Università degli Studi di Padova

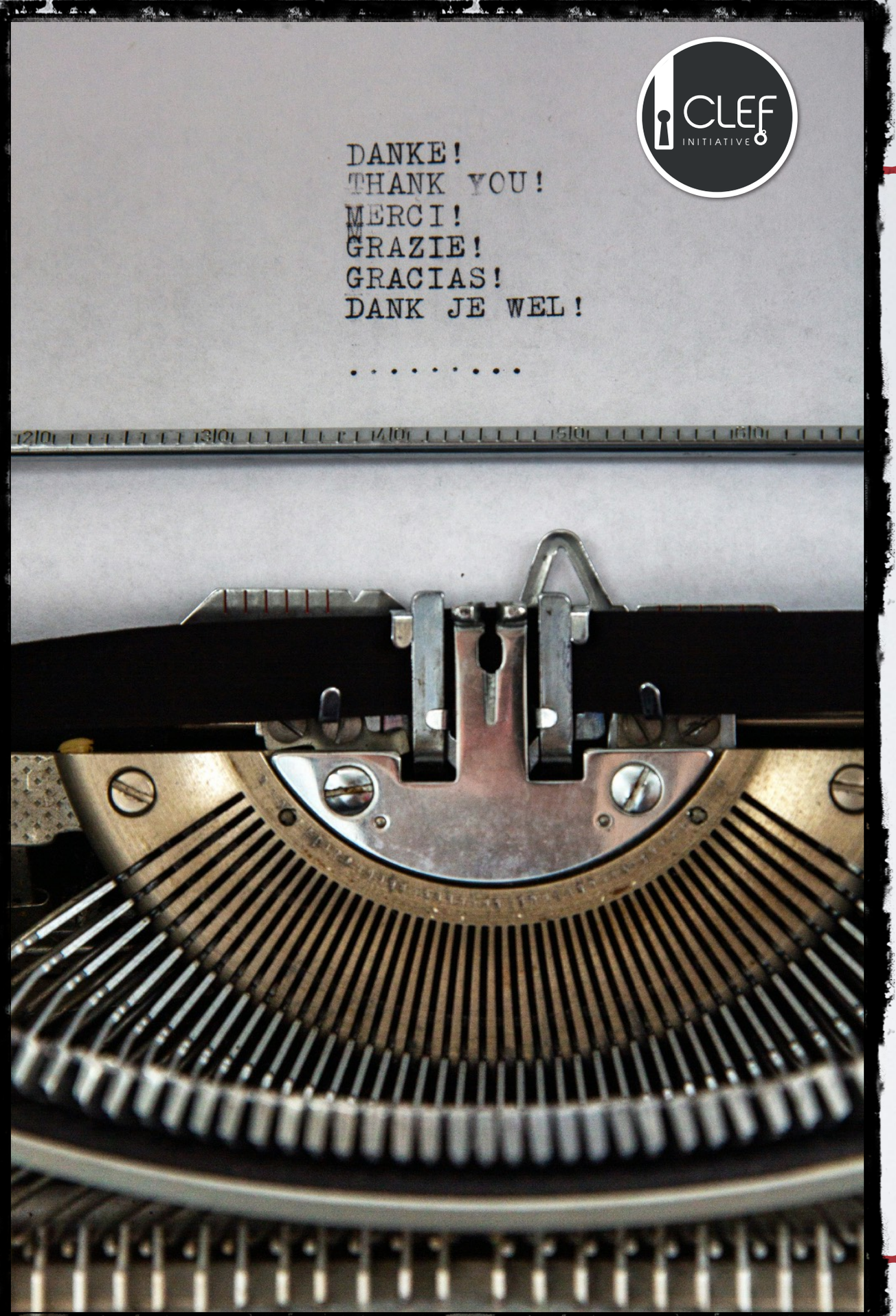
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Growing up in Cranfield, Maturing in Generative IR

Nicola Ferro

Intelligent Interactive Information Access (IIIA) Hub
Department of Information Engineering
University of Padua

Tony Kent Strix Memorial Award Lecture
9th January 2025, Online

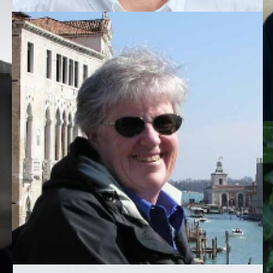

DANKE!
THANK YOU!
MERCI!
GRAZIE!
GRACIAS!
DANK JE WEL!
.....








DANKE!
THANK YOU!
MERCI!
GRAZIE!
GRACIAS!
DANK JE WEL!
.....




DANKE!
THANK YOU!
MERCI!
GRAZIE!
GRACIAS!
DANK JE WEL!
.....


DANKE!
THANK YOU!
MERCI!
GRAZIE!
GRACIAS!
DANK JE WEL!
.....
CLEF
INITIATIVE


Cranfield

The Beginning

- **Cranfield Paradigm** by Cyril W. Cleverdon

- Defines the use of **experimental collections**
  - **documents** (corpora)
  - **topics**, which are a surrogate for information needs
  - **relevance judgments** (binary or graded) also called relevance assessment or ground-truth (or qrels)

- Ensures **comparability** and **repeatability** of the experiments

Cyril W. Cleverdon

Cleverdon, C. W. (1962). *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK.
Cleverdon, C. W. (1997). *The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P., editors, Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

The Beginning

- **Cranfield Paradigm** by Cyril W. Cleverdon

- Defines the use of **experimental collections**
  - **documents** (corpora)
  - **topics**, which are a surrogate for information needs
  - **relevance judgments** (binary or graded) also called relevance assessment or ground-truth (or qrels)

- Ensures **comparability** and **repeatability** of the experiments

Cyril W. Cleverdon

Cleverdon, C. W. (1962). *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK.
Cleverdon, C. W. (1997). *The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P., editors, Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

The “Ideal Test Collection”

Corpora → (not historical) corpora are typically OK

- < 500 documents, no real value
- 1-2,000 documents, minimally acceptable
- > 10,000 documents, actually needed

Topics → typical size is still 50 topics

- < 75 topics, no real value
- 250 topics, minimally acceptable
- > 1,000 topics, actually needed

Relevance Judgements → binary is still most common option, diversity only recently

- multi-graded (highly and fairly relevant)
- types (novel, stimulating, …)
- need for **pooling** (still open research issue)

Karen Spärck Jones

C. J. “Keith” van Rijsbergen

Spärck Jones, K. and van Rijsbergen, C. J. (1975). *Report on the need for and provision of an ‘ideal’ information retrieval test collection*. British Library Research and Development Report 5266, University Computer Laboratory, Cambridge.

