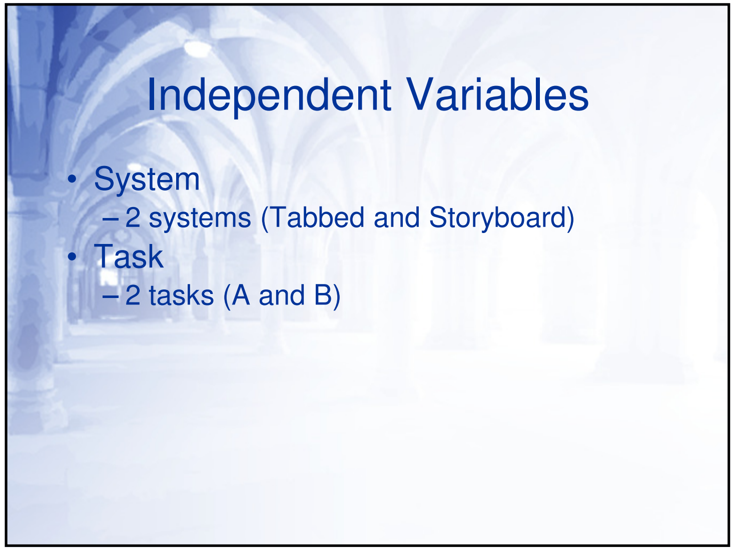
- Your task is to find, using the system, shots which reflect the important political events and people during the end of 2005. Material to find may include shots of politicians, speeches, interviews, panel discussions and in particular shots linking the different people and events together. For instance, searches may include famous leaders such as George Bush or Tony Blair, and include thematic situations in which they are involved together (for example, the in war in Iraq is of common relevance to both of the above leaders). Other international organisations such as the UN and EU, and shots illustrating events involving these organisations are also of significance to your video report

Experimental design

- Within subject design
  - Each user used both systems

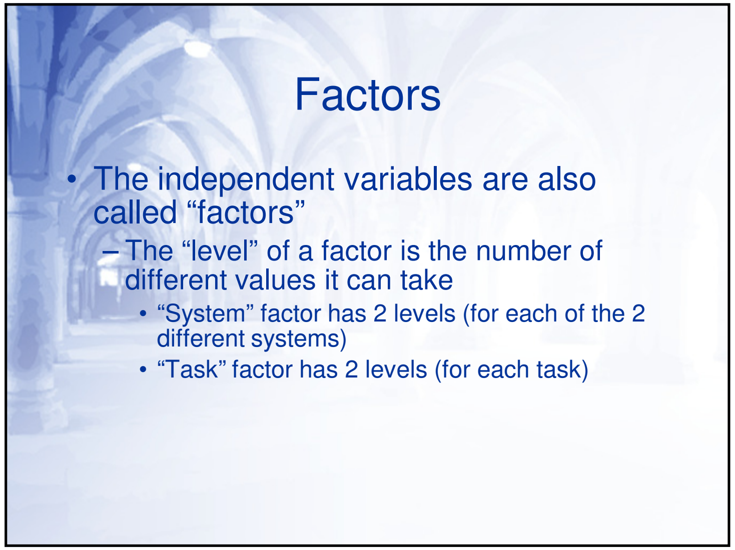
Dependent Variable

- Degree of exploration in the video collection
  - With four measures:
    - Number of searches carried out
    - Number of panels or tabs opened over the course of the task
    - Number of shots marked relevant
    - Number of shots played



