

IR in Context of the User: Interactive IR Evaluation

Peter Ingwersen

Royal School of LIS

Denmark

pi@iva.dk – <http://www.iva.dk/pi>

Oslo University College, Norway

Agenda - 1

- **Introduction** (20 min)
 - Research Frameworks vs. Models
 - Central components of Interactive IR (IIR)
 - The Integrated Cognitive Research Framework for IR
- **From Simulation to 'Ultra-light' IIR** (20 min)
 - Short-term IR interaction experiments
 - Sample study – Diane Kelly (2005/2007)

Agenda - 2

- **Experimental Research Designs with Test persons** (25 min)
 - Interactive-light session-based IR studies
 - Request types
 - Test persons
 - Design of task-based simulated search situations
 - Relevance and evaluation measures in IIR
 - Sample study – Pia Borlund (2000; 2003b)

Agenda - 3

- **Naturalistic Field Investigations of IIR** (20 min)
 - Integrating context variables
 - Live systems & (simulated) work tasks
 - Sample Study – Marianne Lykke (Nielsen) (2001; 2004)
- **Wrapping up** (5 min)

Questions are welcome during the sessions

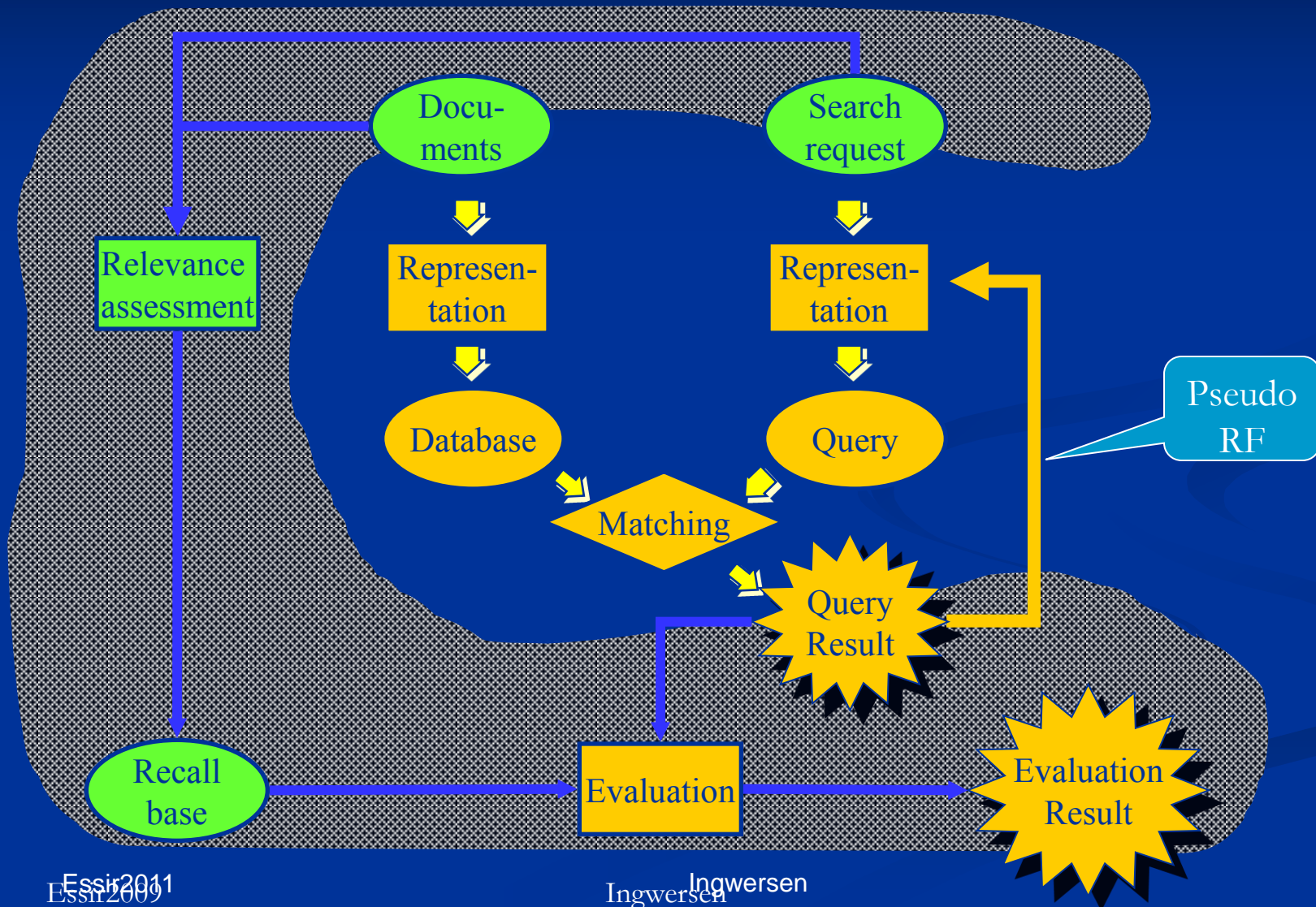
Advanced Information Retrieval / Ricardo Baetza Yeates & Massimo Melucci (eds.). Springer, 2011, p. 91-118.

Frameworks & Models – difference?

- Frameworks describe
 - Essential objects to study
 - The relationships of objects
 - The changes in the objects / relationships that affect the functioning of the system
 - Promising goals and methods of research
- Frameworks contain (tacit) shared assumptions
 - ontological, conceptual, factual, epistemological, and methodological
- The concept model
 - A precise (often formal) representation of objects and relationships (or processes) *within* a framework
 - Modeling may also in principle encompass human actors and organizations
- **Frameworks may lead to**
 - Research Designs, incl.
 - Research Questions; Experimental Setting; Methodology

The Lab. Research Framework

– cave with central variables (The Turn, 2005)

