# User-oriented Evaluation in IR

Kal Jarvelin

UNIVERSITY
OF TAMPERE

INFIM
Information Studies and Interactive Media
SCHOOL OF INFORMATION SCIENCES

INFIM

UNIVERSITY OF TAMPERE

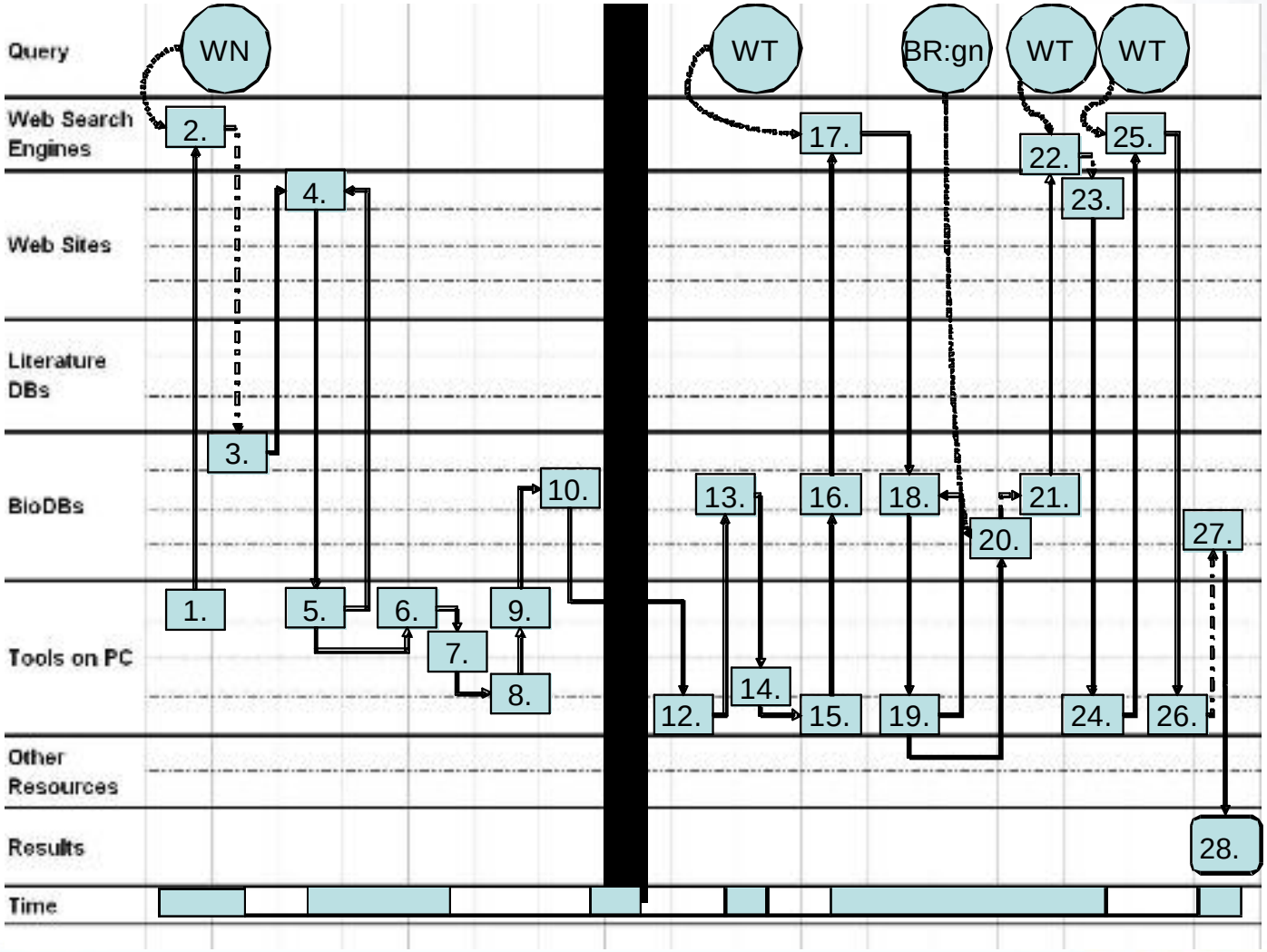
# Outline

- What is evaluation?
- IR evaluation landscape
- Test collection based evaluation
- User-Centered Evaluation
- Operational Systems Evaluation
- Beyond evaluation?

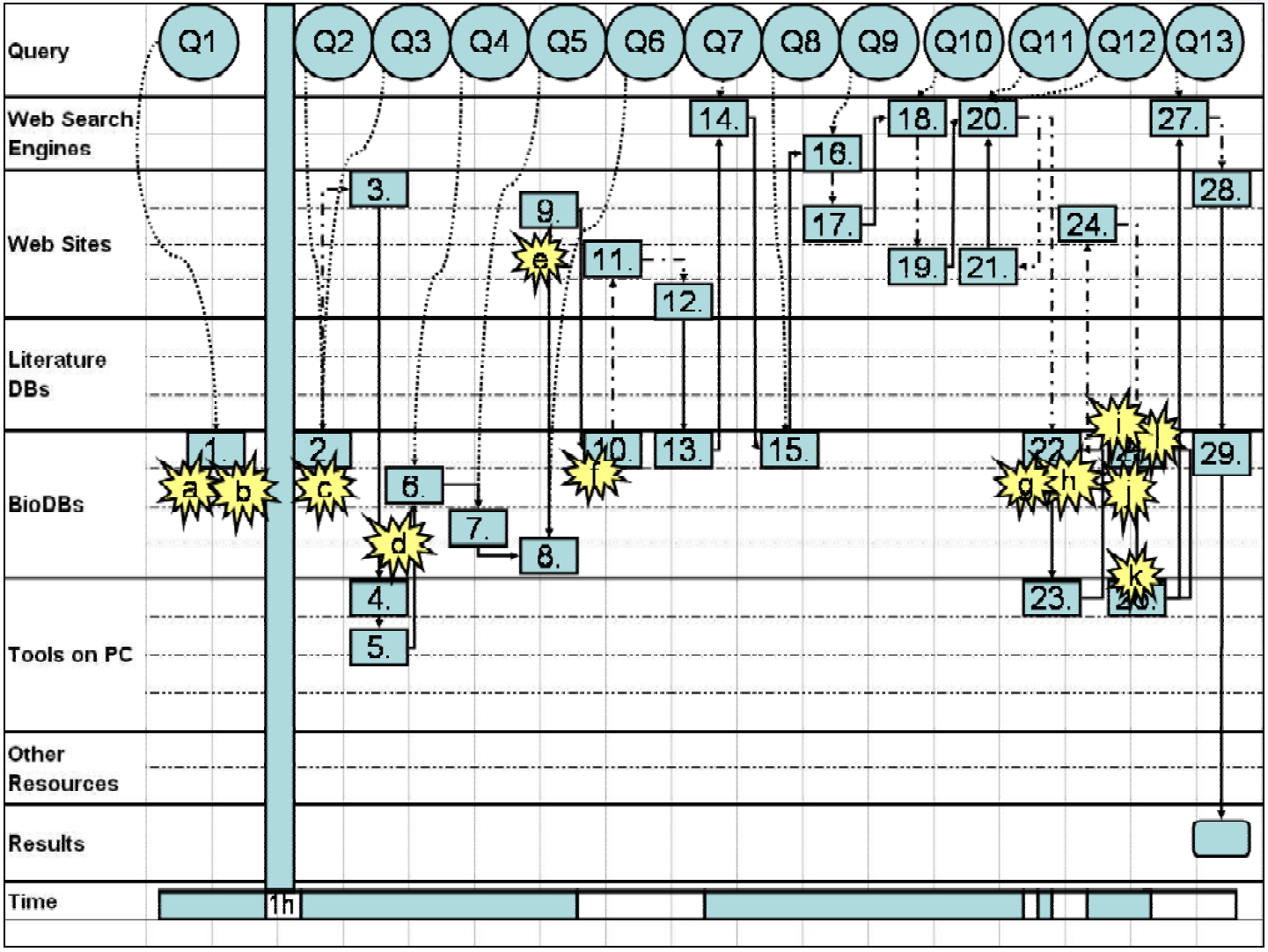
# 1. Evaluation

- Systematic determination of merit of an object using some criteria
- In IR evaluation typically focuses on an IR system or a component
- The criteria typically focus on the quality of IR system output, the ranked list
- However, there are alternatives