The “Ideal Test Collection”

Corpora → (not historical) corpora are typically OK

- < 500 documents, no real value
- 1-2,000 documents, minimally acceptable
- > 10,000 documents, actually needed

Topics → typical size is still 50 topics

- < 75 topics, no real value
- 250 topics, minimally acceptable
- > 1,000 topics, actually needed

Relevance Judgements → binary is still most common option, diversity only recently

- multi-graded (highly and fairly relevant)
- types (novel, stimulating, …)
- need for **pooling** (still open research issue)

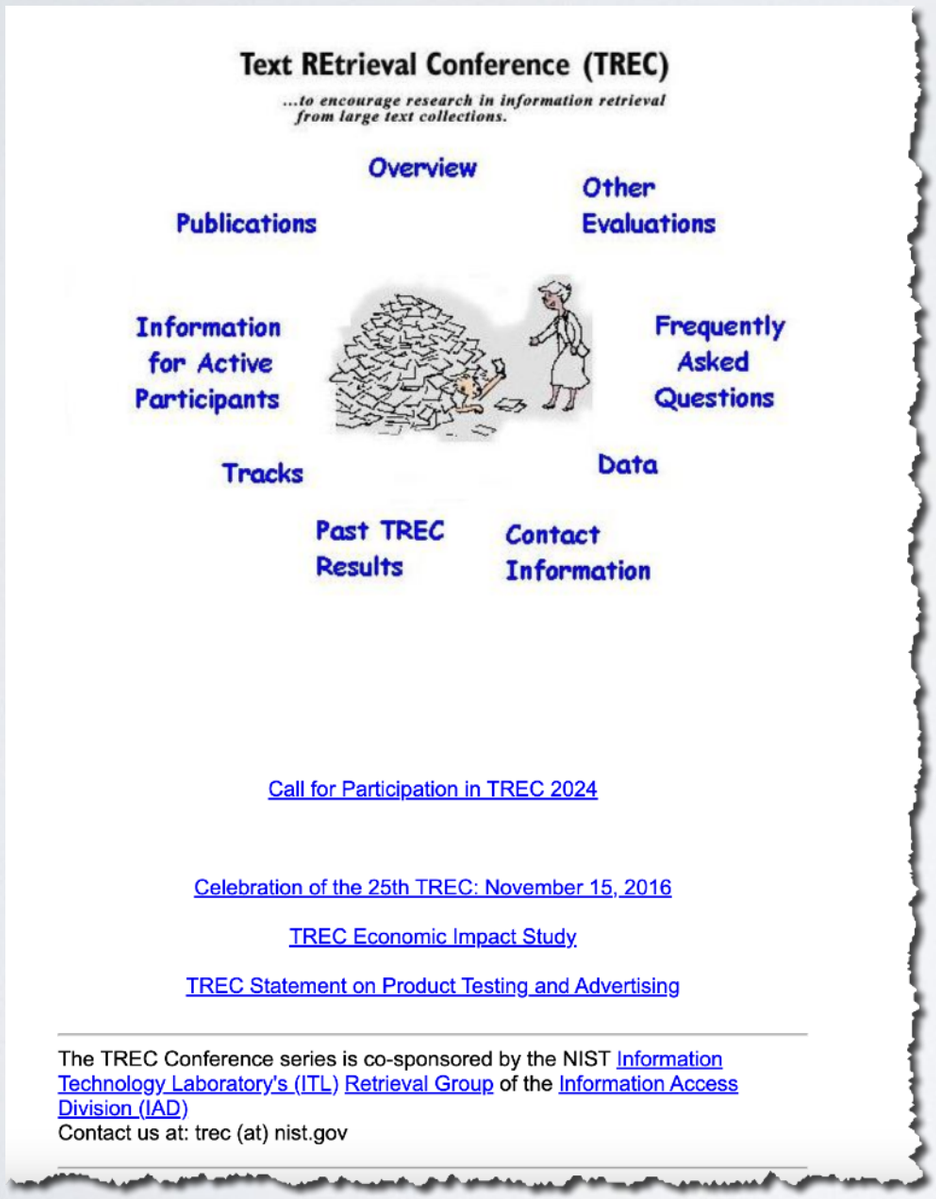
Karen Spärck Jones

C. J. “Keith” van Rijsbergen

Spärck Jones, K. and van Rijsbergen, C. J. (1975). *Report on the need for and provision of an ‘ideal’ information retrieval test collection*. British Library Research and Development Report 5266, University Computer Laboratory, Cambridge.

Large-scale Evaluation Initiatives: TREC

- **TREC** (Text REtrieval Conference), USA, since 1992
  - https://trec.nist.gov/



Ian Soboroff

Donna Harman

Ellen M. Voorhees

Harman, D. K. and Voorhees, E. M., editors (2005). *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, USA.

Large-scale Evaluation Initiatives: NTCIR

- **NTCIR** (NII Testbeds and Community for Information access Research), Japan, since 1999
  - http://research.nii.ac.jp/ntcir/index-en.html

Noriko Kando

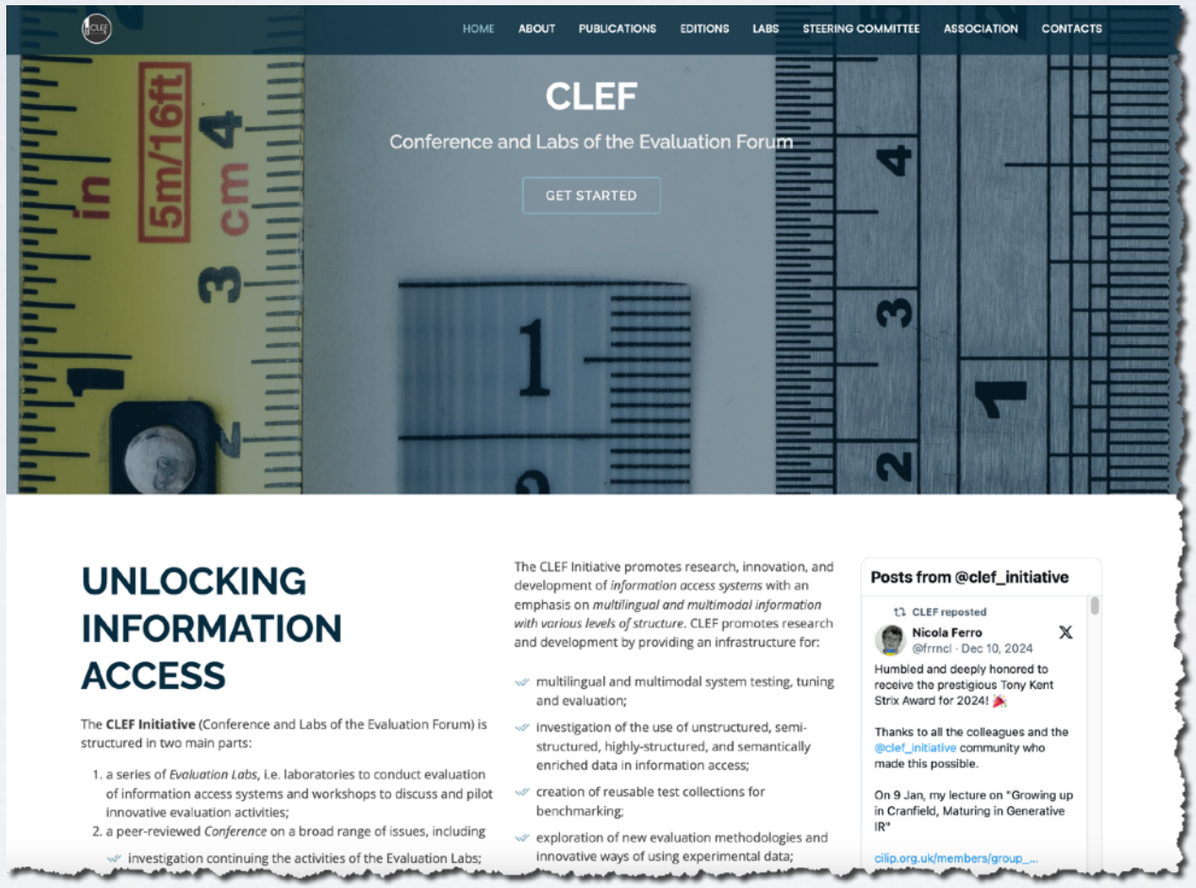
Sakai, T., Oard, D. W., and Kando, N., editors (2021). *Evaluating Information Retrieval and Access Tasks – NTCIR's Legacy of Research Impact*, volume 43 of *The Information Retrieval Series*. Springer International Publishing, Germany.

Large-scale Evaluation Initiatives: CLEF

- **CLEF** (Conference and Labs of the Evaluation Forum), Europe, since 2000
  - https://www.clef-initiative.eu/

Alberto Barrón-Cedeño

Alba García Seco de Herrera



Carol Ann Peters

Nicola Ferro

Ferro, N. and Peters, C., editors (2019). *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*. Springer International Publishing, Germany.

