We are Gathered Here Together: Why? Introduction to the NII Workshop on **Whole-Session Evaluation of Interactive Information Retrieval Systems**

Nicholas J. Belkin
School of Communication & Information
Rutgers University
belkin@rutgers.edu

Who are “We”?

- IR researchers concerned with proper evaluation of the performance of IR systems
- IR researchers dissatisfied with current criteria, measures, and methods of evaluation of the performance of IR systems
- IR researchers who are particularly concerned with user-centered evaluation of the performance of IR systems

What's Meant by "User-Centered" Evaluation?

- Understanding the goals/tasks that lead people to engage with IR systems
- Understanding the behaviors that people engage in during information seeking
- Devising criteria and measures of performance that correspond to those goals/tasks, and to those behaviors
- Developing methods for applying those criteria and measures

What is Meant by "Whole-Session Evaluation"?

- Understanding that the information-seeking activities related to motivating goals/tasks may not be limited to IR system responses to single queries, but may require support over a number of *different activities* during a *search session*, or over *multiple* search sessions.
- Developing criteria, measures and methods which are appropriate to evaluating IR system support with respect to the *process* of a search session, the *outcome* of a search session, and the outcomes of *multiple* search sessions

What can “We” do Here to Address Whole Session Evaluation

- Complain about other evaluation criteria, measures and methods
  - Not a starter
- Identify the most significant problems in accomplishing whole session evaluation
- Propose some means for addressing these problems
- Suggest a(some) framework(s) for accomplishing whole session evaluation

Possible Goals for the Workshop

- A *document* outlining:
  - Need for whole-session evaluation;
  - Problems facing whole-session evaluation
  - Potential approaches to those problems, with pluses, minuses and contexts of application
  - Proposals of research directions in implementing whole-session evaluation
- A set of possible *frameworks* in which whole-session evaluation could take place
- A *proposal* to TREC (or some other evaluation forum) for a Whole-Session Evaluation Track in 2013
- Establishment of a *research community* to share data and experiences

Ways to Accomplish these Goals

- Draw on the expertise of the group: sharing and comparing your *homework*
- Identify groups of participants sharing similar concerns/solutions wrt whole-session evaluation as “break-out groups”
- Reports from groups as the basis for debate on proposals
- New groups formed as result of debate, iterate
- Small groups assigned to produce the products (and their components) of the workshop

Some Things We're Likely to Argue About

- The value and forms of formal models of search
- What constitutes a "search session"
- How to relate system support to task outcomes, or should we even bother
- Simulation of search sessions, perhaps versus study of "real" goals, tasks, behaviors
- Comparative evaluation, a la TREC
- If "test collections" are thought to be desirable or necessary, what form(s) should they take

Some Things we should Agree On

- We'll be (at least moderately) nice to one another
- We're here to collaborate and create, not to pontificate
- We will devote significant time to breaks from formal work
- We will enjoy Shonan Village, its facilities, foods, and environment
- We will do our best to achieve at least drafts of our goals by Friday afternoon