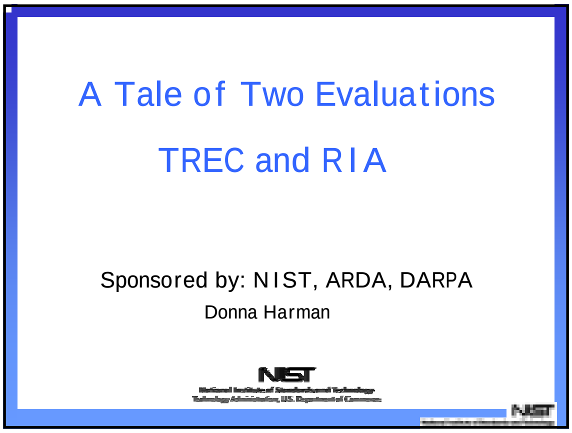
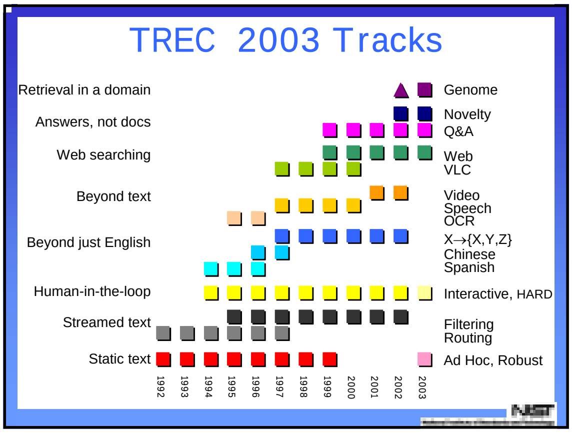


A Tale of Two Evaluations

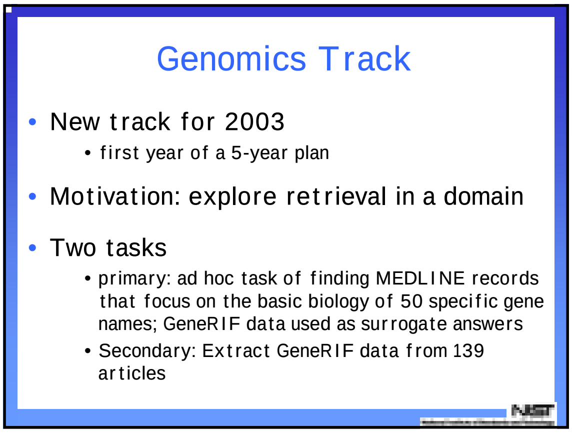
TREC and RIA

Sponsored by: NIST, ARDA, DARPA

Donna Harman

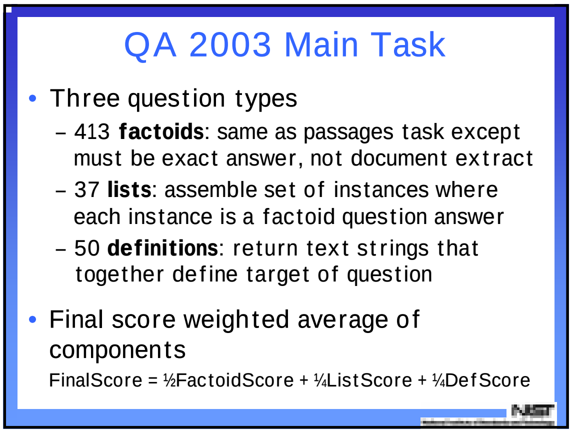


TREC 2003 Tracks



Genomics Track

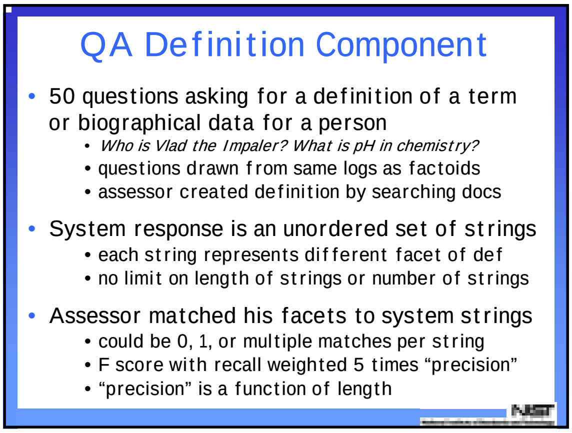
- New track for 2003
  - first year of a 5-year plan
- Motivation: explore retrieval in a domain
- Two tasks
  - primary: ad hoc task of finding MEDLINE records that focus on the basic biology of 50 specific gene names; GeneRIF data used as surrogate answers
  - Secondary: Extract GeneRIF data from 139 articles