Large-scale Evaluation Initiatives: FIRE

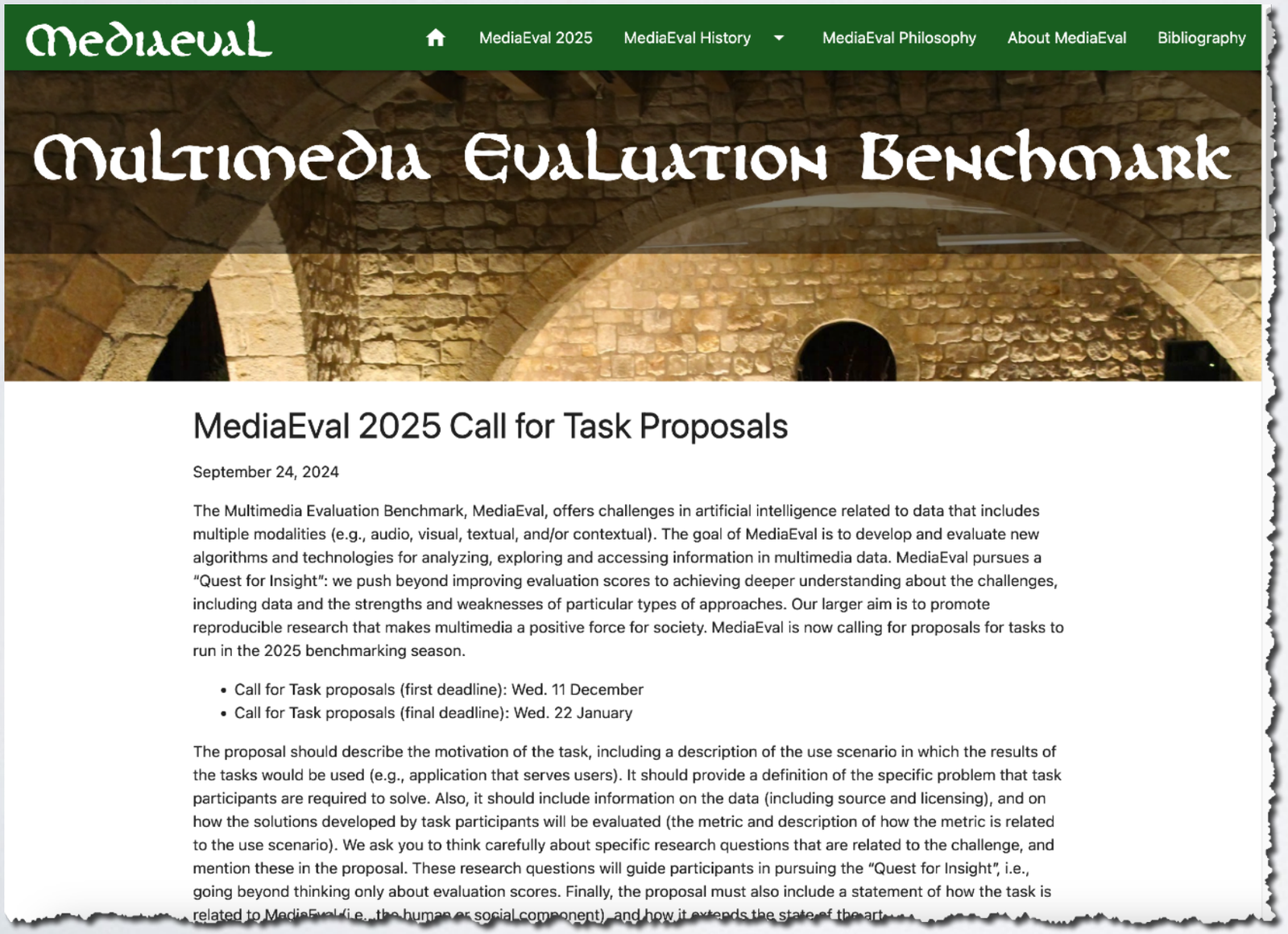
- **FIRE** (Forum for Information Retrieval Evaluation), India, since 2008
  - https://fire.irsi.org.in/

Mandar Mitra

Prasenjit Majumder

Large-scale Evaluation Initiatives: MediaEval

- **MediaEval** (Benchmarking Initiative for Multimedia Evaluation), Europe, since 2010
  - https://multimediaeval.github.io/



Gareth Jones

Martha Larson

Achievements: A Science of Ms

How Valuable is Evaluation?

- The **TREC 2010 Economic Impact** study estimated in about **30 M$** the overall **investment in TREC** by NIST
  - probably much much more if we had a means to estimate also the investment by participants in TREC

- They are the **pillars** for all the subsequent **scientific research** and **technology development**
  - TREC estimated the **return on investment** in the range of **3-5** for each invested dollar

Rowe, B. R., Wood, D. W., Link, A. L., and Simoni, D. A. (2010). *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program*. RTI Project Number 0211875, RTI International, USA. http: //trec.nist.gov/pubs/2010.economic.impact.pdf.

Conference and Labs of the Evaluation Forum

TRADE

MARK

Courtesy of Paolo Rosso at CLEF 2013

CLIR: The Grand Challenge

AAAI 1997 Spring Symposium:
Fully multilingual and multimodal information retrieval systems

Doug Oard

- capable of processing a query in any medium and any language
- finding relevant information from a multilingual multimedia collection containing documents in any language and form
- and presenting it in the style most likely to be useful to the user

David Hull

Oard, D. W. and Hull, D. A. (1997). AAAI 1997 Spring Symposium Series Reports. Cross-Language Text and Speech Retrieval. *AI Magazine*, 18(3):82.

CLIR: The Grand Challenge

AAAI 1997 Spring Symposium:

Fully multilingual and multimodal information retrieval systems

Doug Oard

- capable of processing a query in any medium and any language
- finding relevant information from a multilingual multimedia collection containing documents in any language and form
- and presenting it in the style most likely to be useful to the user

David Hull

Oard, D. W. and Hull, D. A. (1997). AAAI 1997 Spring Symposium Series Reports. Cross-Language Text and Speech Retrieval. *AI Magazine*, 18(3):82.

Classic CLEF: Years+ of History

- 1997 – First CLIR system evaluation campaigns in US and Japan: TREC and NTCIR
  - CLEF actually began life in 1997 as a track for Cross Language Information Retrieval (CLIR) within TREC. Mainly, English centered tasks (EN -> X, X -> EN).
- 2000-2009 – CLIR evaluation in Europe: CLEF (Cross-Language Evaluation Forum)
  - Fully multilingual, multimodal information retrieval systems capable of processing a query in any medium and any language finding relevant information from a multilingual multimedia collection containing documents in any language and form, and presenting it in the style most likely to be useful to the user

Classic CLEF: Years+ of History

- 1997 – First CLIR system evaluation campaigns in US and Japan: TREC and NTCIR
  - CLEF actually began life in 1997 as a track for Cross Language Information Retrieval (CLIR) within TREC. Mainly, English centered tasks (EN -> X, X -> EN).
- 2000-2009 – CLIR evaluation in Europe: CLEF (Cross-Language Evaluation Forum)
  - Fully multilingual, multimodal information retrieval systems capable of processing a query in any medium and any language finding relevant information from a multilingual multimedia collection containing documents in any language and form, and presenting it in the style most likely to be useful to the user