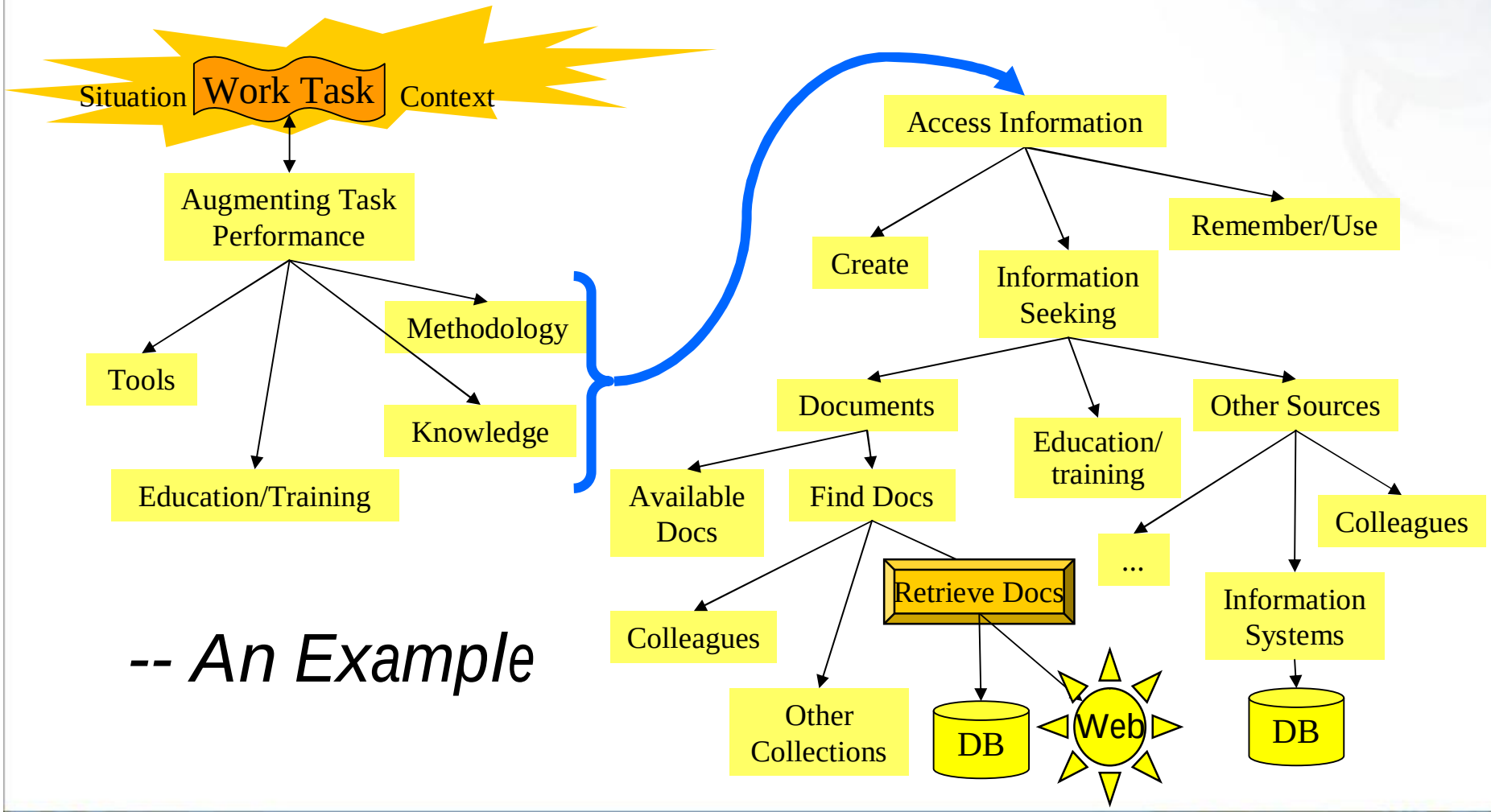
# Sample Work Task Process in Biotech




(Kumpulainen & Järvelin 2010)


Query
Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10
Q11
Q12
Q13
Web Search Engines
Web Sites
Literature DBs
BioDBs
Tools on PC
Other Resources
Results
Time
1h
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.
16.
17.
18.
19.
20.
21.
22.
23.
24.
27.
28.
29.
a
b
c
d
e
f
g
h
i
k

# Task-based IR: Means - Ends




*-- An Example*

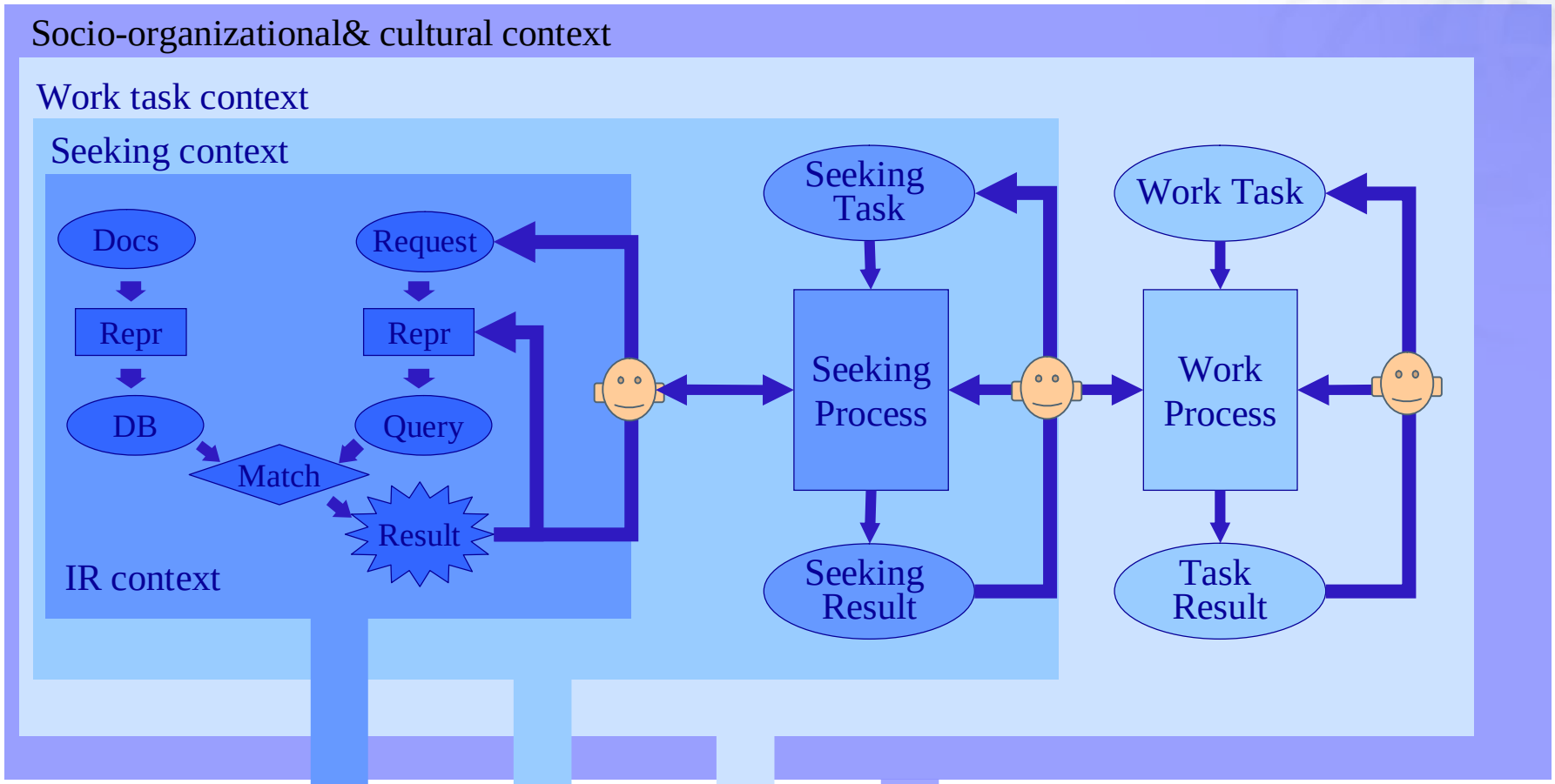
# Back to Evaluation …

- What do we want to evaluate - and why?
  - practical life challenging
  - surrogate objects in evaluation
  - one's task may be the development of a system component … or … of organizational work
- Experimental vs. naturalistic evaluation
  - experimental heuristics vs. practical meaning

# 2. IR evaluation landscape

- Three views on the landscape
  - Nested contexts: IR - seeking - tasks - organization
  - Barriers in nested contexts
  - Kelly's evaluation study continuum

# Frameworks for IR (Evaluation)




Evaluation Criteria:

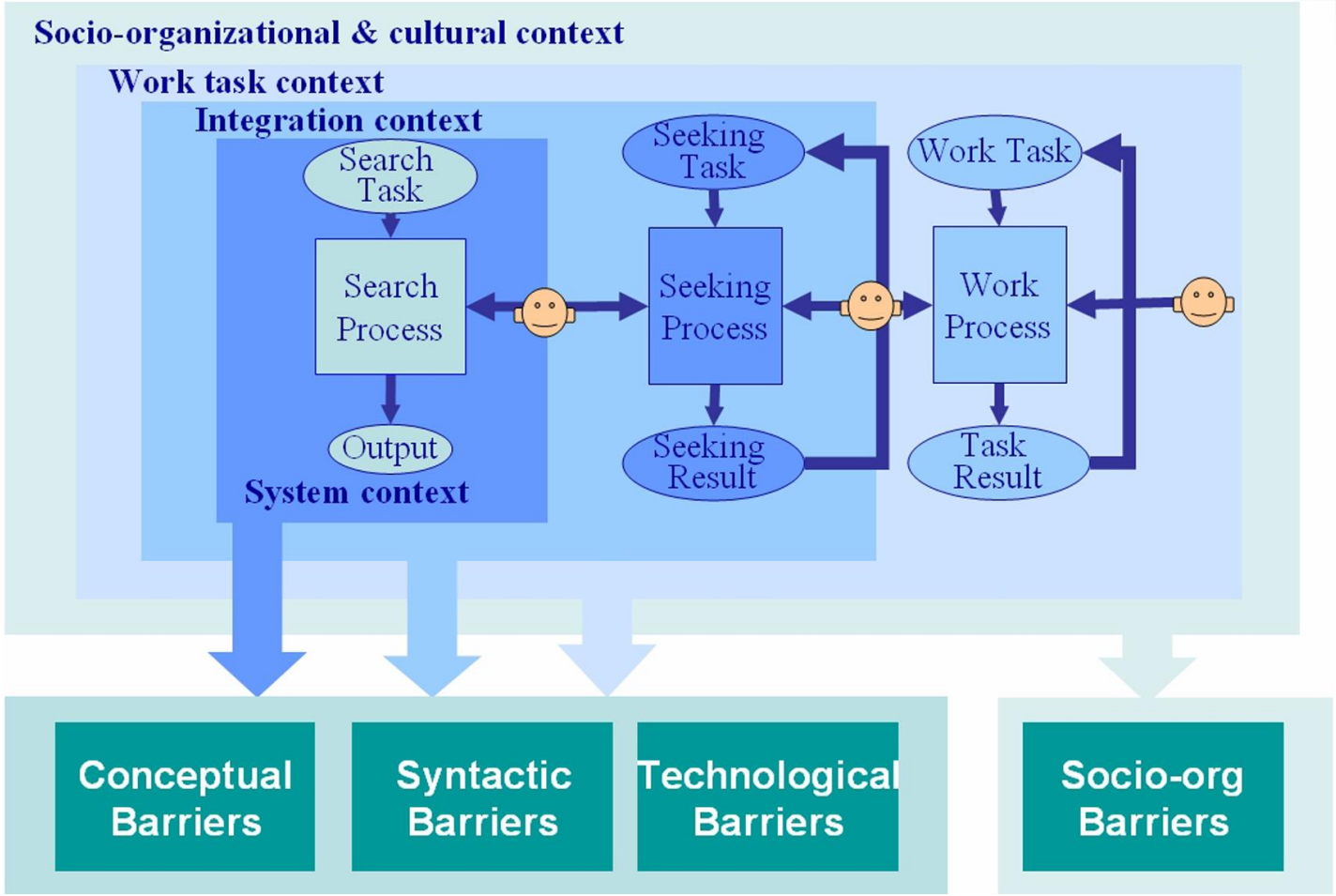
D: Socio-cognitive relevance; quality of work task result