QA 2003 Main Task

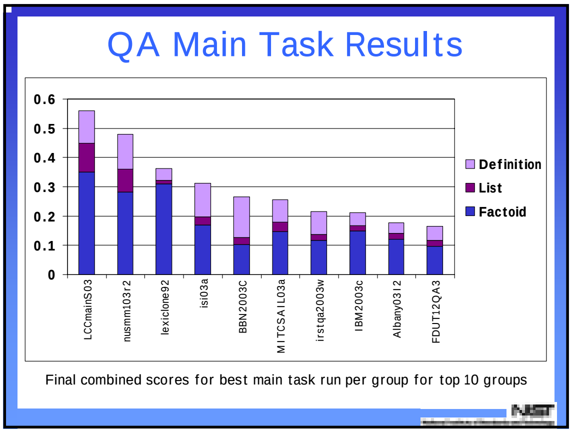
- Three question types
  - 413 **factoids**: same as passages task except must be exact answer, not document extract
  - 37 **lists**: assemble set of instances where each instance is a factoid question answer
  - 50 **definitions**: return text strings that together define target of question
- Final score weighted average of components

FinalScore = ½FactoidScore + ¼ListScore + ¼DefScore



QA Definition Component

- 50 questions asking for a definition of a term or biographical data for a person
  - *Who is Vlad the Impaler? What is pH in chemistry?*
  - questions drawn from same logs as factoids
  - assessor created definition by searching docs
- System response is an unordered set of strings
  - each string represents different facet of def
  - no limit on length of strings or number of strings
- Assessor matched his facets to system strings
  - could be 0, 1, or multiple matches per string
  - F score with recall weighted 5 times "precision"
  - "precision" is a function of length



QA Main Task Results

Final combined scores for best main task run per group for top 10 groups

HARD track

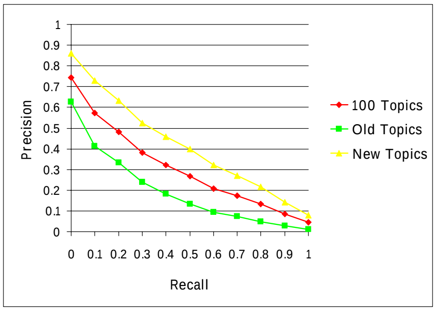
- Goal: improve ad hoc retrieval by customizing the search to the user using:

1) Metadata from topic statements
    1) the purpose of the search
    2) the genre or granularity of the desired response
    3) the user's familiarity with the subject matter
    4) biographical data about user (age, sex, etc.)
2) Clarifying forms
    1) assessor (surrogate user) spends at most 3 minutes/topic responding to topic-specific form
    2) example uses: sense resolution, relevance judgments

Robust Retrieval Track

- New track in 2003
- Motivations:
    - focus on poorly performing topics since average effectiveness usually masks huge variance
    - bring traditional ad hoc task back to TREC
- Task
    - 100 topics
        - 50 old topics from TRECs 6-8
        - 50 new tropics created by 2003 assessors
    - TREC 6-8 document collection: disks 4&5 (no CR)
    - standard trec_eval evaluation plus new measures

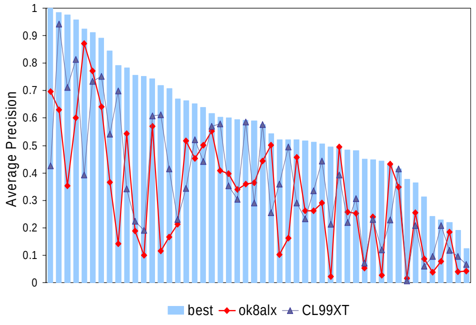
2003 Robust Retrieval Track



Retrieval Methods

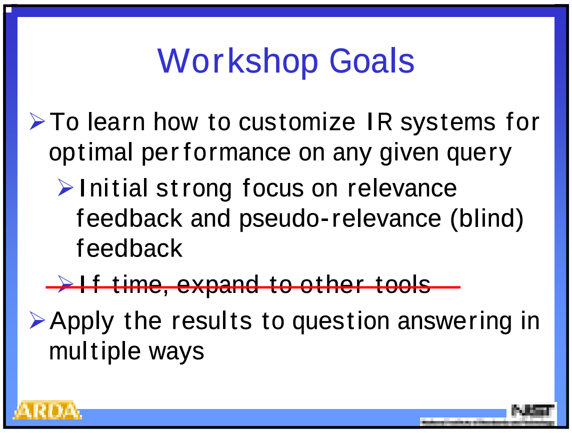
- CUNY and Waterloo expanded using the web (and possibly other collections)
    - effective, even for poor performers
- QE based on target collection generally improved mean scores, but did not help poor performers
- Approaches for poor performers
    - predict when to expand
    - fuse results from multiple runs
    - reorder top ranked based on clustering of retrieved set