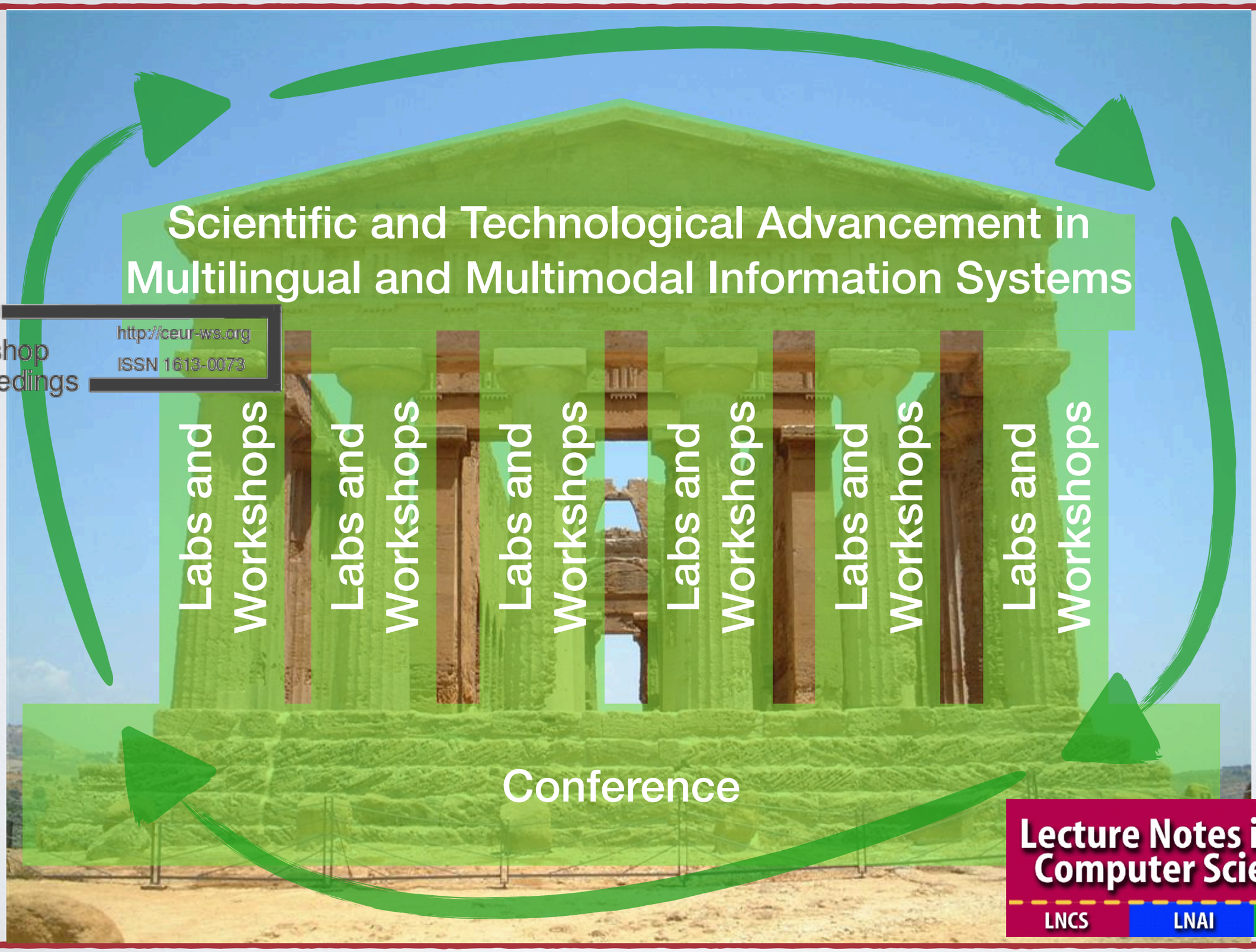
Carol: The True Name Behind CLEF

Where To Go Next? The CLEF Initiative

The direction depended on the Fellowship

The **CLEF Initiative** is a self-organized body whose main mission is to promote research, innovation, and development of information access systems with an emphasis on **multilingual** and **multimodal information** with **various levels of structure**.

Approach (since 2010)



CEUR Workshop Proceedings
http://ceur-ws.org
ISSN 1613-0073

Lecture Notes in Computer Science
LNCS LNAI LNBI

CLEF and ECIR (since 2019)

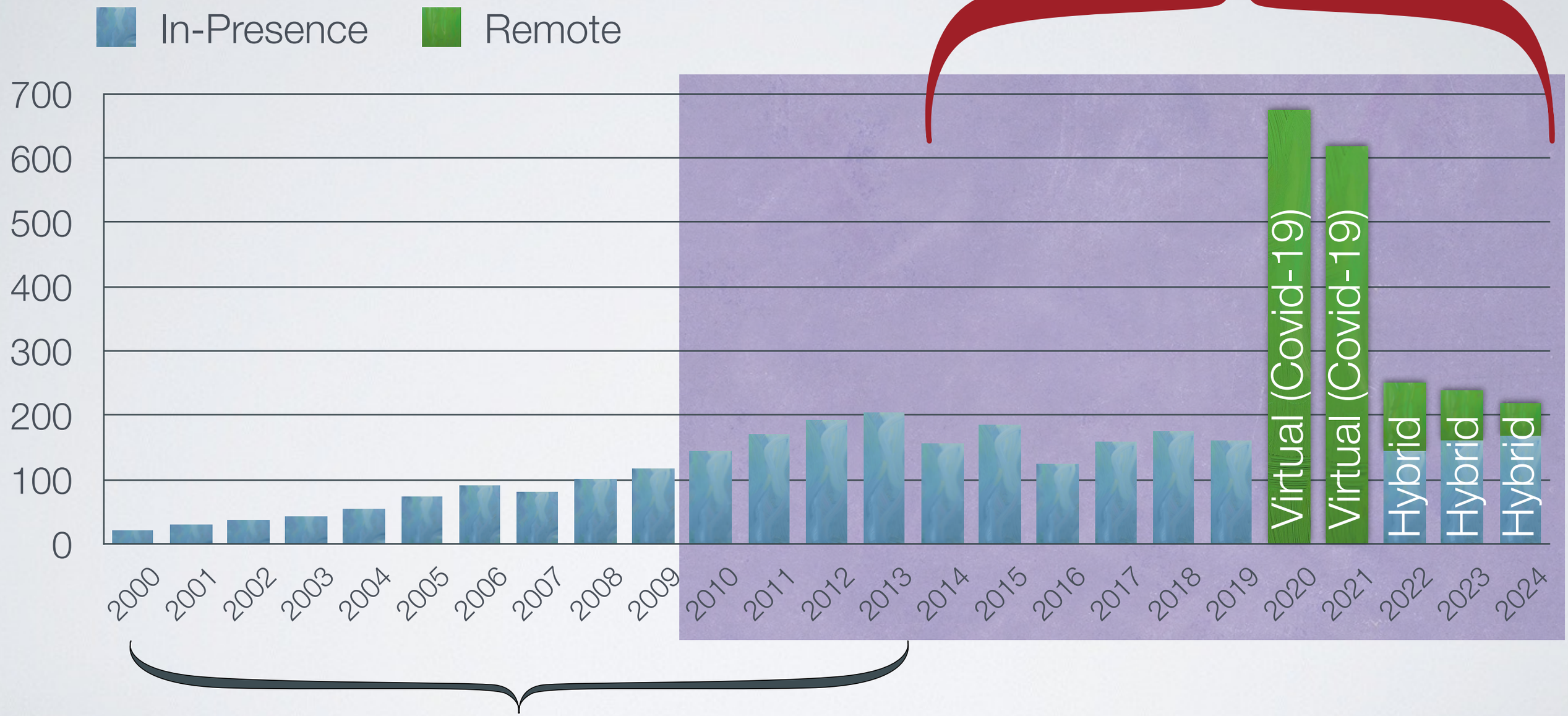
CLEF Resultless Review Process (since 2023)

Dagstuhl Seminar 23031, January 2023
Frontiers of Information Access Experimentation for Research and Education

Bauer, C., Carterette, B. A., Ferro, N., Fuhr, N., and Faggioli, G., editors (2023). *Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education*, Dagstuhl Reports, Volume 13, Number 1. Schloss Dagstuhl–Leibniz-Zentrum f'ur Informatik, Germany.

