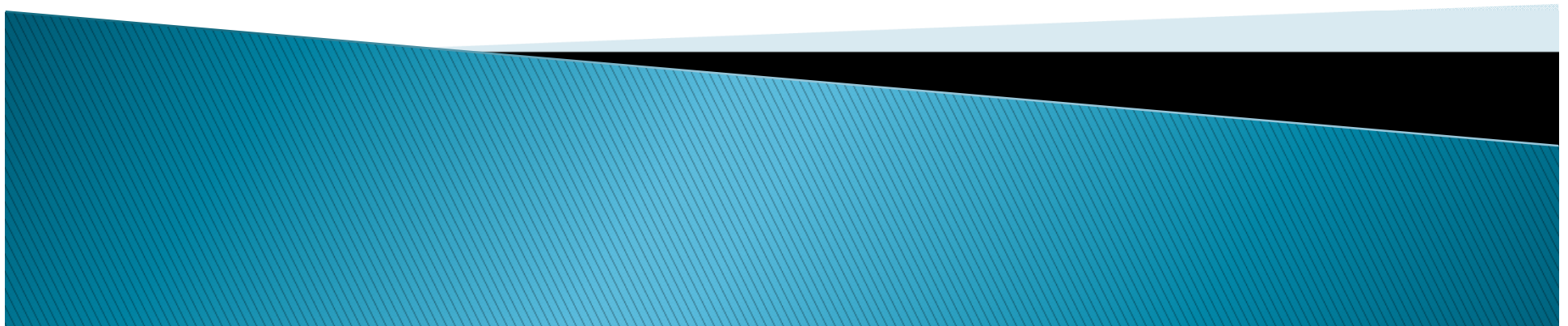# TREC Session Track

Ben Carterette
Paul Clough
Evangelos Kanoulas
Mark Sanderson

Vision

Vision

Cranfield Paradigm
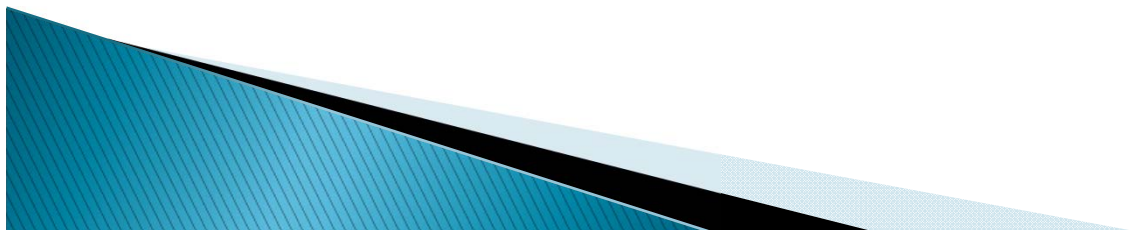- Simple user model
- Controlled experiments
- Reusable but static test collections

Session Track

Interactive Evaluation/ Live Labs
- Full user participation
- Many degrees of freedom
- Unrepeatable experiments

▸ Put user in the evaluation loop

Vision

- Simulate an interactive experiment
  - Compare IIR systems
    - **by controlling user interactions**
  - Build a collection that is **portable** and **reusable**
  - Devise measures to evaluate the utility a user obtains throughout a session

- Test systems in a large variety of cases that lead to sessions

Sessions

- Categorization on "who's to blame" basis?
  - Corpus
    - There is no composite document that fulfills the users need
  - User
    - Users cannot express their need by the appropriate query
   - Users learn about their need throughout the session
  - System
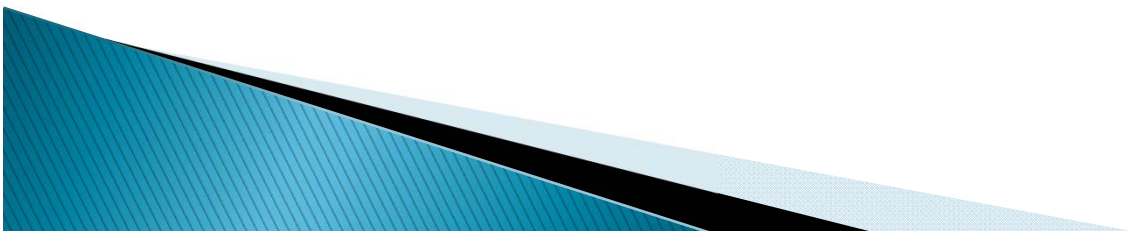
IIR systems for sessions

- Accumulate information through the session

  - Learn from the past
    - Improve on the current query

  - Look into the future

Goals

(**G1**) Test the ability of retrieval systems to use the history of user interactions (past queries,…) to improve performance on the current query

(**G2**) Evaluate system performance over an entire session instead of a single query
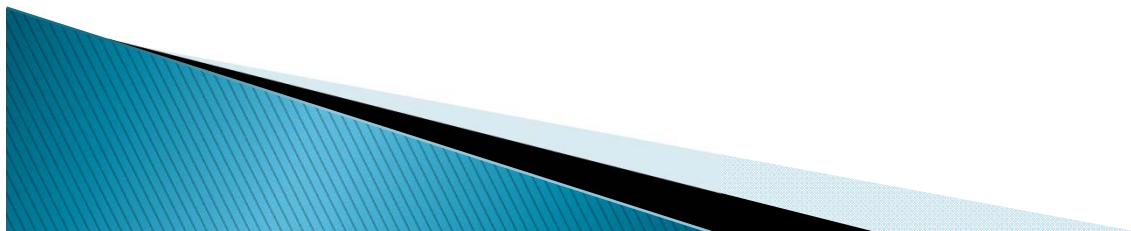
Reality

Task 1: If history can teach us anything...