C: Quality of info & work process/result, learning

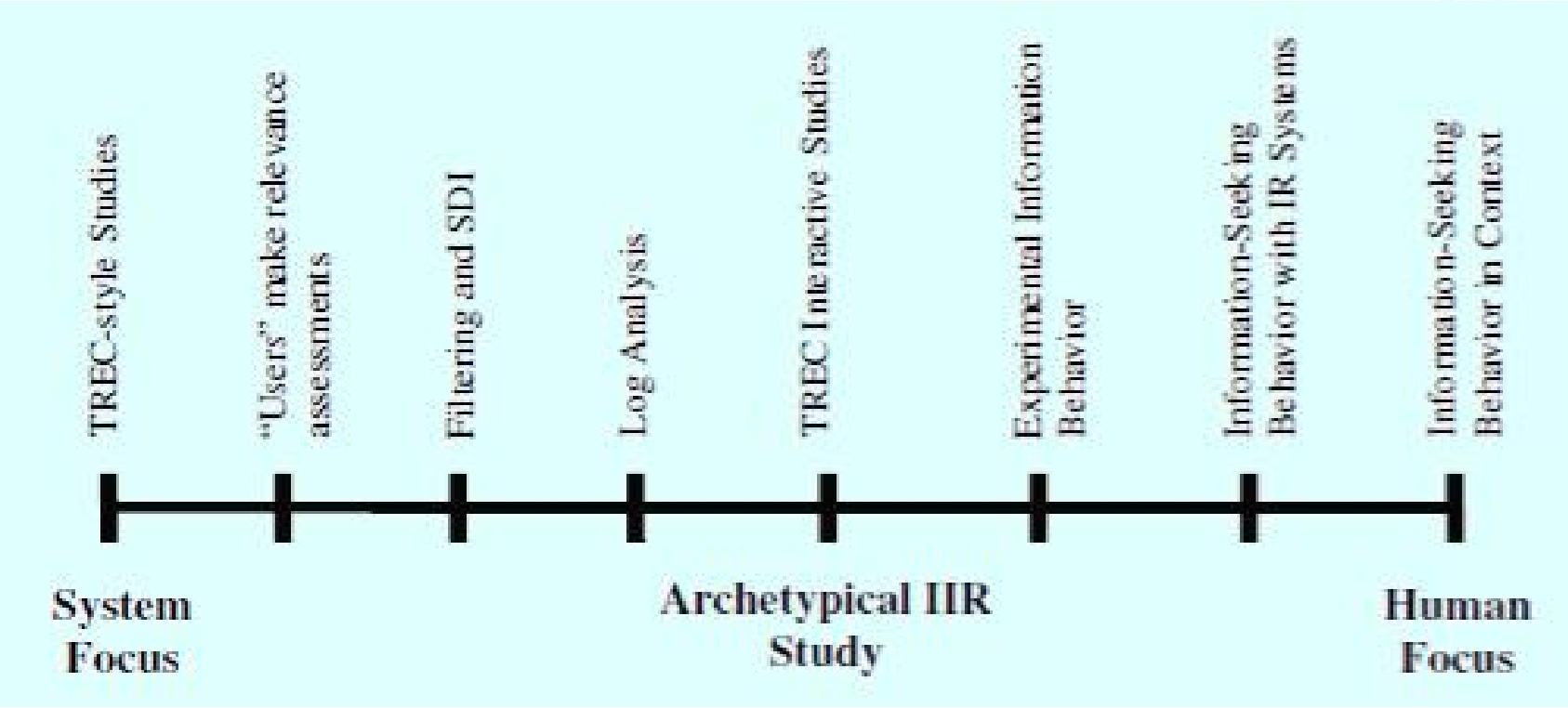
B: Usability, quality of information/process, learning

A: Recall, precision, efficiency, quality of process

# Access and Barriers



# The continuum of IR evaluation studies




(Kelly 2009 FnTIR)

INFIM

UNIVERSITY OF TAMPERE

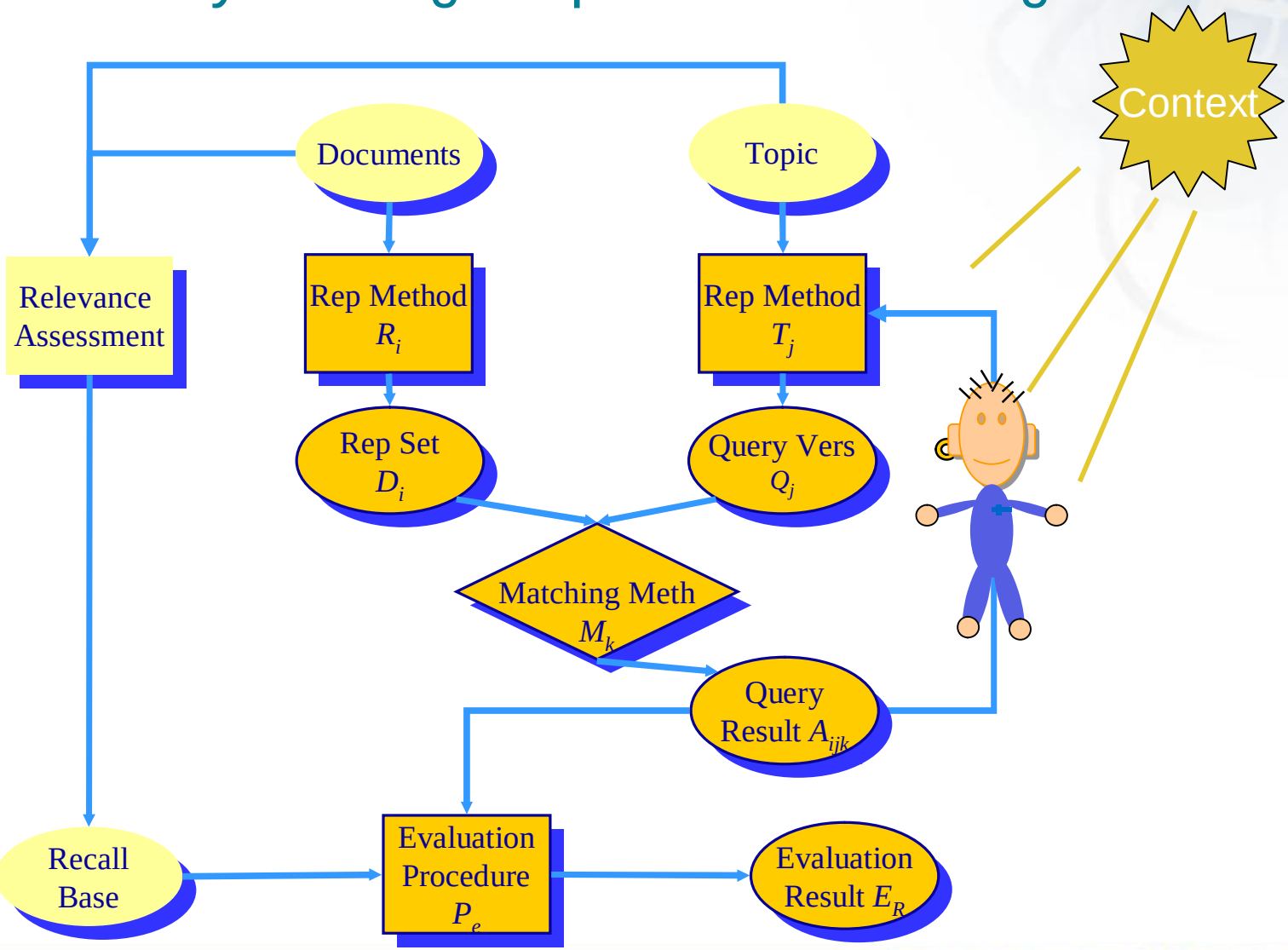
# 3. Test collection based evaluation

- Other talks at Winter School:
  - TREC-style evaluation (DH)
  - IR metrics and statistics (SR)
- TREC-style evaluation tries to abstract away much users' individual variability
  - achieving controllability of experiments
  - achieving comparability of experiments
- Nevertheless, there is a user / task model

# TREC User/Task Model

- Test collection studies have a simple model
  - an individual searcher
  - a single query per info need, long scanning,
  - exhaustive need, overlapping information valuable
  - matches a part of real life
- Such evaluation means *simple simulation of a searcher* interacting with an IR system
- Test collections allow more …

# The Laboratory Setting: Experimental Design



INFIM

UNIVERSITY OF TAMPERE

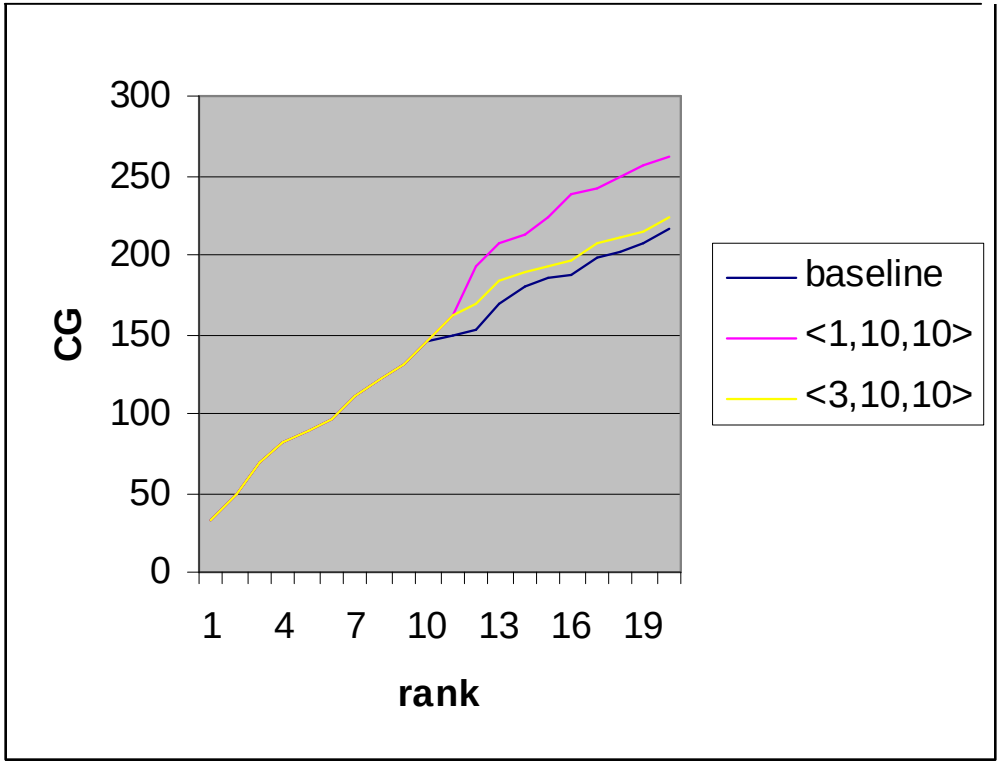
# Session Simulation for Evaluation

- Evaluation of more general sessions based on test collections requires
  - generation of behaviors: queries, reformulations, results scans (seen docs), decisions including stopping
  - handling of previously seen documents and other content duplicates in evaluation
  - reconsideration of evaluation goals and metrics
- Such simulation abstracts away the interface and link following