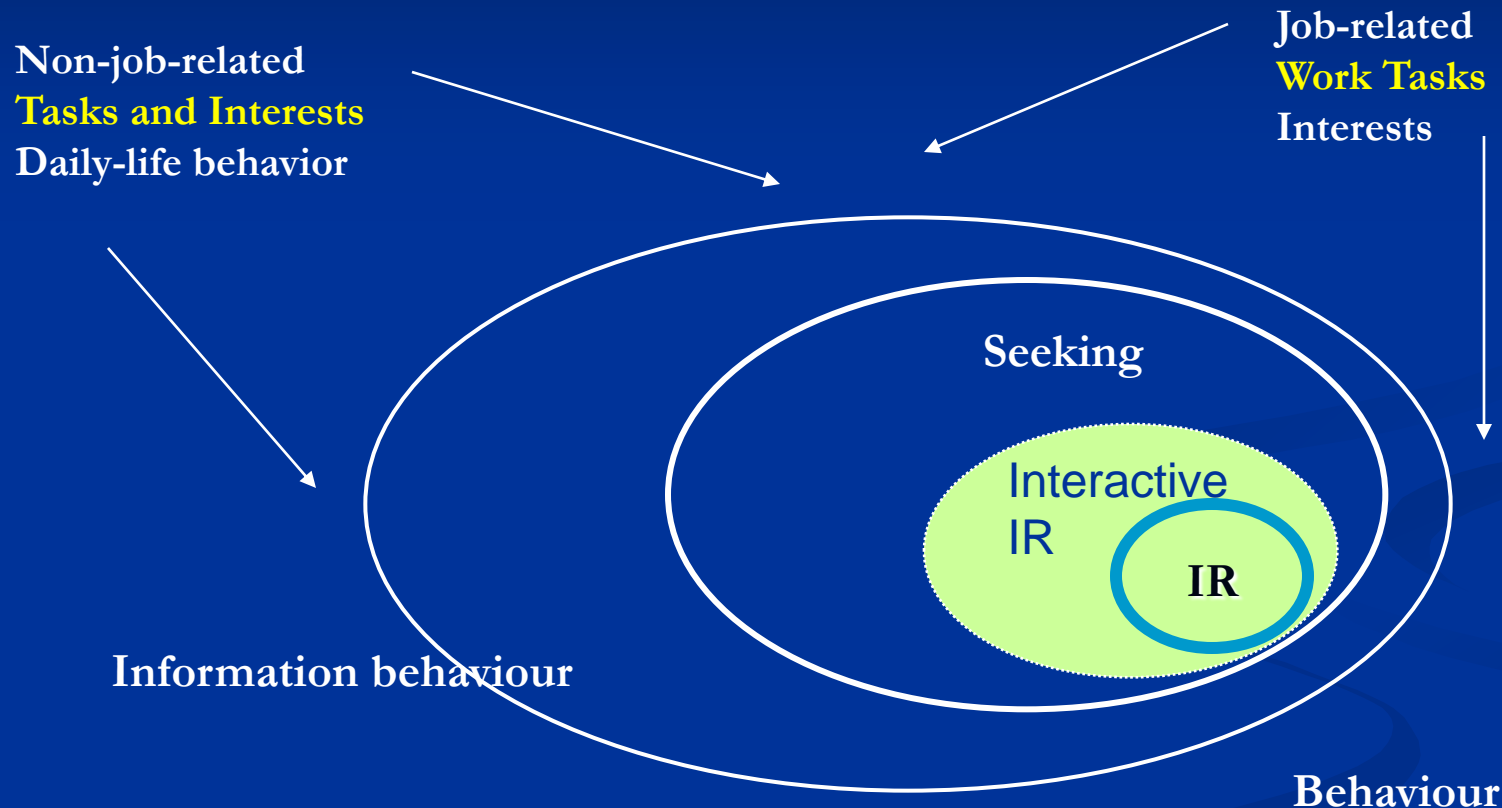


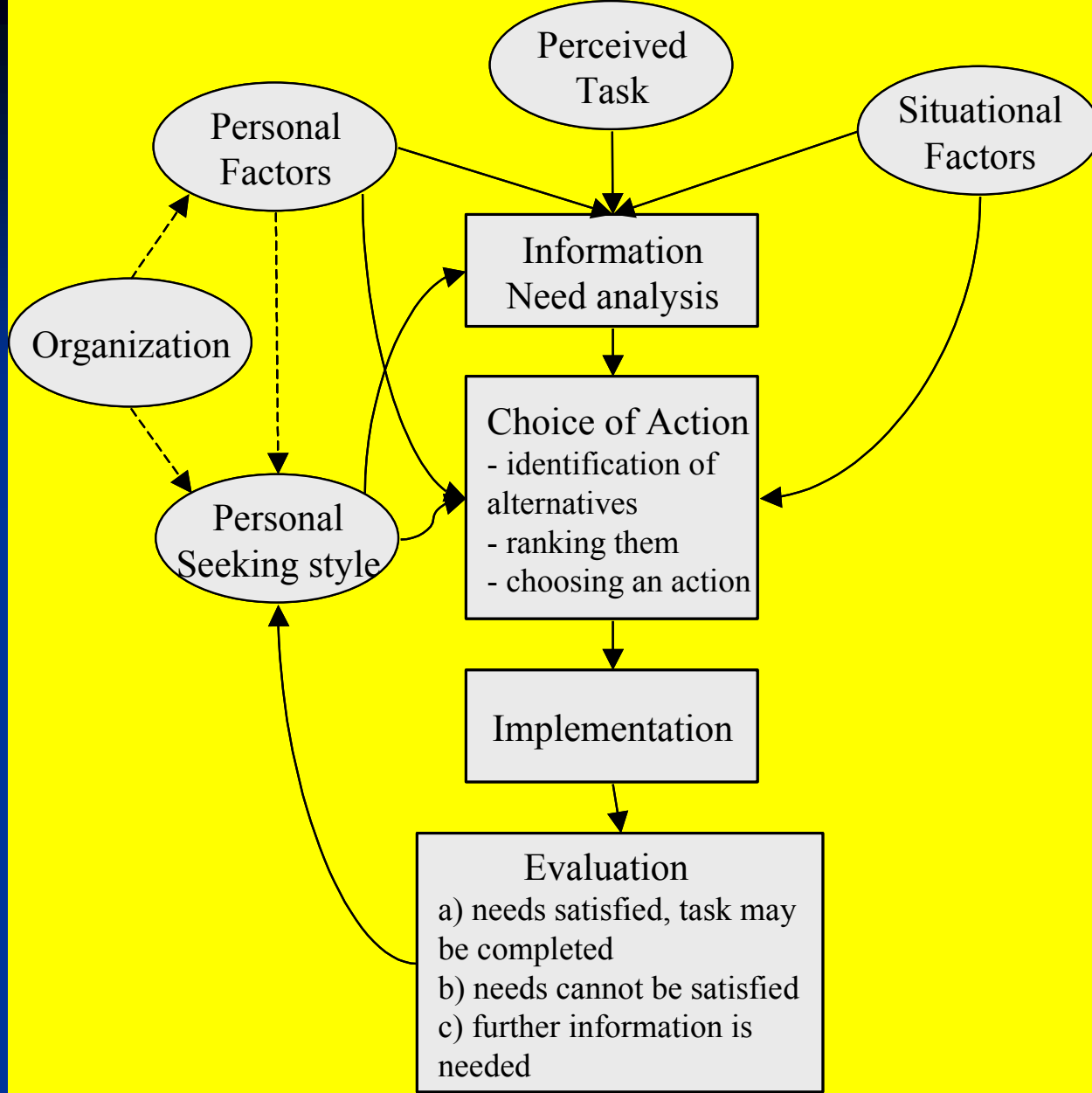
User-centered (contextual) MODELS

- **Examples** (in display order)
- **Wilson, 1999** (conceptual: Info. Behavior; Seek; IR)
- **Byström & Järvelin, 1995** (flow chart: Info. Seek)
- **Saracevic, 1996** (conceptual, stratified: IR)
- **Vakkari, 2000** (flow chart, Online Search; Relevance)
- **Wang & Soergel, 1998** (conceptual: Relevance Assessment Process & Criteria)

Information behaviour and IR

T. Wilson's Onion Model, 1999 - extended:

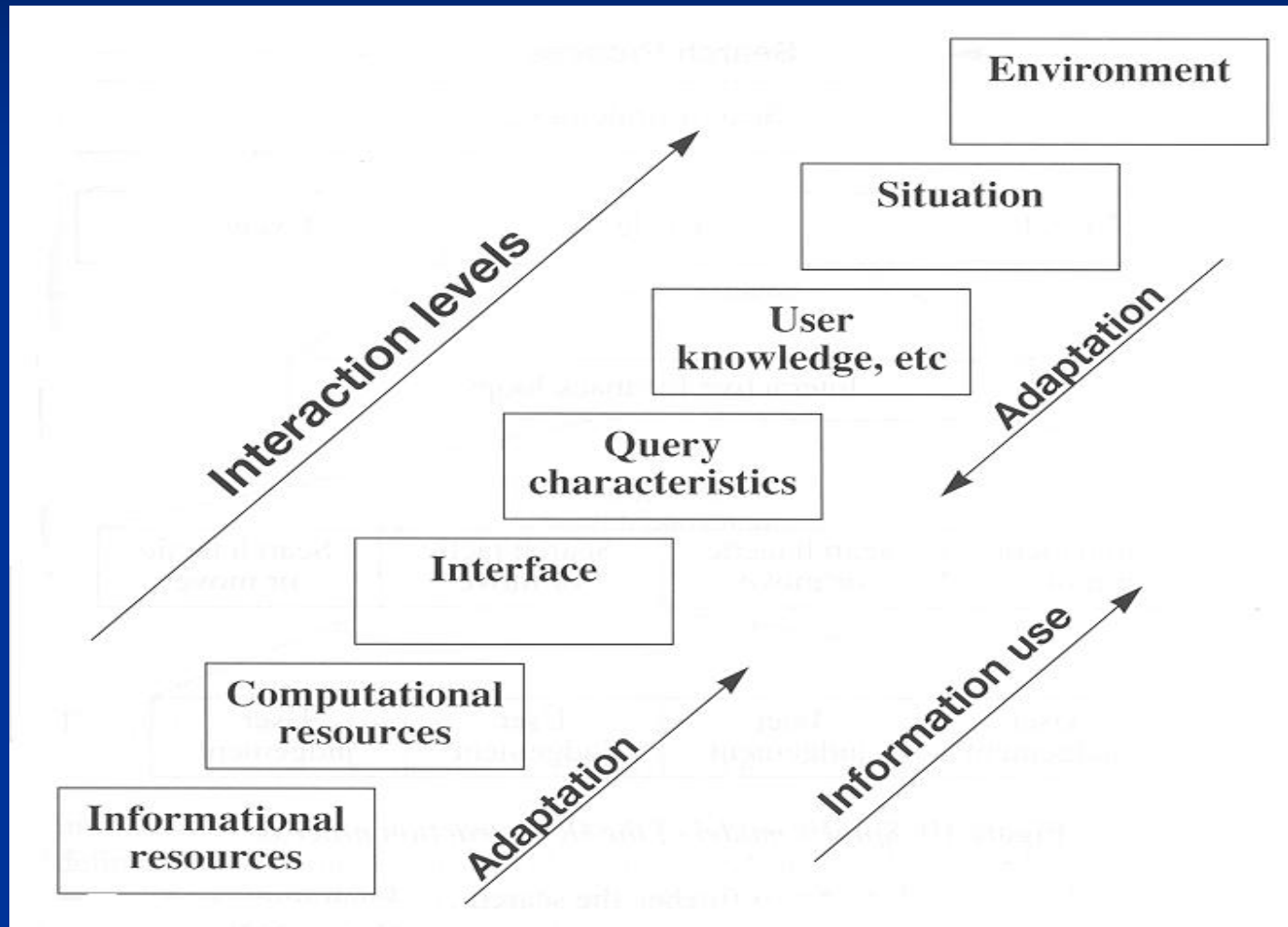




IS&R model,
1995: Bystöm &
Järvelin, fig. 2

(From: *The Turn*, p. 69)

Saracevic' stratified model for IIR (1996)



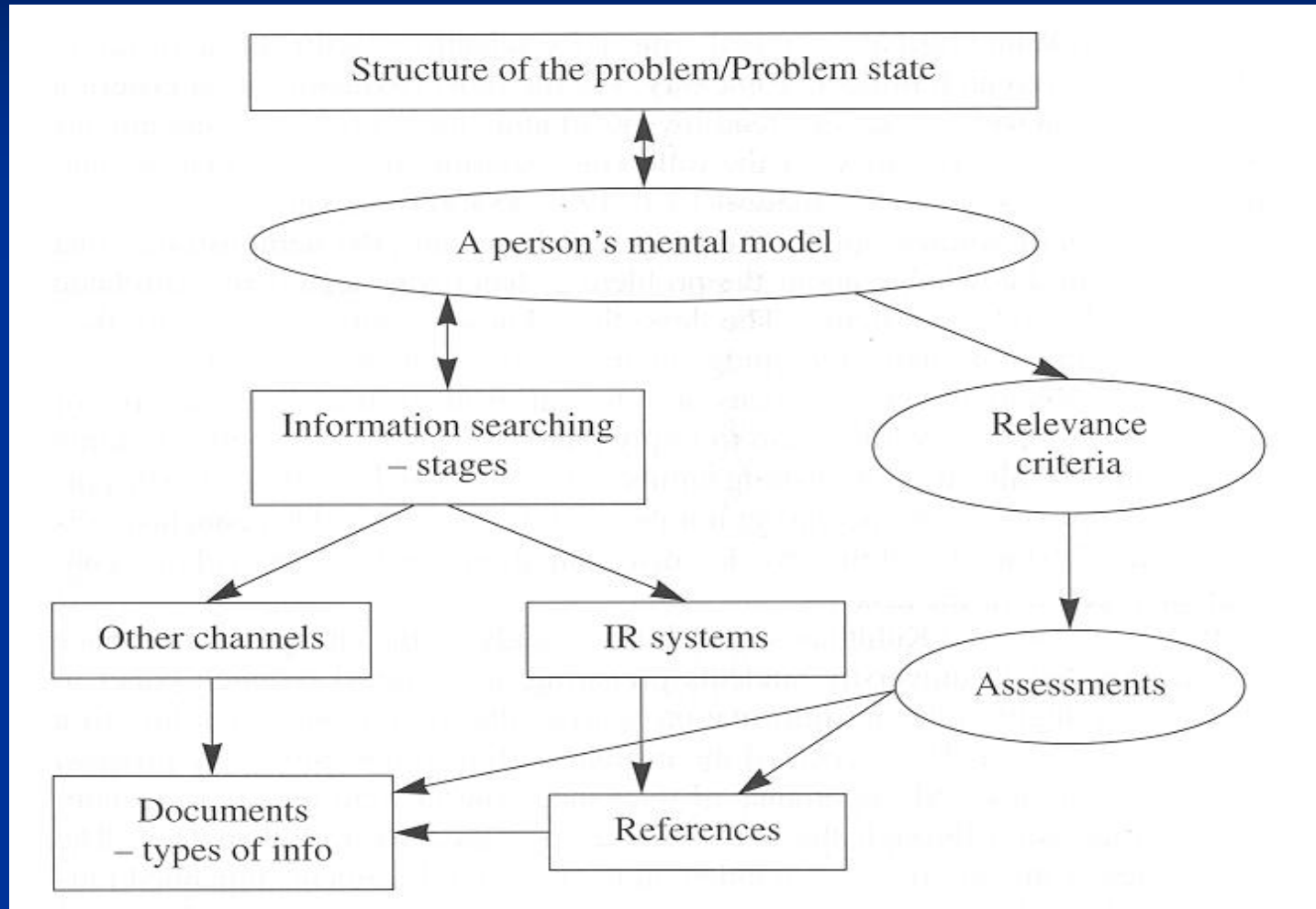
Wang & Soergel 1998



DIEs: Document Information Elements
Values: Document Values/Worth

(From: *The Turn*, p. 201)