The Problem



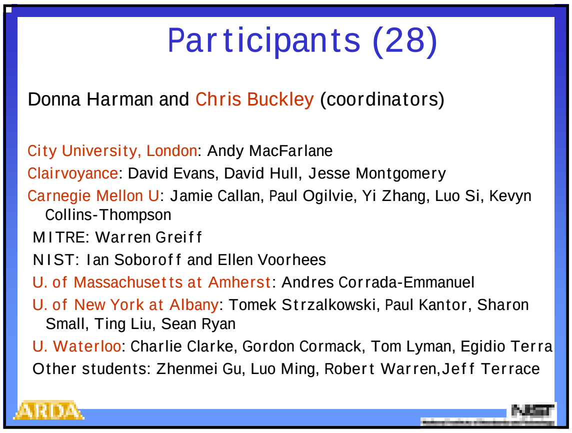
RIA Workshop

- In the summer of 2003, NIST organized a 6-week workshop called Reliable Information Access (RIA)
- RIA was part of the Northeast Regional Research Center summer workshop series sponsored by the Advanced Research and Development Activity of the US Department of Defense



Workshop Goals

- To learn how to customize IR systems for optimal performance on any given query
  - Initial strong focus on relevance feedback and pseudo-relevance (blind) feedback
  - ~~If time, expand to other tools~~
- Apply the results to question answering in multiple ways



Participants (28)

Donna Harman and Chris Buckley (coordinators)

City University, London: Andy MacFarlane

Clairvoyance: David Evans, David Hull, Jesse Montgomery

Carnegie Mellon U: Jamie Callan, Paul Ogilvie, Yi Zhang, Luo Si, Kevyn Collins-Thompson

MITRE: Warren Greiff

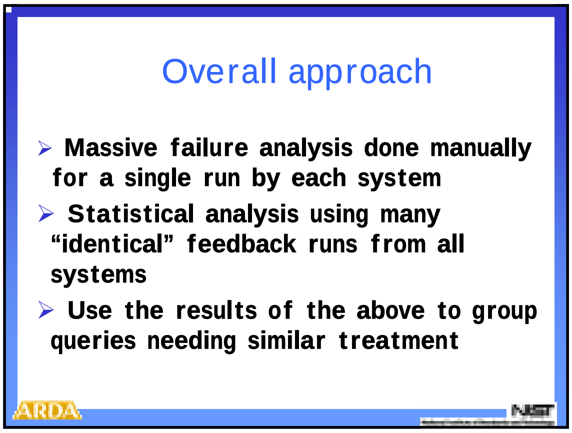
NIST: Ian Soboroff and Ellen Voorhees

U. of Massachusetts at Amherst: Andres Corrada-Emmanuel

U. of New York at Albany: Tomek Strzalkowski, Paul Kantor, Sharon Small, Ting Liu, Sean Ryan

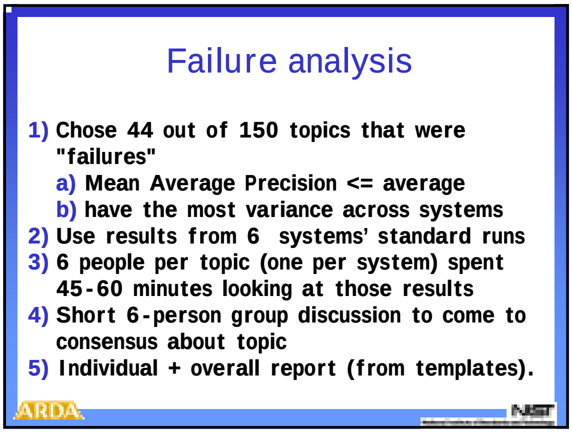
U. Waterloo: Charlie Clarke, Gordon Cormack, Tom Lyman, Egidio Terra