Participation: Attendees

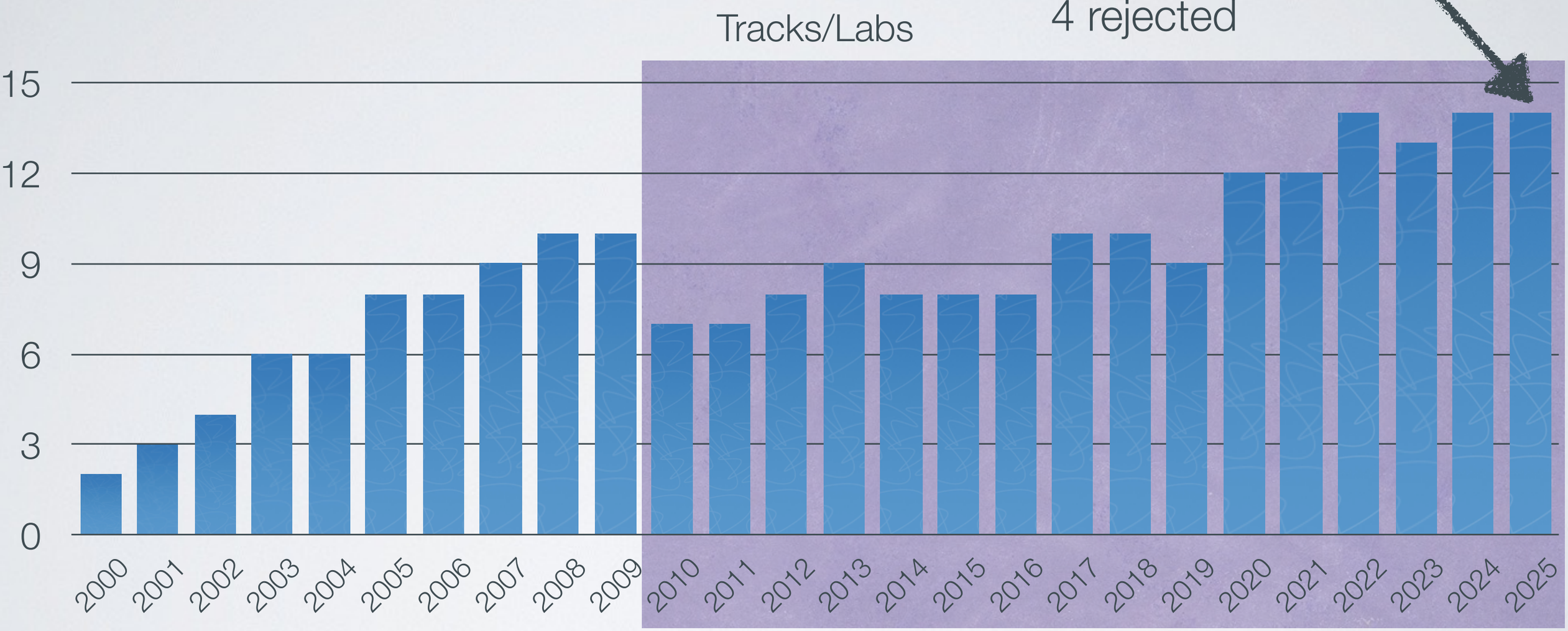
100% Voluntary Effort Based



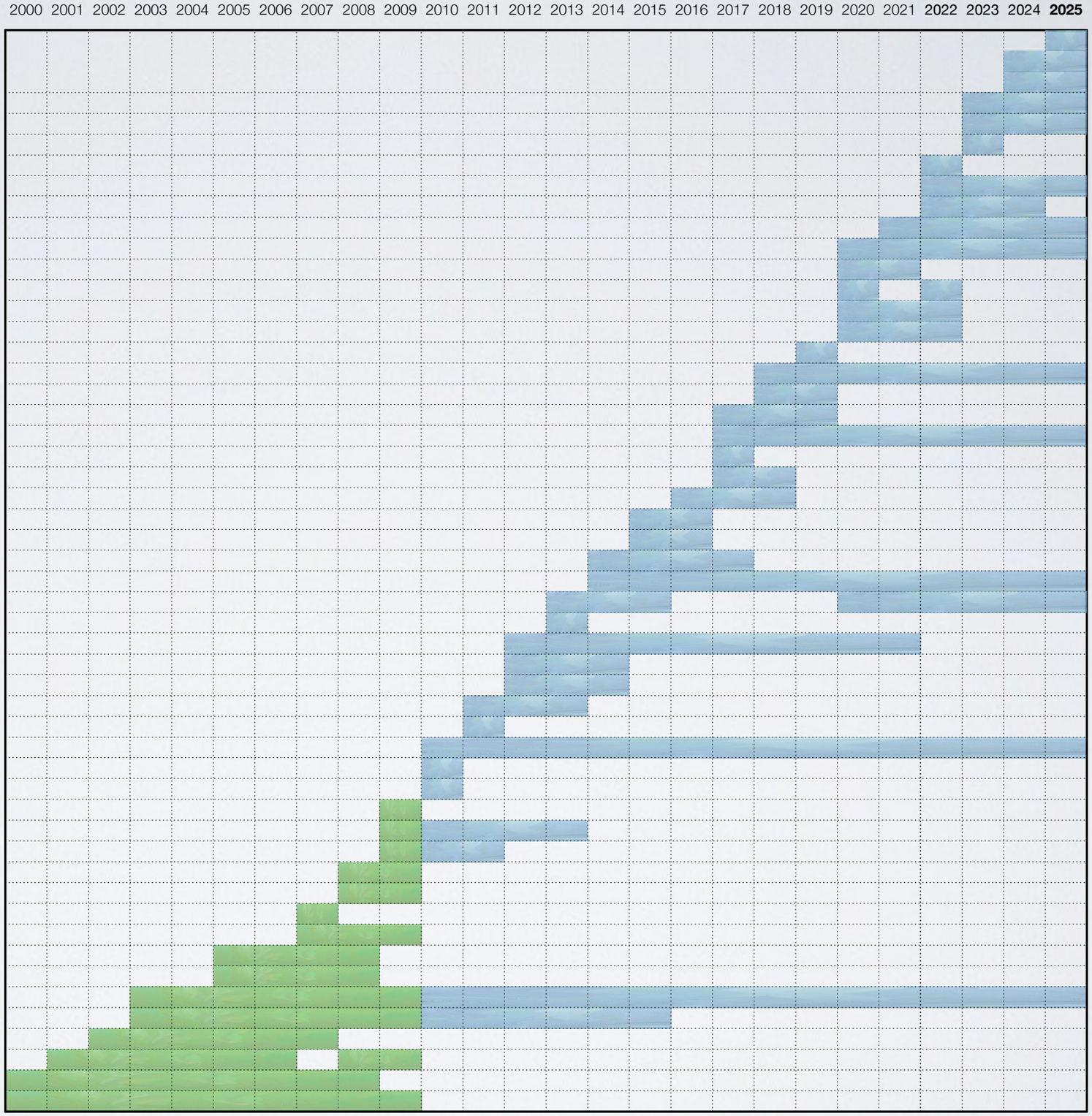
Mainly voluntary effort + project funding

