

Multimedia IR Evaluation Initiatives

Alan F. Smeaton

IR has an evaluation history

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- (Classical) Information Retrieval roots in information science & computer science;
- Always, it has been empirical ... about experimentation, testing, evaluation, assessment, measurement;
- Why ?
- In the early days this was easy though closed & narrow;
- Then, IR came centre-stage, with pressure to grow collection sizes and heterogeneity, multiple types of user applications, multiple media;
- Yet IR, and MMIR, has a fundamental, ingrained tendency to evaluate, measure, etc.
- Evaluation campaigns, pooling resources, have addressed issues of scale and cost;

Evaluation in IR

- There is much to actual user information seeking ...
 - what information do I want,
 - what will I do with it,
 - what information will be useful,
 - how will I get it,
 - what words should I use to formulate a query,
 - what documents to save for later,
 - how is this document different from the other 15 million that Google has retrieved for me,
 - is there better information,
 - do I trust this information,
 - have I missed anything,
 - ...
- As we develop new components for information seeking, e.g. analysis, indexing, matching, etc., it is important to measure performance and contribution

Information seeking & People

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Greatest variable in information seeking is users ...

- experiences
- knowledge
- tasks
- verbal skills
- tolerances
- genders
- preferences
- personalities
- beliefs
- motivations
- ages
- opinions
- cultures
- cognitive abilities

ASPECTS OF Evaluation

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- **System EVALUATION**

- tests quality of IR system/engine
- high volume of queries
- no user involvement
 - *simulate* user
- cheap and popular
- highly controlled !

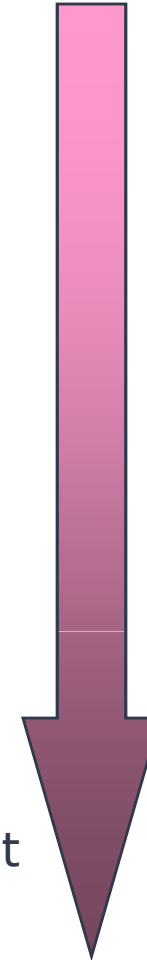
- **User EVALUATION**

- tests quality of IR system + interface
- (usually) low volume of queries
- direct user involvement
- artificial test

- **Operational EVALUATION**

- similar to user but in real situations
- expensive and difficult to run but very good test

Cost, time, effort,
experienced
needed



IR Evaluation - the Emerging View

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Two approaches to improving information access in general
 - computer science - system evaluation
 - propose new (well-founded) solutions
 - evaluate them in evaluation campaigns to uncover
 - what benefits searchers and in what way
 - new questions for investigation information science
 - investigate searching behaviour from a human perspective
 - user evaluation
 - identify generalities amongst searchers or search behaviour
 - identify meaningful differences between searchers or search behaviour
 - make recommendations to system designers

User-oriented evaluations

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- In practice, USER EVALUATIONS do evaluate the whole system
 - algorithms plus interfaces
 - mostly comparative
 - e.g. two interfaces to same system
 - objective measures
 - e.g. number of relevant documents found, time to search
 - subjective measures
 - easy to search, easy to learn, popular
 - qualitative and quantitative analysis
 - also proposed are things like cost, quality of information, search satisfaction
 - less controlled than test collection evaluation
- .. but this is too expensive to do, so we do system evaluations instead and we do them in benchmark evaluation campaigns

Cranfield Evaluation Paradigm

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- In 1960s, Cranfield College of Aeronautics wanted to test indexing techniques for text abstracts;
- Created test queries on a static document collection and each was judged as R/Non-R for each Q;
- That was the first experimental IR evaluation, and it continues today;
- Fundamental is “relevance”, quite subjective, but pillar of IR evaluation;
- We assume that relevance = {topicality, user satisfaction, pertinence, system & user relevance, usefulness, and everything else}

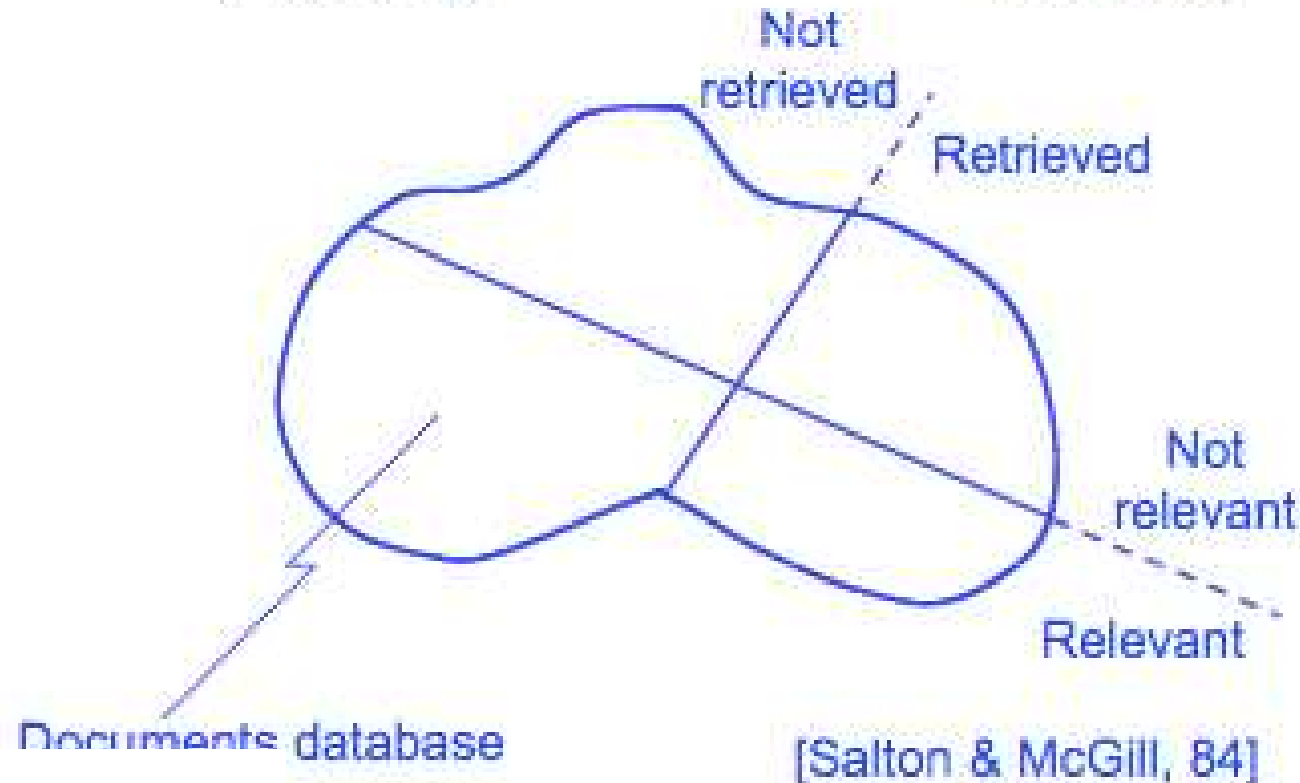
Pooling

- For small collections, manual judgment of R/NR of all docs was feasible;
- For sizable collections it is not;
- Pooling accepts N document rankings and pools top-X ranked from each, eliminates duplicates, judges set of top-X documents;
- Incomplete, but expectation is that reldocs rise to the top;
- Shown to work best when pooling very different IR approaches, which find different relevant documents;
- Basis for most evaluation campaigns (see later)

Precision and Recall

$$P = \frac{|\text{relevant \& retrieved}|}{|\text{retrieved}|}$$

$$R = \frac{|\text{relevant \& retrieved}|}{|\text{relevant}|}$$



Precision and Recall

	Retrieved	Not Retrieved	
Relevant	a	b	$n_1 = a + b$
Non relevant	c	d	
	$n_2 = a + c$		$N = a + b + c + d$

$$P = \frac{a}{n_2}$$

$$R = \frac{a}{n_1}$$

- Binary relevance, binary retrieval, IR ranks the documents by $P(\text{Rel})$

Worked example of P & R

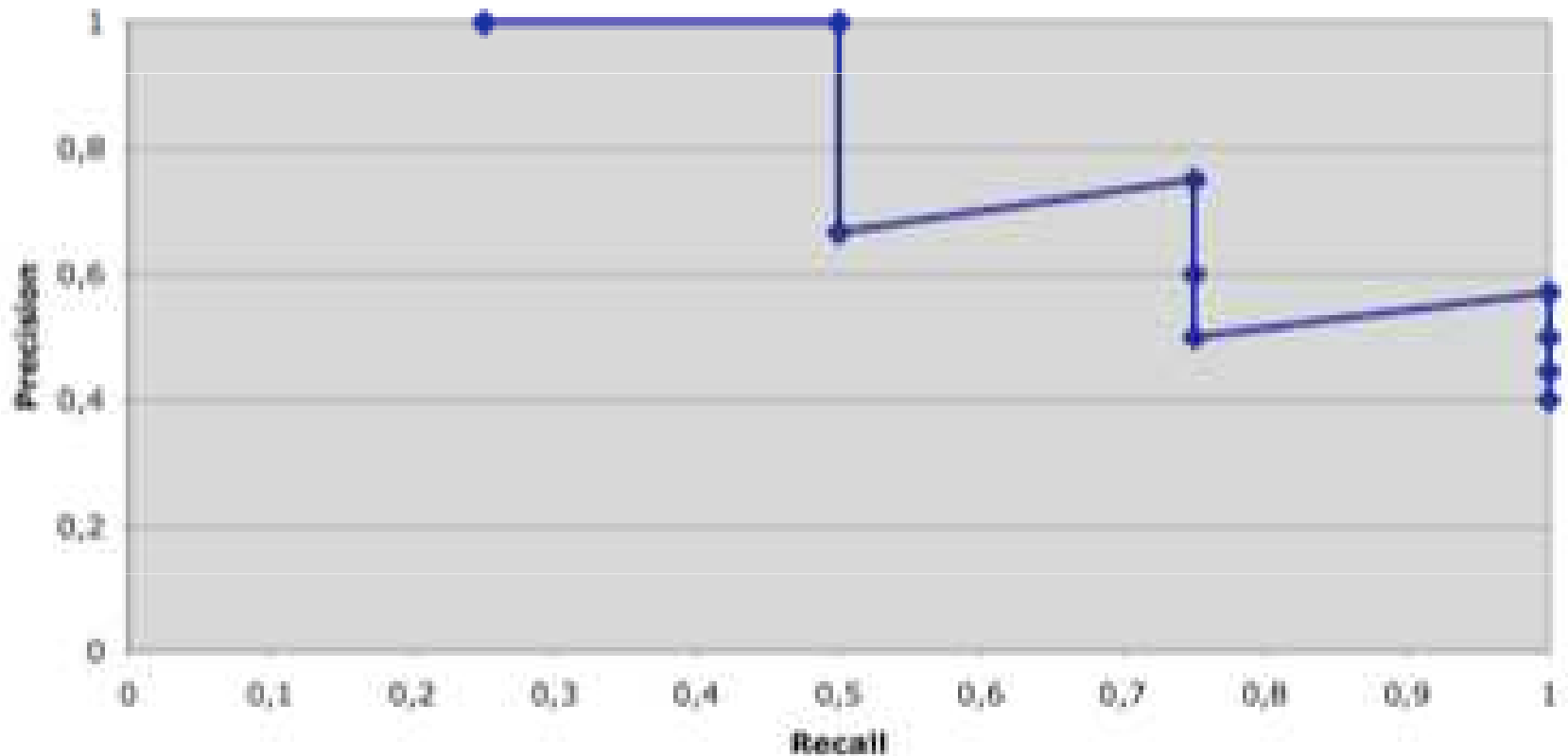
- 1 = Rel
- 0 = Non Rel
- 4 Reldocs in the collection

(following example from Mizzaro)

Rank	Rel?	R	P
1	1	0,25	1
2	1	0,5	1
3	0	0,5	0,67
4	1	0,75	0,75
5	0	0,75	0,6
6	0	0,75	0,5
7	1	1	0,57
8	0	1	0,5
9	0	1	0,44
10	0	1	0,4