# Simulation of very short sessions

- Very short sessions: single reformulation
  - TREC Session Track 2010
    - initial query too broad, narrow, or off-target
  - Relevance Feedback
    - initial query and one round of feedback
    - intellectual (vs. automatic)
    - what if human searcher err?
- Straightforward: freezing seen documents, removing duplicates, traditional metrics

# Effect of RF Amount and Quality




RFB with CG evaluation with scenarios *baseline, <3,10,10>, <1,10,10>, weighting 0-1-10-100*

Keskustalo & al (2008) IRJ

# User modeling for RF simulation

| Fallibility Scenario | Real Doc Rel Grade | Human Judgment Probabilities n | m | f | h |
|---|---|---|---|---|---|
| **1.00** | n | **1.0** | 0.0 | 0.0 | 0.0 |
| | m | 0.0 | **1.0** | 0.0 | 0.0 |
| | f | 0.0 | 0.0 | **1.0** | 0.0 |
| | h | 0.0 | 0.0 | 0.0 | **1.0** |
| **0.75** | n | **0.75** | 0.125 | 0.075 | 0.05 |
| | m | 0.10 | **0.75** | 0.10 | 0.05 |
| | f | 0.05 | 0.10 | **0.75** | 0.10 |
| | h | 0.05 | 0.075 | 0.125 | **0.75** |
| **0.50** | n | **0.50** | 0.25 | 0.15 | 0.10 |
| | m | 0.20 | **0.50** | 0.20 | 0.10 |
| | f | 0.10 | 0.20 | **0.50** | 0.20 |
| | h | 0.10 | 0.15 | 0.25 | **0.50** |
| **0.25** | n | **0.25** | 0.25 | 0.25 | 0.25 |
| | m | 0.25 | **0.25** | 0.25 | 0.25 |
| | f | 0.25 | 0.25 | **0.25** | 0.25 |
| | h | 0.25 | 0.25 | 0.25 | **0.25** |

| Fallibility Scenario | Human Judgment Probabilities relevance | n | m | f | h |
|---|---|---|---|---|---|
| **0.50-0.80** | n | **0.5** | 0.4 | 0.1 | 0.0 |
| | m | 0.4 | **0.5** | 0.1 | 0.0 |
| | f | 0.0 | 0.1 | **0.8** | 0.1 |
| | h | 0.0 | 0.0 | 0.2 | **0.8** |

(n):nonrelevant
(m):marginal
(f):fair
(h):highly relevant documents

Baskaya & al. (ECIR 2011)

INFIM

UNIVERSITY OF TAMPERE

# Simulation of Complex Sessions

- More realism in simulation involves
  - varying search tasks
  - time / effort constraints
  - multiple queries, complex reformulations
  - variable scanning depth
  - interaction of task, effort, seen results, consulted documents (learning)
- Simulation gets complex but not impossible

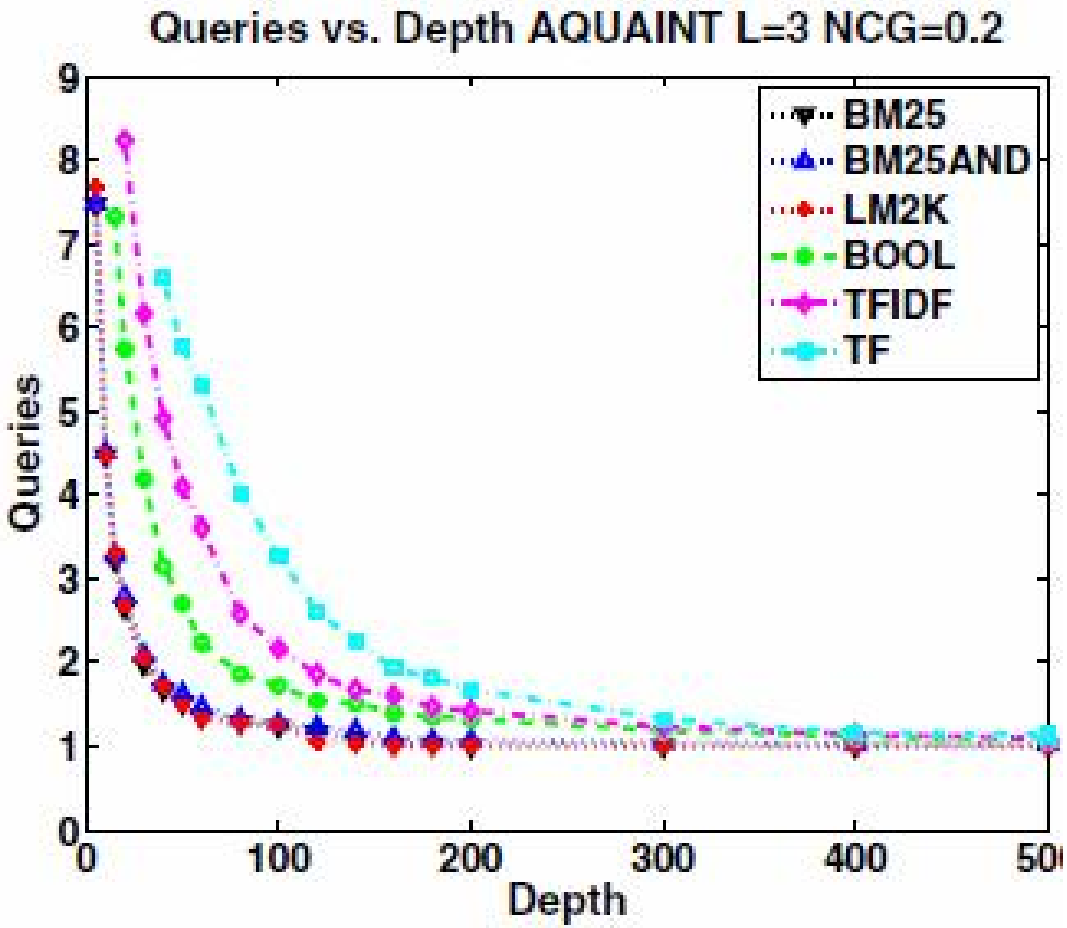
INFIM

UNIVERSITY OF TAMPERE

# Azzopardi's simulation

- Normalized gain goals in searching, optimal sessions that minimize costs for given output
- Time in trading between queries vs. scanning
- Variable numbers of reformulations, 3w queries
- Variable scanning depths
- Query generation from relevant documents
- Six ranking methods (tf … BM25)
- Interaction of costs (q=2.6 s, s=2.3 s), and retrieval methods under gain goal

# Azzopardi's simulation

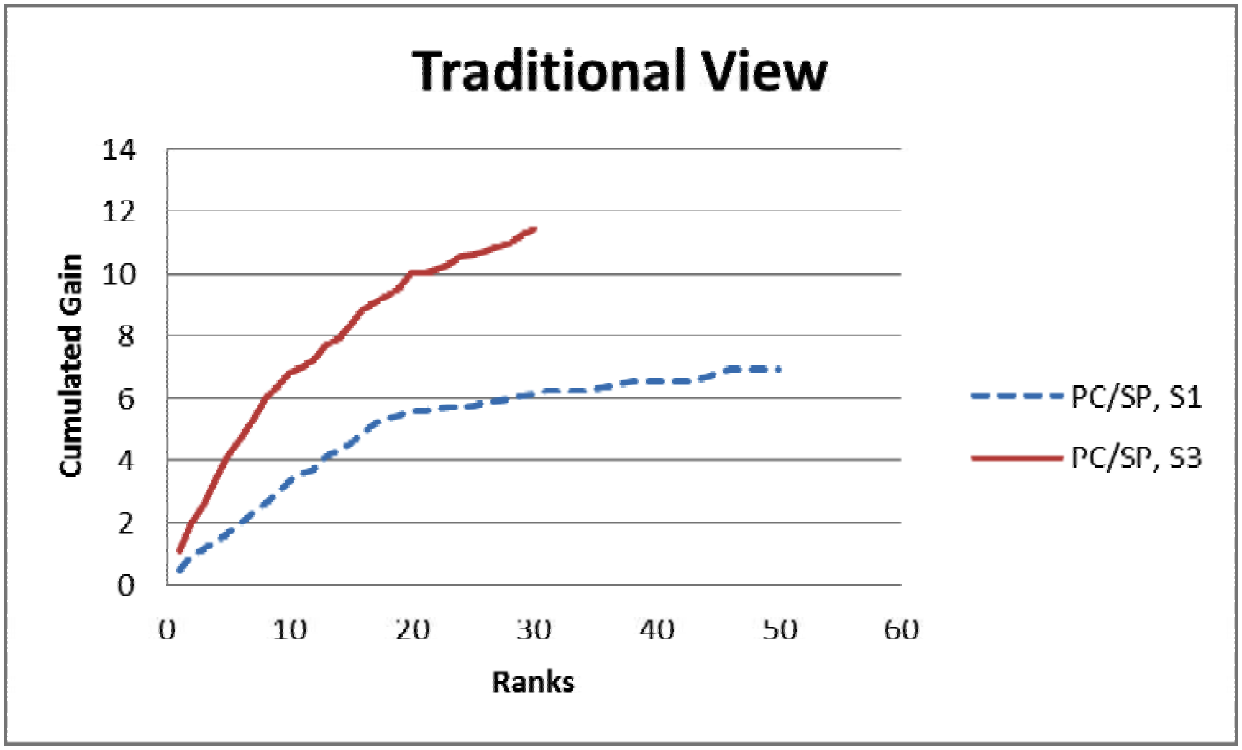



(Azzopardi, Sigir 2011)

# UTA Simulation

- max gain search tasks, session properties
- time constraints
- 5 strategies in reformul ries
- variable scanning dept
- one engine
- interaction of device props, effort, seen results