IR and relevance in Seeking context – Seeking into IS&R: Vakkari 2000



Research Setting Types

- **Laboratory experiments** – no test persons, but
 - Simulations – Log analyses (not treated in presentation)
- **Laboratory study** – with test persons:
- **‘Ultra-light’** (short-term interaction: 1-2 retrieval runs)
– or **‘Interactive light’** (session-based multi-run interaction)
- **Field experiment** – *experimental* (artificial) *situation* in *natural setting* with test persons
- **Field study** – study of *natural* performance or behavior in *natural setting* with test persons
 - Longitudinal studies
- **Case study** – (qualitative) study with few test persons

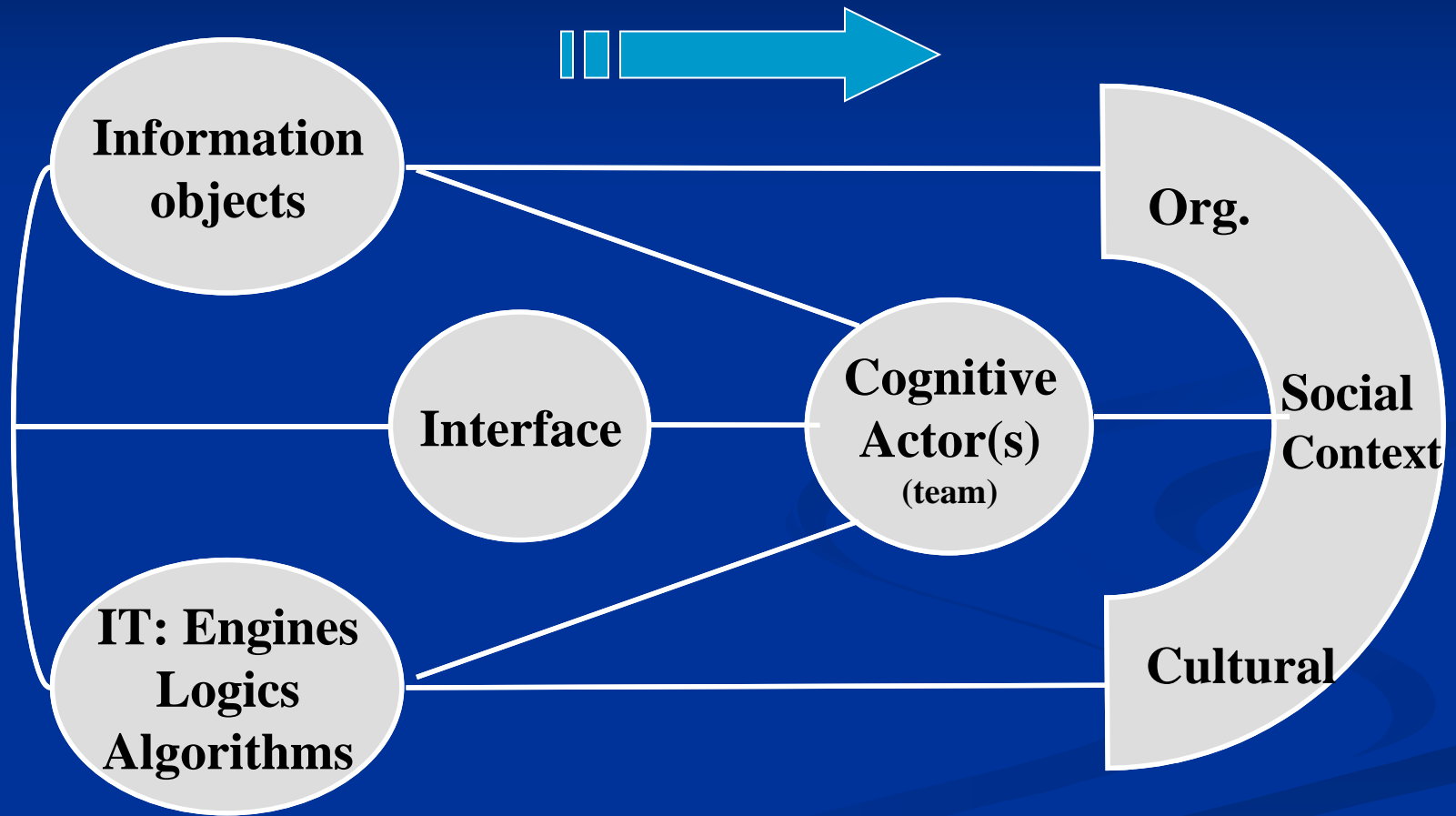
Variables involved in a test:

- **Independent** (the ‘cause’), e.g.,
 - Interface functionalities; Different IR models; Searcher knowledge
- **Dependent** (the ‘effect’), e.g.,
 - Performance measures of output (recall/prec.; CumGain; usability)
- **Controlled** (held constant; statistically neutralized; randomized):
 - Database; Information objects
 - Search algorithms
 - Simulated work task situations – Assigned TREC topics
 - Test persons
- **Hidden variables (Moderating or Intervening)**, e.g.,
 - Variation of test persons’ levels of experience ...!!! – see the Integrated Research Framework for IR

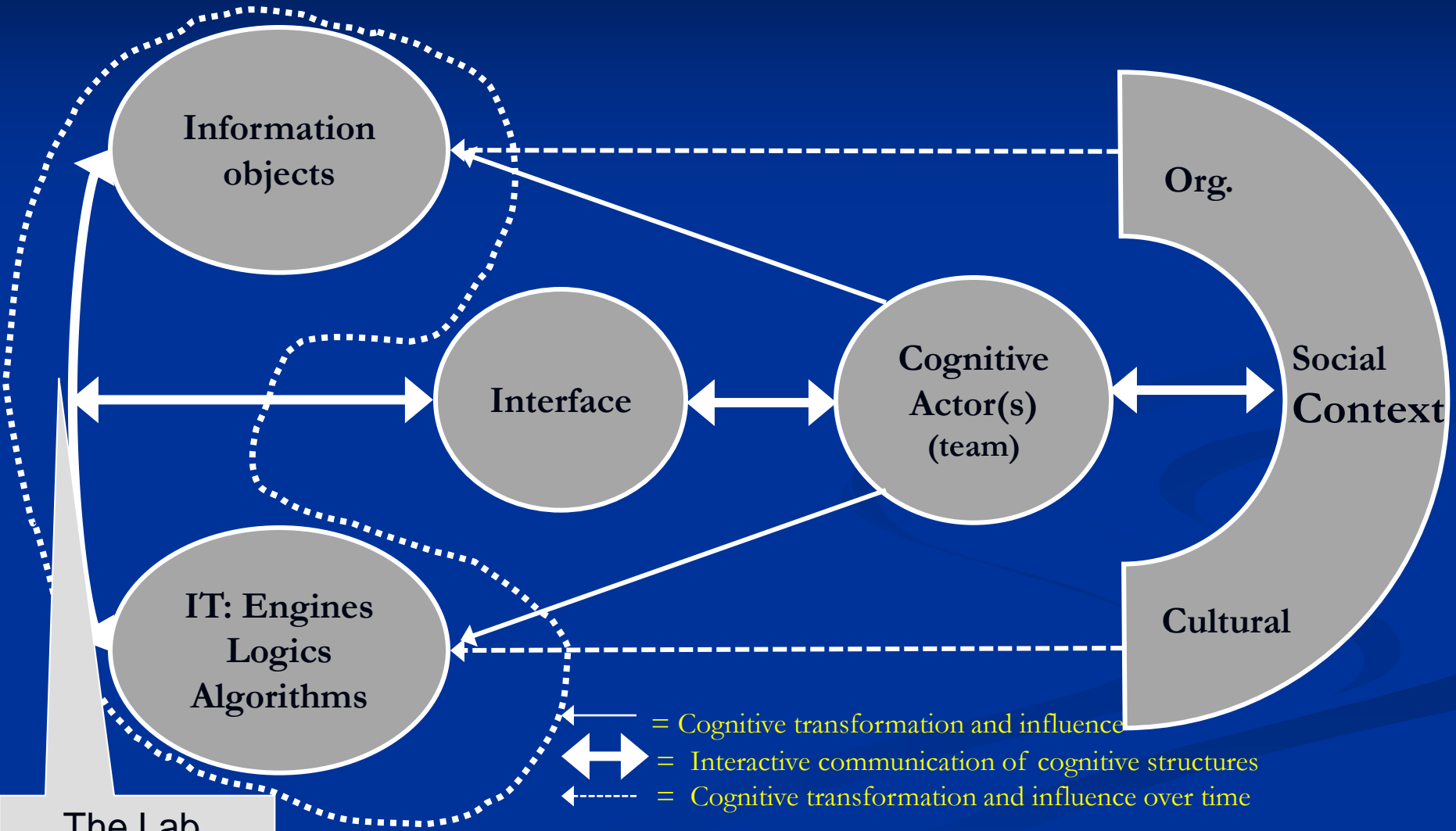
Agenda - 1

- ✓ **Introduction to Tutorial** (20 min)
 - ✓ Research Frameworks vs. Models
 - Central components of Interactive IR (IIR)
 - The Integrated Cognitive Research Framework for IR
- **From Simulation to ‘Ultra-light’ IIR** (20 min)
 - Short-term IR interaction experiments
 - Sample study – Diane Kelly (2005/2007)