Labs

20 proposals
14 accepted
2 asked to merge
4 rejected



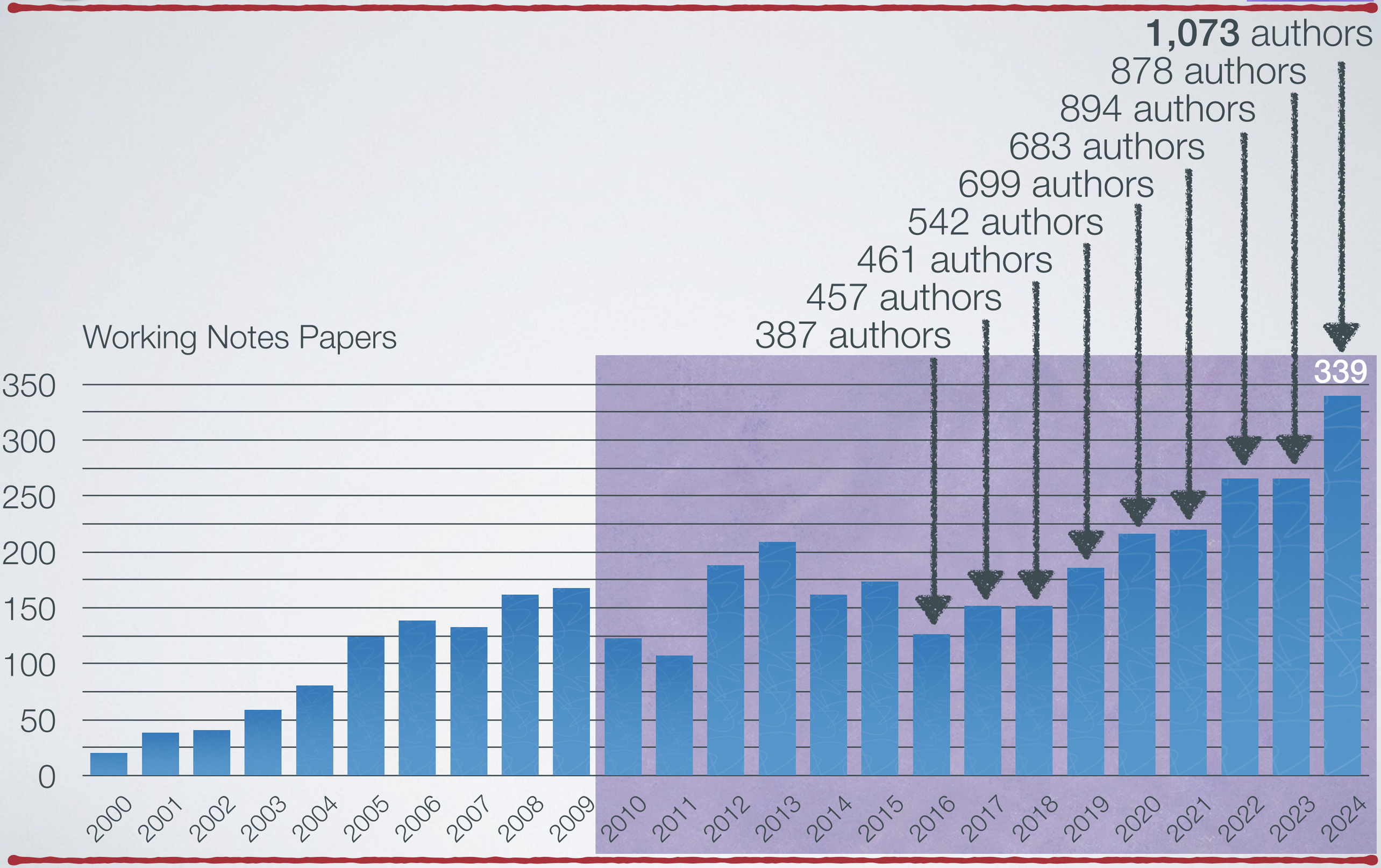
CLEF Labs over Time



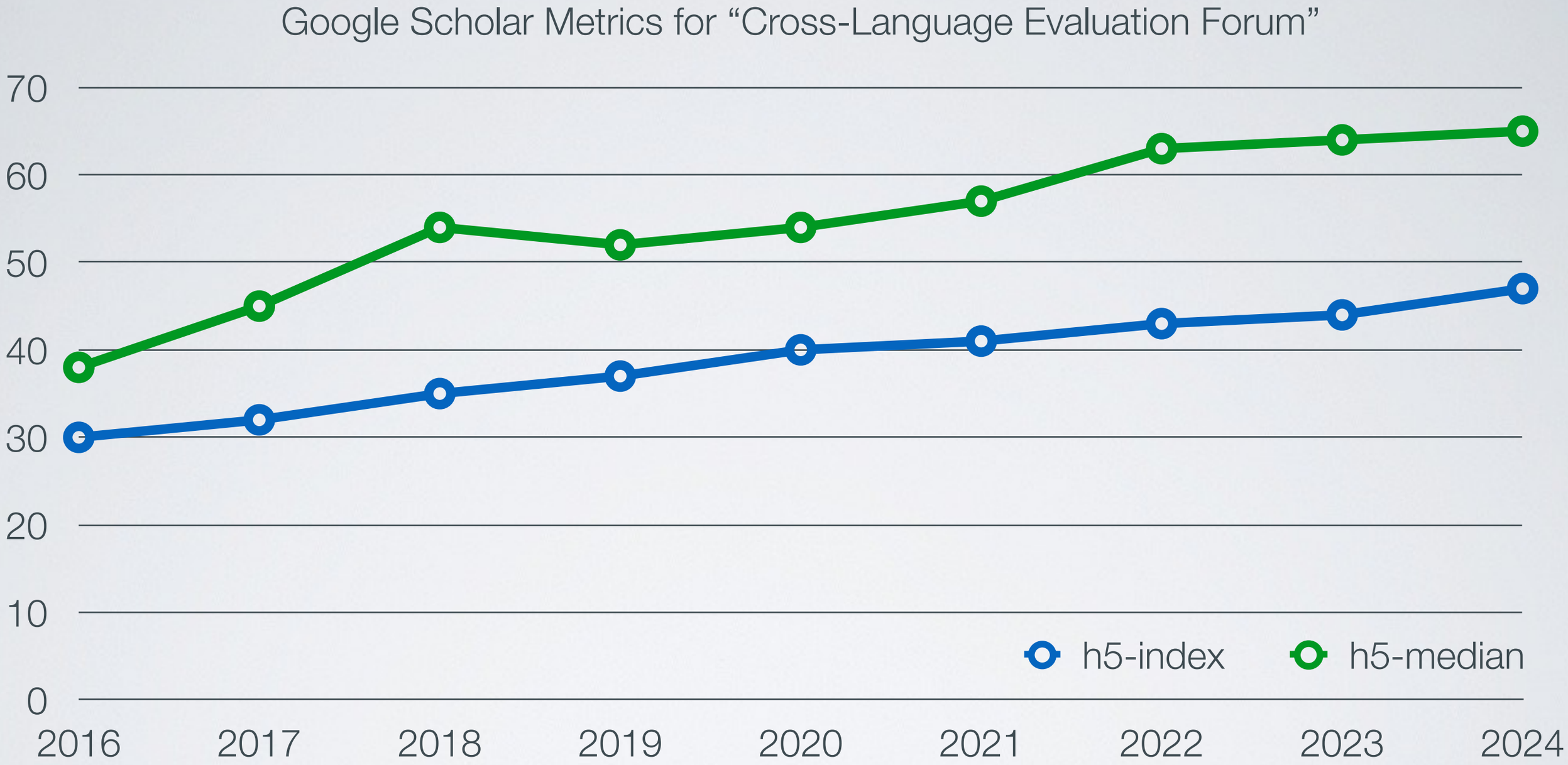
Growing up in Cranfield, Maturing in Generative IR
Tony Kent Strix Memorial Award Lecture, 9th January 2025, Online

Nicola Ferro

Working Notes



Publication “Universe”