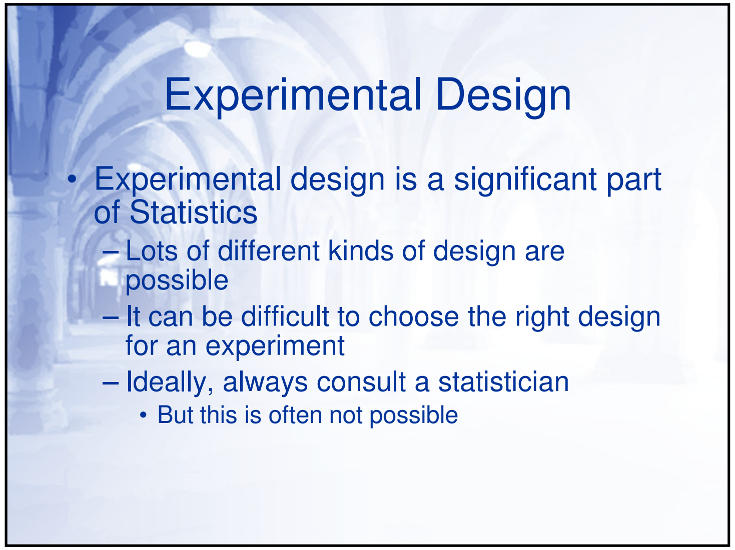
Independent Variables

- System
  - 2 systems (Tabbed and Storyboard)
- Task
  - 2 tasks (A and B)



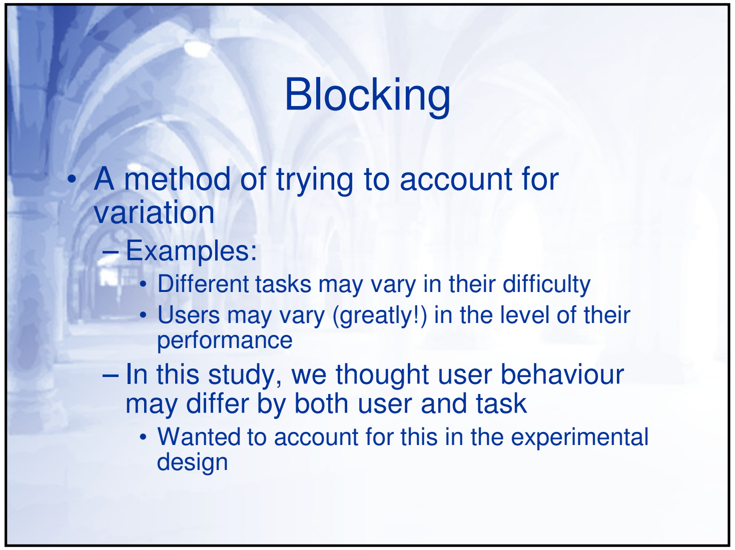
Factors

- The independent variables are also called “factors”
  - The “level” of a factor is the number of different values it can take
    - “System” factor has 2 levels (for each of the 2 different systems)
    - “Task” factor has 2 levels (for each task)



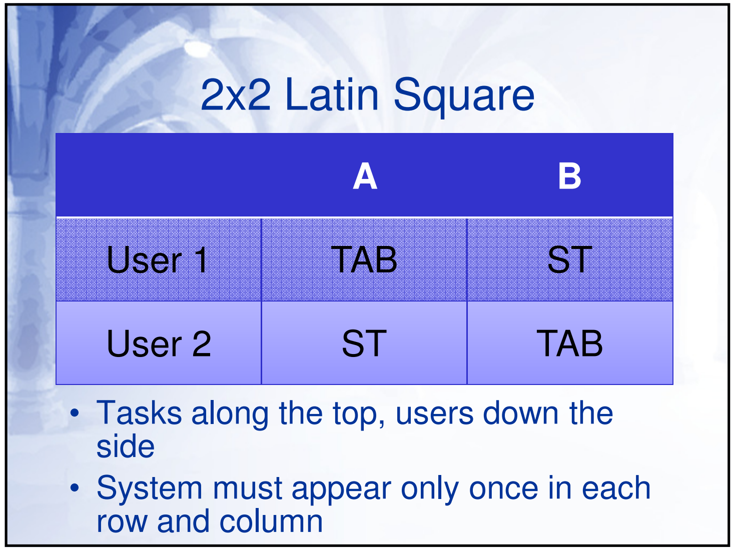
Experimental Design

- Experimental design is a significant part of Statistics
  - Lots of different kinds of design are possible
  - It can be difficult to choose the right design for an experiment
  - Ideally, always consult a statistician
    - But this is often not possible