Central Components of Interactive IR – the basic integrated framework



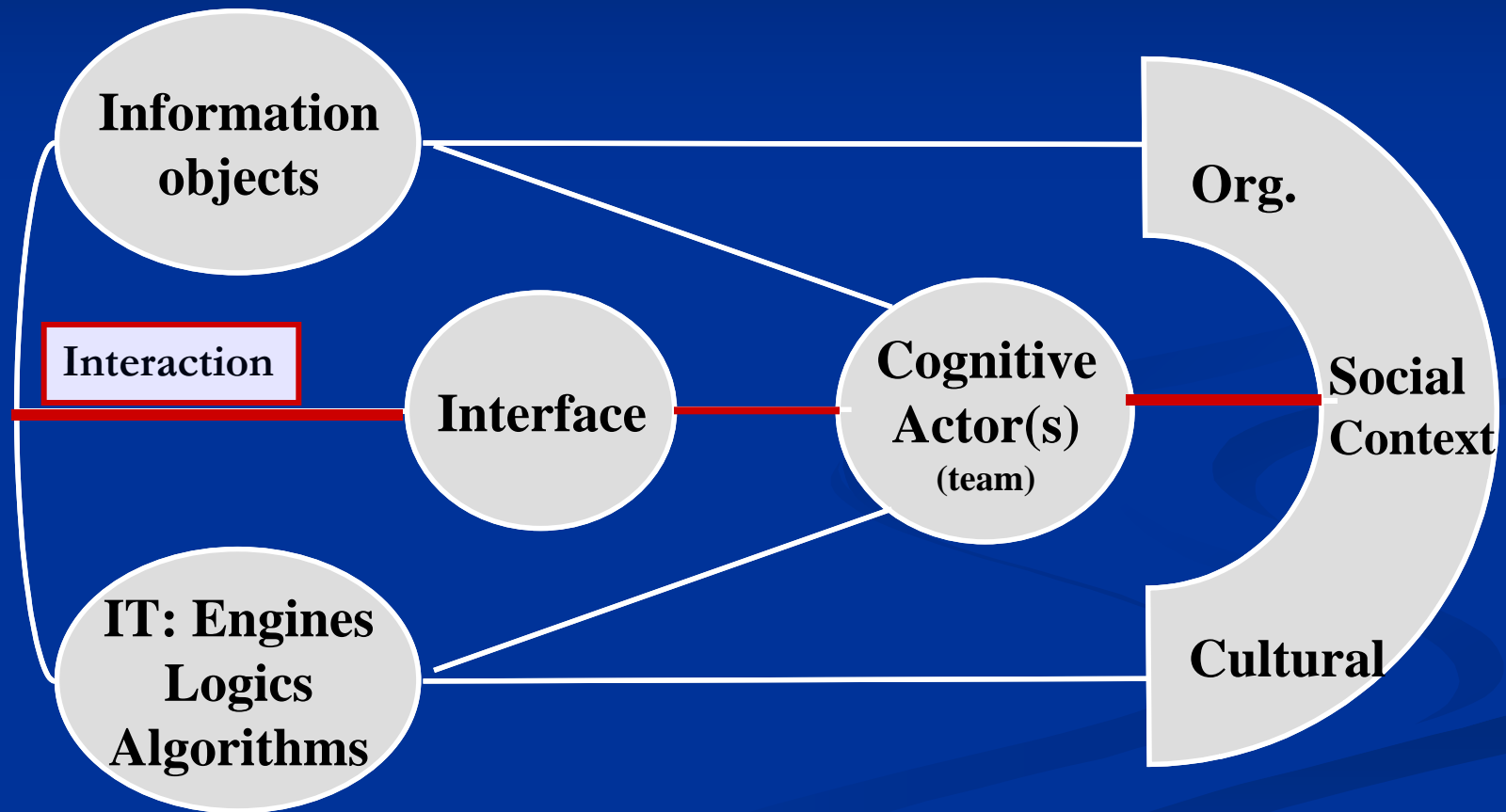
Central Components of Interactive IR – the basis of the integrated framework



The Lab.
Framework

Dimensions and Range of Variables in the Integrated IIR framework:

9 dimensions from 6 components



Categories of Dimensions in the Cognitive Research Framework

1. Natural work task dimension
 2. Natural search task dimension
 3. Actor characteristics dimension
 4. Perceived work task dimension
 5. Perceived search task
 6. Document dimension
 7. Algorithmic search engine dimension
 8. Algorithmic interface dimension
 9. Access and interaction dimension
- Socio-org. task dimensions
- Actor dimensions
- Each containing multiple variables
- Algorithmic dimensions
-
- The diagram illustrates the categorization of nine dimensions into three groups. Dimensions 1 and 2 are grouped as 'Socio-org. task dimensions'. Dimensions 3, 4, 5, and 6 are grouped as 'Actor dimensions', with a callout box pointing to this group stating 'Each containing multiple variables'. Dimensions 7, 8, and 9 are grouped as 'Algorithmic dimensions'.

Natural Work Tasks (WT) & Org	Natural Search Tasks (ST)	Actor	Perceived Work Tasks	Perceived Search Tasks
WT Structure	ST Structure	Domain Knowledge	Perceived WT Structure	Perceived Information Need Content
WT Strategies & Practices	ST Strategies & Practices	IS&R Knowledge	Perceived WT Strategies & Practices	Perceived ST Structure/Type
WT Granularity, Size & Complexity	ST Granularity, Size & Complexity	Experience on Work Task	Perceived WT Granularity, Size & Complexity	Perceived ST Strategies & Practices
WT Dependencies	ST Dependencies	Experience on Search Task	Perceived WT Dependencies	Perceived ST Specificity & Complexity
WT Requirements	ST Requirements	Stage in Work Task Execution	Perceived WT Requirements	Perceived ST Dependencies
WT Domain & Context	ST Domain & Context	Perception of Socio-Org. Context	Perceived WT Domain & Context	Perceived ST Stability
		Sources of Difficulty		Perceived ST Domain & Context
		Motivation & Emotional State		

Variables with values

Natural Work Tasks (WT) & Org	Natural Search Tasks (ST)	Actor	Perceived Work Tasks	Perceived Search Tasks
WT Structure	ST Structure	Domain Knowledge	Perceived WT Structure	Perceived Information Need Content
WT Strategies & Practices	ST Strategies & Practices	IS&R Knowledge	Perceived WT Strategies & Practices	Perceived ST Structure/Type
WT Granularity, Size & Complexity	ST Granularity, Size & Complexity	Experience on Work Task	Perceived WT Granularity, Size & Complexity	Perceived ST Strategies & Practices
WT Dependencies	ST Dependencies	Experience on Search Task	Perceived WT Dependencies	Perceived ST Specificity & Complexity
WT Requirements	ST Requirements	Stage in Work Task Execution	Perceived WT Requirements	Perceived ST Dependencies
WT Domain & Context	ST Domain & Context	Perception of Socio-Org. Context	Perceived WT Domain & Context	Perceived ST Stability
		Sources of Difficulty		Perceived ST Domain & Context
		Motivation & Emotional State		

Document and Source	IR Engines IT Component	IR Inter-faces	Access and Interaction
Document Structure	Exact Match Models	Domain Model Attributes	Interaction Duration
Document Types	Best Match Models	System Model Features	Actors or Components
Document Genres	Degree of Doc. Structure and Content Used	User Model Features	Kind of Interaction and Access
Information Type in Document	Use of NLP to Document Indexing	System Model Adaption	Strategies and Tactics
Communication Function	Doc. Metadata Representation	User Model Building	Purpose of Human Communication
Temporal Aspects	Use of Weights in Doc. indexing	Request Model Builder	Purpose of System Communication
Document Sign Language	Degree of Req. Structure and Content Used	Retrieval Strategy	Interaction Mode
Layout and Style	Use of NLP to Request Indexing	Response Generation	Least effort Factors
Document Isness	Req. Metadata Representation	Feedback Generation	-
Document Content	Use of Weights in Requests	Mapping ST History	
Contextual Hyperlink Structure		Explanation Features	
Human Source (see Actor)		Transformation of Messages Scheduler	

Number of variables using the Framework

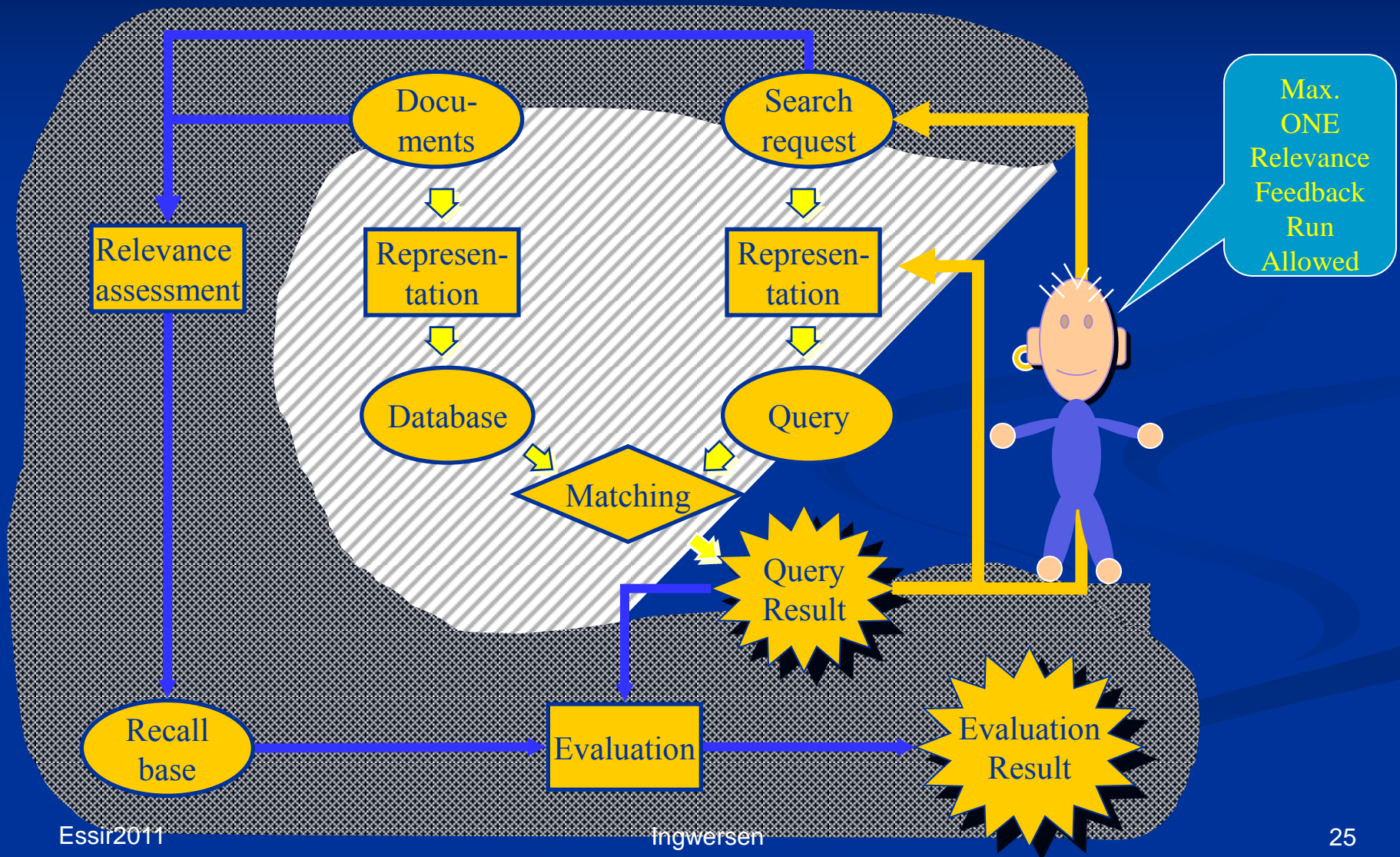
- Maximum application of **three independent** variables simultaneously!
- Can be done **in pairs** – and by total control of binary values of variables, e.g.
 1. **Interface function X**, value a/b
 2. **Personal IR expertise**, values none/much
 3. **In domain Z, work task type**: *routine* – but Rich/Poorly defined

There are many relevant combinations made from the Framework!