Other students: Zhenmei Gu, Luo Ming, Robert Warren,Jeff Terrace



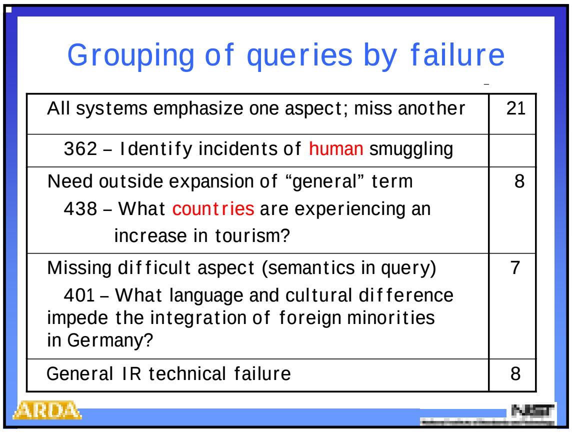
Overall approach

- **Massive failure analysis done manually for a single run by each system**
- **Statistical analysis using many “identical” feedback runs from all systems**
- **Use the results of the above to group queries needing similar treatment**



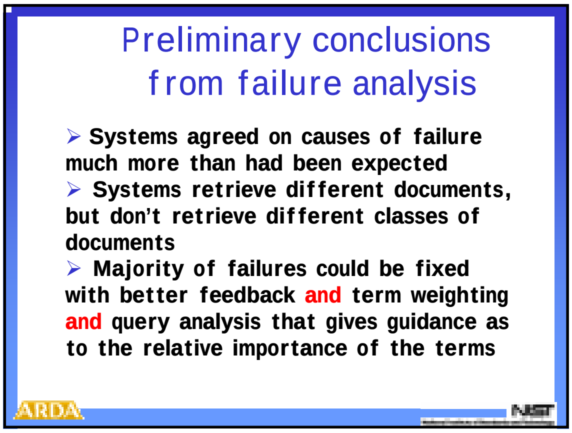
Failure analysis

1) **Chose 44 out of 150 topics that were "failures"**
   a) **Mean Average Precision <= average**
   b) **have the most variance across systems**
2) **Use results from 6 systems’ standard runs**
3) **6 people per topic (one per system) spent 45-60 minutes looking at those results**
4) **Short 6-person group discussion to come to consensus about topic**
5) **Individual + overall report (from templates).**



Grouping of queries by failure

All systems emphasize one aspect; miss another	21
362 – Identify incidents of human smuggling	
Need outside expansion of “general” term 438 – What countries are experiencing an increase in tourism?	8
Missing difficult aspect (semantics in query) 401 – What language and cultural difference impede the integration of foreign minorities in Germany?	7
General IR technical failure	8



Preliminary conclusions from failure analysis

- **Systems agreed on causes of failure much more than had been expected**
- **Systems retrieve different documents, but don’t retrieve different classes of documents**
- **Majority of failures could be fixed with better feedback and term weighting and query analysis that gives guidance as to the relative importance of the terms**

(Blind) Relevance Feedback

What are new methods of producing steel?

* FBIS4-53871 title1 ....
FT923-9006 title2 ....
* FBIS4-27797 .
* FT944-1455 .
FBIS3-24678 .
FT923-9281 .
* FT923-10837 .
FT922-11827 .
FT941-11316 .
.

List of experiments run

bf_base: base runs for all systems both using blind feedback (bf) and no feedback

bf_numdocs: vary #docs used for bf from 0-100

bf_numdocs_relonly: same but only use relevant

bf_numterms: vary #terms added from 0-100

bf_pass_numterms: same but use passages as source instead of documents

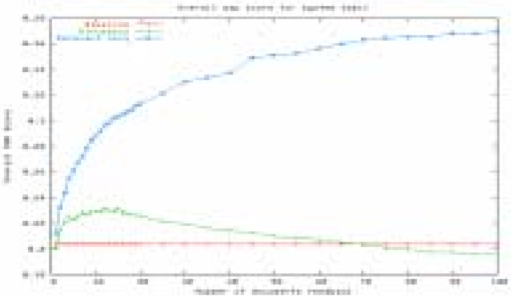
bf_swap-doc: use documents from other systems

bf_swap_doc_term: expand using docs and terms

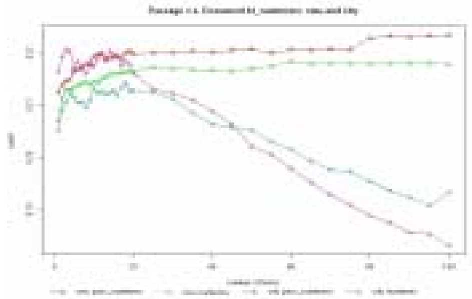
bf_swap_doc_cluster: use CLARIT clusters