▸ All participants should be provided with a **fixed** history of user interactions

Participants Task

▶ For each session participants submitted 4 ranked lists (RLs)

(**RL1**) current query

(**RL2**) current query & past queries in the session

(**RL3**) current query & past queries in the session and ranked URLs

(**RL4**) current query & past queries in the session, ranked URLs, clicks and dwell times
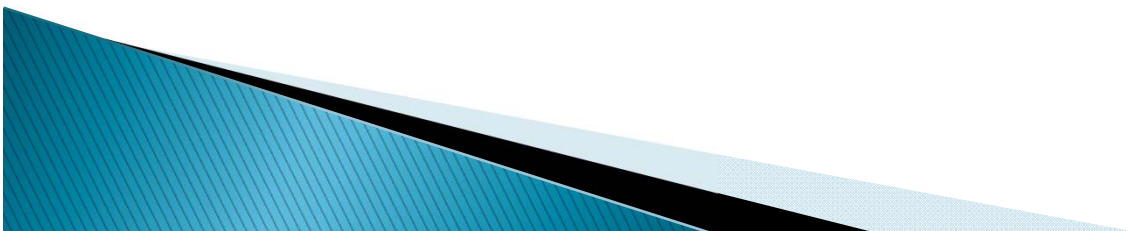
Challenges

▸ Come up with topics that can lead to sessions with more than one query.

  ◦ In 2011 we generated about 100 topics, collected about 1000 sessions:
    • 90% were single session queries.

  ◦ In 2012 we got about the same percentage of single-query sessions
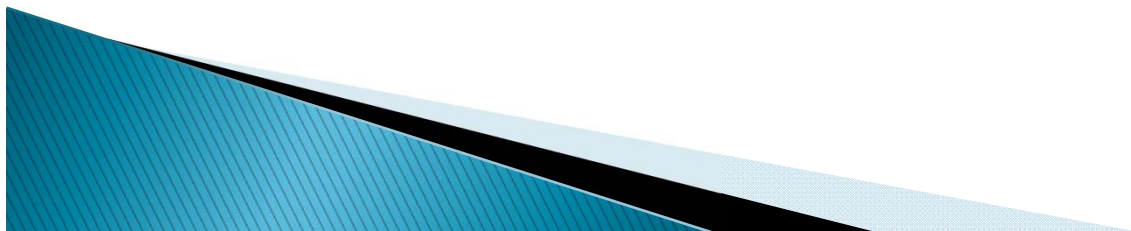
Challenges

- Come up with topics that can lead to different search task types.

  - In 2011 we generated factual tasks with specific goal(s)

  - In 2012 we generated a variety of task types
    - factual / intellectual tasks
    - specific / amorphous goals
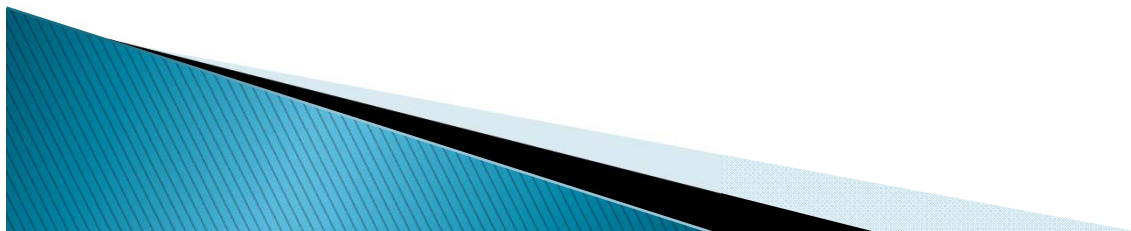
More challenges

- Evaluating retrieval systems

  - The notion of relevance may change throughout the session

  - Novelty needs to be considered

  - Different levels of diversification throughout the session should be applied

Task 2 : Looking into the future ...
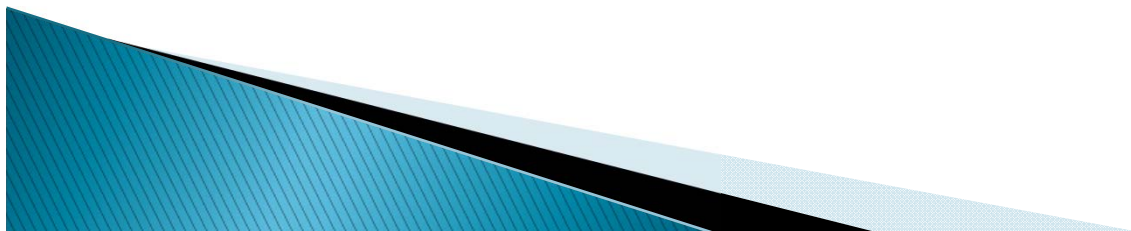
▸ Whole-session evaluation

Task 2 : Looking into the future ...

- Whole-session evaluation

- Simulate user interactions
  - Given a topic, a ranked list, and relevance information
    - Simulate browsing behavior
    - Simulate clicks, dwell times, etc.
    - **Simulate query reformulation**

- Quantify utility

Conclusions

▸ Built static test collection

▸ Evaluate retrieval performance when session information is available

▸ Still far from a whole-session evaluation
  ◦ Is it possible to build reusable and portable but dynamic test collections?
  ◦ Is it worth it?