Agenda - 1

- ✓ **Introduction to Tutorial** (20 min)
 - ✓ Research Frameworks vs. Models
 - ✓ Central components of Interactive IR (IIR)
 - ✓ The Integrated Cognitive Research Framework for IR
- **From Simulation to ‘Ultra-light’ IIR** (20 min)
 - Short-term IR interaction experiments
 - Sample study – Diane Kelly (2005/2007)

IR interaction 'Ultra-light' – *short-term IIR*



Lab IR - 'ultra light' interaction

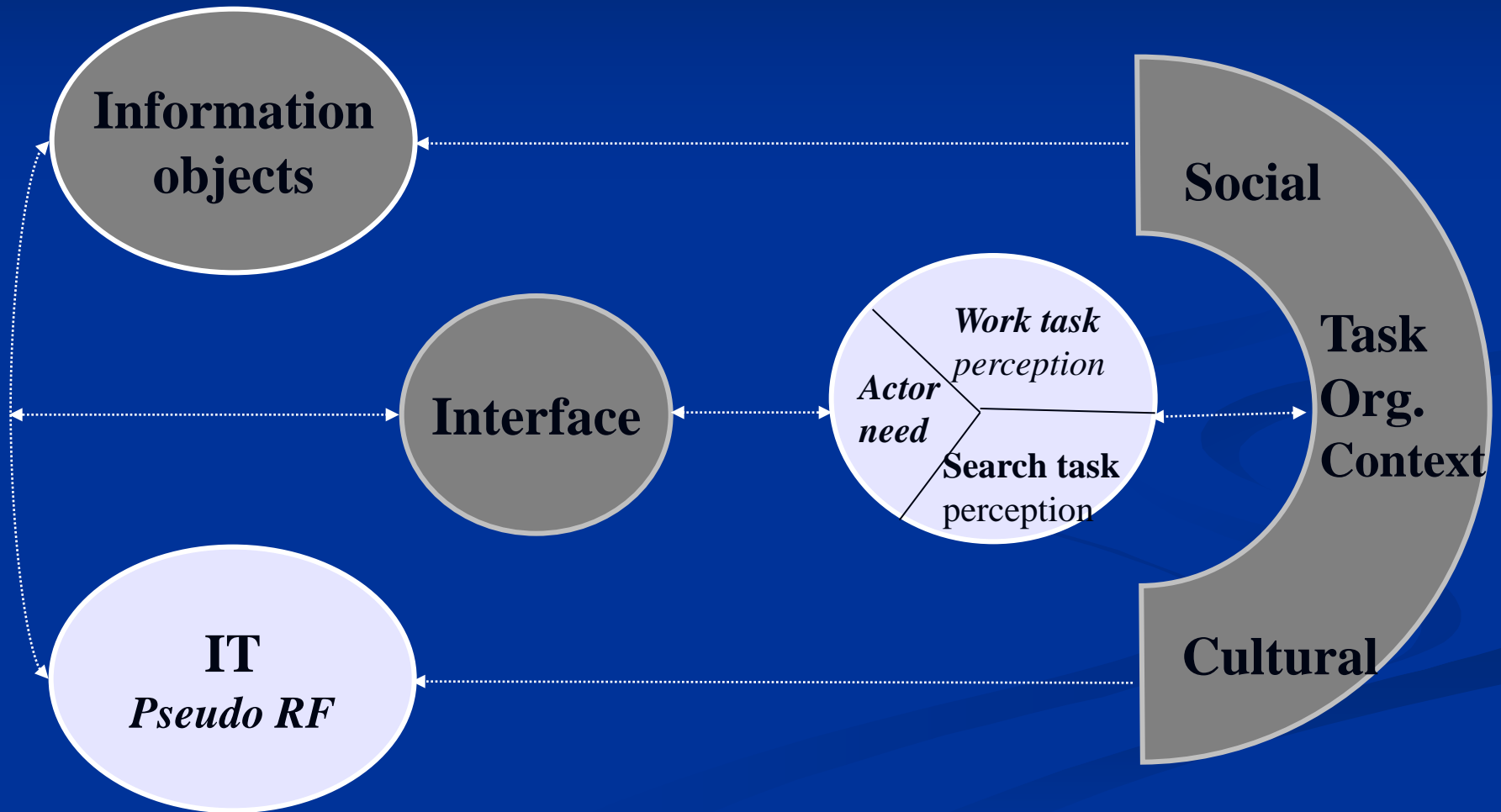
- In this 1-2 run setting we have **two ways** of measuring performance:
 1. By **assessor in pre-existing test collection** (as in TREC with unrealistically long delay between run and assessment – but equal to all).

Assessments may be applied to judging **second run results** (made by the test persons) – like using pseudo RF after first run

Lab IR - 'ultra-light' interaction

2. By **all test searchers** of the **first run results** of the same query session (Secondary run assessments of results used if first run done by pseudo RF).
 - One may pool performance scores across the set of **assigned requests** (topics) or **simulated tasks** given in experiment – because of the max. two runs:
 - **Good: No learning effects can influence the experiment**
 - or the same effects appear as when using TREC assessors
 - That is why this setting is 'interactive ultra-light'
 - Graded relevance assessments possible
 - Can be used **OUTSIDE** traditional test collections!
 - **Bad: Quite few documents are commonly assessed for relevance per test searcher** (or vary much!)
 - The setting is **limited in realism (only 2 runs)**

Interactive 'Ultra-light' experiment. Research question concerned with **variables** from **actor & IT**



IIR Interactive ‘Ultra-light’ sample

- **Kelly, D., Dollu, V.D. & Xin Fu** (2005). The loquacious user: A document-independent source of terms for query expansion. In: *Proceedings of the 28th Annual ACM-SIGIR Conference on Research and Development in Information retrieval*: 457-464. (also as IP&M article in 2007, 43(1): 30-46) – extended in IPM, 2007.
- **RQ:** Does multi-evidence of users’ information need situation improve retrieval performance through query expansion, compared to initial request and pseudo relevance feedback?

Research setting

- 13 test persons supplied ...
 - 45 natural 'topics' to HARD TREC (title and description) and
 - Relevance assessments
- HARD TREC collection; Lemur system (BM25)
 1. Topic title+description run by Lemur (bag-of-words) one run; **serves as baseline (BL)**.
 2. Pseudo RF modes (top-5; top-10;...) run on top of BL
 3. Each test person asked 4 questions via a form:

Research setting 2

- (Q1) state the times in the past he/she had searched that topic;
- (Q2) describe what he/she *already knows* about the topic (knowledge state);
- (Q3) state *why* he/she wants to know about the topic; and
- (Q4) add *any keywords* that further describe the topic.

Research setting 3

- Controlled variables: BM 25; 45 topics; HARD coll.
- Independent variables:
 1. Pseudo RF variations – on top of baseline (BL)
 2. Q2-Q4 combinations (term weights) – on top of BL
- Dependent var.: MAP – statistical significance test
- **RESULTS, yield of different words (mean):**
- Baseline : 9.33 – Q2: 16.18 – Q3: 10.67 – Q4: 3.3
- **Single Q-forms outperform BL**
- **Q2-Q3 (and Q2-Q4) combined outperform BL plus any pseudo RF**
- **Performance increases with query length.**

Summary: IIR 'Ultra-Light'

■ Strength:

- Easy to apply existing test collections, with ...
- Relevance assessments existing *a priori* (as in TREC or INEX)
- New relevance assessments possible – with graded assessments and over many assessors (test persons): weighted assessments
- Can lead to more solid interactive investigations in later studies

■ Weakness:

- Are all variable values known?? (people means hidden ones!)
- 'Ultra-light' IIR is limited in realism (1-2 iterations; context features hardly in play)
- Limited number of documents assessed (per test person)