Blocking

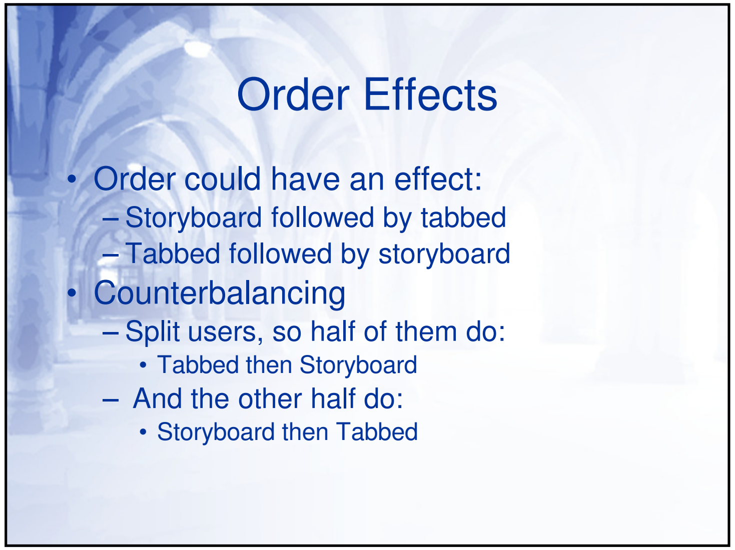
- A method of trying to account for variation
  - Examples:
    - Different tasks may vary in their difficulty
    - Users may vary (greatly!) in the level of their performance
  - In this study, we thought user behaviour may differ by both user and task
    - Wanted to account for this in the experimental design



2x2 Latin Square

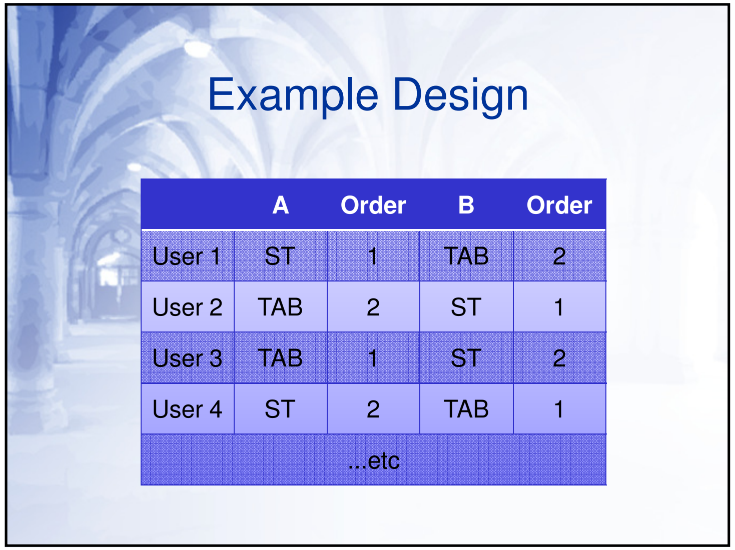
	A	B
User 1	TAB	ST
User 2	ST	TAB

- Tasks along the top, users down the side
- System must appear only once in each row and column