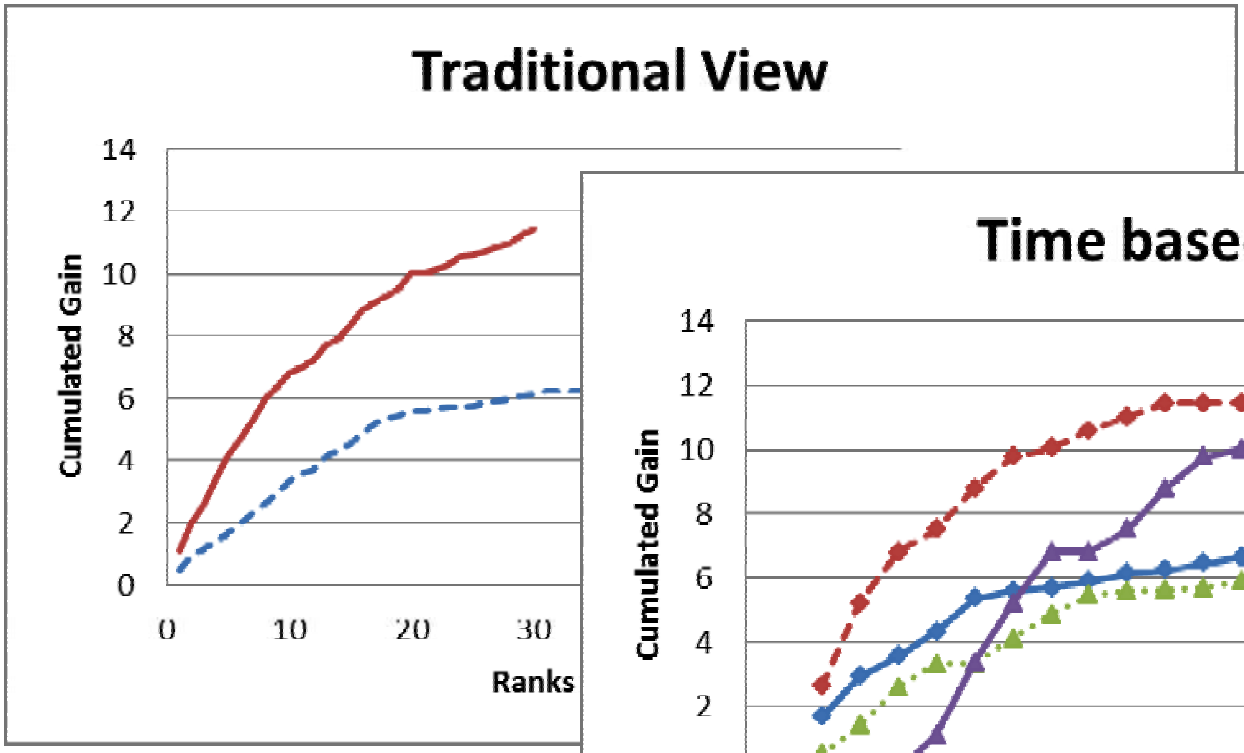
# Evaluation of Ranking is Traditional …