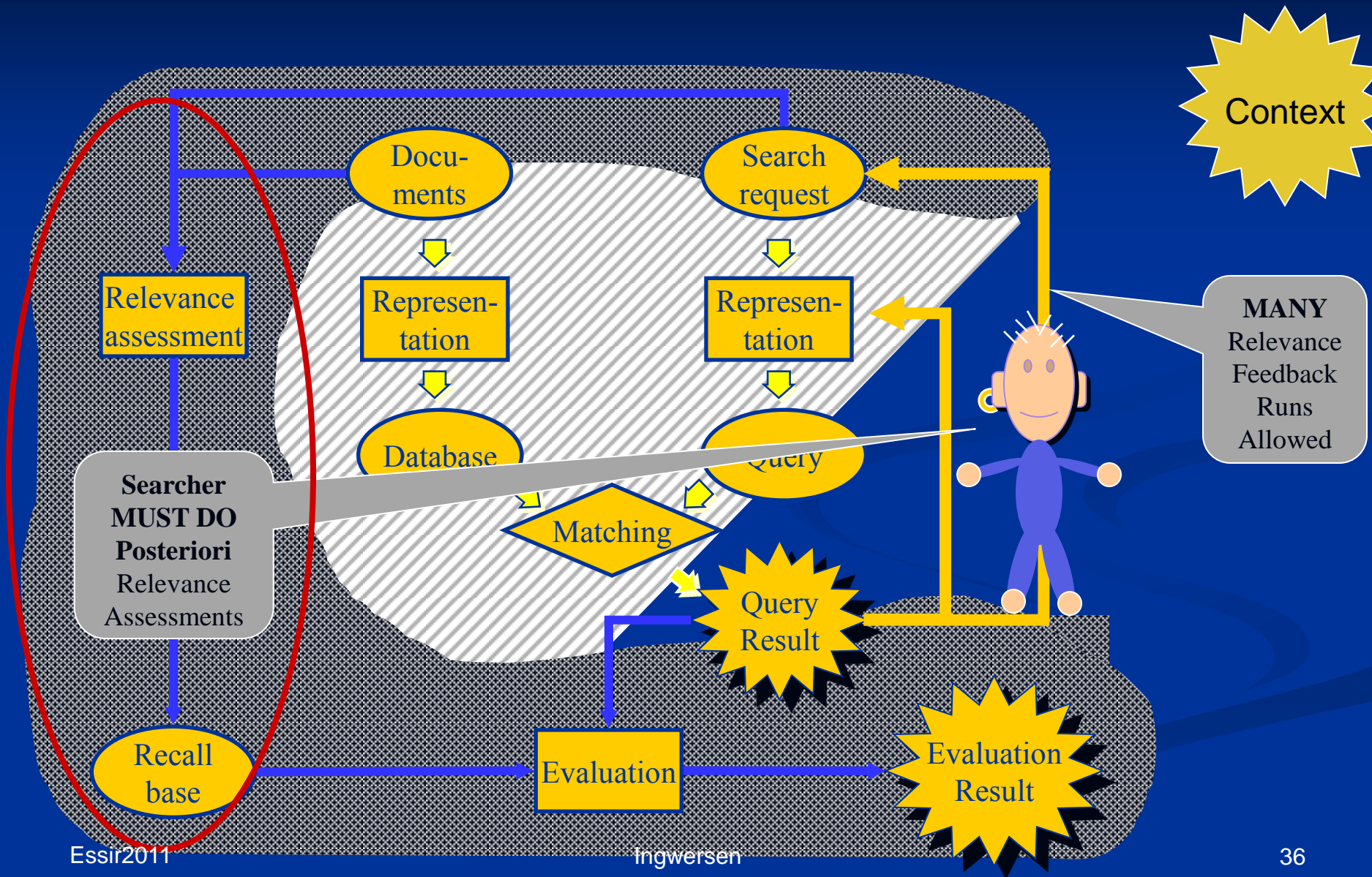
Agenda - 1

- ✓ **Introduction to Tutorial** (20 min)
 - ✓ Research Frameworks vs. Models
 - ✓ Central components of Interactive IR (IIR)
 - ✓ The Integrated Cognitive Research Framework for IR
- ✓ **From Simulation to 'Ultra-light' IIR** (20 min)
 - ✓ Short-term IR interaction experiments
 - ✓ Sample study – Diane Kelly (2005/2007)

Agenda - 2

- **Experimental Set-ups with Test Persons** (25 min)
 - Interactive-light session-based IR studies
 - Request types
 - Test persons
 - Design of task-based simulated search situations
 - Relevance and evaluation measures in IIR
 - Sample study – Pia Borlund (2000; 2003b)

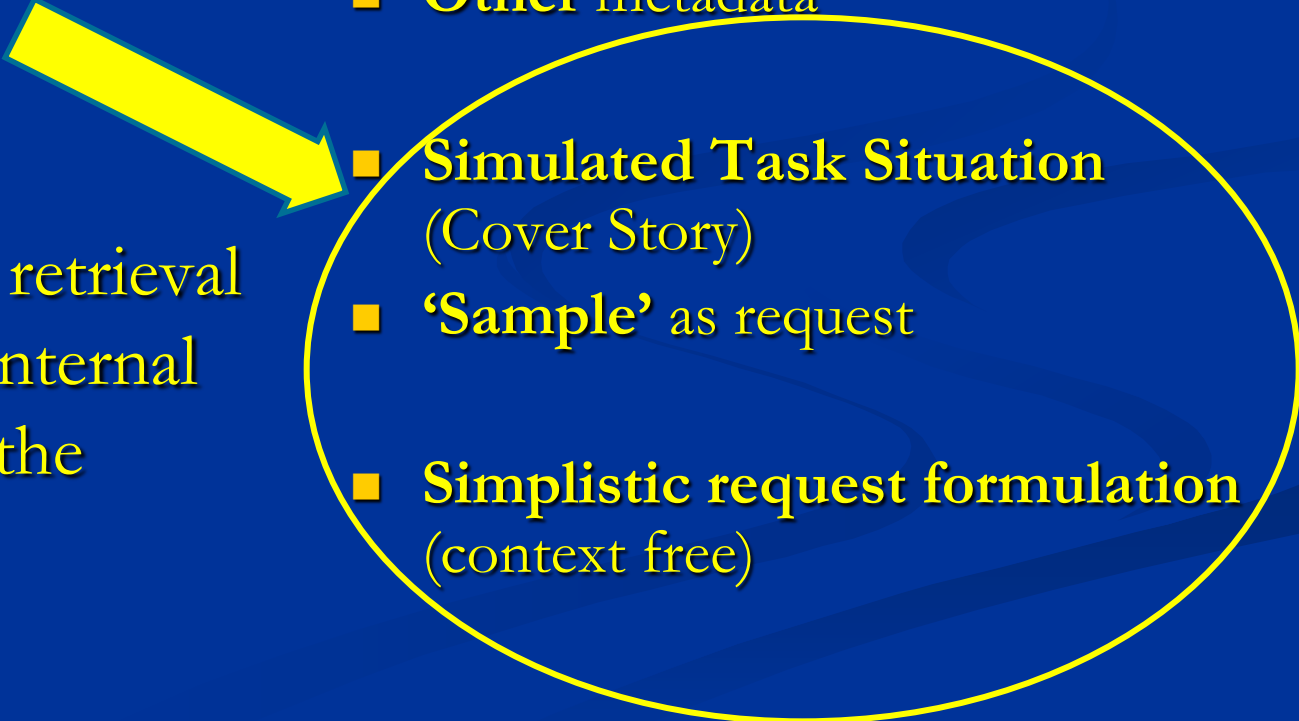
IR interaction 'Light'



Data Collection Means

- Observation
- Thinking (talking) aloud - Introspection
- Eye-tracking
- Critical incidence
- Questionnaires
- Interviews (structured; open-ended; closed)
 - Post or/and pre-interviews
- Focus groups
- Diaries – Self reporting
- Logging and recording of behavior (system/client logs)
 - Assessments of relevance

Request Types in ('ultra') 'light' IIR

- **Natural request/** real need of test person - or
 - **Assigned** to test person
 - **Topical** (as TREC 'topics')
 - **Factual**
 - **'Known Item'**
 - **Other** metadata
 - **'Query'** is the retrieval mechanism's internal translation of the **REQUEST**
 - **Simulated Task Situation** (Cover Story)
 - **'Sample'** as request
 - **Simplistic request formulation** (context free)
- 

Number of Test Persons

- Number depends on **goal** of research & no. of **variables**:
- **Behavioral** field study/experiment: **many persons** (>30 test persons required – and **some** (2-3) search jobs, to be statistically valid)
- **Performance-like** field experiment: **many search jobs** per person (4-10) – but **less** (~ 15) test persons required.
 - Note: Sanderson et al. paper: IIX 2005 on no. of topics necessary for statistical validity:
> 60!! (if applying MAP on top-15)
- The best design: always > 25 persons

Test Persons ...

- In order to be statistically significant, or really indicative, each cell in the cross tabulation result matrix should contain **25-30 units** (rule of thumb).
- Example (performance/evaluation goal with **3 independent (binary) variables, done in pairs:**
 **$2 \times 2 \times 2 = 8$ cells \times 30 units = 240 units in total):
 - You have 2 x 10 test persons (doctors & med. stud.)
 - They need to do **12 search jobs per person** = 120 units per group over 2×2 additional variable values, for reasons of cross tabulations = $120 \times 2 = 240$ jobs!
 - or 2 x 20 persons doing 6 jobs each.**

Latin Square research design – *The Turn*, p. 253-254

system **X**

system **Y**

1: A, B, C

4: D, E, F

1: D, F, E

4: A, C, B

2: C, B, A

5: F, E, D

2: E, F, D

5: B, C, A

3: C, A, B

6: F, D, E

3: E, D, F

6: B, A, C

6 test persons (1-6);

6 real / simulated work tasks/ or assigned topics (A-F)

Agenda - 2

- ✓ **Experimental Set-ups with Test Persons** (25 min)
 - ✓ Interactive-light session-based IR studies
 - ✓ Request types
 - ✓ Test persons
 - Design of task-based simulated search situations
 - Relevance and evaluation measures in IIR
 - Sample study – Pia Borlund (2000; 2003a)

Simulated Work Task Situations

– or 'cover stories' – to trigger natural information needs (& requests)

Example from study on relevance assessments on the Web (See 'Ultra-Light' in Bibliography: Papaeconomou, 2008):

Beijing is hosting in 2008 (8th-24th August) the Olympic Games. A friend of yours, who is a big fan of the Olympic Games, wants to attend the events and asks you to join in this trip. You find this invitation interesting. You are not a big fan of the games but you always wanted to visit China, therefore you want to find information about the sightseeing in the city and the activities that the Chinese will offer during the games. Find for instance places you could visit, activities you could do in relation to the Chinese culture or in the spirit of the games.

Borlund IIR (2003b) evaluation package

Simulated situation: sim A

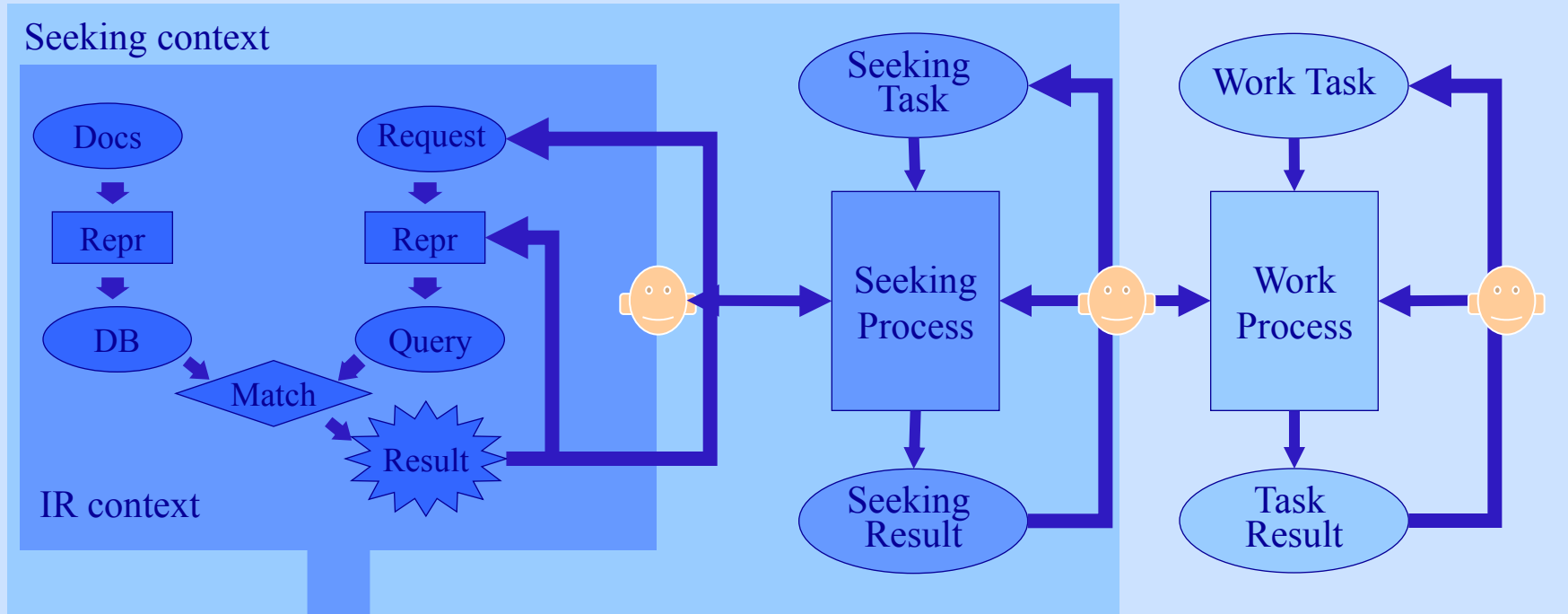
Simulated work task situation: After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications 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.