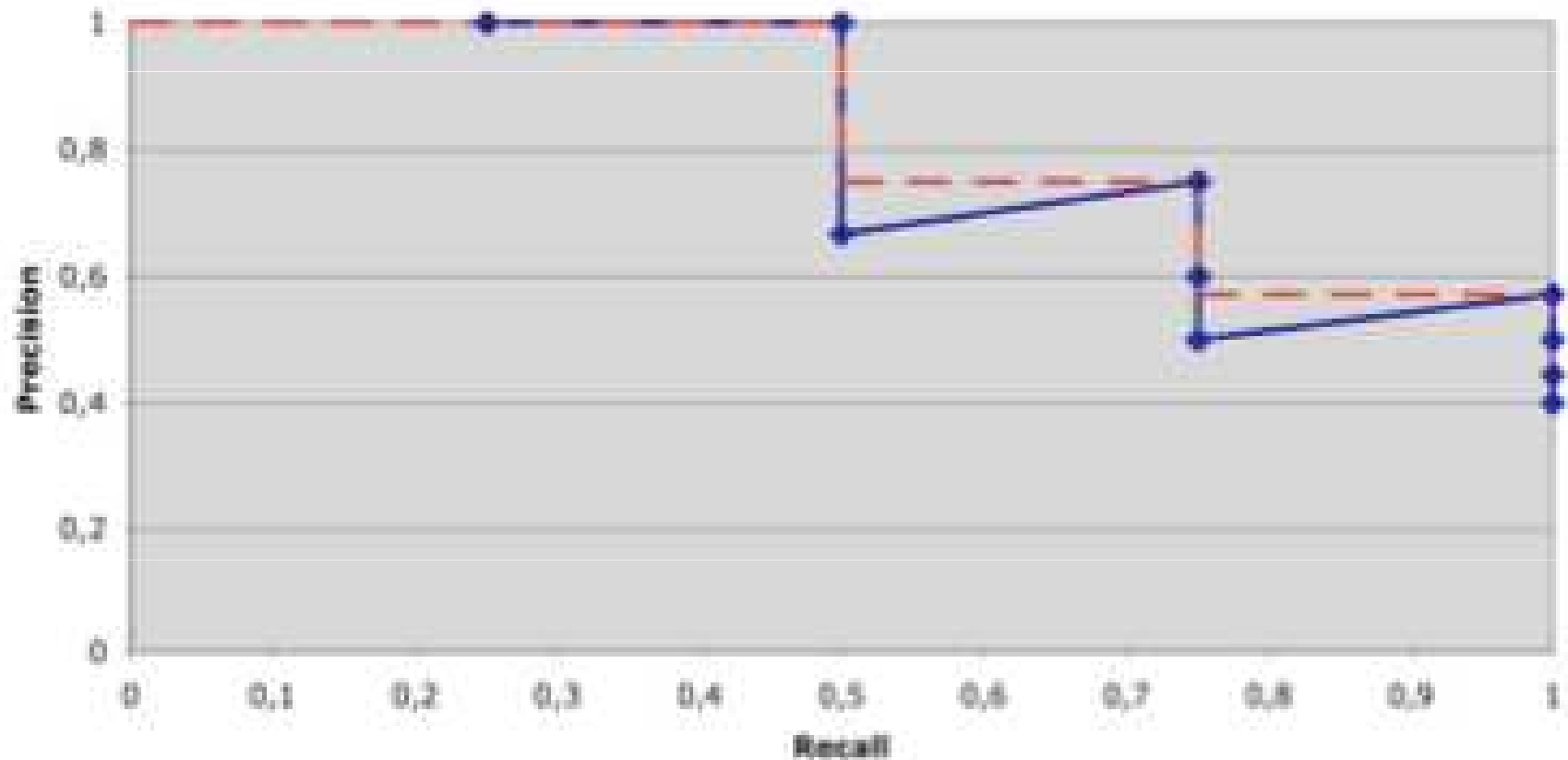
PR Graph

- Actual PR graph for 1 query



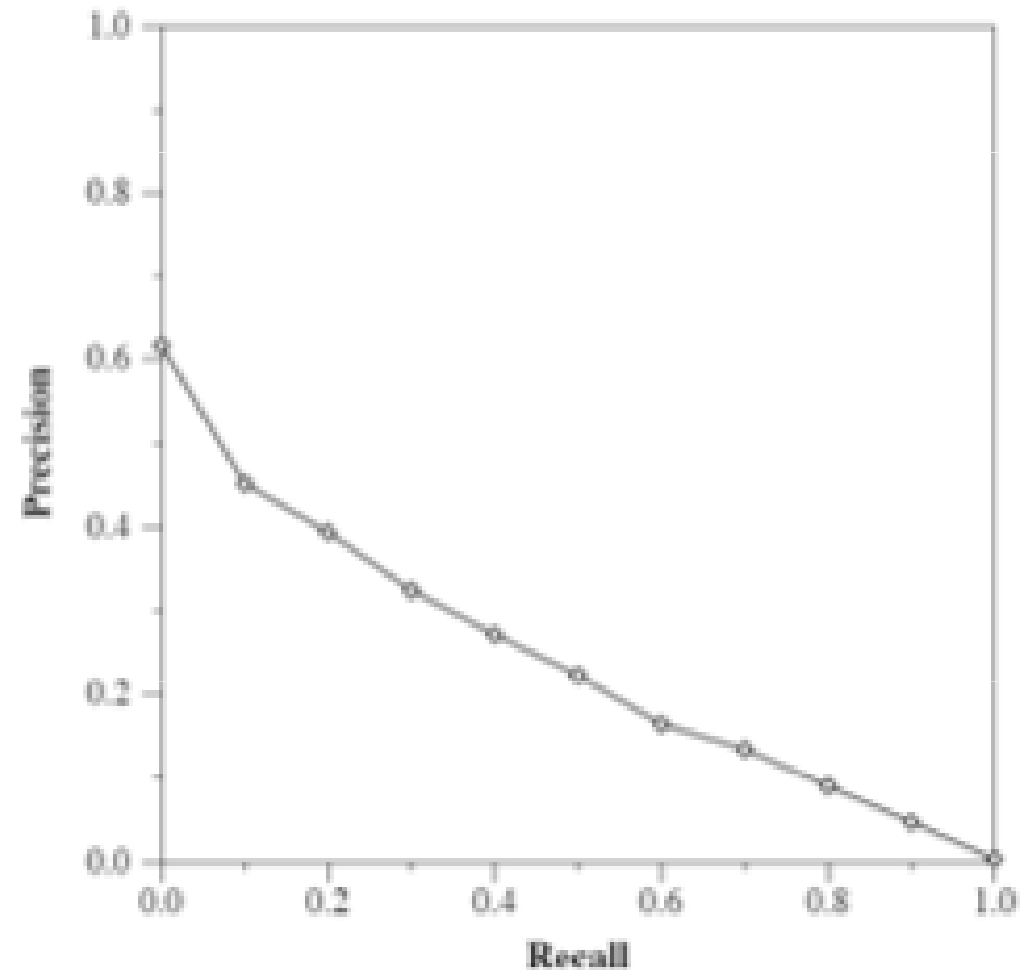
Interpolating PR

- Interpolated PR curve



PR Curve

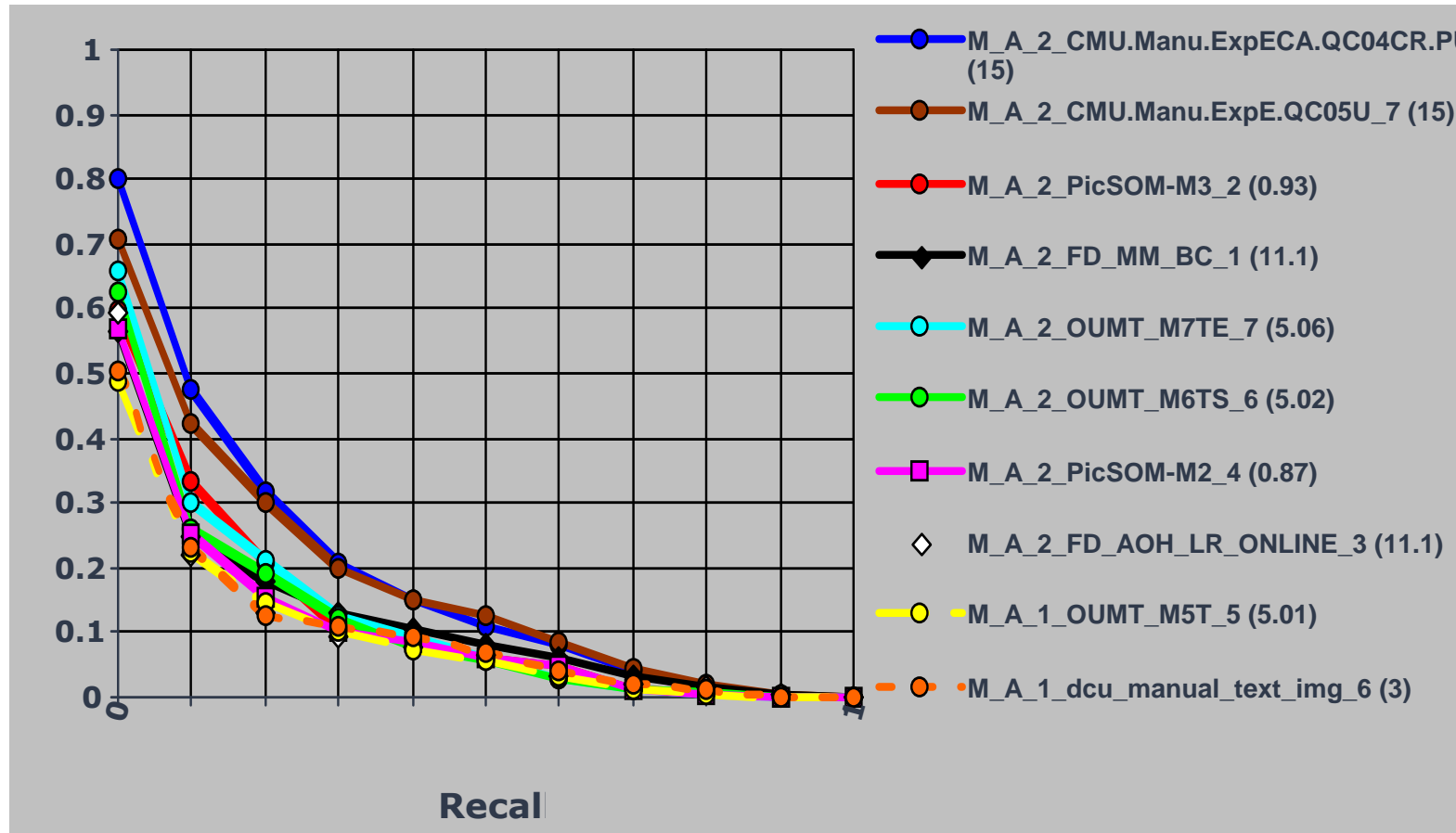
- PR curve then averaged over set of queries;
- # steps depends on # queries;
- 11 recall levels 0, 0.1, 0.2, ... 1;



17

Comparison of several curves

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G



- PR curve is a golden standard;
- Often recall can't be computed exactly
- Comparisons can be difficult, single measure ?

Single Measure - AvgP

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Average of the precision values after each relevant document is retrieved;
 - NOT average of P at 11 recall points
 - If not retrieved, precision - 0

Rank	Rel?	R	P
1	1	0,25	1
2	1	0,5	1
3	0	0,5	0,67
4	1	0,75	0,75
5	0	0,75	0,6
6	0	0,75	0,5
7	1	1	0,57
8	0	1	0,5
9	0	1	0,44
10	0	1	0,4

Single Valued Metrics

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Mean Average Precision (MAP)
 - Average P is for one query
 - MAP is the mean across set of queries
- Interpolated MAP = average of the average of precisions at 11 standard recall points;
- P@5, P@10, P@20, P@N
 - P@10 often used for web search
 - P@1 for “I’m Feeling Lucky” searches
- R-Precision
 - P@R, R=# relevant documents
 - Precision after R documents

Other Measures

- ESL - Expected Search Length
 - $ESL(x) = \#$ documents to be read from ranked list, to have x relevant documents
 - Average over queries, not a single value, $f(x)$;
- DCG - Discounted Cumulative Gain
 - Assumes N relevance levels and measures gain that a doc gives to user
- Others
 - ADM Average Distance Measure

trec-eval

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- trec-eval is a publicly available program, developed by Chris Buckley, used extensively by TREC and other evaluation campaigns, which computes many usable metric values based on standardised file input formats;
- Its available, multi-platform, easy to use, so use it !

TREC

- the Text Retrieval Conference (TREC) initiative began in 1991 as a reaction to small collection sizes and the need for a more coordinated evaluation among researchers
- over the intervening decade and a half spawned over a dozen IR-related tasks
- 2005 (14th) had 117 research groups !
- following the success of TREC, many evaluation campaigns have been launched;
- most follow the TREC model, which in turn follows Cranfield model ... acquire data and distribute it, formulate search topics, accept and pool submissions, manually assess pools, calculate metrics;
- In no particular order ...

TREC Tracks

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- After initial search task there was strong interest in diversification;
- This led to the development of “tracks”;
 - See next slides
- Bonus is that TREC test collections are large enough so that they realistically model operational settings.
- Most of today's commercial search engines include technology first developed in TREC.

2006 TREC Tracks

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- **Blog Track**
A new track in 2006 to explore information seeking behavior in the blogosphere.
- **Enterprise Track**
A new track in 2005 to study enterprise search: satisfying a user who is searching the data of an organization to complete some task.
- **Genomics Track**
To study retrieval tasks in the domain of genomics data (broadly construed to include not just gene sequences but also supporting documentation such as research papers, lab reports, etc.) The genomics track first ran in 2003
- **Legal Track**
A new track in 2006 to develop search technology that meets the needs of lawyers to engage in effective discovery in digital document collections
- **Question Answering Track**
To take a step closer to information retrieval rather than document retrieval
- **SPAM Track**
New in 2005 to provide a standard evaluation of current and proposed spam filtering approaches, thereby laying the foundation for the evaluation of more general email filtering and retrieval tasks.
- **Terabyte Track**
First run in 2004 to investigate whether/how the IR community can scale traditional IR test-collection-based evaluation to significantly larger document collections

Past TREC Tracks

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- **Cross-Language Track**
Investigates the ability of retrieval systems to find documents that pertain to a topic regardless of the language in which the document is written.
- **Cross-Language Track**
CL was run in 2002, now studied in both CLEF and the NTCIR workshops.
- **Filtering Track**
User's information need is stable (and some relevant documents are known) but there is a stream of new documents and the system must make a binary decision as to whether the document should be retrieved
- **HARD**
To achieve High Accuracy Retrieval from Documents by leveraging additional information about the searcher and/or the search context, through techniques such as passage retrieval and using very targeted interaction with the searcher.
- **Interactive Track**
Studies user interaction with text retrieval systems, carry out studies with real users using a common collection and set of user queries.
- **Novelty Track**
Investigate systems' abilities to locate new (i.e., non-redundant) information.
- **Robust Retrieval Track**
Traditional ad hoc retrieval but with the focus on individual topic effectiveness rather than average effectiveness
- **Video Track**
Video track devoted to research in retrieval of digital video.
- **Web Track**
Featuring search on a document set that is a snapshot of the World Wide Web.

Number of participants per track and total number of distinct participants in each TREC

Track	TREC											
	92	93	94	95	96	97	98	99	00	01	02	03
Ad hoc	18	24	26	23	28	31	42	41	—	—	—	—
Routing	16	25	25	15	16	21	—	—	—	—	—	—
Interactive	—	—	3	11	2	9	8	7	6	6	6	—
Spanish	—	—	4	10	7	—	—	—	—	—	—	—
Confusion	—	—	—	4	5	—	—	—	—	—	—	—
Database merging	—	—	—	3	3	—	—	—	—	—	—	—
filtering	—	—	—	4	7	10	12	14	15	19	21	—
Chinese	—	—	—	—	9	12	—	—	—	—	—	—
NLP	—	—	—	—	4	2	—	—	—	—	—	—
Speech	—	—	—	—	—	13	10	10	3	—	—	—
Cross-language	—	—	—	—	—	13	9	13	16	10	9	—
High precision	—	—	—	—	—	5	4	—	—	—	—	—
Very large corpus	—	—	—	—	—	—	7	6	—	—	—	—
Query	—	—	—	—	—	—	2	5	6	—	—	—
Question answering	—	—	—	—	—	—	—	20	28	36	34	33
Web	—	—	—	—	—	—	—	17	23	30	23	27
Video	—	—	—	—	—	—	—	—	—	12	19	—
Novelty	—	—	—	—	—	—	—	—	—	—	13	14
Genome	—	—	—	—	—	—	—	—	—	—	—	29
HARD	—	—	—	—	—	—	—	—	—	—	—	14
Robust	—	—	—	—	—	—	—	—	—	—	—	16
Total participants	25	31	33	36	38	51	56	66	69	87	93	93

TREC Contribution

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- TREC was the original, it defined the modality, focused attention on evaluation campaigns and their usefulness, had real impact on the quality of (text) IR, reared a generation of IR researchers and spawned nearly 000's of papers;
- TREC also spun off a large number of copycat evaluation campaigns;
- TREC continues now, 15 years later, as strong as ever (c.100 participants)
- TREC Overview at <http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=10667&mode=toc>

"Copycat" Evaluation Campaigns

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

In 2006 alone ...

- **CLEF** - 74 groups, 6th year, mono/x-lingual retrieval, 12 languages !
- **NTCIR** - Asian CLEF, 6th year, Korean, Chinese, Japanese
- **INEX** - XML element retrieval, 6th year, 80 groups
- **VACE** - US w/ international - tools to assist human analysts monitor and annotate video - TV news/surveillance/UAV meetings
- **ETISEO** - French, 23 groups, visual processing for surveillance
- **PETS** - 7th year - object detection and tracking
- **AMI** - analysis from instrumented meeting rooms
- **ImagEval** - French, CBIR, see C. Fluhr presentation
- **Benchathlon** and **CLEAR** = VACE \cap CHIL

- All are metrics-based, use XML for data submission and exchange;

Cross-Language Evaluation Forum

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Tests aspects of mono- and cross-lingual IR;
- 2006 was the 7th in the annual series, DELOS funded, 74 groups;
- Proceedings online; 8 tracks in 2005
 - Mono-, bi- and ML doc retrieval on news;
 - Mono- and cross-lingual retrieval on structured scientific data;
 - Interactive CL information retrieval;
 - ML question-answering;
 - CL retrieval in image collections;
 - CL spoken document retrieval;
 - CL geographic retrieval;

Cross-Language Evaluation Forum

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- 7 different document collections including 2M news documents, in 12 languages;
- Bulgarian, Dutch, English, Finnish, French, German, Hungarian, Italian, Portuguese, Russian, Spanish & Swedish;
- Herculean effort in securing permission for data provision;
- Task-specific details very much driven by the participants with loose control from the funders;
- Big team effort;

NTCIR

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Like CLEF, except Asian, and not as big !
- 2005 was 6th running;
- Like CLEF, NTCIR follows TREC quite faithfully;
- 4 tasks
 - ML, Bilingual and single language IR on Chinese, Korean and Japanese
 - Cross-lingual question-answering seeking named entities in Chinese, English, Japanese pairs
 - Patent retrieval and classification using Japanese and US patents
 - Question-answering beyond factoids and asking “why” - on Japanese

Initiative for the Evaluation of XML Retrieval (INEX)

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Started in 2002, 2006 was the 5th running, 80 participants;
- Aim is to establish infrastructure, collection & scoring for IR which exploits available structural information (XML elements) to yield more focused retrieval;
- XML IR may retrieve mixture of paragraphs, sections, etc.
- Collection is 659,300 Wikipedia articles from 113,483 categories with average 161 XML nodes from 5000 tagset - previously it was IEEE articles;
- Participants create candidate topics as content only or content-and-structure, gain access to document collection only then;
- Main task is ad hoc retrieval plus tasks in NL query, heterogeneous documents, interactive, document mining and Multimedia

Video Analysis and Content Extraction (VACE)

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- VACE is a US DTO funding program for US groups, 14 funded participants;
- Address lack of tools to assist human analysts monitor and annotate video for indexing;
- Data is broadcast TV news, surveillance, UAV, meetings, ground reconnaissance;
- Tasks are detection a/o tracking of people, faces, vehicles and text in that data;
- VACE includes open evaluations with international participation
 - Increases competition, thus increases progress
 - Encourages peer review and information exchange, minimises wheel reinvention, focuses on common problems
 - Open evaluations include VACE, CLEAR, and TRECVID;


ETISEO

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Started Jan 2005, funded by French government, 23 participants;
- Aim to evaluate vision techniques for video surveillance applications;
- Video data used is single and multi-view surveillance (airport, car park, corridor, subway);
- Ground truth is annotations and classifications of persons, vehicles, groups;
- Tasks are detection, localisation, classification and tracking of physical objects, and event recognition

Performance Evaluation of Tracking & Surveillance

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- 2006 was the 7th PETS cycle, funded by an EU FP6 project ISCAPS;
- Evaluates object detection and tracking for video surveillance, metrics based 
- Data is multi-view surveillance video (4 cameras);
- Task is event detection - left luggage in public place

Augmented Multi-Party Interaction (AMI)

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- EU FP6 project covering meeting analysis from instrumented meeting rooms;
- Run by IDIAP, CH;
 - 2D multi-person tracking, head bounding box for each frame;
 - Head pose estimation;
 - Estimation of focus-of-attention (FoA) in meetings as table, documents, screen, or people using gaze

ImagEval

- French national evaluation campaign now open to other Europeans, 14? participants;
- 5 tasks related to content based image retrieval:
 - Recognition of (geometric) image transformations like rotation, projection, etc.;
 - Retrieval based on combining text and image;
 - Detect and extract text regions from images;
 - Detect objects in images - cars, planes, flowers, cats, churches, Eiffel tower, table, PC or TV, US flag;
 - (Semantic) feature detection - indoor, outdoor, people, night, day, etc.;
- Various data (image) sources, O(000's);

Benchathlon and CLEAR

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Benchathlon is a clearinghouse for data, annotations, evaluation measures, tools and architecture for CBIR;
- CLEAR is a cross-campaign collaboration concerned with getting consensus and crossover on evaluation of event classification evaluation;
- $\text{CLEAR} = \text{VACE} \cap \text{CHIL}$

TRECvid: Video IR Evaluation

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- In 2001, “video retrieval” started as a TREC track;
- Usual TREC mode of operation (data-topics-search submissions-pooling-evaluation by metrics-workshop) but additional, related tasks besides search;
- In 2003 TRECvid separated from TREC because it was sufficiently different, and had enough participation, though TREC and TRECvid workshops are co-located;
- Started small, grew rapidly;
- TRECvid 2006 tasks featured shot bound detection, concept detection and 3 kinds of search - automatic, manual and interactive;

Major responsibilities

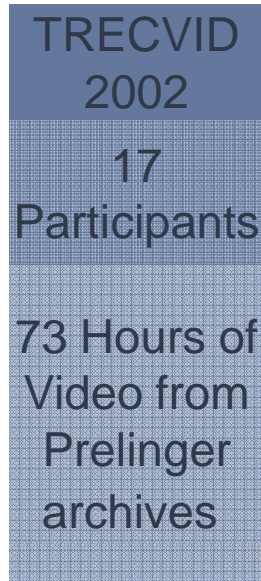
C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- ❑ **NIST:** Organize data, tasks, and other resources of interest with input from sponsors and participating researchers
 - ❑ Select and secure available data for training and test
 - ❑ Define user and system tasks, submission formats, etc.
- ❑ **LDC:** Collect, secure IPR, prepare, distribute data
- ❑ **NIST:** Define and develop evaluation infrastructure
 - ❑ Create shot boundary ground truth
 - ❑ Create and support interactive judging software for features and search
 - ❑ Create and support the scoring software for all tasks
- ❑ **Researchers:** create common resources
 - ❑ Master shot definition (Fraunhofer Institute, Berlin)
 - ❑ Common keyframes (Dublin City University)
 - ❑ Annotation tools (IBM, CMU)
 - ❑ Feature training data annotation (20+ groups)
- ❑ **Researchers:** Develop systems
- ❑ **Researchers:** Run systems on the test data and submit results to NIST
- ❑ **NIST:** Evaluate submissions
 - ❑ Run automatic shot boundary scoring software
 - ❑ Manage the manual judging by contractors viewing a sample of system output (~76,000 shots for features, ~78,000 shots for search)
- ❑ **NIST, Researchers:** Analyze and present results
- ❑ **NIST:** Organize and run annual workshop in mid-November at NIST

Evolution: data, tasks, participants

C E N T E R F O R D I G I T A L V I D E O P R O

Growing Participation



Growing Data Sets

Thanks to J. Smith (IBM) for earlier slide

TRECVID Data

- TRECVID acquires video data, sorts usage permissions and distributes to 70+ groups;
- In preparation we are always working on possible new data sources, with lots of cul-de-sacs;
- For 2007 we're working on ...
 - informational, news magazine video from the Sound and Vision archive (the Netherlands) - we have 400h.
 - 150 hours of BBC rushes
 - possibly some produced news and some news rushes from Spanish TV
- We have over 500 hours of new materials ... is that impressive ?