- Google Scholar for “*CLEF evaluation*”
  - **96,300 hits**

Publication “Universe” (2024)

Categories > Engineering & Computer Science > **Databases & Information Systems**

	Publication	h5-index	h5-median
1.	International World Wide Web Conferences (WWW)	112	169
2.	IEEE Transactions on Knowledge and Data Engineering	107	185
3.	ACM SIGIR Conference on Research and Development in Information Retrieval	103	149
4.	Information Processing & Management	96	157
5.	ACM International Conference on Information and Knowledge Management	91	133
6.	Journal of Big Data	79	191
7.	International Conference on Very Large Databases	79	122
8.	ACM International Conference on Web Search and Data Mining	77	130
9.	ACM SIGMOD International Conference on Management of Data	73	109
10.	International Conference on Data Engineering	69	94
11.	International Conference on Web and Social Media (ICWSM)	56	88
12.	IEEE International Conference on Big Data	54	79
13.	ACM Conference on Recommender Systems	53	81
14.	Information Systems	49	77
15.	World Wide Web	49	71
16.	ACM Transactions on Information Systems (TOIS)	48	98
17.	Knowledge and Information Systems	48	85
18.	ACM Transactions on Intelligent Systems and Technology (TIST)	47	85
19.	Workshop of Cross-Language Evaluation Forum	47	65
20.	ACM Transactions on Internet Technology (TOIT)	44	69

"AIRS" OR "WWW" OR "information retrieval" OR "Information and Knowledg

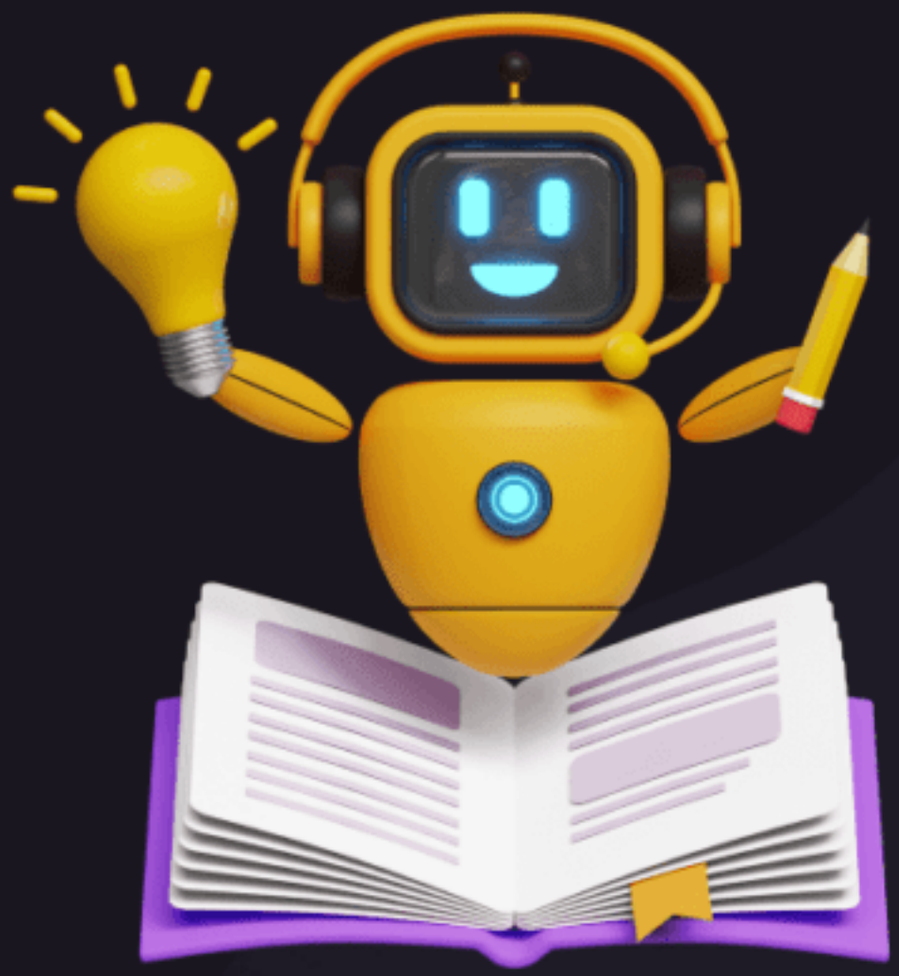
Top 20 publications matching *"AIRS" OR "WWW" OR "information retrieval...*

	Publication	h5-index	h5-median
1.	International World Wide Web Conferences (WWW)	112	169
2.	IEEE Transactions on Knowledge and Data Engineering	107	185
3.	ACM SIGIR Conference on Research and Development in Information Retrieval	103	149
4.	Information Processing & Management	96	157
5.	ACM International Conference on Information and Knowledge Management	91	133
6.	ACM International Conference on Web Search and Data Mining	77	130
7.	ACM Conference on Recommender Systems	53	81
8.	Journal of the Association for Information Science and Technology	49	69
9.	ACM Transactions on Information Systems (TOIS)	48	98
10.	Workshop of Cross-Language Evaluation Forum	47	65
11.	International Society for Music Information Retrieval Conference	43	69
12.	European Conference on Advances in Information Retrieval	42	60
13.	VINE Journal of Information and Knowledge Management Systems	38	61
14.	Forum for Information Retrieval Evaluation	30	39
15.	ACM SIGIR International Conference on Theory of Information Retrieval	24	42
16.	ACM Transactions on the Web (TWEB)	24	39
17.	International Journal of Multimedia Information Retrieval	24	36
18.	International ACM/IEEE Joint Conference on Digital Libraries	23	33
19.	Romanian Journal of Information Science and Technology	19	32
20.	International Journal on Digital Libraries	19	30


AI
Prompt...

R.I.P. Information Retrieval?

Who is Helping Who?