Integrated Framework and Relevance Criteria

Socio-organizational & cultural context

Work task context



Evaluation Criteria:

A: Recall, precision, efficiency, (quality of information/process)

B: Usability, Graded rel., CumGain; Quality of information/process

C: Quality of info & work task result; Graded R.

D: Socio-cognitive relevance; Social utility: rating; citations; inlinks;

Relevance & Evaluation Measures

The measurement of performance by use of **non-binary (graded, scaled or gliding)** based performance measures (generalized by Järvelin & Kekäläinen, 2002)

- Realistic assessment behaviour
- Indication of users' subjective impression of system performance and satisfaction of information need: **usability** (Hornbæk, 2006; Nielsen, 2006)
- Other measurements to be used on **Interaction Process**:
 - **Display time; No. of requests/queries; Visits & Downlods**
 - **Selection patterns; Views & clicks; Social utility assessments;**
 - **No. of documents assessed; Perceived ease of process; ...**

Agenda - 2

- ✓ **Experimental Set-ups with Test Persons** (40 min)
 - ✓ Interactive-light session-based IR studies
 - ✓ Request types
 - ✓ Test persons
 - ✓ Design of task-based simulated search situations
 - ✓ Relevance and evaluation measures in IIR
 - Sample study – Pia Borlund (2000; 2003a)

The Borlund Case (2000; 2003b)

- Research questions
 - 1) **Can simulated information needs substitute real information needs?**
 - **Hypothesis is: YES!**
 - 2) What makes a 'good' simulated situation with reference to semantic openness and types of topics of the simulated situations?

Experimental Setting:

- Data collection:** Financial Times (TREC data) and The Herald (current)
- Test system:** Full-text online system
Probabilistic based retrieval engine
- Test persons:** 24 university students
(undergraduates and graduates)
- From:** Computing, engineering, psychology, geography, English history, etc.
- Info. needs:** **24 real needs** (1 real need per test person)
96 simulated information needs
(4 simulated task situations per test person)
- Location of tests:** **IR Laboratory at Glasgow University**

Natural Work Tasks (WT) & Org	Natural Search Tasks (ST)	Actor	Perceived Work Tasks	Perceived Search Tasks
WT Structure	ST Structure	Domain Knowledge	Perceived WT Structure	Perceived Information Need Content
WT Strategies & Practices	ST Strategies & Practices	IS&R Knowledge	Perceived WT Strategies & Practices	Perceived ST Structure/Type
WT Granularity, Size & Complexity	ST Granularity, Size & Complexity	Experience on Work Task	Perceived WT Granularity, Size & Complexity	Perceived ST Strategies & Practices
WT Dependencies	ST Dependencies	Experience on Search Task	Perceived WT Dependencies	Perceived ST Specificity & Complexity
WT Requirements	ST Requirements	Stage in Work Task Execution	Perceived WT Requirements	Perceived ST Dependencies
WT Domain & Context	ST Domain & Context	Perception of Socio-Org. Context	Perceived WT Domain & Context	Perceived ST Stability
		Sources of Difficulty	Independent Variables	Perceived ST Domain & Context
		Motivation & Emotional State		

Document and Source	IR Engines IT Component	IR Inter-faces	Access and Interaction
Document Structure	Exact Match Models	Domain Model Attributes	Interaction Duration
Document Types	Best Match Models	System Model Features	Actors or Components
Document Genres	Degree of Doc. Structure and Content Used	User Model Features	Kind of Interaction and Access
Information Type in Document	Use of NLP to Document Indexing	System Model Adaption	Strategies and Tactics
Communication Function	Doc. Metadata Representation	User Model Building	Purpose of Human Communication
Temporal Aspects	Use of Weights in Doc. indexing	Request Model Builder	Purpose of System Communication
Document Sign Language	Degree of Req. Structure and Content Used	Retrieval Strategy	Interaction Mode
Layout and Style	Use of NLP to Request Indexing	Response Generation	Least effort Factors
Document Isness	Req. Metadata Representation	Feedback Generation	-
Document Content	Use of Weights in Requests	Mapping ST History	<div style="border: 2px solid blue; padding: 10px; display: inline-block;"> Controlled Variables </div>
Contextual Hyperlink Structure		Explanation Features	
Human Source (see Actor)		Transformation of Messages	
		Scheduler	

Agenda - 3

- **Naturalistic Field Investigations of IIR** (20 min)
 - Integrating context variables
 - Live systems & (simulated/real) work tasks
 - Sample Study – Marianne Lykke Nielsen (2001; 2004)
- **Wrapping up of Tutorial** (5 min)

Questions are welcome during the tutorial sessions

Keep things simple!

- If you can isolate one (or two) variables as independent – then stick to that.
- Real-life studies are much more uncertain and complex than laboratory tests
- **A robust research setting is crucial**
- **Natural search jobs (e.g. exploratory) mixed with simulated ones (but must be realistic!)**
- **Test persons do relevance assessments!**

Agenda - 3

- ✓ **Naturalistic Field Investigations of IIR** (20 min)
 - ✓ Integrating context variables
 - ✓ Live systems & (simulated) work tasks
 - Sample Study – Marianne Lykke Nielsen (2001; 2004)
- **Wrapping up** (5 min)

Questions are welcome during the tutorial sessions

Natural IR Interaction KMS Sample

- **Marianne L. Nielsen (2004):** Task-based evaluation of associative thesaurus in real-life environment.
Proceedings of the ASIST 2004 Annual Meeting; Providence, Rhode Island, November 13 - 18, 2004. 437 - 447.
- **Research setting:** Danish Pharmaceutical Company
- **Goal:** To observe if a company thesaurus (ontology) based on *human conceptual associations* affects searching behavior, retrieval performance and searcher satisfaction **different** from a **domain-based thesaurus**.

Associative Thesaurus - ASSO

- Nielsen, M.L. (2001). A framework for work task based thesaurus design. *Journal of Documentation*, 57 (6), 774-797.
- Made from several association tests with 35 employees from the company, supplying synonyms, narrow and broader concepts, based on the "**company vocabulary**" (task/product-based)
- This thesaurus was larger in number of entries (379 more) and associative terms than the "**control thesaurus**" – made by domain expert and based on the "**scientific vocabulary**".

Research Design - 1

- **20 test persons** from the basic and clinical researchers, including marketing employees (also with scientific background)
- **3 simulated search task situations** (next slide) per test person, all having same structure and **based on real work tasks** observed by recently logged requests to company retrieval system.
- **“Blind testing”** of the two thesaurus types: test persons were told that the investigation was part of the system design process. Only the research team knew who searched which thesaurus type!

Search Job A

You are Product Manager working for Lundbeck Pharma.

A physician who wants to know if the combination of Citalipram and Lithium leads to approve therapeutic effect on Bipolar Disorders, has consulted you.

You need to find reports or articles investigating interaction and effect of the two drugs.

Research Design - 2

- **Steps in the field study** (2 hours per test person):
 1. Capture search skills (e-mail questionnaire)
 2. Explanation session
 3. **Pre-search interview** of searcher's mental model concerning each search job / expectations
 4. **Search session** with **relevance assessments** (logging and structured observation of each job)
 5. **Post-search interview** of motivation & satisfaction for each search job.

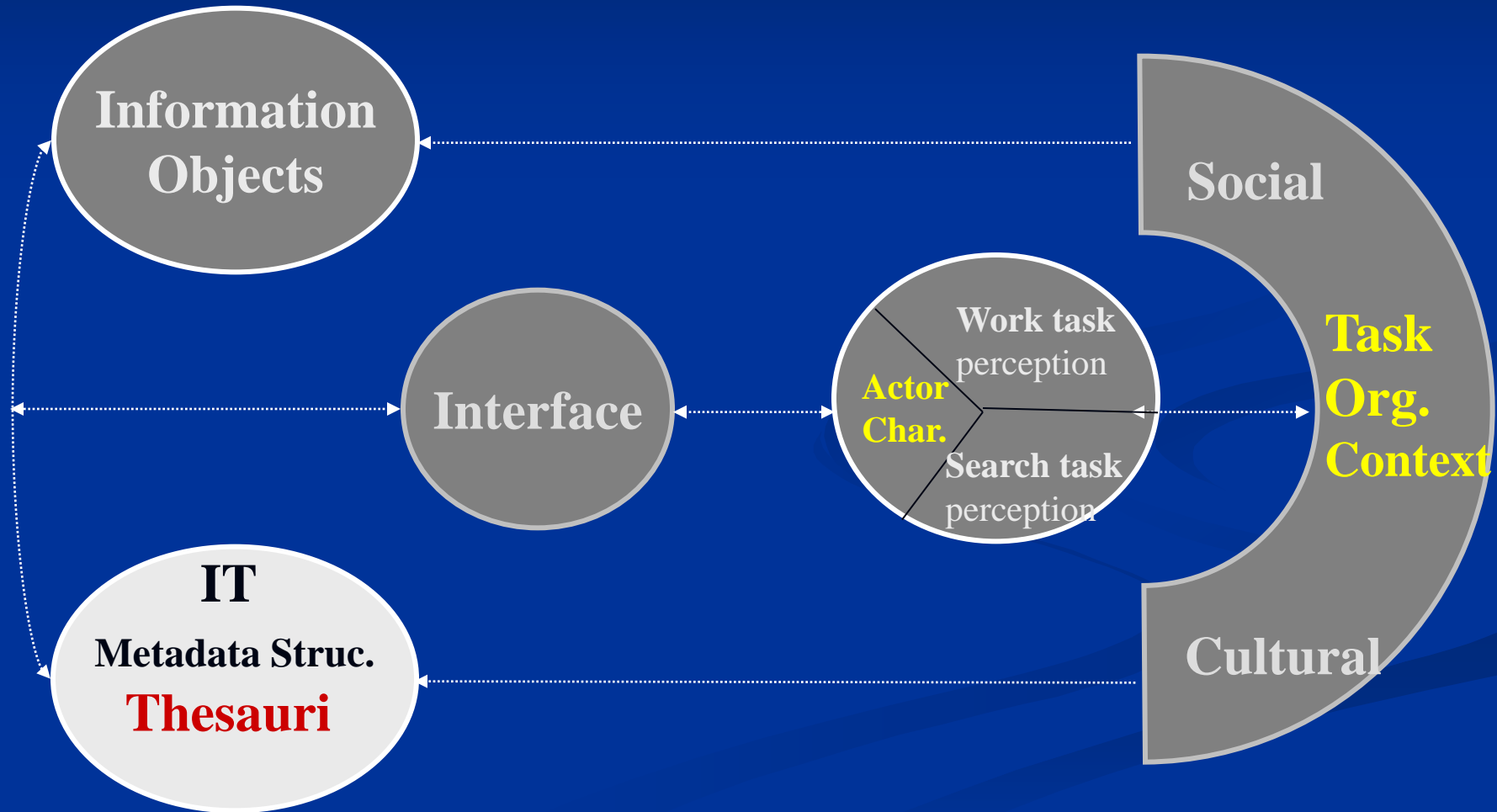
Research Design - 3

- **Latin square execution** (slide 59) to avoid learning effects & all search jobs are tried out on both thesauri:
- 10 persons x 3 search jobs in ASSO = 30 units
- 10 persons x 3 search jobs in DOMAIN = 30 units (in reality there were only 2 x 29, due to error)
- **Relevance assessments: three-graded:** Highly relevant; Partially relevant; Not relevant.
- **Measures:** Recall/Precision; Behavior; Satisfaction