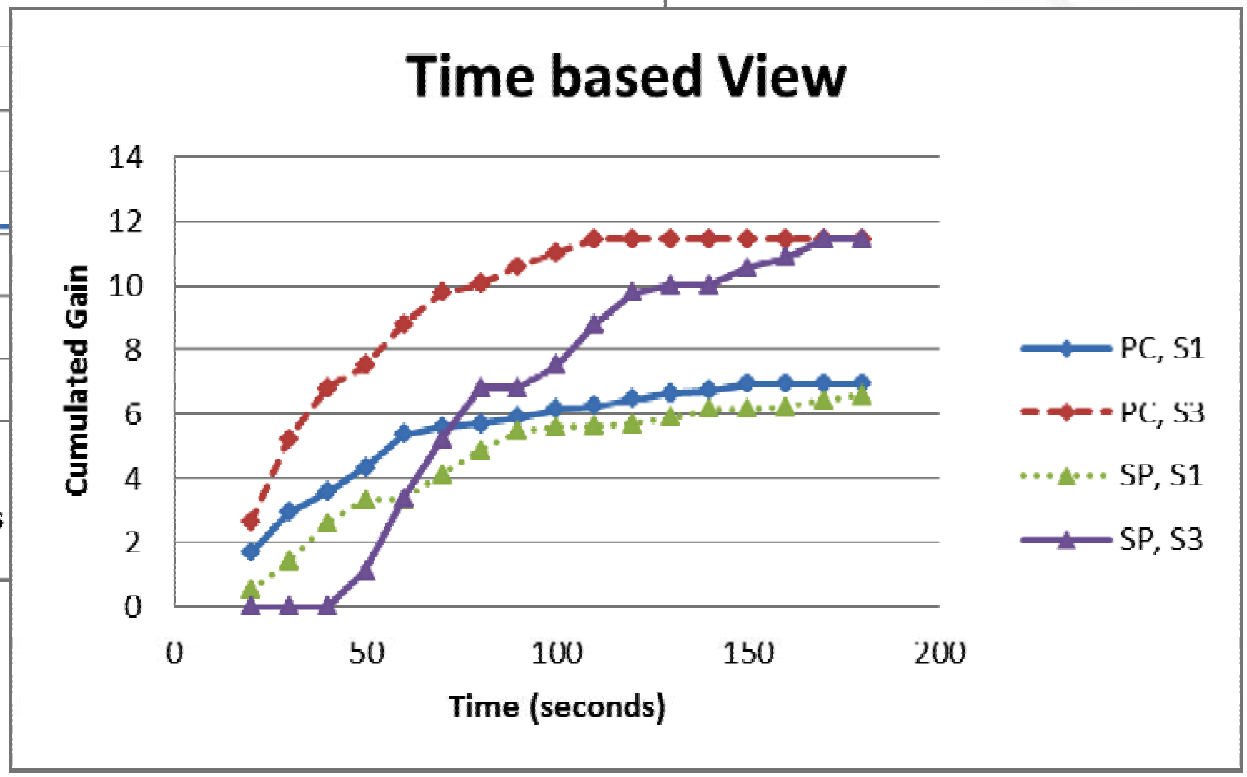


INFIM

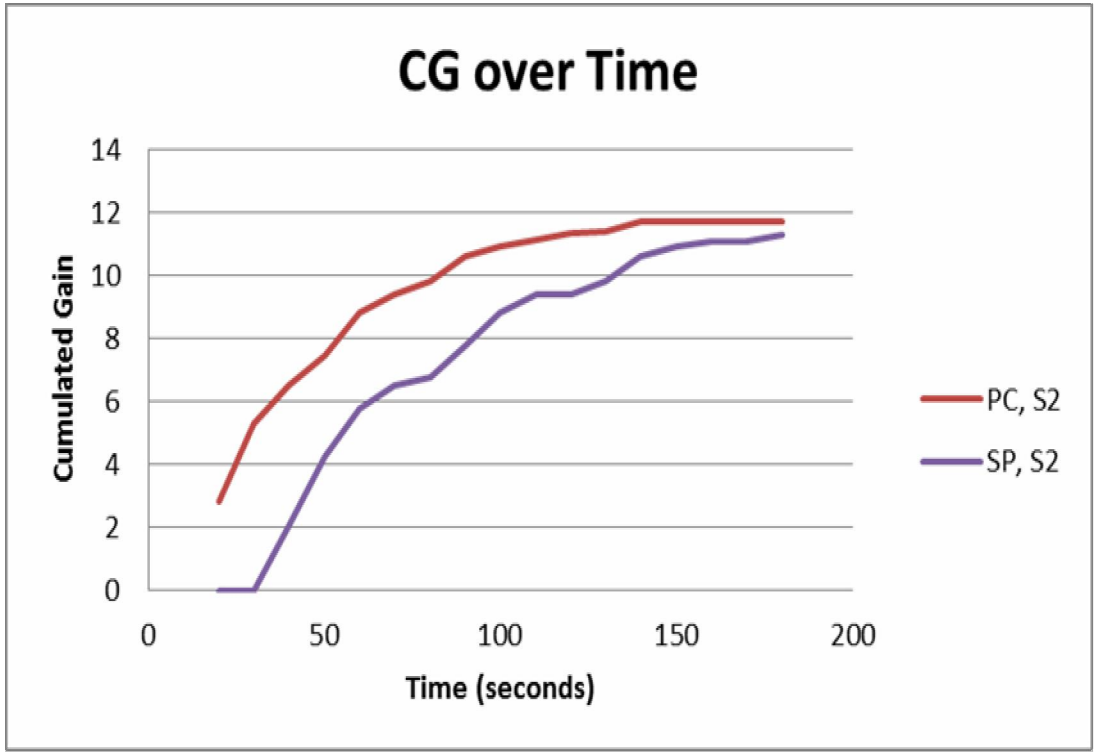
UNIVERSITY OF TAMPERE

# … but Evaluation with Time is Tricky

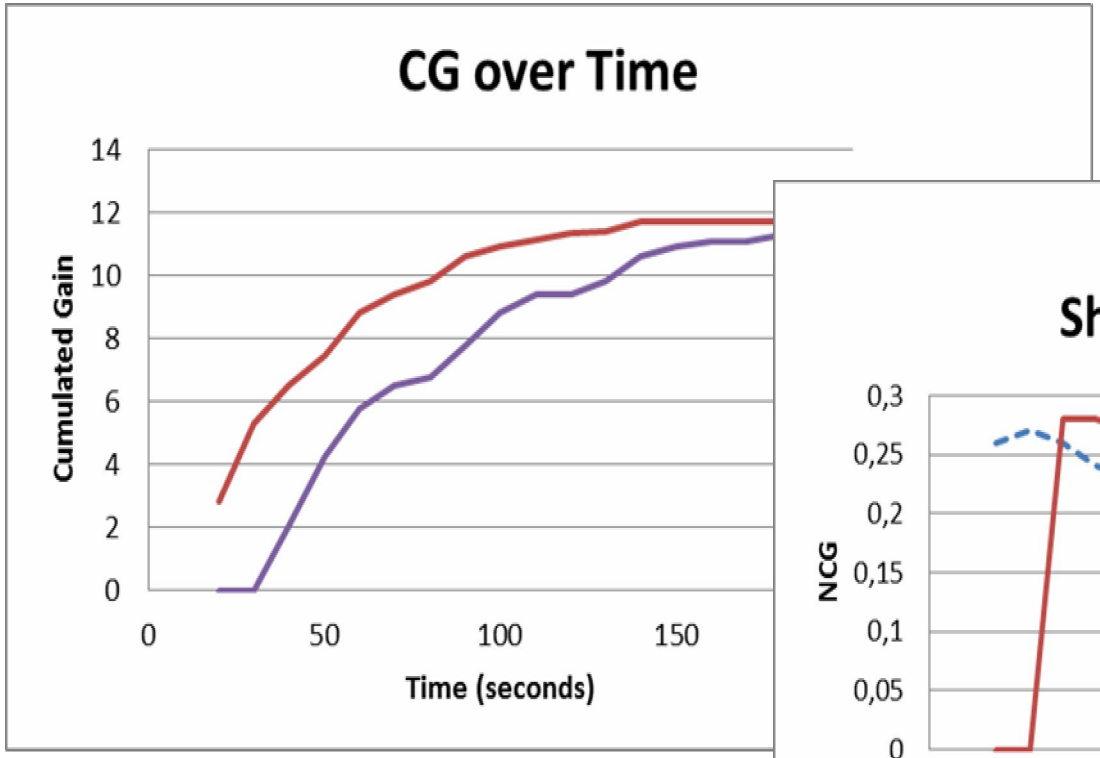


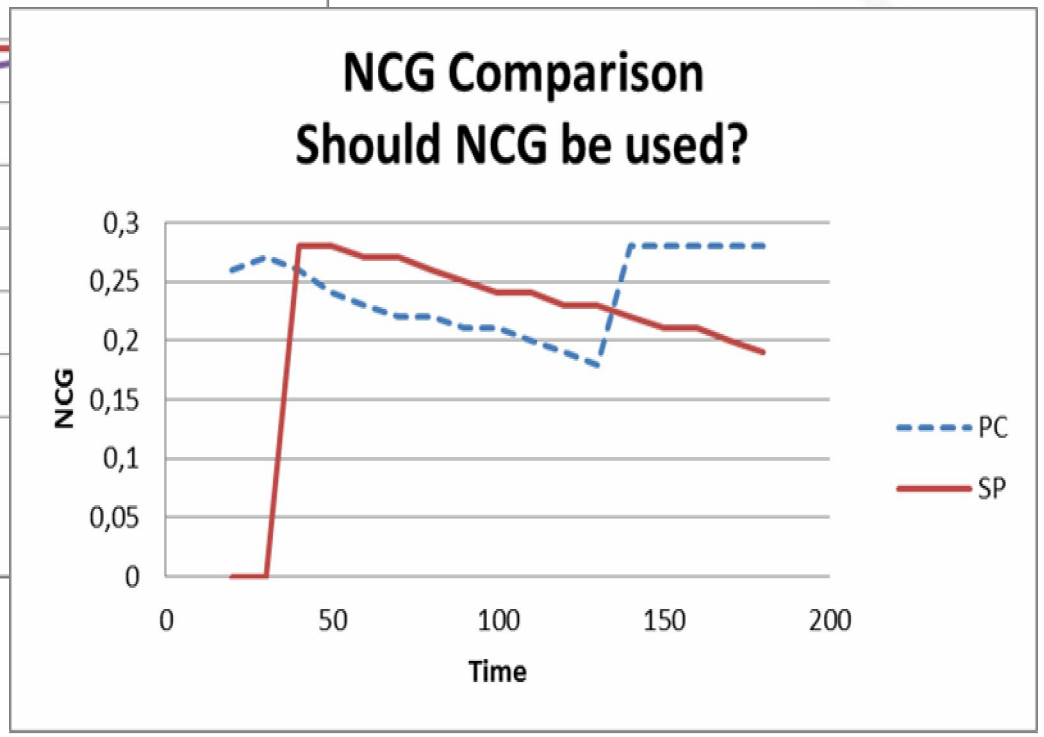



# Gain over Time can be Understood …



# .... but it is Tricky with Normalization



# Challenges in Session Simulation

- Setting the search task
- Simulating an inherently stochastic process
  - query formulation, browsing, stopping
- Handling duplicates produced by different queries
  - when are retrieved documents “seen”
- Simulating uncertainty or errors in relevance assessments / feedback
- Evaluating - results (ranking) vs. effort (time)
- The big why

# 4. User-Centered Evaluation

- A human user in the IR setting:
  - standardization of evaluation vanishes
  - no single experimental design to follow
  - need to define the system, its goals, and the evaluation criteria
- At the one end, the system consists of an IR system, a test collection, and a human operator of the IR system
  - goals: high-quality ranked list of documents; diversity
  - constraints: time, resources used
  - metrics: traditional, clock time, diversity

# User-Centered Evaluation, 2

- At the other end, the system consists of a document collection, an IR system, the human actor and a work/search task to perform
  - no topics or relevance assessments other than which the human actor individually creates
  - the primary goal of the system is to perform the task
  - additional goals / constraints: time, result quality
  - evaluation criteria: quality of the task result, clock time, the task performer's experience and satisfaction

# Variables in Study Designs

1. Work task variables: work task, its outcome and context
2. Search task variables: search task, its outcome and context
3. Actor characteristics: physical, emotional and cognitive features
4. Perceived work task: actor's perception of the work task
5. Perceived search task: actor's perception of the search task
6. Document variables: all document features and representations
7. Search engine variables: features of the IT component
8. Interface variables, dealing with interface functionalities
9. Access and interaction variables: features of IR and social interaction

# User-based Designs

- Dependent variables typically
  - process (duration, number of query reformulations),
  - output (MAP, nDCG, searcher satisfaction), or
  - outcomes variables (work task completion and quality).
- Independent variables differ
  - systems in the focus:
    - search engine and interface features are often systematically varied, and
    - actor and task variables controlled
  - searchers in the focus:
    - their characteristics (like knowledge) are varied, and
    - search engine and interface features often fixed

# User-based Designs, 2

- Need a model of the system being evaluated, indicating variables and their interactions in attaining the goals
- If this is done poorly,
  - the performance of the system cannot be properly evaluated,
  - the role of each contributing factor remains unclear
- Rational evaluation of interactive IR requires many study designs for assessing
  - the contribution of each variable affecting the process, and
  - their interaction

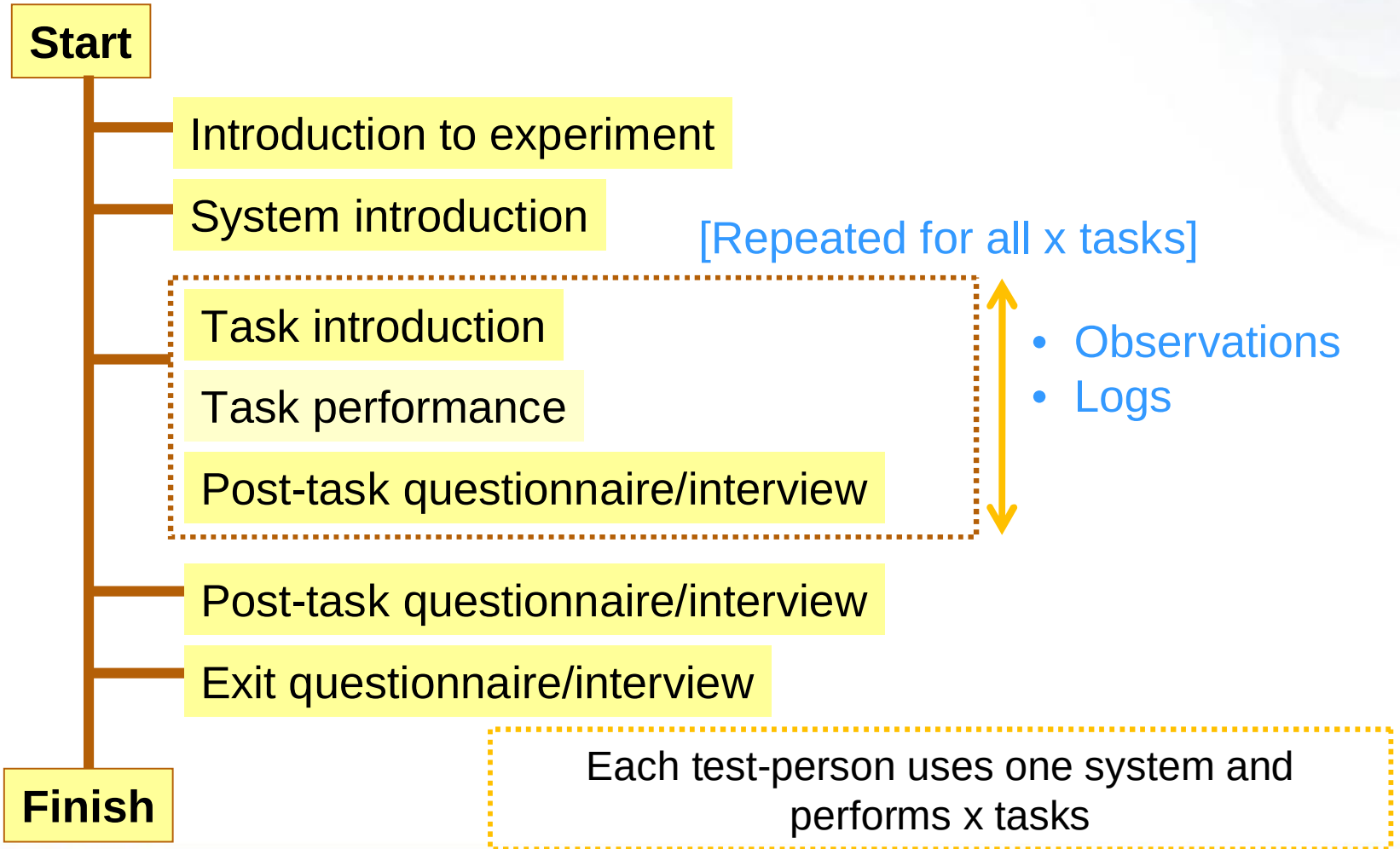
# Design Issues, 1

- Variables - nonstandard - careful operationalization
  - e.g., how to measure searching knowledge?
- Direct and indirect observables
  - users' explicit behaviors are directly measurable while their thoughts and feelings need indirect means
- Data collection methods
  - think-aloud; self-report and diaries; observation and logging; questionnaires and interviews
  - data collection may require days and months.
- Baselines and calibration
  - run the interactions with an experimental system and with a standard baseline system, which every research group uses