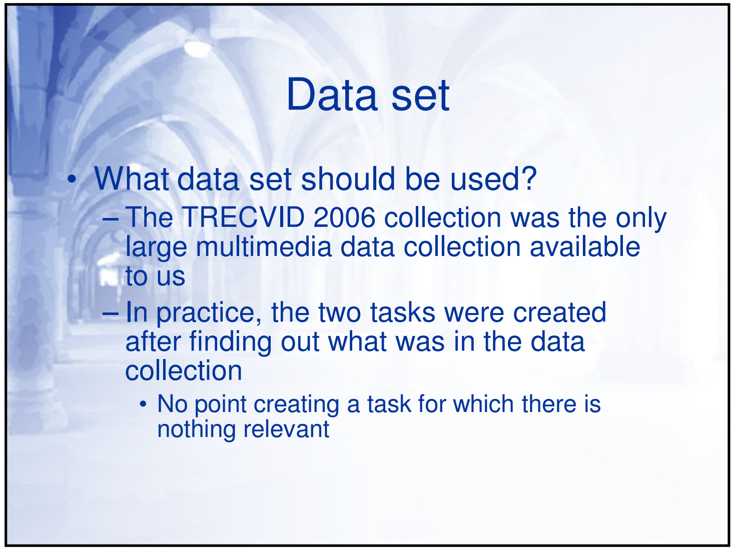
Order Effects

- Order could have an effect:
  - Storyboard followed by tabbed
  - Tabbed followed by storyboard
- Counterbalancing
  - Split users, so half of them do:
    - Tabbed then Storyboard
  - And the other half do:
    - Storyboard then Tabbed



Example Design

	A	Order	B	Order
User 1	ST	1	TAB	2
User 2	TAB	2	ST	1
User 3	TAB	1	ST	2
User 4	ST	2	TAB	1
...etc				



Data set

- What data set should be used?
  - The TRECVID 2006 collection was the only large multimedia data collection available to us
  - In practice, the two tasks were created after finding out what was in the data collection
    - No point creating a task for which there is nothing relevant

Procedure

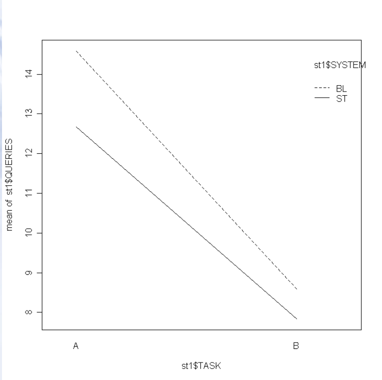
- Entry questionnaire + consent form
- Training of first system
  - Pre-created tutorial
  - Allow user to search with system
- Present first task and start
- Post-search questionnaire
- *Repeat for second task and system*
- Exit questionnaire + payment

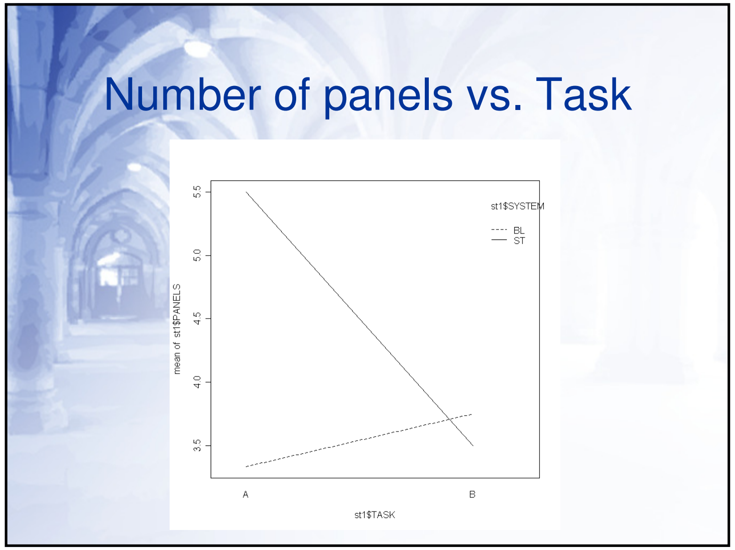
Users

- 24 users were recruited
  - Mostly students (not a surprise!)
- All were paid 20 pounds
- It took roughly 2 hours per user
  - Need to be careful with timing
  - Things always take longer than you expect