TRECvid data

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- 2001 - **11** hours, 74 topics, ~ 8000 shots
- 2002 - **40** hours, 25 topics, ~14000 shots
- 2003 - **56** hours, 25 topics, ~32000 shots
 - Broadcast TV news, common ASR, common annotations
- 2004 - **61** hours, 24 topics, ~33000 shots
 - More broadcast TV news
- 2005 - **140** hours, 24 topics, ~120000 shots
 - Arabic, Chinese & English broadcast TV news
 - Common ASR, translation, annotations
- 2006 - **156** hours, 24 topics, ~140000 shots
 - Same broadcast TV news, ASR, translation & annotations
- 2007 - **100** hours, 24 topics, Dutch TV shows
 - Additionally, 100 hours of BBC rushes video

2006 Evaluated tasks: 54 finishers 1

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

Accenture Technology Labs	USA	-- -- -- RU
AIIA Laboratory	Greece	SB -- -- --
AT&T Labs - Research	USA	SB -- SE RU
Beijing Jiaotong U.	China	-- -- SE --
Bilkent U.	Turkey	-- FE SE --
Carnegie Mellon U.	USA	-- FE SE --
Chinese Academy of Sciences (CAS/MCG)	China	-- -- -- RU
Chinese Academy of Sciences (CAS/JDL)	China	SB -- -- --
Chinese U. of Hong Kong	China	-- FE SE --
City University of Hong Kong (CityUHK)	China	SB FE SE --
CLIPS-IMAG	France	SB FE SE --
Columbia U.	USA	-- FE SE --
COST292 (www.cost292.org)	***	SB FE SE RU
Curtin U. of Technology	Australia	SB -- -- RU
DFKI GmbH	Germany	-- -- -- RU
Dokuz Eylul U.	Turkey	SB -- -- --
Dublin City U.	Ireland	-- -- SE RU
Florida International U.	USA	SB -- -- --
Fudan U.	China	-- FE SE --

2006: Extended teams

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- COST292
 - LABRI, Bordeaux
 - Delft University of Technology, Netherlands
 - Bilkent University
 - Dublin City University
 - National Technical University of Athens
 - Queen Mary, University of London
 - ITI, Thessaloniki, Greece
 - University of Belgrade
 - University of Zilina
 - University of Bristol

2006 Evaluated tasks: 54 finishers 2

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

FX Palo Alto Laboratory Inc	USA	SB	FE	SE	--
Helsinki U. of Technology	Finland	SB	FE	SE	--
Huazhong U. of Science and Technology	China	SB	--	--	--
IBM T. J. Watson Research Center	USA		--	FE	SE RU
Imperial College London / Johns Hopkins U.	UK/USA		--	FE	SE --
Indian Institute of Technology at Bombay	India	SB	--	--	--
NUS / I2R	Singapore	--	FE	SE	--
IIT / NCSR Demokritos	Greece	SB	--	--	--
Institut EURECOM	France	--	FE	--	RU
Joanneum Research Forschungsgesellschaft	Austria	--	--	--	RU
KDDI/Tokushima U./Tokyo U. of Technology	Japan		SB	FE	-- --
Kspace (kspace.qmul.net)	***	--	FE	SE	-
Laboratory ETIS	Greece		SB	--	-- --
LIP6 - Laboratoire d'Informatique de Paris 6	France	--	FE	--	--
Mediamill / U. of Amsterdam	the Netherlands	--	FE	SE	--
Microsoft Research Asia	China	--	FE	--	--
Motorola Multimedia Research Laboratory	USA	SB	--	--	--
National Taiwan U.	Taiwan	--	FE	--	--

2006: Extended teams

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- K-SPACE
 - Queen Mary University of London
 - Koblenz University
 - Joanneum Research Forschungsgesellschaft mbH
 - Informatics and Telematics Institute
 - Dublin City University
 - Centrum voor Wiskunde en Informatica
 - Groupe des Ecoles des Telecommunications
 - Institut National de l'Audiovisuel
 - Institut Eurecom
 - University of Glasgow
 - German Research Centre for Artificial Intelligence (DFKI/LT)
 - Technische University Berlin
 - Ecole Polytechnique Federale de Lausanne
 - University of Economics, Prague

2006 Evaluated tasks: 54 finishers 3

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

NII/ISM	Japan	-- FE -- --
RMIT U. School of CS&IT	Australia	SB -- SE --
Tokyo Institute of Technology	Japan	SB FE -- --
Tsinghua U.	China	SB FE SE RU
U. of Bremen TZI	Germany	-- FE -- --
U. of California at Berkeley	USA	-- FE -- --
U. of Central Florida	USA	-- FE SE --
U. of Electro-Communications	Japan	-- FE -- --
U. of Glasgow / U. of Sheffield	UK	-- FE SE --
U. of Iowa	USA	-- FE SE --
U. of Marburg	Germany	SB -- -- RU
U. of Modena and Reggio Emilia	Italy	SB -- -- --
U. of Ottawa / Carleton U.	Canada	SB -- -- --
U. of Oxford	UK	-- FE SE --
U. of Sao Paolo	Brazil	SB -- -- --
U. Rey Juan Carlos	Spain	SB -- SE RU
Zhejiang U.	China	SB FE SE --

BBC rushes task 2006: 12 finishers

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

Accenture Tecnology Labs / Siderean Software, Inc.	USA
AT&T Labs Research	USA
Chinese Academy of Sciences (CAS/MCG)	China
COST292	...
Curtin Univ.	Australia
DFKI Kaiserslautern	Germany
Dublin City Univ. / Univ. Rey Juan Carlos	Ireland / Spain
IBM	USA
Institut Eurecom	France
Joanneum Research	Austria
Marburg Univ.	Germany
Tsinghua Univ.	China

TRECVID 2007 tasks

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Shot boundary detection - ya-ya-ya
- Detecting semantic concepts/features (39)
- Searching based on topics
 - Automatic, manual, interactive
- ... and now ... automatic summarisation

- Summarisation is very 'fresh' ... decided to do this only in late Jan, already we have results.

SBD summary

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- SBD is ... hard cuts & GTs
- Can be easy, or can be hard;
- Can work on compressed, or uncompressed domains;
- Can be fast, or it can be slow;
- Can be good, or can be very good;
- ... and it might now be a solved problem;

TRECVID Feature Detection

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- (Semantic) concepts, (Semantic) features ...
- For (semantic) feature/concept identification - this is useful in retrieval and used within TRECVID for search task, and as a challenge itself;
- Initially this was done solo by groups
 - Annotate a small corpus;
 - Train a SVM as a feature detector;
- Problem was that this was not scalable to 000's of features;
- 2006 specified 39 features, and (manual) evaluation of 20 of these;
- Now, the task is much more collaborative among participants with shared resources;

2006 TRECVID Features

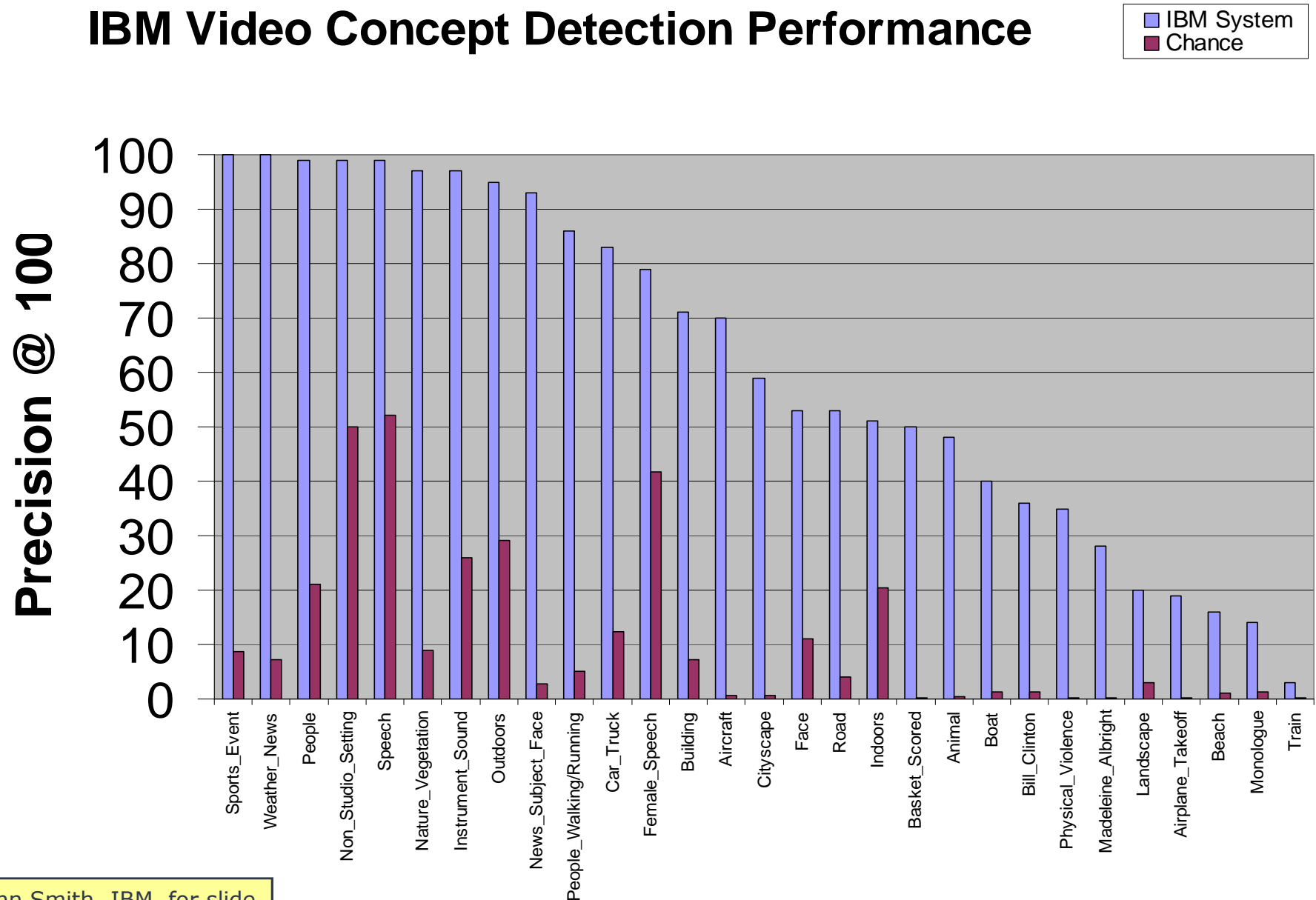
C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

1. Sports:
2. Entertainment:
3. Weather:
4. Court:
5. Office:
6. Meeting:
7. Studio:
8. Outdoor:
9. Building:
10. Desert:
11. Vegetation:
12. Mountain:
13. Road:
14. Sky:
15. Snow:
16. Urban:
17. Waterscape_Waterfront:
18. Crowd:
19. Face:
20. Person:
21. Government-Leader:
22. Corporate-Leader
23. Police_Security:
24. Military:
25. Prisoner:
26. Animal:
27. Computer_TV-screen:
28. Flag-US:
29. Airplane:
30. Car:
31. Bus:
32. Truck:
33. Boat_Ship:
34. Walking_Running:
35. People-Marching:
36. Explosion_Fire:
37. Natural-Disaster:
38. Maps:
39. Charts:

IBM Feature Detection Performance

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

IBM Video Concept Detection Performance

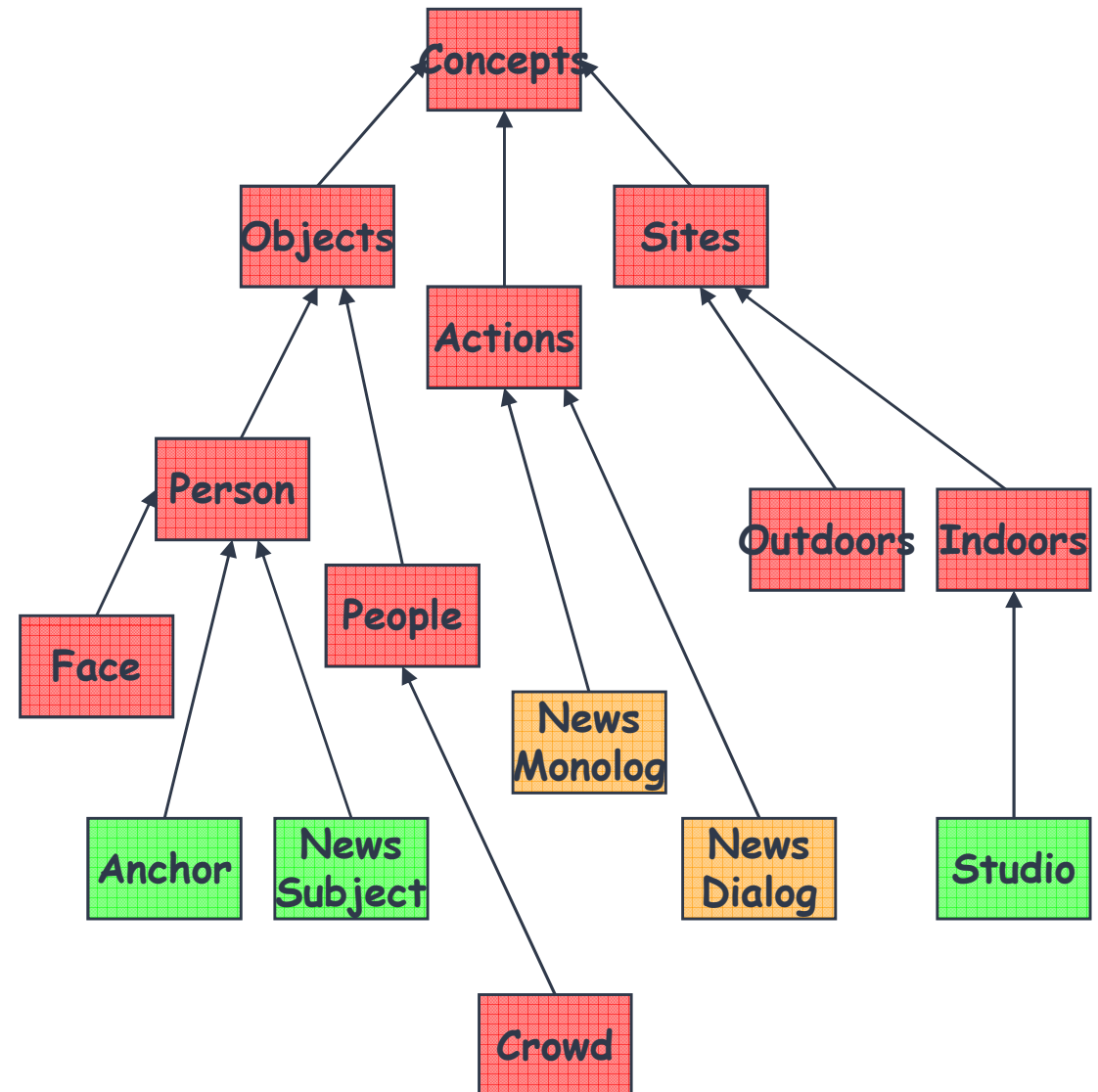


Thanks to John Smith, IBM, for slide

Feature Ontologies

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Previous work on feature detection treated features as independent;
- Now emphasis is on collections of features;
- LSCOM is a c.1000-concept ontology where concepts are related, dependent, and automatically computable;



Search Task Definition

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Goal: given a test collection, a topic and a common shot boundary reference, return a ranked list of at most 1,000 shots which best satisfy the need;
- NIST now creating more topics asking for general (vs. specific)
- In 2006 NIST created 10 of 24 topics to ask for video of events – encouraging exploration beyond *one-keyframe-per-shot*
- How were topics created ? Videos were viewed by NIST personnel, notes taken on content, and candidates emerging were chosen;

Search Task Definition

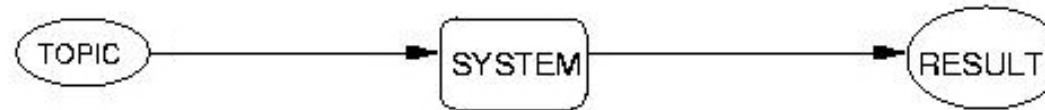
C E N T R E F O R D I G I T A L V I D E O P R O C E S S I N G

- Per-search measures: average precision, elapsed time
- Per-run measure: mean average precision (MAP)
- Interactive search participants were asked to have their subjects complete pre, post-topic and post-search questionnaires;
- Each result for a topic can come from only 1 user search; same searcher does not need to be used for all topics.