# Design Issues, 2

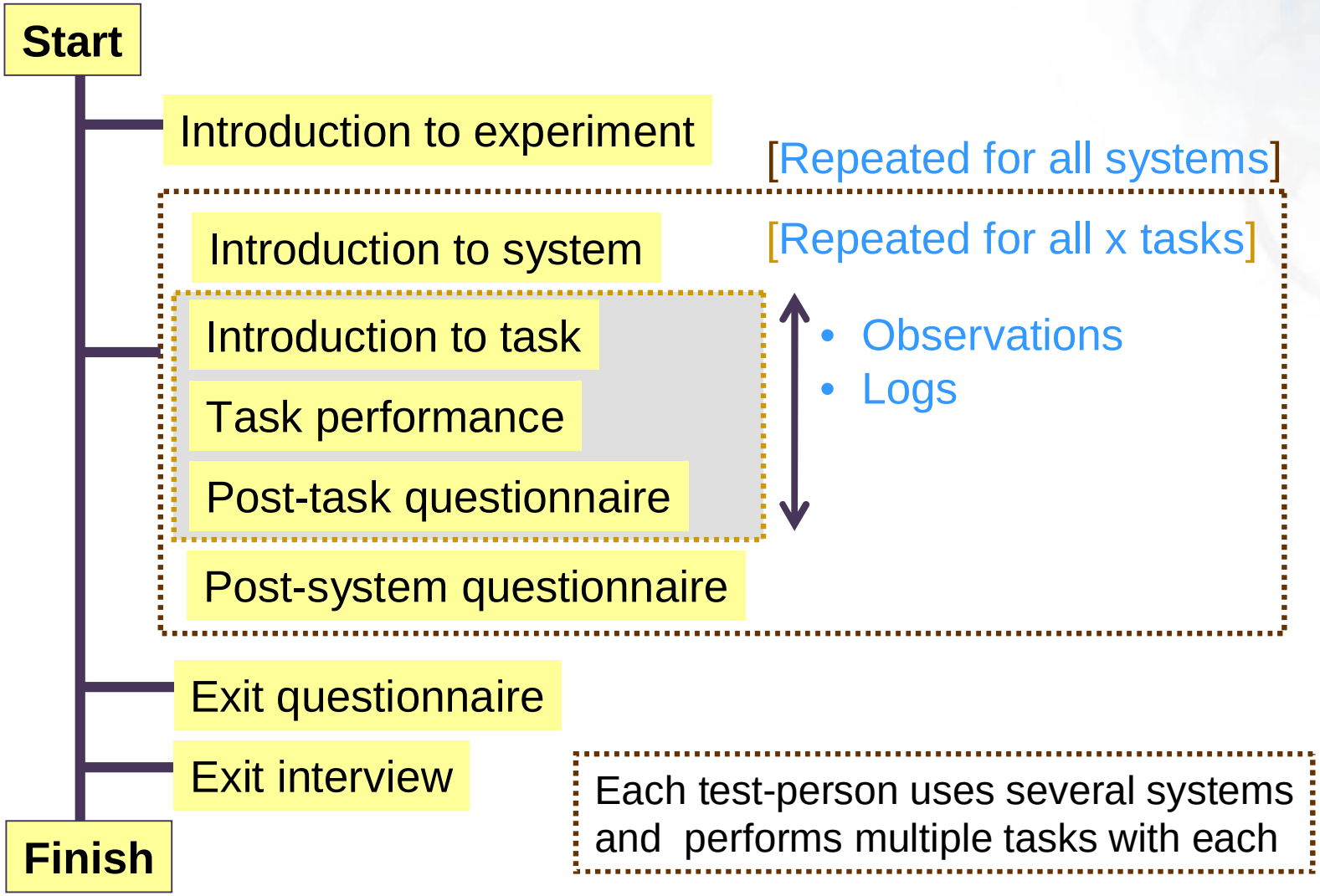
- Factorial designs for testing multiple conditions
  - a factorial design with 8 conditions: {sysA, sysB} x {noexp, expert} x {WorkT1, WorkT2}
  - many test persons required
- Neutralizing variation and learning effects
  - no repetition of the same test task within a short time frame
  - Latin Square design
- Timing – fatigue – protocols

# Data Collection - Protocol 1




Kelly (2009)

# Data Collection - Protocol 2




Kelly (2009)

INFIM

UNIVERSITY OF TAMPERE

# Limitations, Challenges

- Open and multi-faceted -> many different kinds of studies can be performed
- Practical limitations and costs
  - need to control several factors
  - the number of test persons
  - complexity of test protocols
- Grand challenge: evaluation of interaction
  - the contribution of various factors to the overall goals

# 5. Operational Systems Evaluation

- Real-life IR settings:
  - standardization of evaluation disappears
  - the control of evaluation designs increasingly difficult
  - no single experimental design to follow
  - one may have to *give up experimentation* in evaluation entirely
  - but one needs to define the system being evaluated, its goals, and the evaluation criteria

# Operational Systems Evaluation, 2

- At the *system* end of evaluation
  - TREC/Cranfield approach possible in the collection of test requests, defining the collection, and obtaining relevance assessments
  - system = search engine deployed in some environment,
  - document collections indexed for the engine,
  - real users generating queries and often providing relevance assessments
  - goal -- find relevant documents? -- additionally efficiency?
  - obtaining relevance assessments may require new solutions (click data, crowd-sourcing)

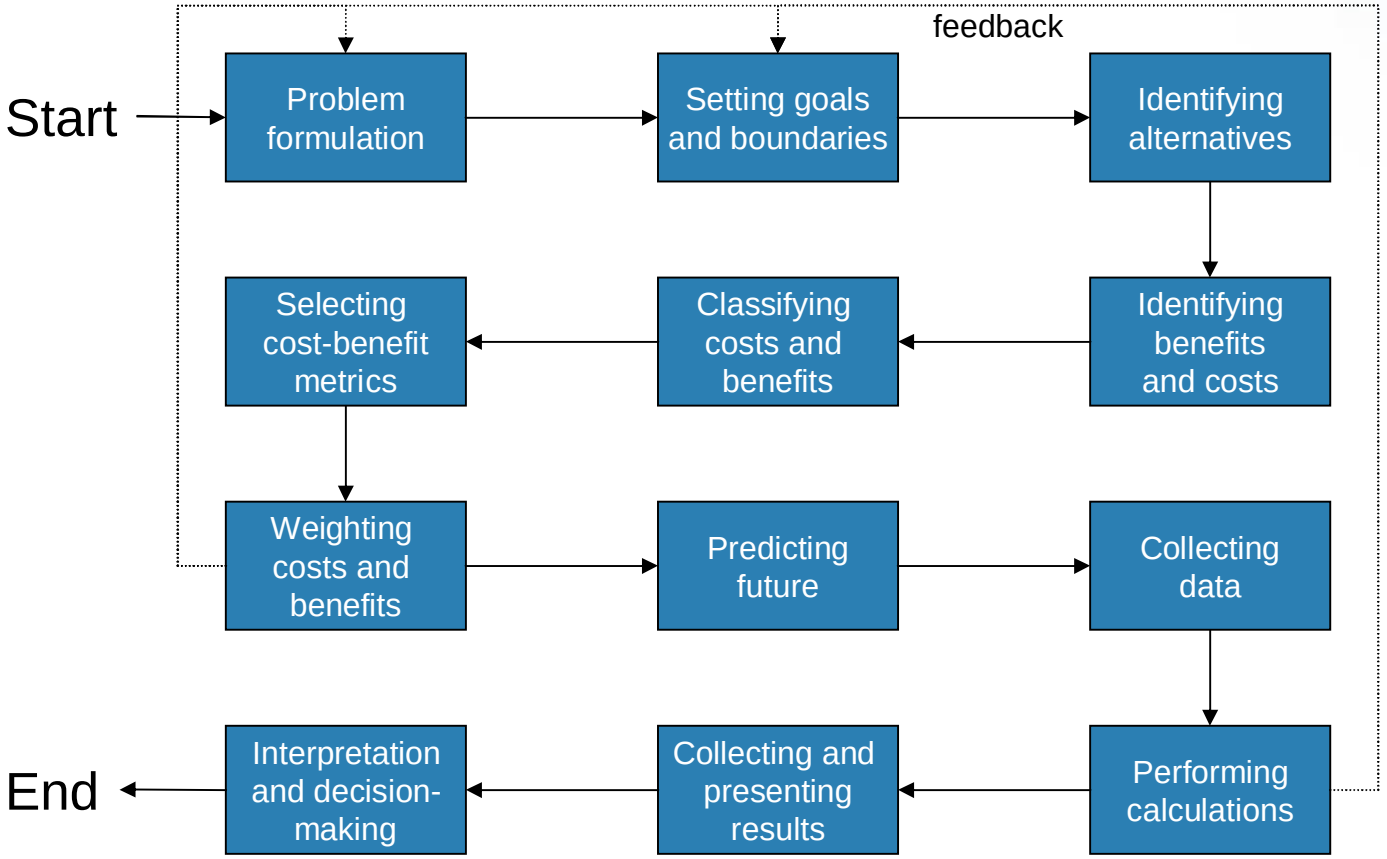
# Operational Systems User Evaluation, 2

- Real tasks, real users, and real needs in natural contexts
  - study goal -- understanding human behavior in searching?
  - system evaluated = databases, search engine(s), and humans using them due to their tasks and in their socio-organizational and system context
  - system goal -- get work (or leisure activity) done (well)?
  - additionally user satisfaction, efficiency (time), and the quality of results?
  - increasing openness allows a great number of evaluation designs but also puts many challenges on the generalizability of the findings

# Operational System Evaluation Designs

- Often a need to evaluate costs and benefits of alternative IR systems *comprehensively*
- Cost-benefit analysis process as a way to design operational IR system evaluation
  - serves both system and user oriented evaluation

# Comprehensive Evaluation Procedure



# Metrics, Data Collection

| Sample eva | |
|---|---|
| **Measure category** | **Sa** |
| Timelines | inc |
| Coverage | by |
| Document quality | co |
| Filtering capability | qu |
| Effort in using | co |
| Cost | su |
| Availability | by |

| Sample evaluation measures for **search engines** | |
|---|---|
| **Measure category** | **Sample measures** |
| Efficiency | throughput, response time |
| Filtering capability | query types and lengths, index types, MAP |
| Effort in using | comprehensibility of interface and query language, support in query formulation, presentation ways, tutoring |
| Effort in deploying | costs and efforts in setting up servers, connections, software, É |
| Engine flexibility | capability to scale up |
| Output quality | compliance to standards (XML), |
| Reliability | fault tolerance, recovery, back-ups |
| Cost | list price, pay-as-you-go policy |
| Availability | by time, geography, technology |

INFIM

UNIVERSITY OF TAMPERE

# Limitations, Challenges

- Operational systems evaluation for real life decision-making requires
  - greater effort
  - more varied data
- ... than user or system evaluations in the lab
- More costly, but more realistic, putting ‘IR science’ in perspective
- Generalizability? Controllability? Repeatability?