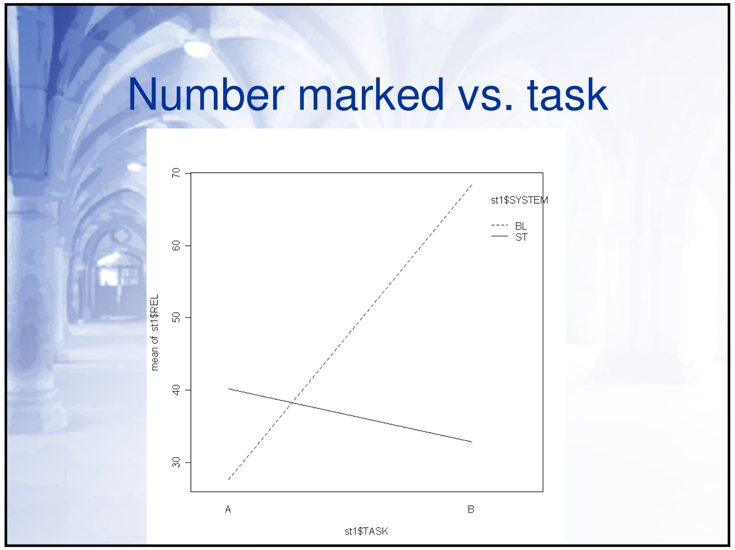
Some results

Number of queries vs. task

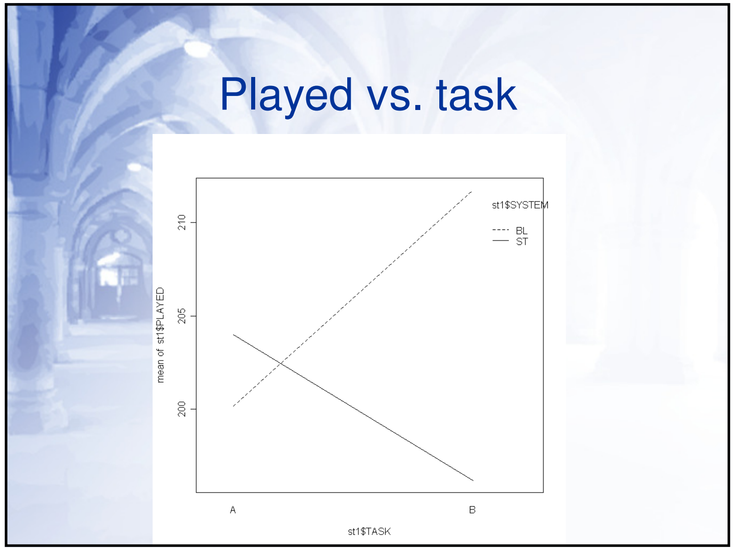




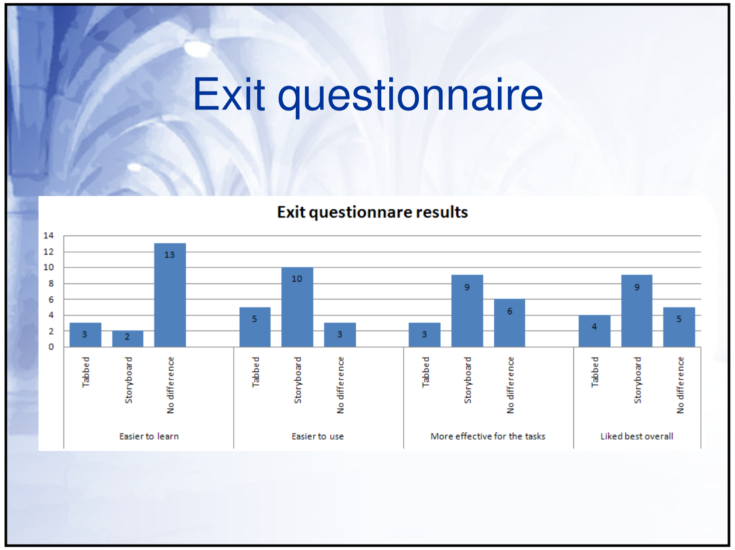
Number of panels vs. Task



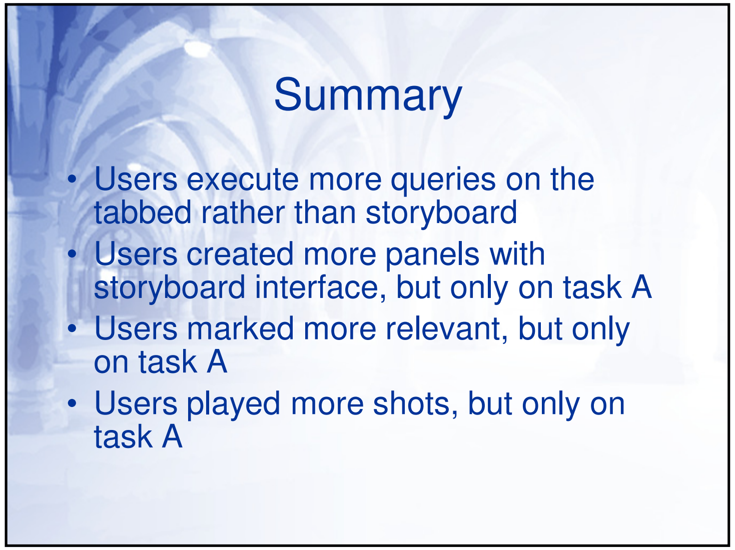
Number marked vs. task



Played vs. task

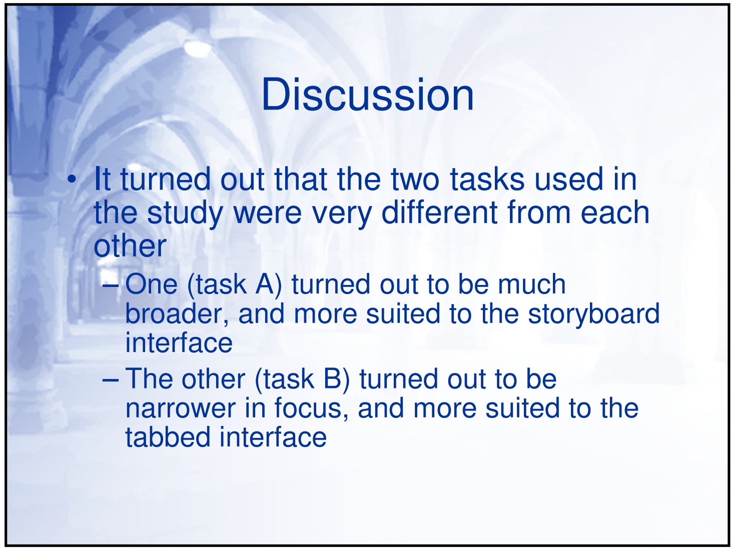


Exit questionnaire



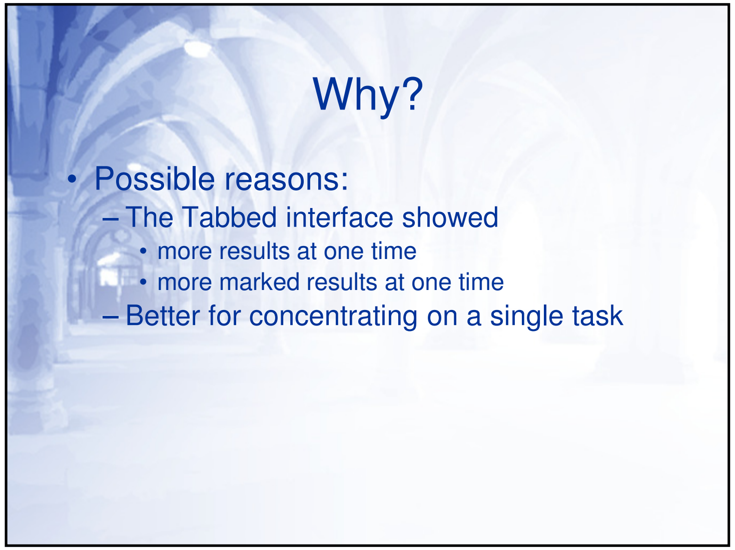
Summary

- Users execute more queries on the tabbed rather than storyboard
- Users created more panels with storyboard interface, but only on task A
- Users marked more relevant, but only on task A
- Users played more shots, but only on task A



Discussion

- It turned out that the two tasks used in the study were very different from each other
  - One (task A) turned out to be much broader, and more suited to the storyboard interface
  - The other (task B) turned out to be narrower in focus, and more suited to the tabbed interface



Why?

- Possible reasons:
  - The Tabbed interface showed
    - more results at one time
    - more marked results at one time
  - Better for concentrating on a single task

End