Interactive, Manual, Automatic

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

AUTOMATIC :



System takes topic as input and produces result without any human intervention

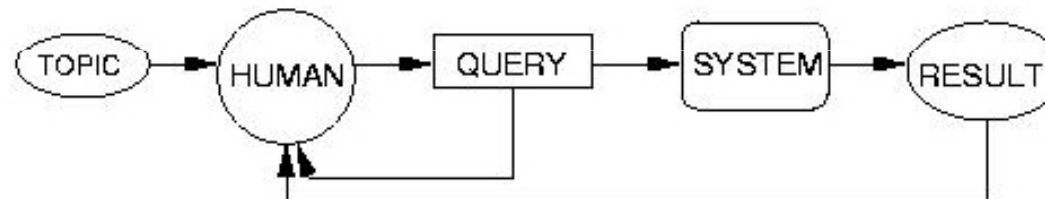
MANUAL :



Human formulates query based on topic and query interface, not on knowledge of collection or search results

System takes query as input and produces result without further human intervention

INTERACTIVE :



Human (re)formulates query based on topic, query, and/or results

System takes query as input and produces result without further human intervention on this invocation

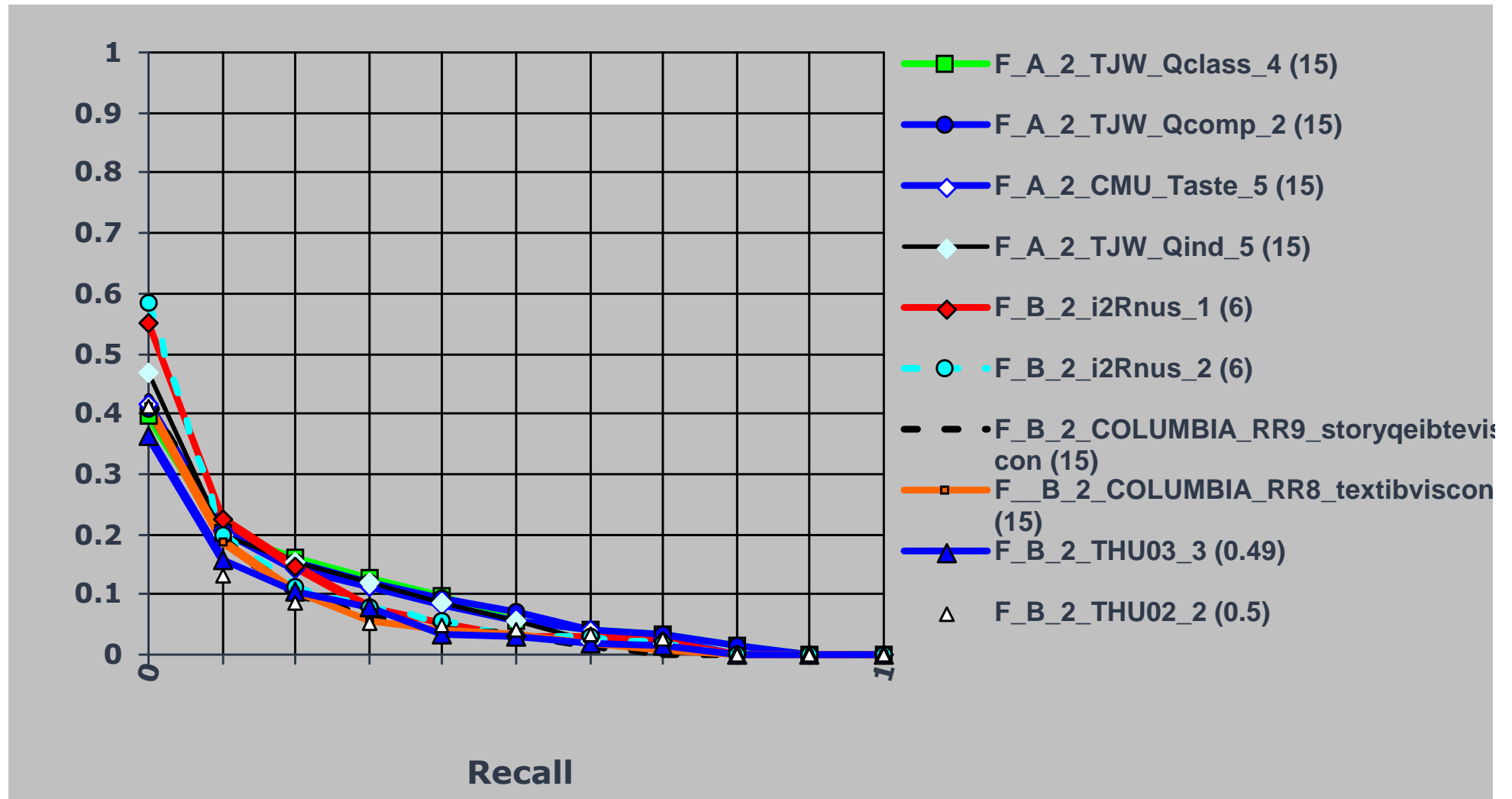
TRECVID Topics

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

149. Find shots of Condoleeza Rice [3, 6, 116]
150. Find shots of Iyad Allawi, the former prime minister of Iraq [3, 6, **13**]
151. Find Find shots of Omar Karami, the former prime minister of Lebannon [2, 5, 301]
152. Find shots of Hu Jintao, president of the People's Republic of China [2, 9, **498**]
153. Find shots of Tony Blair. [2, 4, 42]
154. Find shots of Mahmoud Abbas, also known as Abu Mazen, prime minister of the Palestinian Authority. [2, 9, 93]
155. Find shots of a graphic map of Iraq, location of Bagdhad marked – not a weather map [4, 10, 54]
156. Find shots of tennis players on the court – both players visible at the same time [2, 4, 55]
157. Find shots of people shaking hands [4, 10, **470**]
158. Find shots of a helicopter in flight [2, 8, 63]
159. Find shots of George Bush entering or leaving a vehicle (e.g., car, van, airplane, helicopter, etc), he and vehicle both visible at the same time [2, 7, 29]
160. Find shots of something (e.g., vehicle, aircraft, building, etc.) on fire with flames and smoke visible [2, 9, 169]

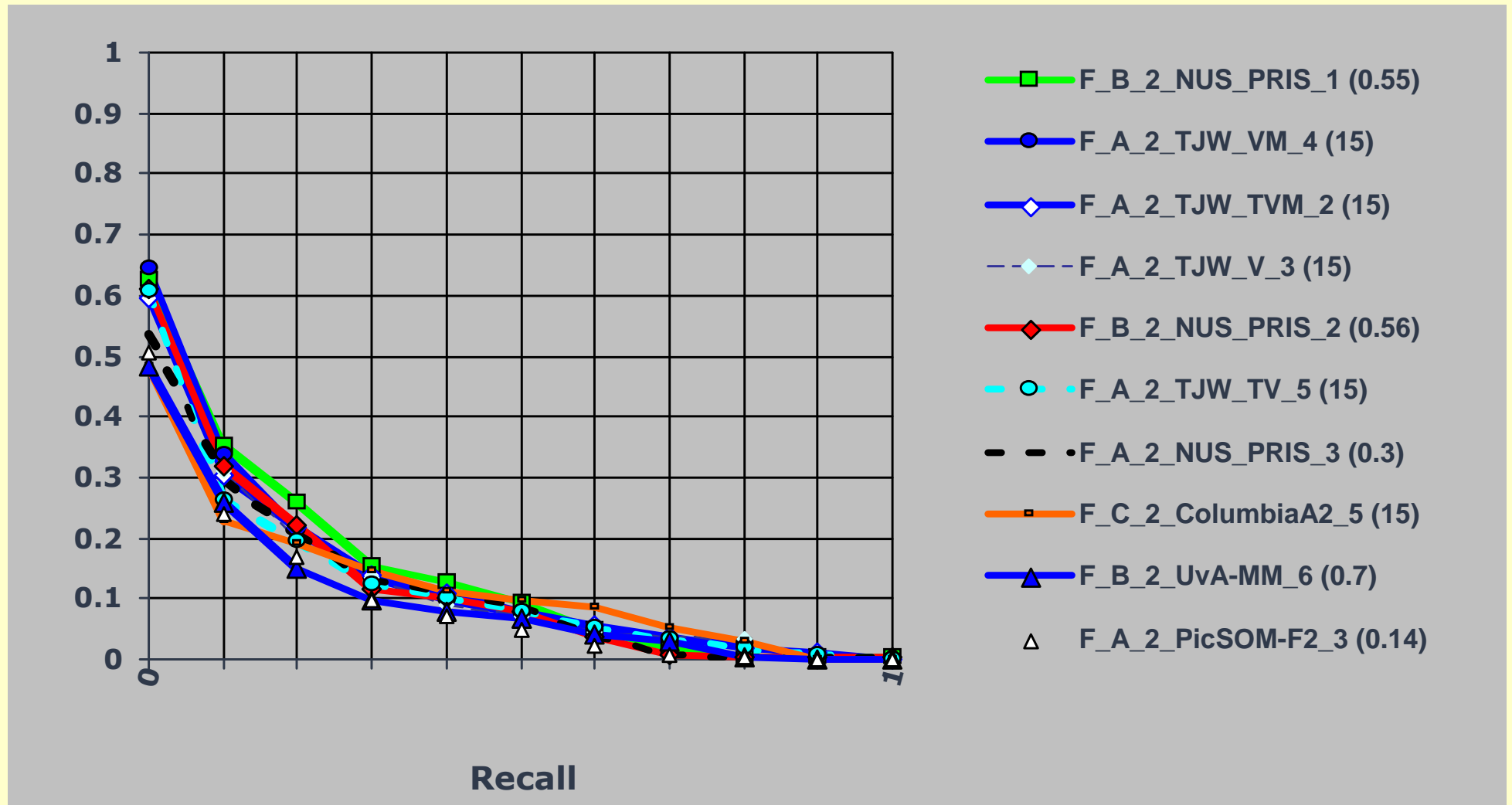
2006: Automatic - top 10 MAP (of 76)

(mean elapsed time (mins) / topic)



2005: Automatic - top 10 MAP (of 42)

(mean elapsed time (mins) / topic)



2006: Manual - top 10 MAP (of 11)

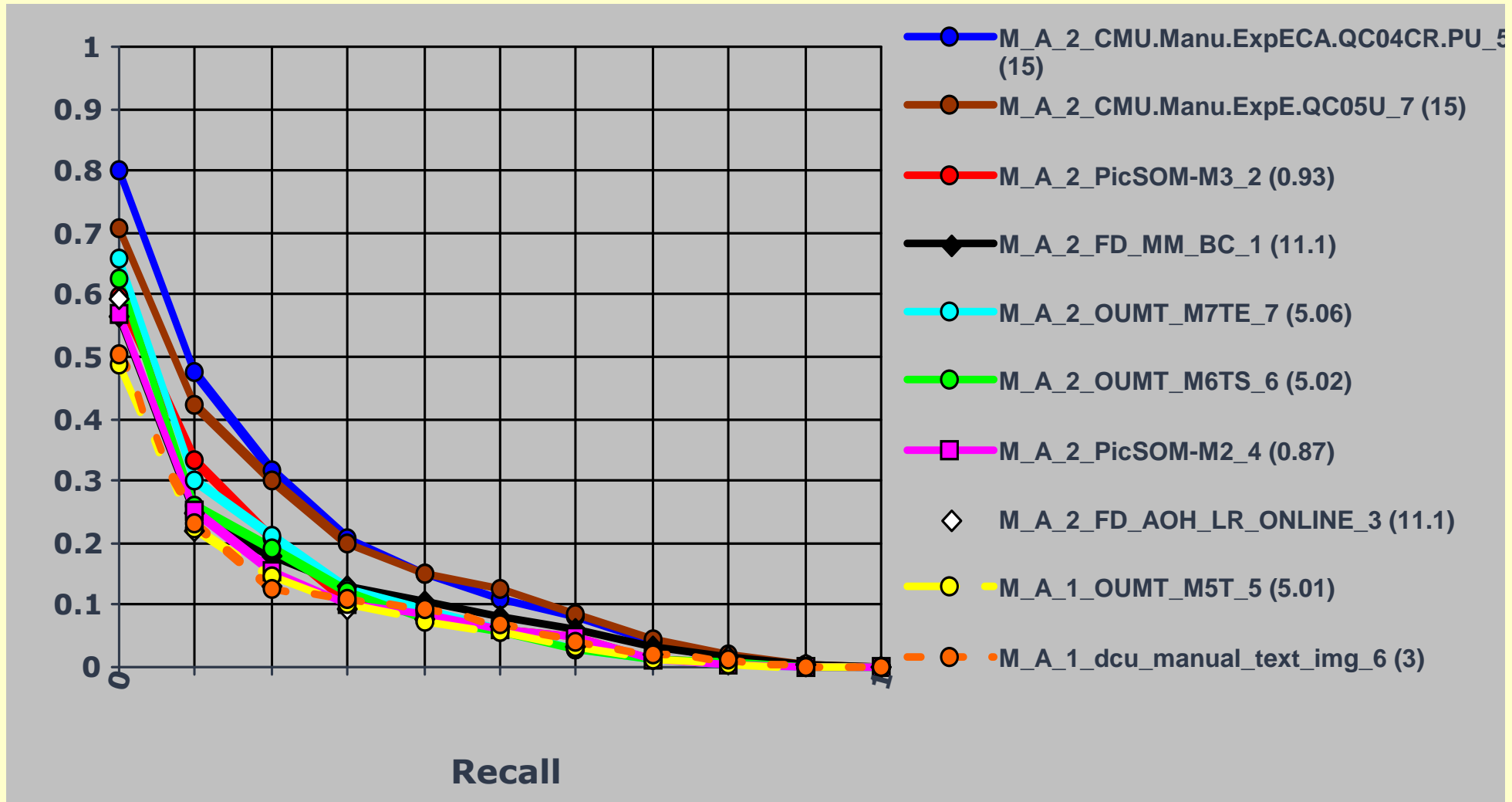
(mean human effort (mins) / topic)



2005: Manual - top 10 MAP (of 26)

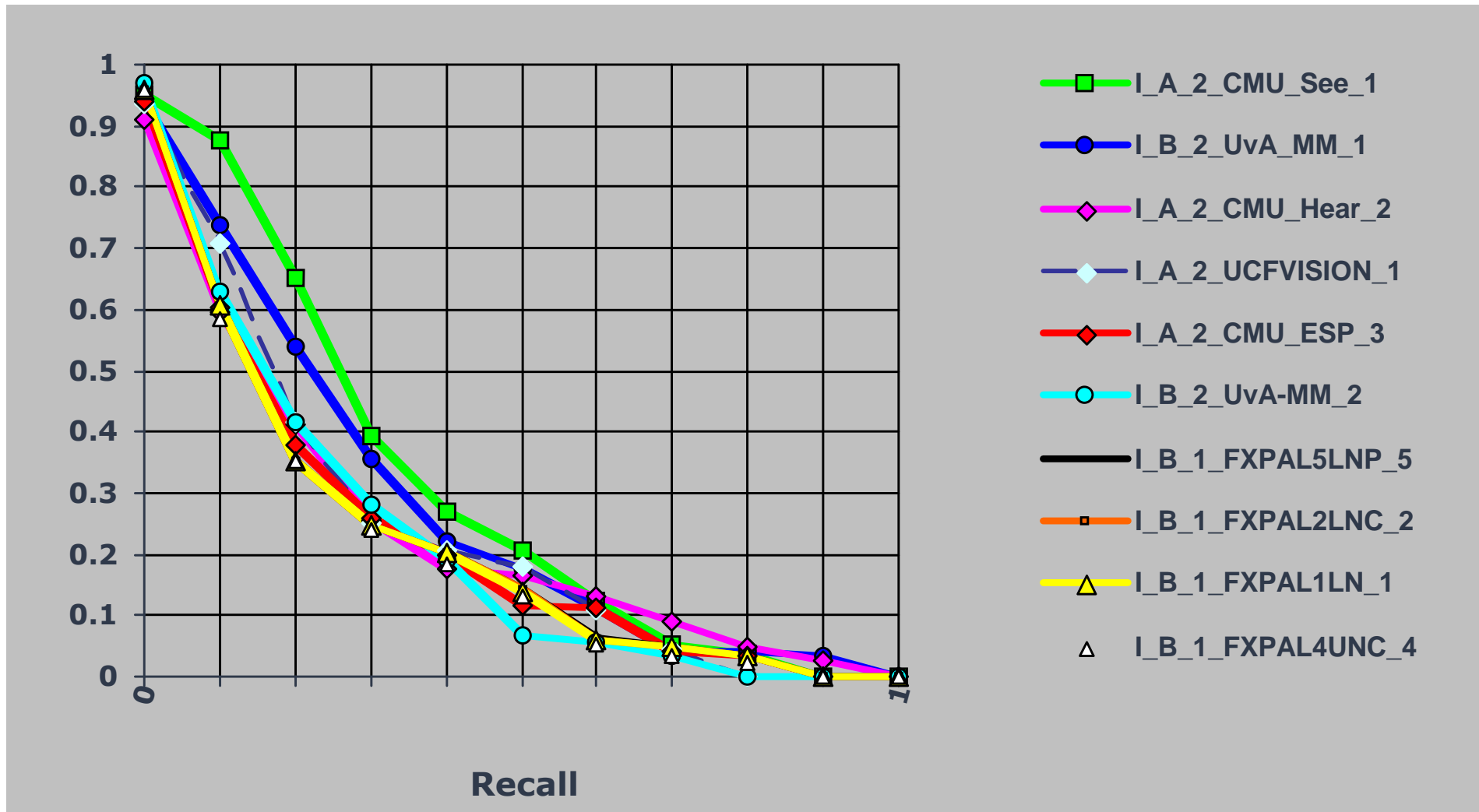
(mean human effort (mins) / topic)