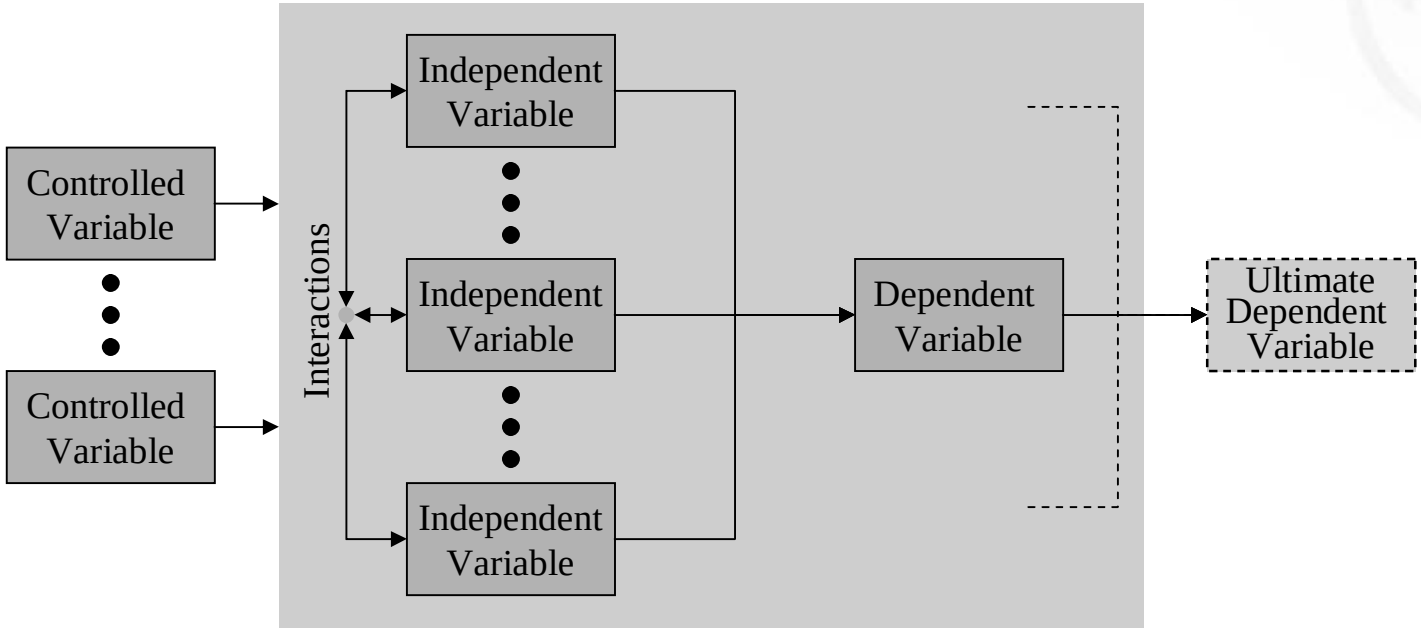
# 6. Beyond evaluation?

- The goals of a research area may be classified as
  - (a) theoretical understanding
  - (b) empirical description and explanation
  - (c) technology development
- Evaluation can serve all of these

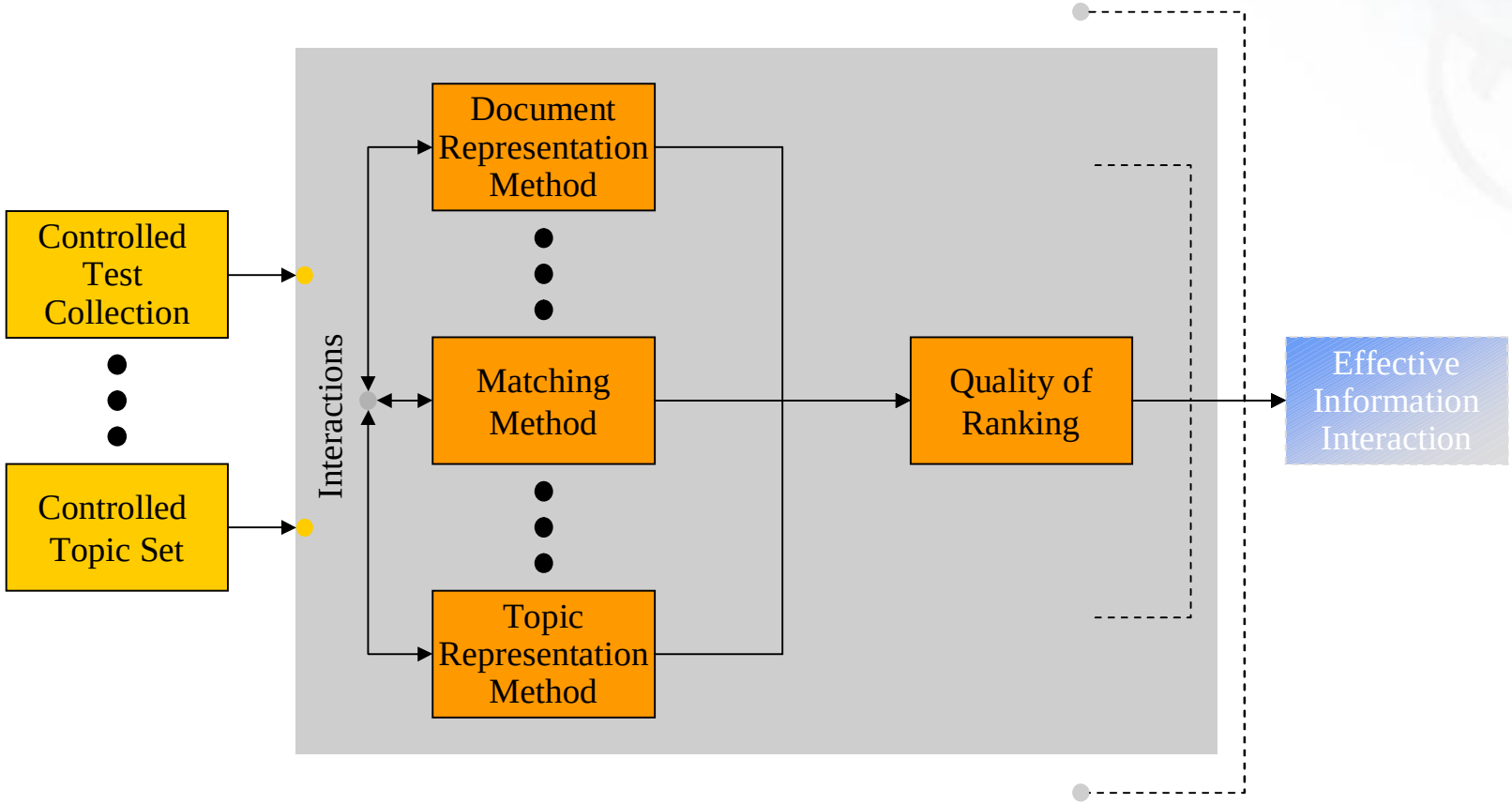
# Theory Development for/with Evaluation

- Evaluation requires a theory of the system being evaluated
  - where are its boundaries
  - which are the factors affecting its functioning and effectiveness
- Evaluation also helps to construct theory of the system
- Theories
  - systematic collections of theoretical and empirical laws
- Scientific laws
  - empirical laws express verified relationships between variables
- Variables -- or metrics
  - represent objects, properties or events
  - are used in hypotheses and laws

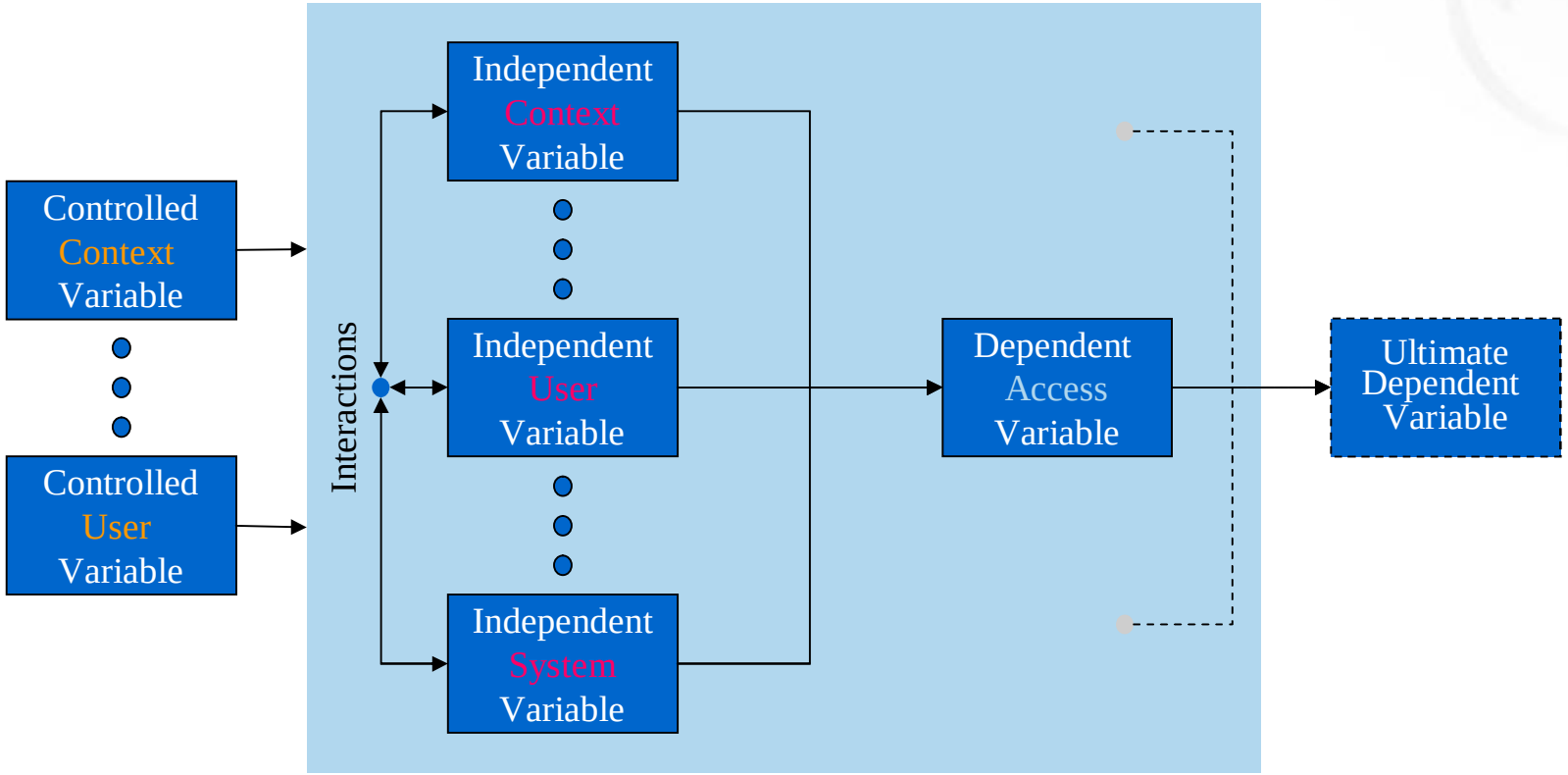
# Structure of Empirical Laws



# Explanation in IR



# User-based Evaluation Necessary



INFIM

UNIVERSITY OF TAMPERE

# When are user studies (not) useful?

- Useful when
  - informing design
  - guiding design
- And otherwise useless?
- Also useful when
  - focusing research on fruitful areas
  - advancing theory / accumulating knowledge on information interaction