Research Design - 4

- **Independent Variable:**
- Document Metadata Representation (two values)
- **Controlled Variables:**
 - Natural Work/Search Task Org. setting
 - Perceived Work Task Structure; Complexity (high)
 - Perceived Information Need
 - Database; Retrieval Engine; Interface
- **Hidden Variables:** Test person characteristics

Naturalistic Field Study (M.L. Nielsen) - variables



Document and Source	IR Engines IT Component	IR Inter-faces	Access and Interaction
Document Structure	Exact Match Models	Domain Model Attributes	Interaction Duration
Document Types	Best Match Models	System Model Features	Actors or Components
Document Genres	Degree of Doc. Structure and Content Used	User Model Features	Kind of Interaction and Access
Information Type in Document	Use of NLP to Document Indexing	System Model Adaption	Strategies and Tactics
Communication Function	Doc. Metadata Representation	User Model Building	Purpose of Human Communication
Temporal Aspects	Use of Weights in Doc. indexing	Request Model Builder	Purpose of System Communication
Document Sign Language	Degree of Req. Structure and Content Used	Retrieval Strategy	Interaction Mode
Layout and Style	Use of NLP to Request Indexing	Response Generation	Least effort Factors
Document Isness	Req. Metadata Representation	Feedback Generation	-
Document Content	Use of Weights in Requests	Mapping ST History	
Contextual Hyperlink Structure		Explanation Features	
Human Source (see Actor)		Transformation of Messages	
		Scheduler	

Natural Work Tasks (WT) & Org	Natural Search Tasks (ST)	Actor	Perceived Work Tasks	Perceived Search Tasks
WT Structure	ST Structure	Domain Knowledge	Perceived WT Structure	Perceived Information Need Content
WT Strategies & Practices	ST Strategies & Practices	IS&R Knowledge	Perceived WT Strategies & Practices	Perceived ST Structure/Type
WT Granularity, Size & Complexity	ST Granularity, Size & Complexity	Experience on Work Task	Perceived WT Granularity, Size & Complexity	Perceived ST Strategies & Practices
WT Dependencies	ST Dependencies	Experience on Search Task	Perceived WT Dependencies	Perceived ST Specificity & Complexity
WT Requirements	ST Requirements	Stage in Work Task Execution	Perceived WT Requirements	Perceived ST Dependencies
WT Domain & Context	ST Domain & Context	Perception of Socio-Org. Context	Perceived WT Domain & Context	Perceived ST Stability
		Sources of Difficulty		Perceived ST Domain & Context
		Motivation & Emotional State		

Selected Results

- Both thesauri show **same IR performance** level
- Both thesauri applied to **Query Formulation & Modification** or as **Lead-in Terms**:
 - Finding **synonyms** and /or more **specific** terms
 - Clarifying **meaning** (in task perspective) of terms
- **ASSO** applied slightly more time (used for **Narrow Terms capture**)
- **DOMAIN** applied more in pre-search stage

Selected Results 2

- Recall / Precision:
- ASSO: .14 / .32 – DOMAIN: .11 / .37
- Note: test persons assessed same documents quite differently!
- This was due to two fundamentally different groups of test persons (hidden variable!):
 - Basic researchers (exploring new drugs)
 - Clinical researchers (clinical drug tests)
 - This also concerns the satisfaction of the use of the thesauri for IR (which was quite high)

Agenda - 3

- ✓ **Naturalistic Field Investigations of IIR** (20 min)
 - ✓ Integrating context variables
 - ✓ Live systems & (simulated) work tasks
 - ✓ Sample Study – Marianne Lykke Nielsen (2001; 2004)
- **Wrapping up of Tutorial** (5 min)

Questions are welcome during the tutorial sessions

Step-by-Step into Light!

- In pure ‘laboratory experiments’ **only simulations** of searcher behavior can be done;
- If one wishes to stick to **existing test collections**, with existing sets of relevance assessments and ‘topics’, only **IR interaction ‘ultra-light’** can be done (in order to avoid learning effects by test persons):
 - Requires short-term IR interaction;
 - In the form of ‘laboratory studies’.
 - Number of test persons, search jobs and research setting follow same line as Interactive ‘light’ IR.

Step-by-Step into Context - Light!

- **IR interaction 'light'** entails **session-based IR**, with test persons' relevance assessments and more intensive monitoring (logs; interviews; observation);
 - **Can be carried out as laboratory study or field experiment**
- Like in 'ultra-light' and 'naturalistic' IR, **number of test persons and search jobs** must assure that '*statistically enough*' data is present in the result matrices when cross tabulating independent variables (see slides 40-41).
- **IR interaction 'light':** assigned realistic requests, simulated task situations and **searcher relevance assessments**
- **Naturalistic IR interaction** assumes natural tasks (mixed with simulated ones) in natural environments

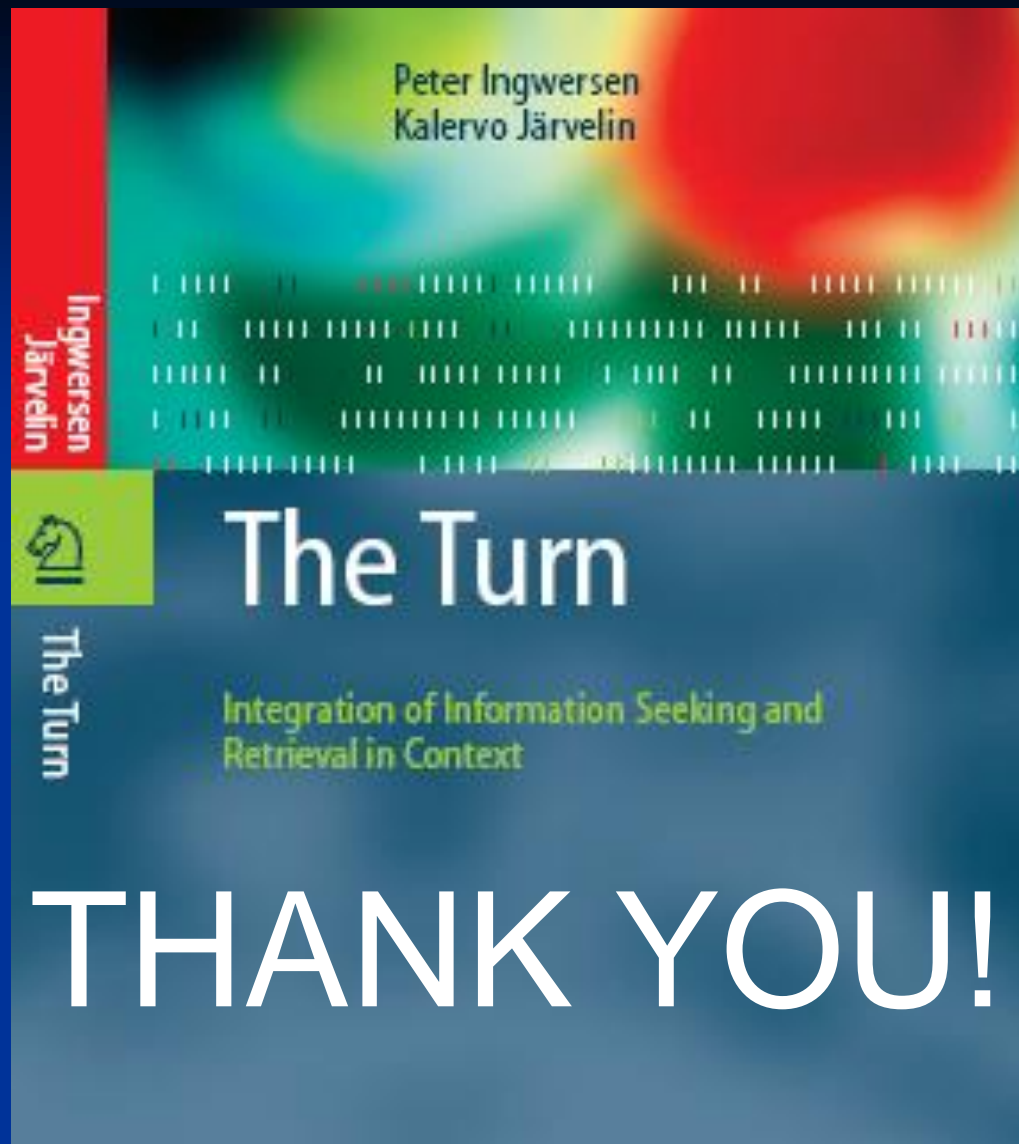
'Ultra-light' and 'Light' IIR

- Assessments can be 4-graded (Vakkari & Sormunen, 2004);
- Realistic but few relevance assessments per person;
- Assessments can be pooled for same search job over all test persons – weighted doc. assessments
- Common recall/precision, MAP, CumGain, P@n, etc. feasible
- You require min. 30 responses per result cell
- Ultra-light lab. studies are effective for tightly controlled IIR experiments (like Kelly et al.)

The Cognitive Research Framework informs about ...

- Central variables to combine as **independent ones**
- Major **variables kept controlled/neutralized** in a setting
- What kind of **variables that are hidden!**
- **Dependent variables** depend on the research goals (the independent variables!)

Novel possible research designs, settings and measures ... there is a lot to do - really!



<http://www.springeronline.com/1-4020-3850-X/>