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G



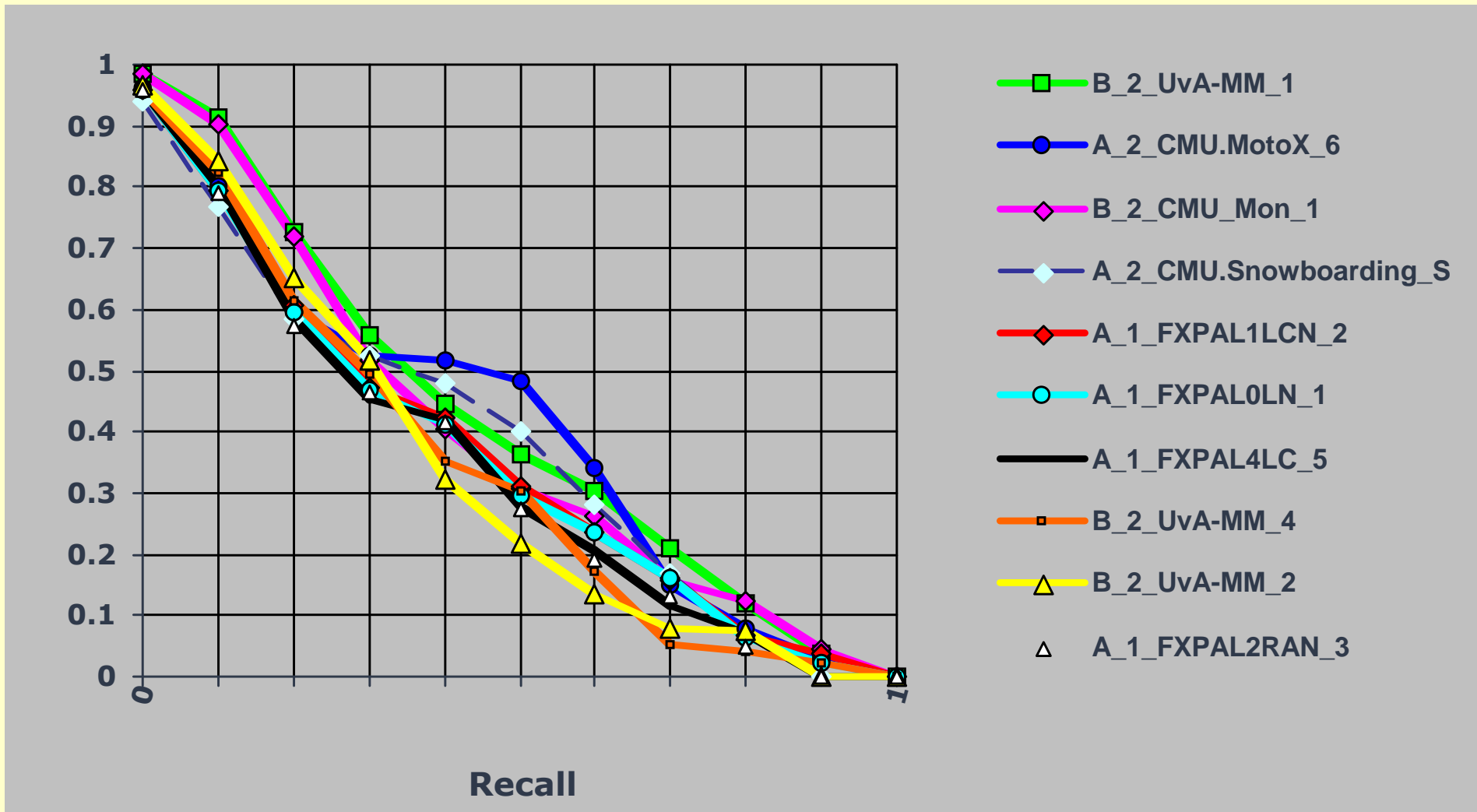
2006: Interactive - top 10 MAP (of 36)

(mean elapsed time for all == ~15 mins/topic)



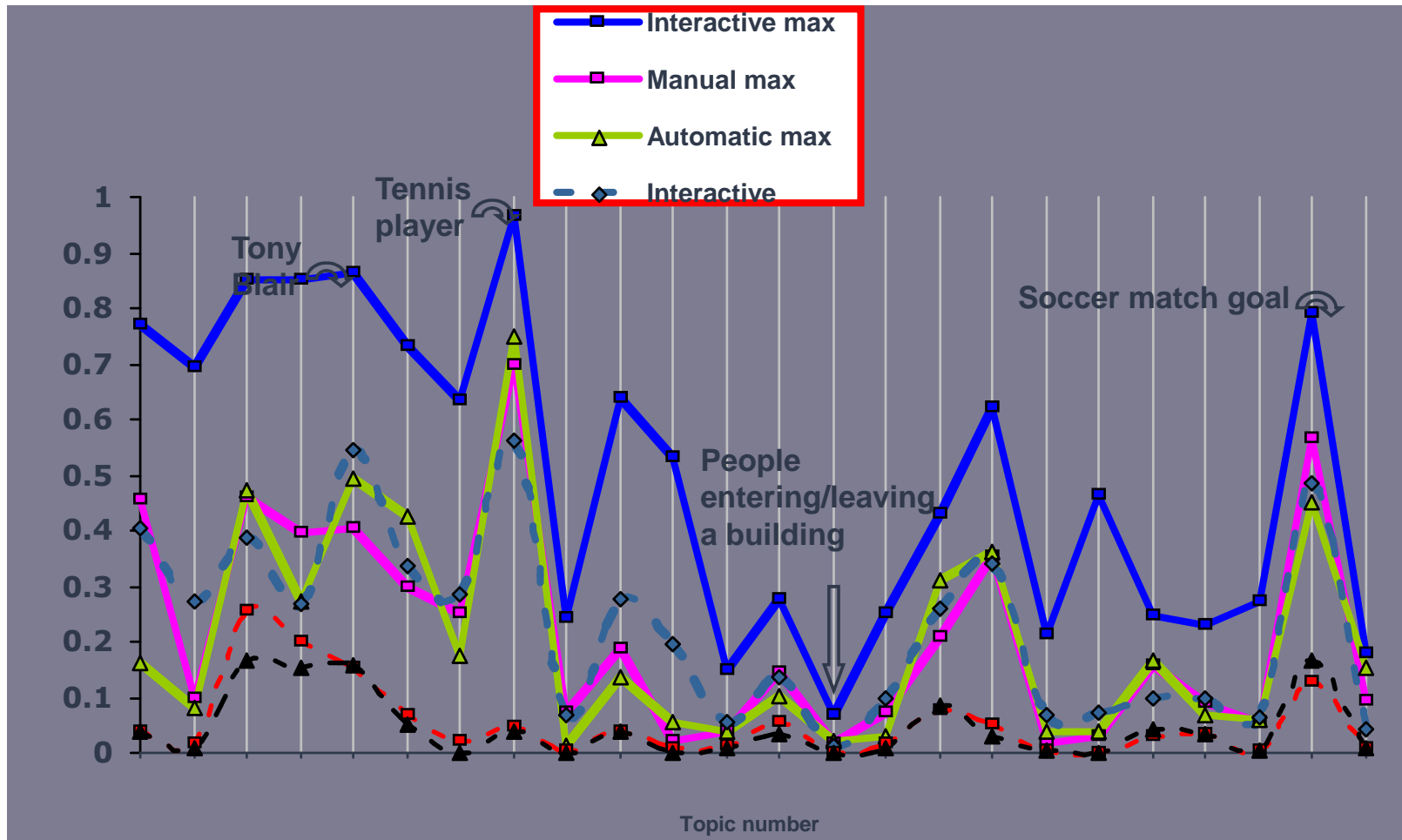
2005: Interactive - top 10 MAP (of 44)

(mean elapsed time for all == ~15 mins/topic)

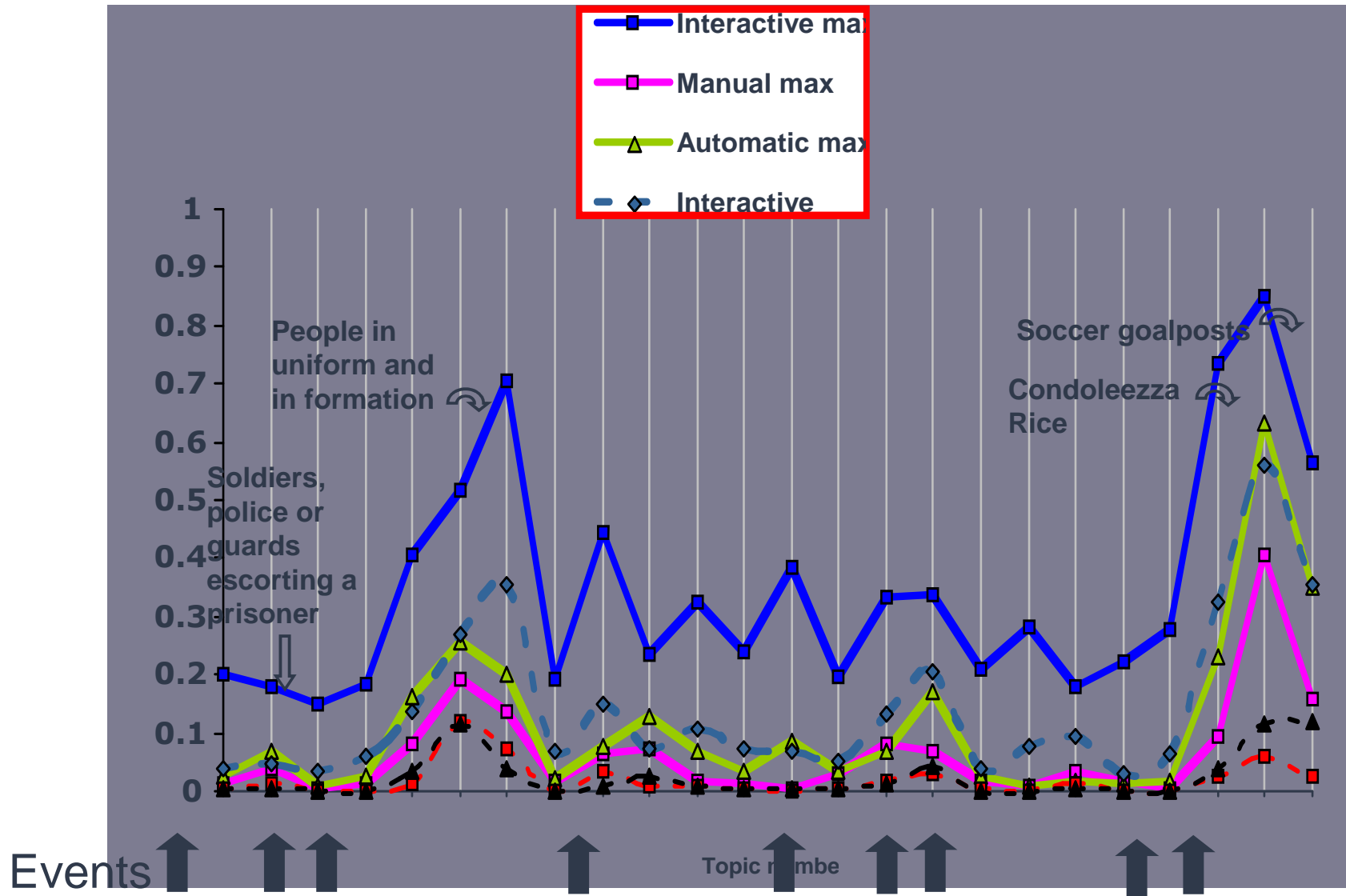


2005: Mean avg. precision

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G



2006: Average precision by topic



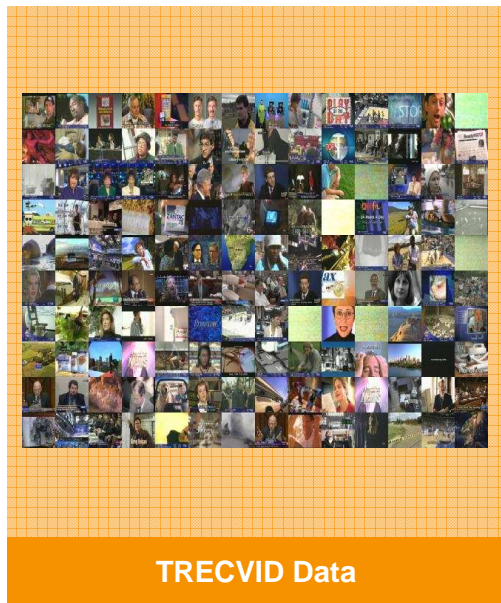
TRECVID 2005 Donations

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

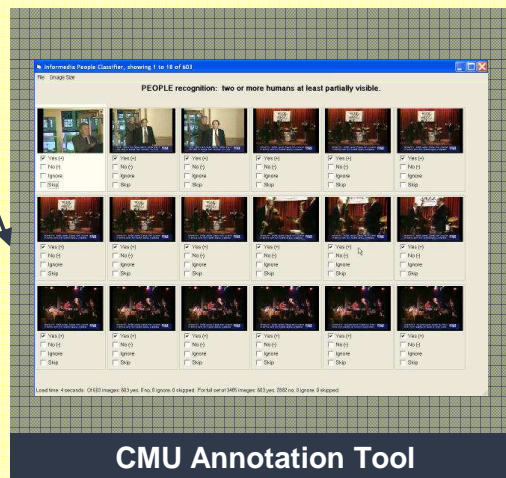
- ... in addition to the video (MPEG-1), closed caption text, MT of CCs, major donations to TRECVID 2005:
 - Master shot reference (Fraunhofer Institute, Berlin)
 - Keyframes for each master shot (DCU)
 - Feature annotation tools (IBM, CMU)
 - Camera motion annotation tool & output (JRS, Austria)
 - Feature annotation (20+ research groups) for 39 features on 50 hours of video
 - Feature detection submissions from all groups for search
 - Low level feature detection output (CMU)
 - Story segmentation output (Columbia University)
- these donations enrich the evaluation, help progress research in the field, and allow easier break-in to the field
- Since then UvA donate MediaMill-101, Columbia donate LSCOM-400 .. features.
- So you could survive profitably on the donations of others, and never even see a frame of video, but participate and do well.

TRECVID-2005 Annotation

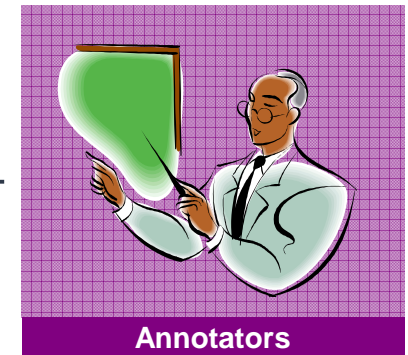
- 80 Hours Video (from development data set)
- 40 Features (from LSCOM-lite)



TRECVID Annotation Task



TRECVID
2005
Common
Training
Data Set



“Volunteers” from
TRECVID
participating
groups

Up to 100%

100%

Up to 100%

35 concepts

X%

100%

100-X%

18 concepts

Thanks to John Smith, IBM, for slide

2007 Annotation

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Coordinated by Georges Quenot @ UJF Grenoble;
- Web-based, concept based, easier/nicer tool to use;
- Annotation used *active learning* to decide what to present to annotator;
- Overall goal was not to (have to) annotate all development data and present shots to annotators in such a way to maximise positives;
- Continuously learning, despite annotation by c.20+ groups;
- Annotation data available to annotators;
- LSCOM annotation is publicly available;

Video Summarisation

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Summary == condensed version of something s.t. judgments about the full thing can be made in less time and effort than the full thing;
- In a world of information overload, summaries have widespread application as surrogates resulting from searches, as previews, as familiarisation with unknown collections;
- Video summaries can be keyframes (static storyboards, dynamic slideshows), skims (fixed or variable speed) or multi-dimensional browsers;
- Literature & previous work shows interest in evaluating summaries, but datasets always small, single-site, closed;

Summarisation Data

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- BBC provided 11 boxes of tapes, 250 ? hours of rushes ... Casualty, House of Elliot, Jonathan Creek, Ancient Greece, Between the Lines & other miscellaneous;
- Rushes ... we digitised 100 hours, each tape -> 1 file of up to 35 minutes duration, average 25 minutes;
- 50 files as development data, 42 files as test data, mixture of sources ... scripted dialogue, environmental sounds, much repeating (==redundancy), wasted shots, clapboards and colourbars;
- System task ... create an MPEG-1 summary of max 4% the original, no interaction, just playback, eliminate redundancy and maximise viewers' efficiency at recognising objects & events as quickly as possible;

Evaluating Summaries

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- How to evaluate a summary ? It is intractable to evaluate a technique to identify all the content of an original video, then do likewise for a summary, and then compare them, in a format which is repeatable and affordable;
- So we approximated, by creating a manual ground truth for the original (42) videos and having assessors view the summaries and assess the groundtruthed content, present or recollectable, in the summary;