bf_swap_doc_fuse: use fusion of other systems

bf_numdocs, relevant only



bf_numterms_passages



Preliminary Lessons Learned

1) Failure analysis

a) systems tend to fail for the same reason

b) getting the right concepts in system query critical

2) Surprises that require more analysis

a) bf_swap_docs: some systems better at providing docs

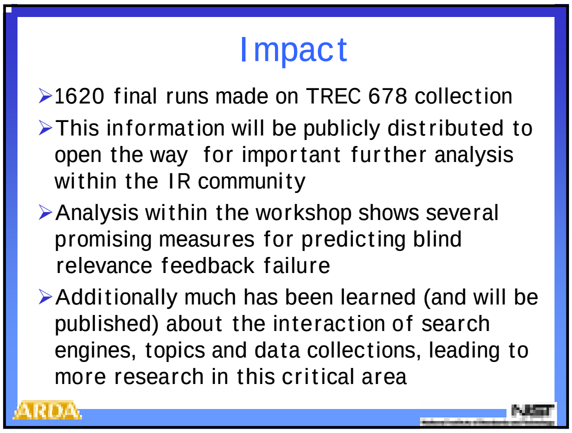
b) some systems more robust during expansion

c) bf_num_docs relevant only: some relevant docs are bad feedback docs

d) no topic in which there were "golden" terms in top 1-4 feedback terms

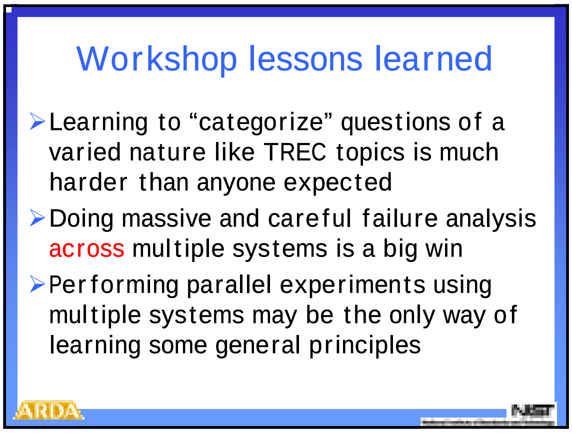
Additional experiments

- topic_analysis: producing & comparing groups of topics using assorted measures
- qa_standard: effect of IR algorithms on QA using docs/passages
- topic_coverage: HITIQA experiment using all systems



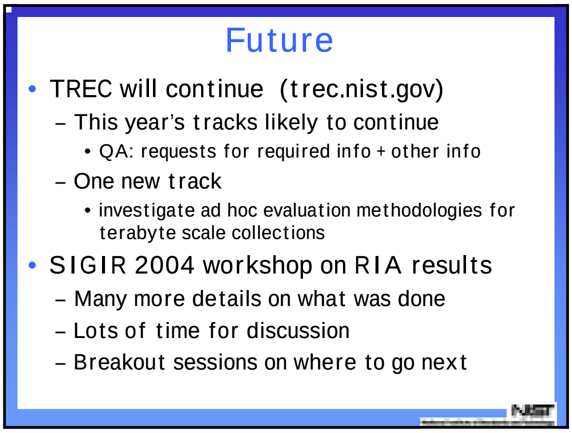
Impact

- 1620 final runs made on TREC 678 collection
- This information will be publicly distributed to open the way for important further analysis within the IR community
- Analysis within the workshop shows several promising measures for predicting blind relevance feedback failure
- Additionally much has been learned (and will be published) about the interaction of search engines, topics and data collections, leading to more research in this critical area



Workshop lessons learned

- Learning to “categorize” questions of a varied nature like TREC topics is much harder than anyone expected
- Doing massive and careful failure analysis across multiple systems is a big win
- Performing parallel experiments using multiple systems may be the only way of learning some general principles



Future

- TREC will continue (trec.nist.gov)
  - This year’s tracks likely to continue
    - QA: requests for required info + other info
  - One new track
    - investigate ad hoc evaluation methodologies for terabyte scale collections
- SIGIR 2004 workshop on RIA results
  - Many more details on what was done
  - Lots of time for discussion
  - Breakout sessions on where to go next