Sample groundtruth #1

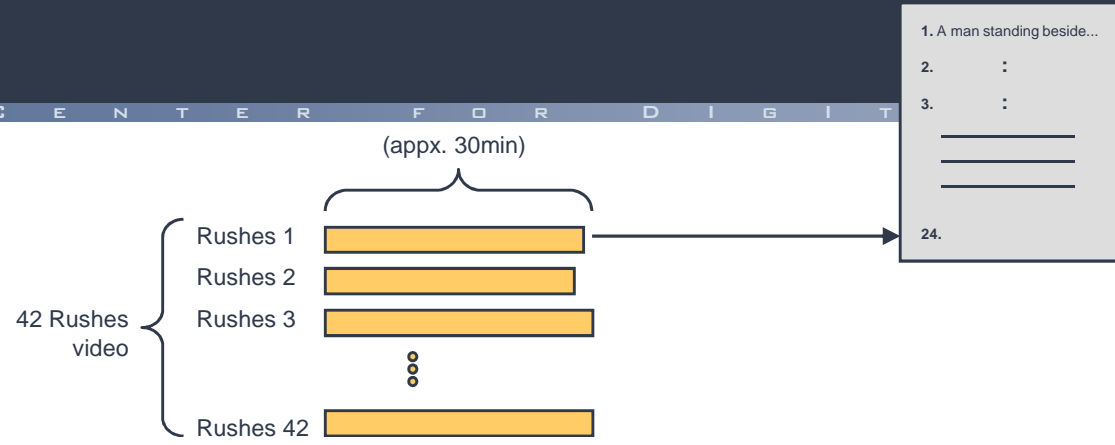
C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

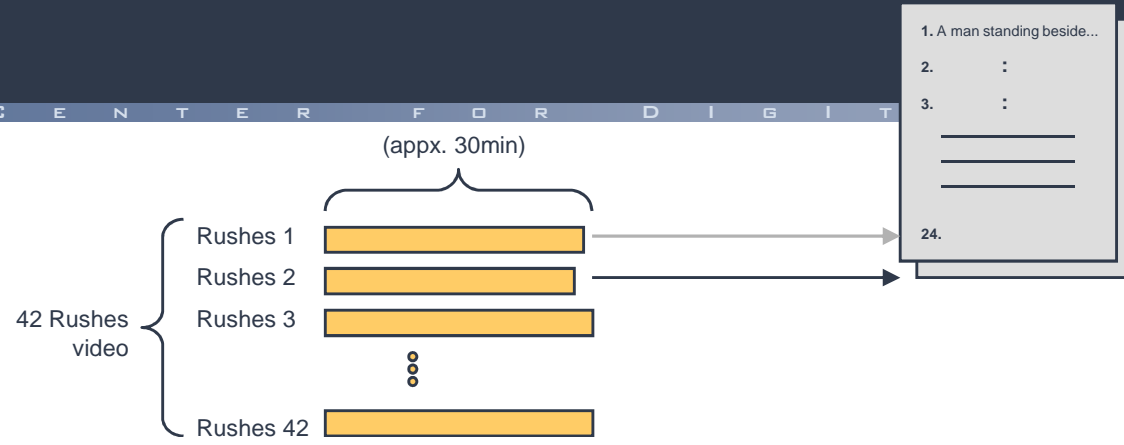
1. two men talk at table on terrace with tree trunk on right
2. two men talk at table, close up, facing man with moustache, shoulder of other
3. Close-up of man with moustache (face and shoulders)
4. two men talk at table, close up, facing man wearing tie, smoking
5. Close-up of man in tie (face and shoulders)
6. man with moustache kneels in garden, talks to men in blue suit behind him.
7. Close-up of man with blue suit (shoulders and head)
8. Close-up of man with moustache (shoulders and head)
9. gates of fortified estate open far away, red car exits.
10. gates of fortified estate open and close far away.
11. 4 people talk, sit around desk

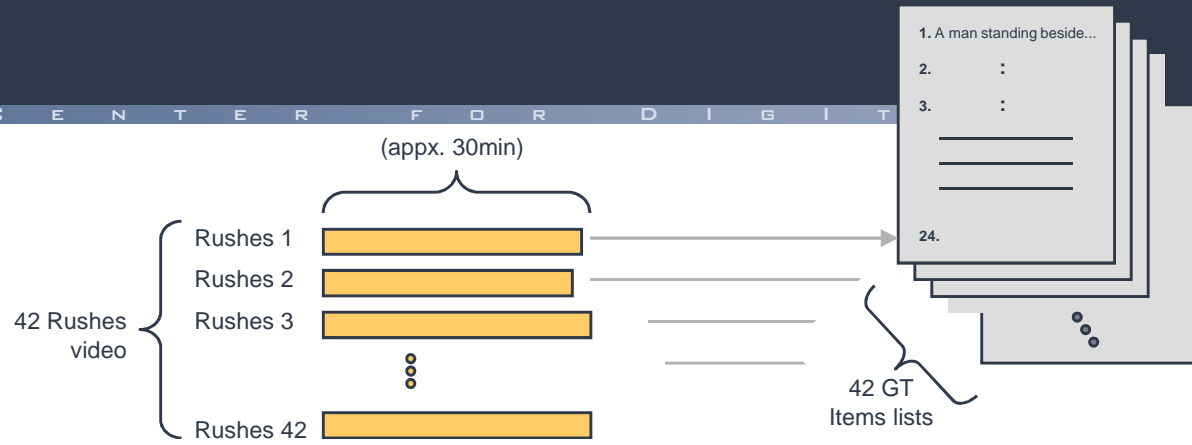
Sample groundtruth #2

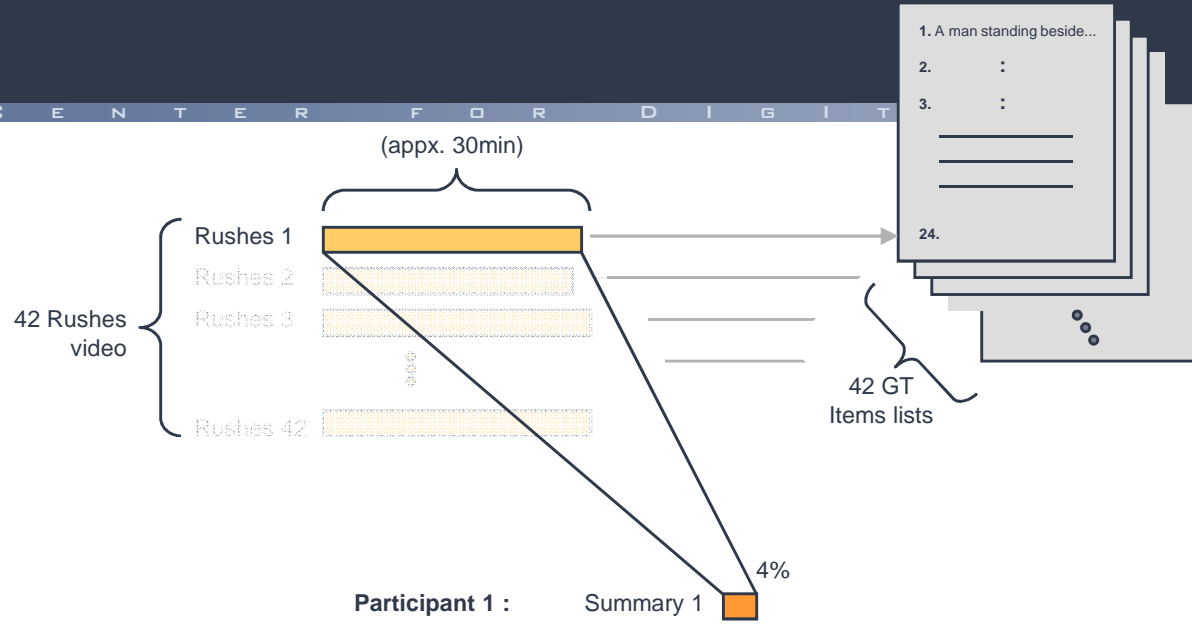
C E N T R E F O R D I G I T A L V I D E O P R O C E S S I N G

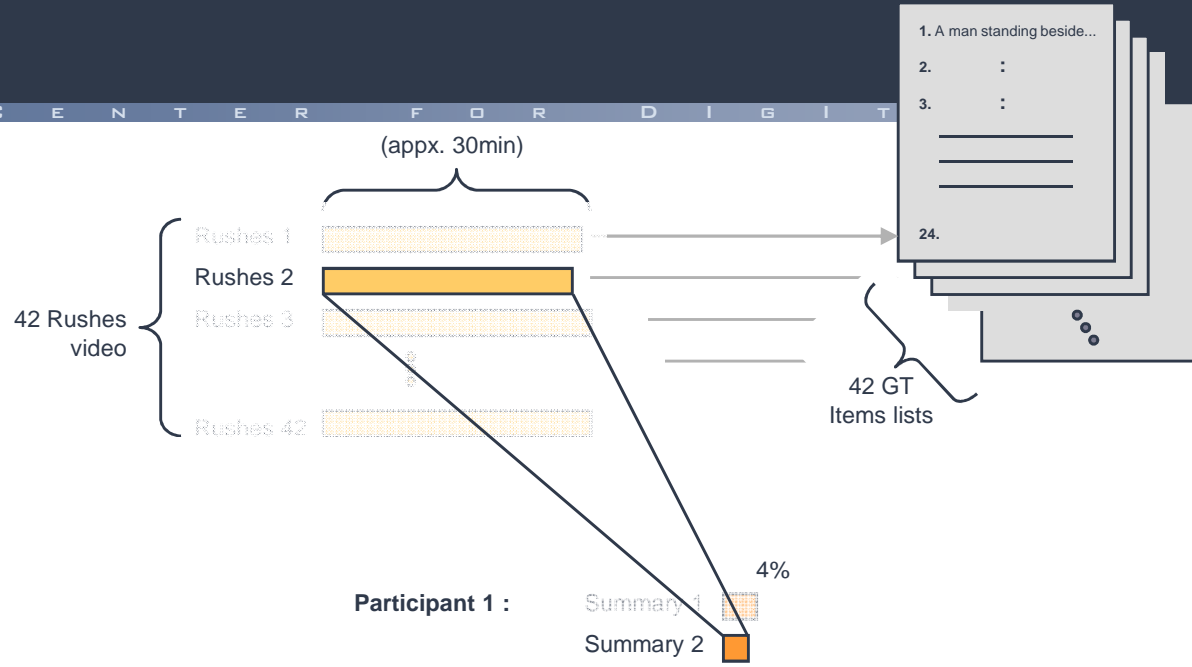
1. closeup of woman, in shadow, rubbing her face
2. man sits on the ground with a monkey on a chair, no one else in the room
3. man sits on the ground with a monkey on a chair, old woman enters scene
4. man sits on the ground with a monkey on a chair, old woman stands at table
5. man sits on the ground with a monkey on a chair, old woman stands at table, woman in red enters scene
6. man sits on the ground with a monkey on a chair, two women stands at table
7. man sits on the ground with a monkey on a chair, old woman stands at table, woman in red exits scene
8. man sits on the ground with a monkey on a chair, old woman exits scene
9. zoom in on man sitting on the ground with a monkey on a chair and old woman standing at table
10. man sits on the ground with a monkey on a chair, old woman standing at table, woman in red sitting
11. man sits on the ground with a monkey on a chair, mans legs visible
12. man sits on the ground with a monkey on a chair, mans legs visible, woman passes in from of them
13. man sits on the ground, monkey exits scene, mans legs visible,
14. man sits on the ground stands up and exits scene
15. closeup of monkey on the chair
16. empty chair
17. closeup of mans, head and shoulders only visible
18. closeup of blonde womans face, head and shoulders only visible
19. closeup of blonde womans face, head and shoulders only visible, old woman visible in the background
20. empty kitchen scene, two doors visible in the background
21. old woman enters kitchen scene, two doors visible in the background
22. woman in red enters kitchen scene, two doors visible in the background
23. two women standing at table in kitchen scene, two doors visible in the background
24. woman in red sits down at the table in the kitchen, two doors visible in the background
25. man walks around kitchen, two doors visible in the background
26. man picks up cup off of kitchen table
27. camera pans left as old woman walks through doorway
28. old woman picks up pad of paper off of cabinet
29. camera pans left as woman in red walks through doorway
30. woman in red reads letter in her hand
31. woman in red fixes her hair in a mirror
32. closeup as woman in red sitting down, no one else is visible
33. woman in red sits picks up coffee jug
34. closeup of woman in red, while woman in blue walks behind her
35. closeup as woman in red drinks, while man walks behind her

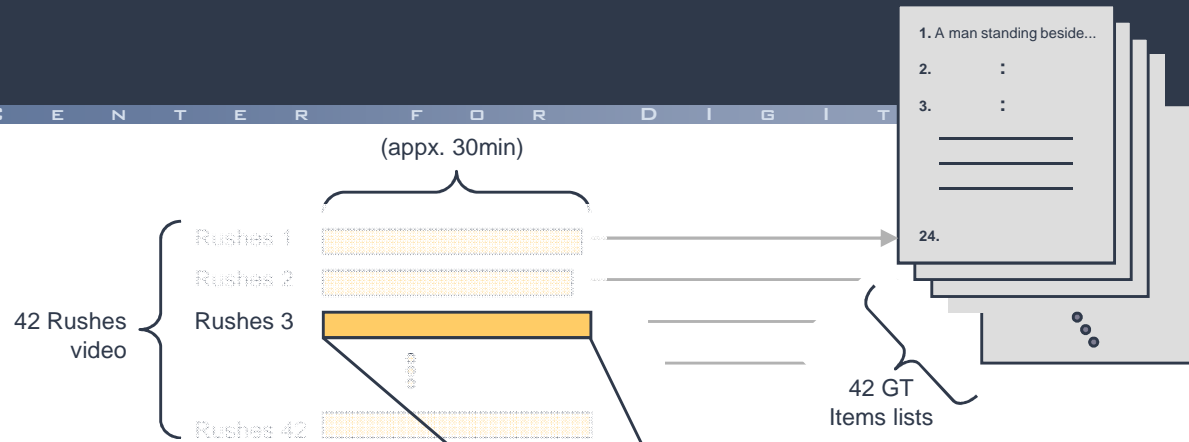


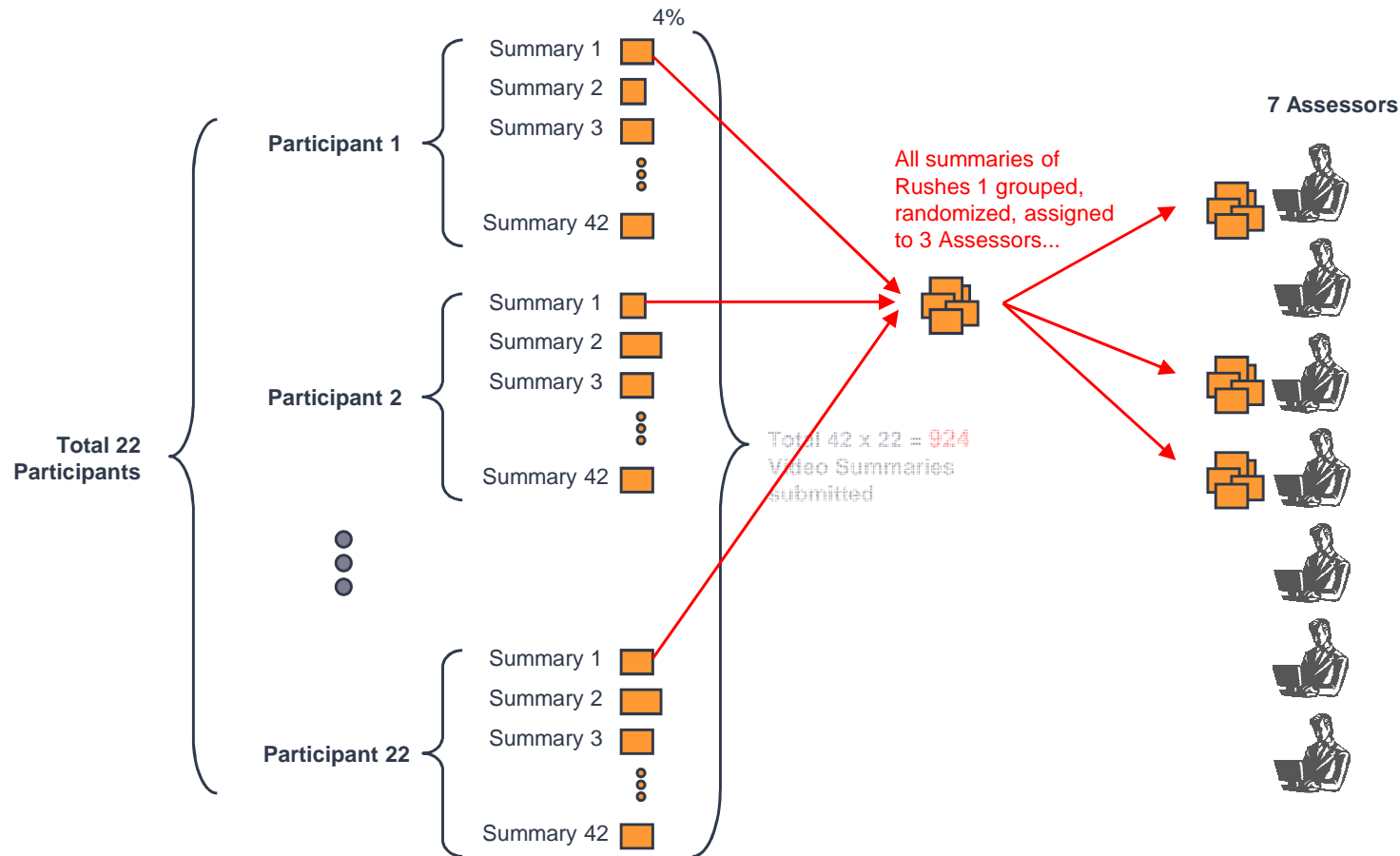
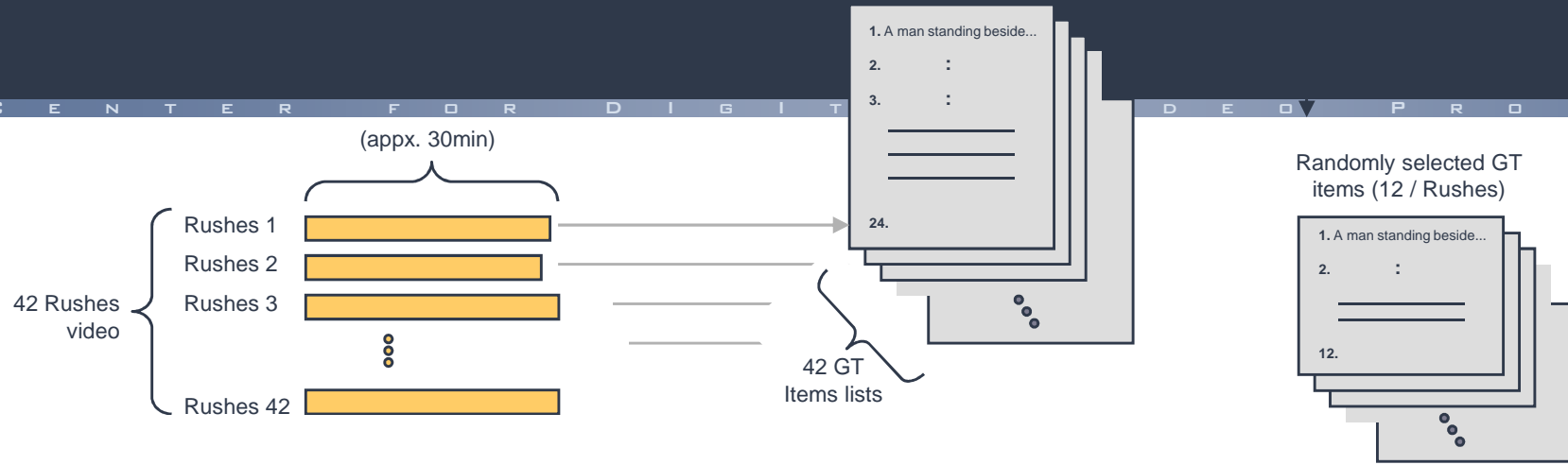


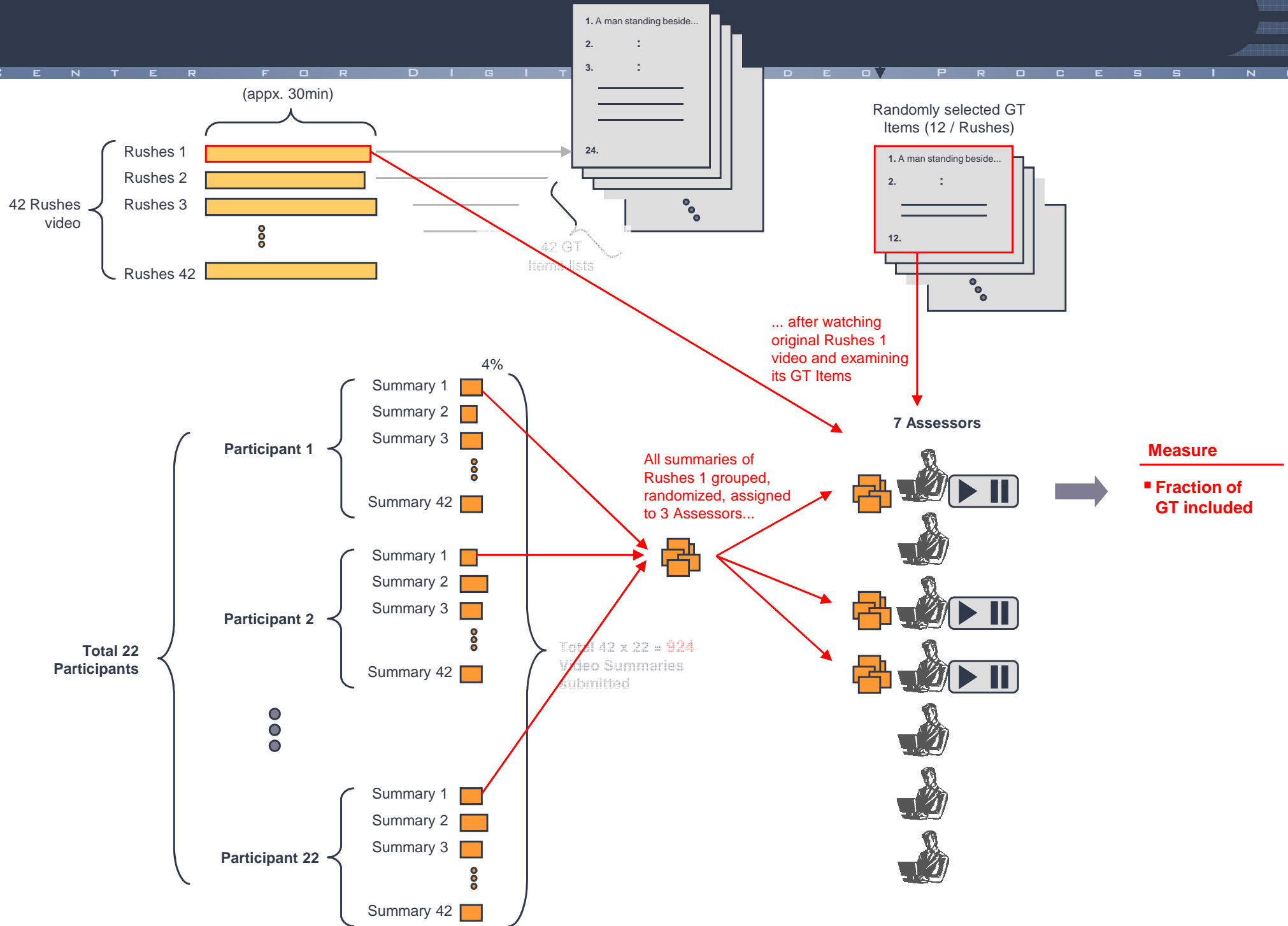


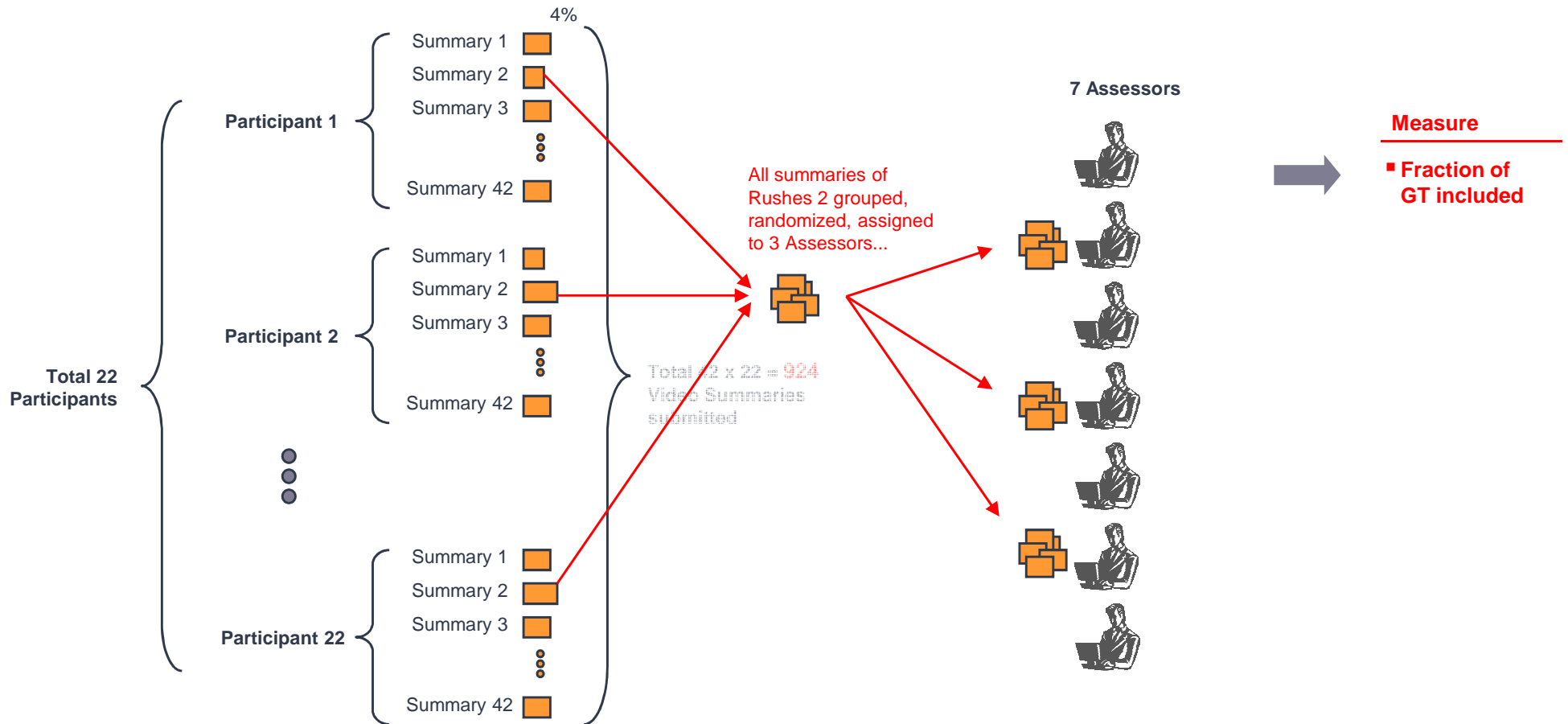
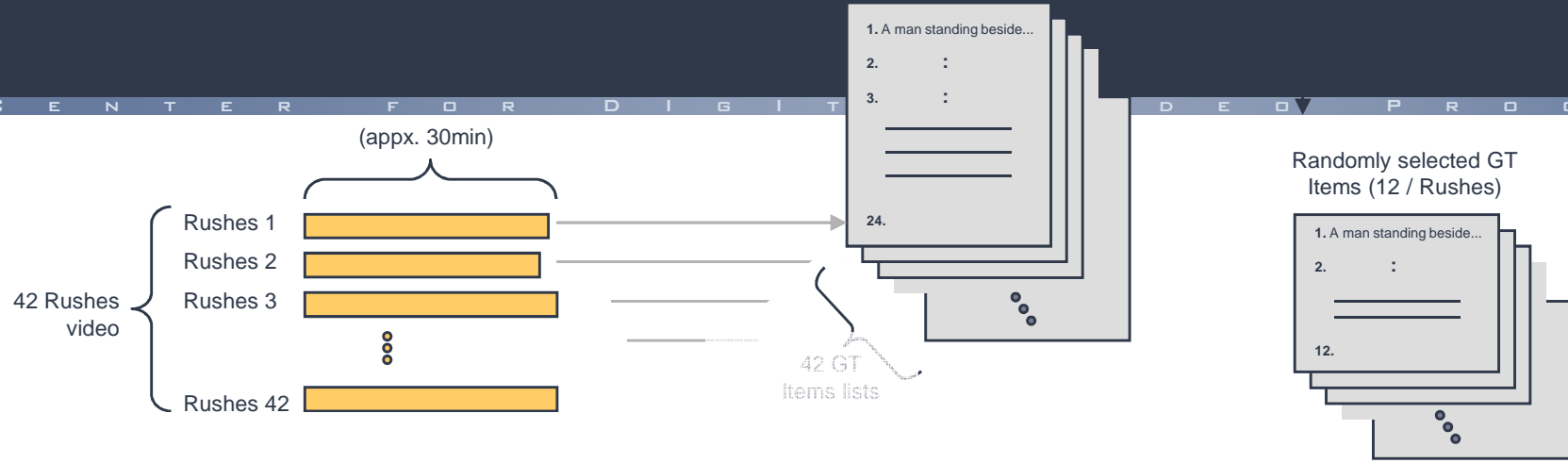


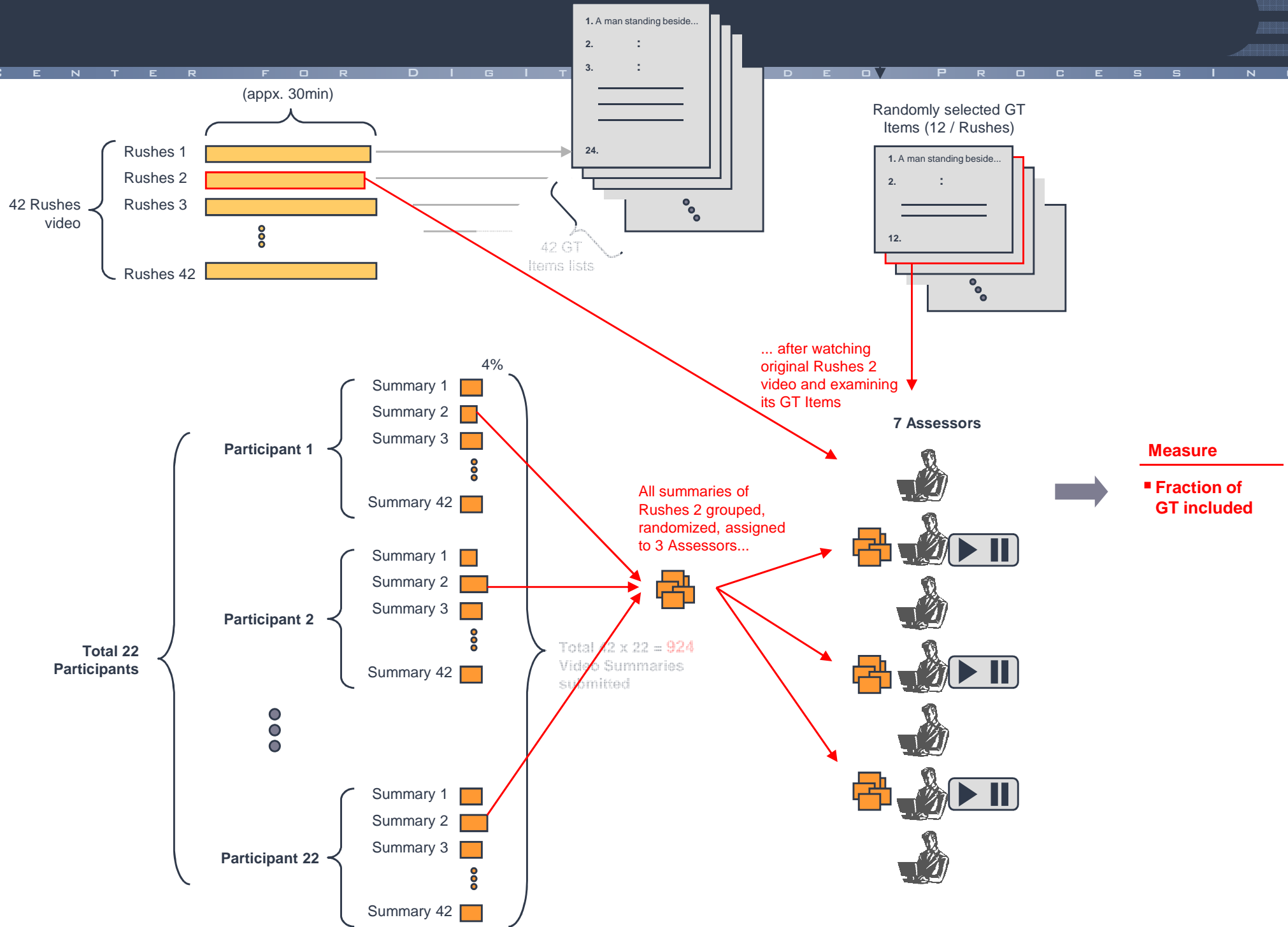


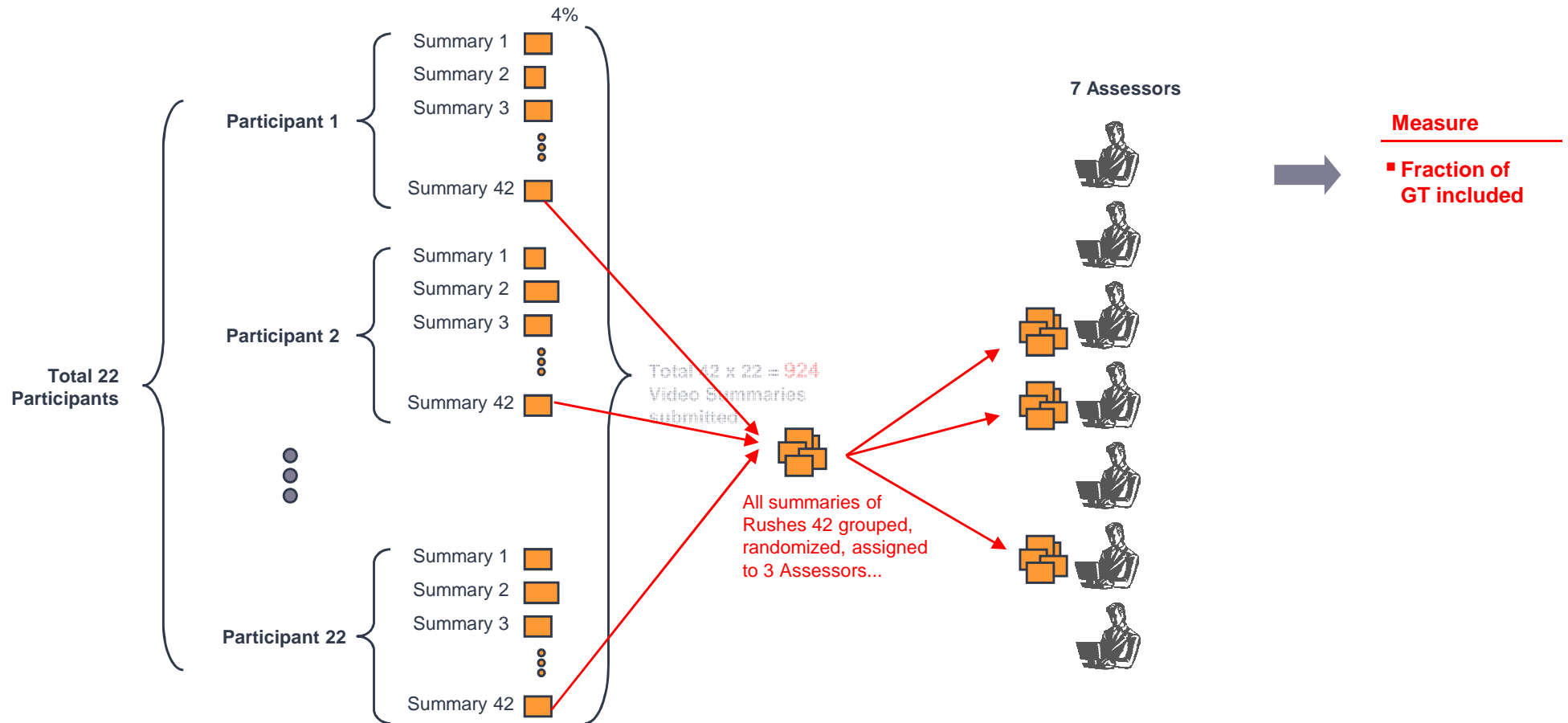
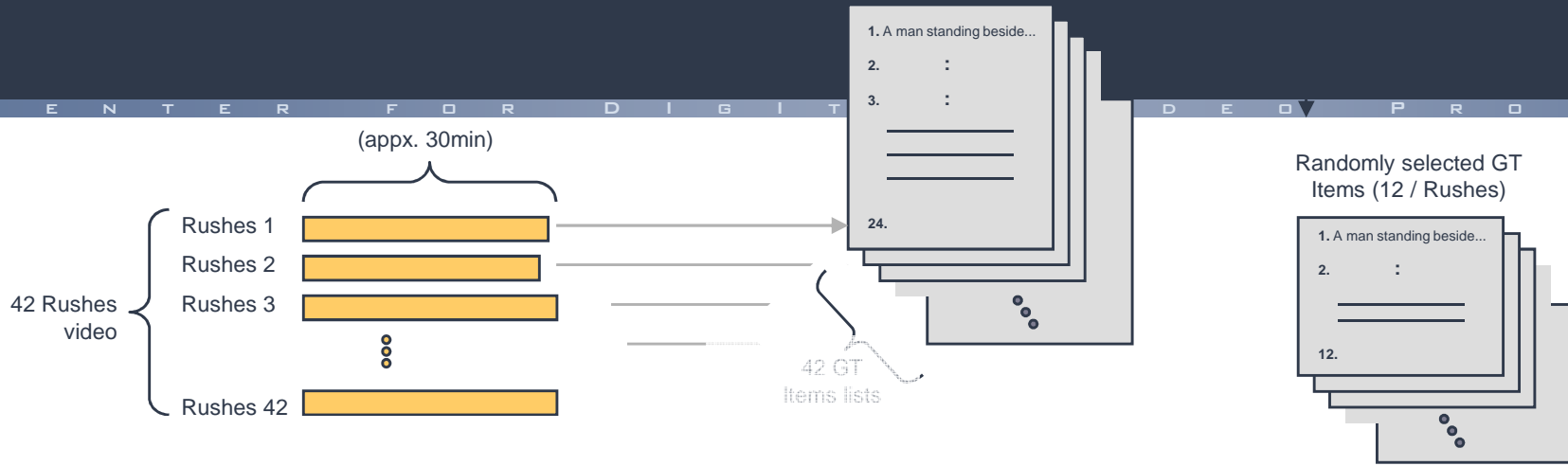


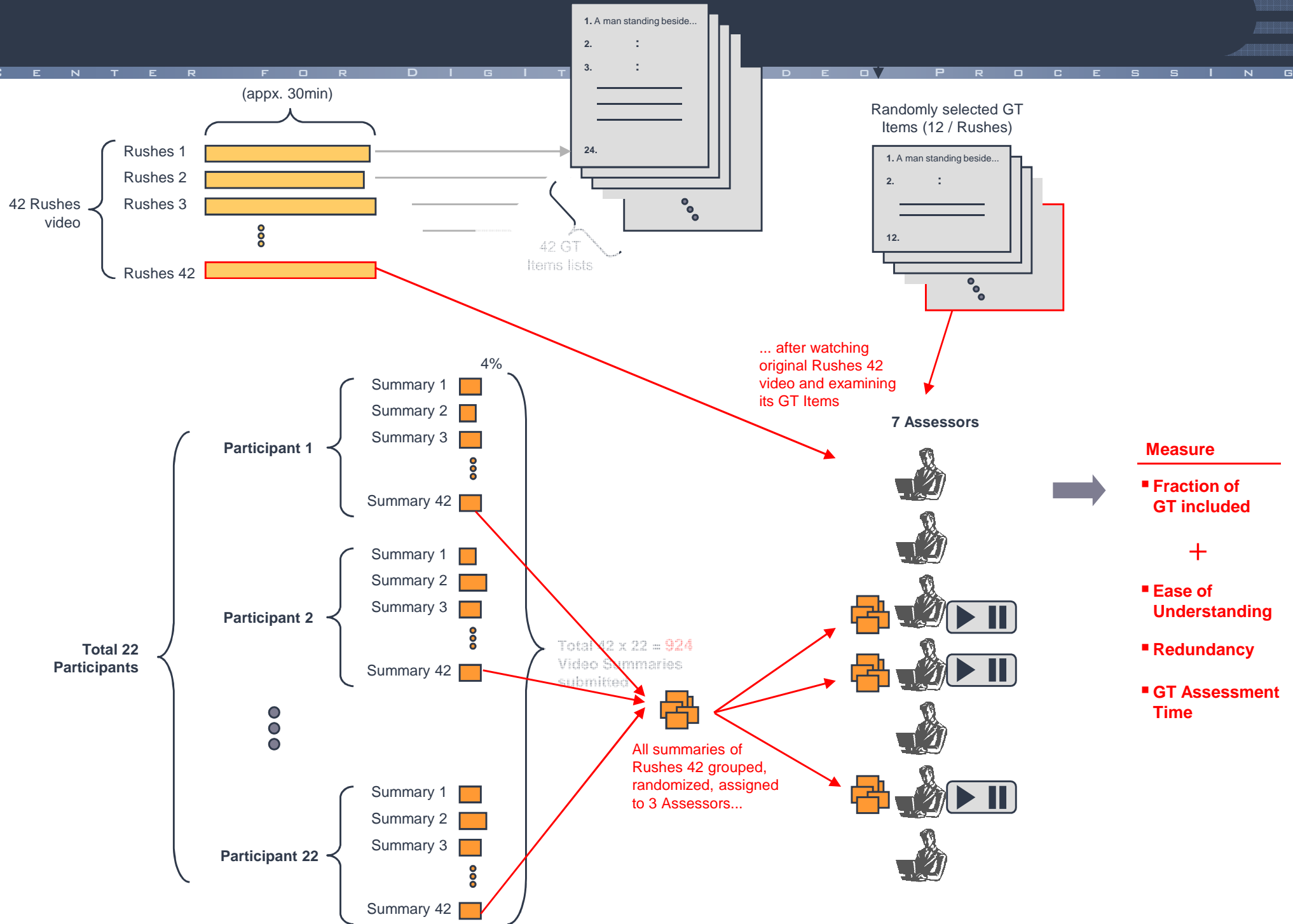












Measures

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Fraction of (12 items of) groundtruth found;
 - Ease of use and amount of near-redundancy, as judged by assessors;
 - Assessment time taken;
 - Summary duration;
 - Summary creation compute time;
-
- 22 groups from 13 countries completed submissions, system papers received end-June!

Participant approaches

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

1. At&T: shot clustering to remove redundancy, use shot with most speech/faces;
2. Brno Univ.: cluster shots using PCA, remove junk shots;
3. CMU: k-means clustering using iterative colour matching, audio coherence;
4. City UHK: obj. detection, camera motion, keypoint matching for repetitive shots;
5. Columbia: duplicate shot detection and ASR;
6. Cost292: face, camera motion, audio excitement;
7. Curtin U: shot clustering using SIFT matching;
8. DCU: amount of motion & faces for keyframe selection;
9. FXPAL: colour distribution, camera motion, for repetition detection;
10. HUT: SOMs for shot pruning to eliminate redundancy;
11. HKPU: junk shot removal, visual & aural redundancy;

Participant approaches

C E N T R E F O R D I G I T A L V I D E O P R O C E S S I N G

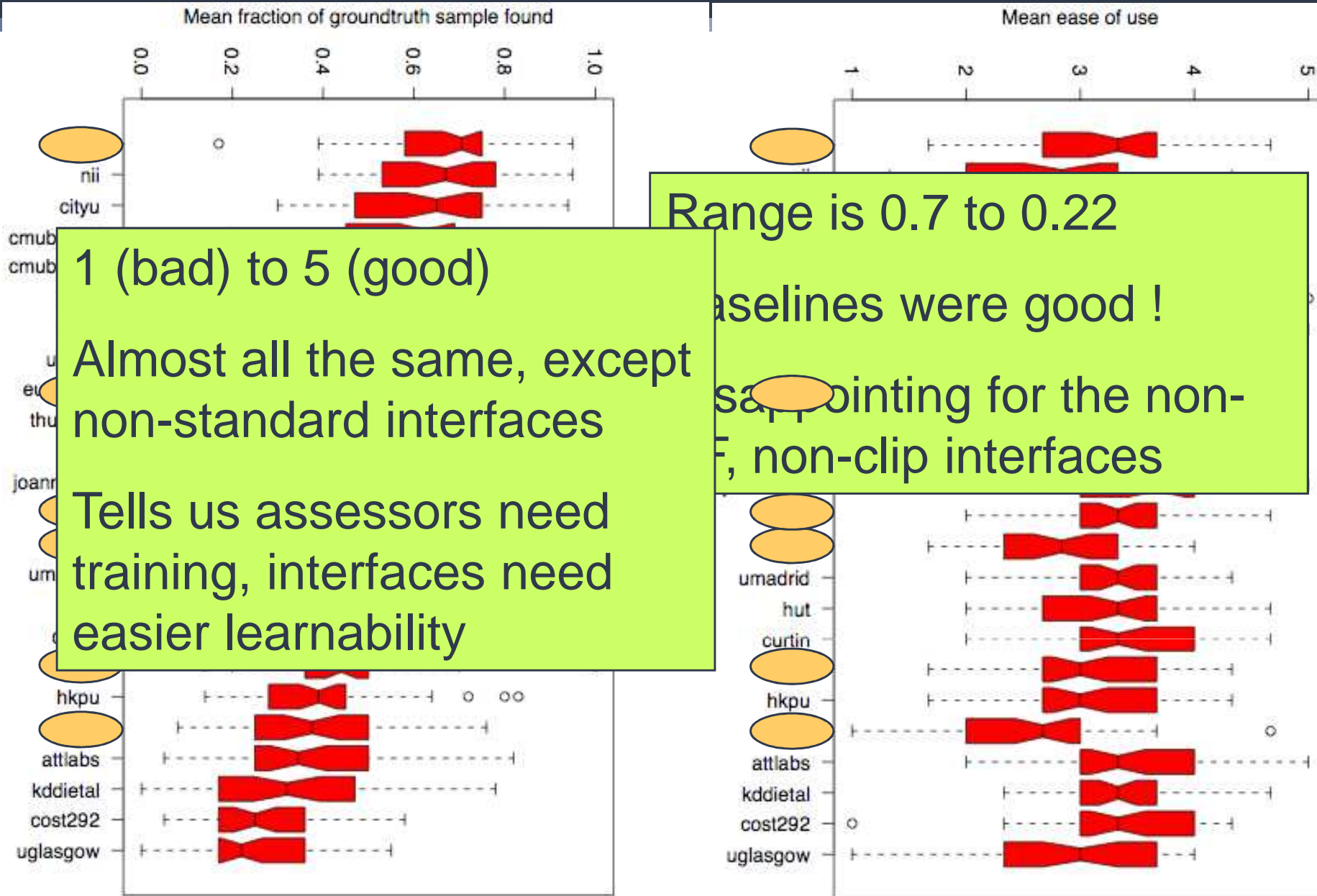
12. Eurecom: determine the most non-redundant shots;
13. Joanneum: variant of LCSS to cluster re-takes of same scene;
14. KDDI: use only low-level features for fast summarisation;
15. LIP6: eliminate repeating shots using 'stacking' technique;
16. NII: feature extraction and clustering;
17. Natl. Taiwan U: LL shot similarity & motion vectors, then cluster;
18. Tsinghua/Intel: keyframe clustering, repetitive segments, main scenes/actors;
19. UCSB: k-means clustering on HL features, speech, camera motion;
20. Glasgow: 0-1 knapsack optimisation problem, shot clustering;
21. UA Madrid: single pass for realtime clustering on-the-fly, colour-based;
22. Sheffield: concatenate some frames from middle of each shot;

Summary formats

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Plain clips: COST292, Curtin, HKPU, KDDI, Madrid, NTU, Sheffield;
- Clips of 1s duration: CMU, CUHK, Helsinki, UCSB;
- Clips FF: NII;
- Main scene/actor, then clips: Tsinghua;
- Plain keyframes: Glasgow;
- Clips with numeric/text indicators of offset/re-takes: AT&T, JRS;
 - **Columbia** - clips w. picture in picture showing repetition & also showing numeric offsets
 - **U Brno** - clips w. picture in picture showing iconic scrollbar offsets, redundancy, scrollbar progress
 - **Eurecom** - clips in 4-windows, FF, clustered, no indicators
 - **LIP6** - clips with VSFF, speed indicator and numeric offsets
 - **FXPAL** - clips with variable speed FF and numeric and iconic offset
 - **DCU** - KFs w/ metadata showing offset, faces, motion

Results: fraction GT/ease of use



Range is 0.7 to 0.22
 baselines were good !
 pointing for the non-
 F, non-clip interfaces

1 (bad) to 5 (good)
 Almost all the same, except
 non-standard interfaces
 Tells us assessors need
 training, interfaces need
 easier learnability

Summary Conclusions

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- This is the first large-scale, multi-participant evaluation of summarisation of video;
- Good agreement on the inclusion of GT in summaries, the most detailed component of evaluation and 4% target could have been even smaller;
- 2007 TRECVID summarisation tied to nature of data - TV series rushes, and techniques not generalisable to other kinds of rushes or non-rushes;
- 2007 concentrated on “did the summary contain all the clear and important material in the original” and less so on issues of redundancy in summary and learnability of summary formats;

Consensus ?

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- Evaluation & benchmarking are important, not threatening, supportive and help to advance the state of the art;
- Multiple domains, not all IR;
- All very metrics-based with agreed evaluation procedures and data formats;
- All have manual self-annotation of ground truth vs. assessment of pooled results;
- All coordinate large volunteer efforts with little sponsorship funding;
- All have growing participation;
- All make their results public and data available to participants, for free;
- All have contributed to raising the profile of evaluation campaigns;
- These evaluation campaigns now exist in many other domains;

TRECvid: **Pros** and Cons (1)

Many good things & some bad things about benchmarking evaluation campaigns

1. secure, prepare, and distribute data, difficult to get ... using the same data, agreed metrics and ground truth allows direct comparisons across and within groups;
2. create critical mass and motivate donations of resources to the campaign from among the participants;

TRECVID: Pros and Cons (2)

C E N T R E F O R D I G I T A L V I D E O P R O C E S S I N G

3. following the known and published guidelines for evaluation, within or outside a formal evaluation campaign, allows direct comparisons with the work of others knowing the methodology is sound and accepted;
4. good performance showcases for funding agencies, industry, promotes the research area;
5. facilitate research groups wanting to move into a new area of research, lowers barriers to entry;
6. groups can easily learn from each other and starting groups can reach better performance, faster;

TRECVID: Pros and Cons (2)

1. everybody addresses the same research challenges using the same measures and so there is no room for diversity, and no scope for novelty or creativity
 - Look at the TRECVID variety, even easier where there is so much sharing
2. outputs easily available but original (video) data can have strings attached because of ©
3. there is a belief that agencies funding these have a stranglehold on the research agenda
 - Participants decide the tasks & metrics
4. dataset defines and restricts problems to be evaluated
 - Story bound detection & over-use of keyframes as shot reps.
5. the set of problems we could address in future work is constrained by the dataset

TRECVID Impact ?

C E N T R E F O R D I G I T A L V I D E O P R O C E S S I N G

- Standardised evaluations and comparisons do improve the underlying science;
- Able to weed out hypotheses from small, idiosyncratic data-dependent phenomena;
- Test on a common, large collection & common metadata;
- Failures are not embarrassing ...
- Unfortunately, virtually all work is done on one extracted KF per shot ... we're fixing this;
- Open participation;
- "TRECVID has been priceless for video analysis and retrieval research"

Conclusions

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

- evaluation campaigns require many choices among competing alternatives
- system oriented evaluations are a fruitful way to concentrate research efforts of a global community
 - Number of published papers and funding agencies supporting TRECVID work
- system evaluations are not user evaluations but they are the level that is achievable when working with c.100 research groups
- previous attempts at user evaluations and cross-site collaborations have not been successful

Conclusion

C E N T E R F O R D I G I T A L V I D E O P R O C E S S I N G

“Many forms of Government have been tried, and will be tried in this world of sin and woe. No one pretends that democracy is perfect or all-wise. Indeed, it has been said that democracy is the worst form of government except all those other forms that have been tried from time to time.”

- Sir Winston Churchill November 11, 1947

“No one pretends that test collections are perfect or all-wise. Indeed, it has been said that test collections are terrible for IR research except that they’re better than current alternatives.”

- Ellen Voorhees, October 2006