Retrieval Augmented Generation

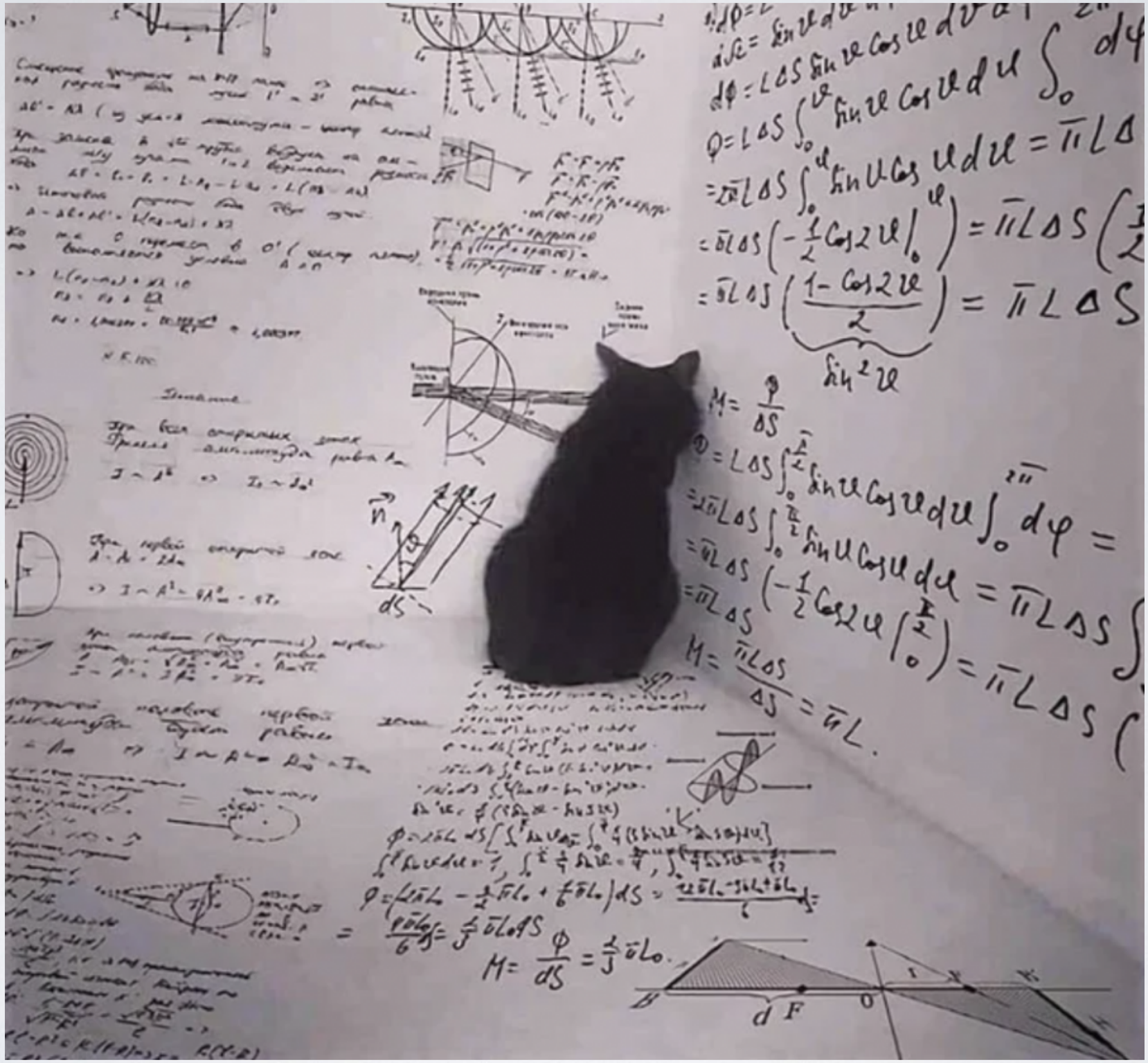
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Proc. 34th Annual Conference on Neural Information Processing Systems (NeurIPS 2020)*, pages 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020.

Do We Always Need Generation?

Generation

Retrieval

How to Adapt Cranfield for Evaluating RAG?

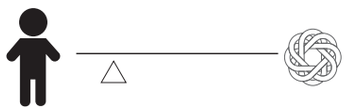
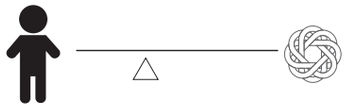
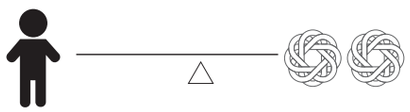
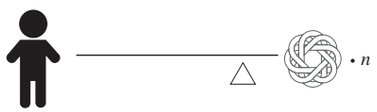


Why Not Just Use LLMs for Relevance Assessment?

- Judgment bias toward a particular LLM
- Bias toward user groups
- Resilience against misinformation
- LLM-based LLM training
- Judging vs. predicting
- Truthfulness and hallucinations

A Spectrum of Human–LLM/AI Collaboration

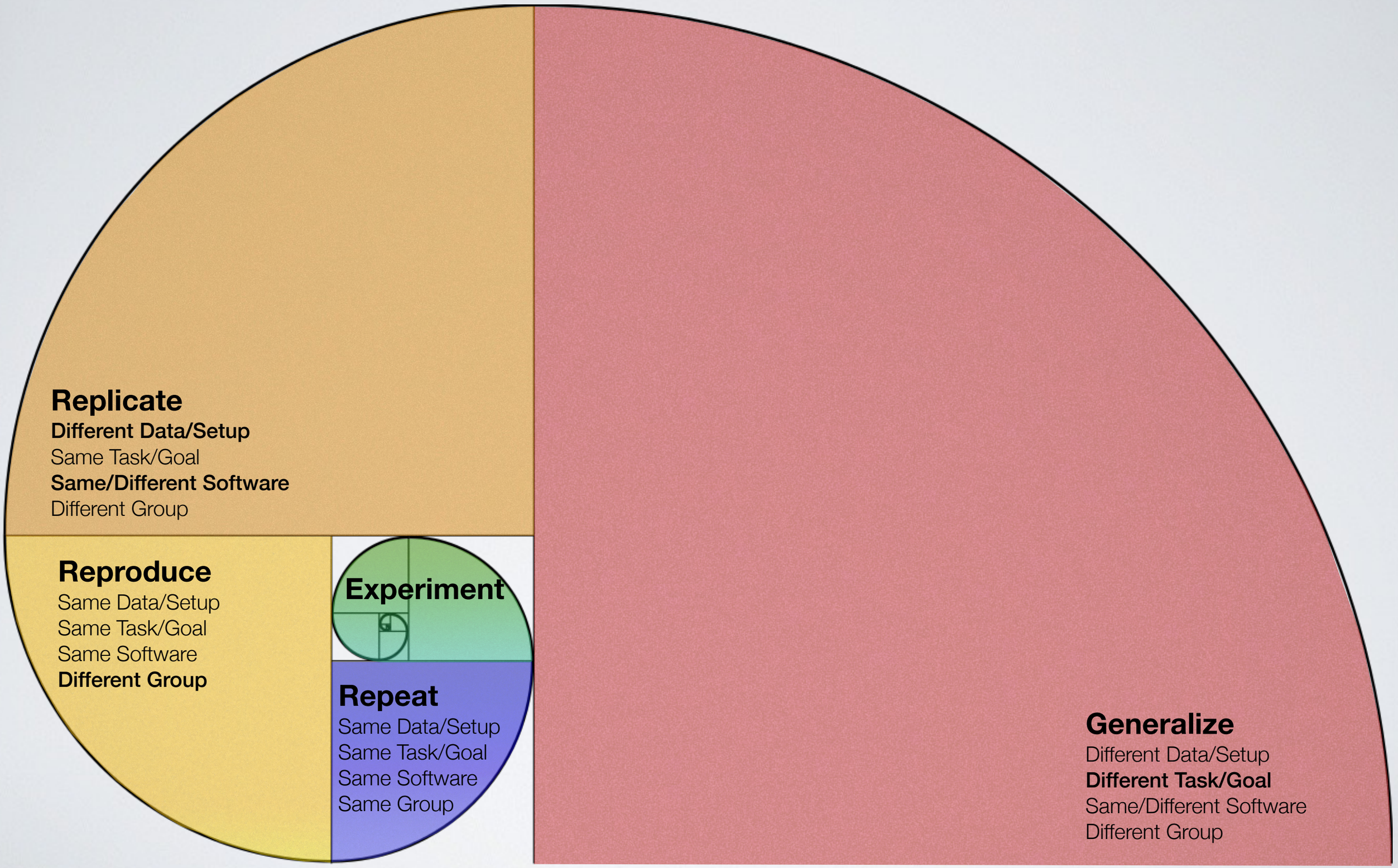
Collaboration perspective: Spectrum of possibilities for collaborative human–machine task organization to make (relevance) decisions. The △ symbol indicates where on the spectrum each possibility falls.

Collaboration Balance	Task Allocation
Human Judgment	
	Humans manually decide what is relevant without any kind of AI support.
	Humans have full control of deciding but are supported by machine-based text highlighting, data clustering, and so forth.
Model in the Loop	
	Humans decide based on LLM-generated summaries needed for the decision.
	Balanced competence partitioning. Humans and LLMs focus on decisions they are good at.
Human in the Loop	
	Two (or more) LLMs each generate a decision and a human selects the best one.
	An LLM makes a decision (and an explanation for it) that a human can accept/reject.
	LLMs are considered crowdworkers varied by specific characteristics, aggregated, and controlled by a human.
Fully Automated	
	Fully automatic decision without humans.

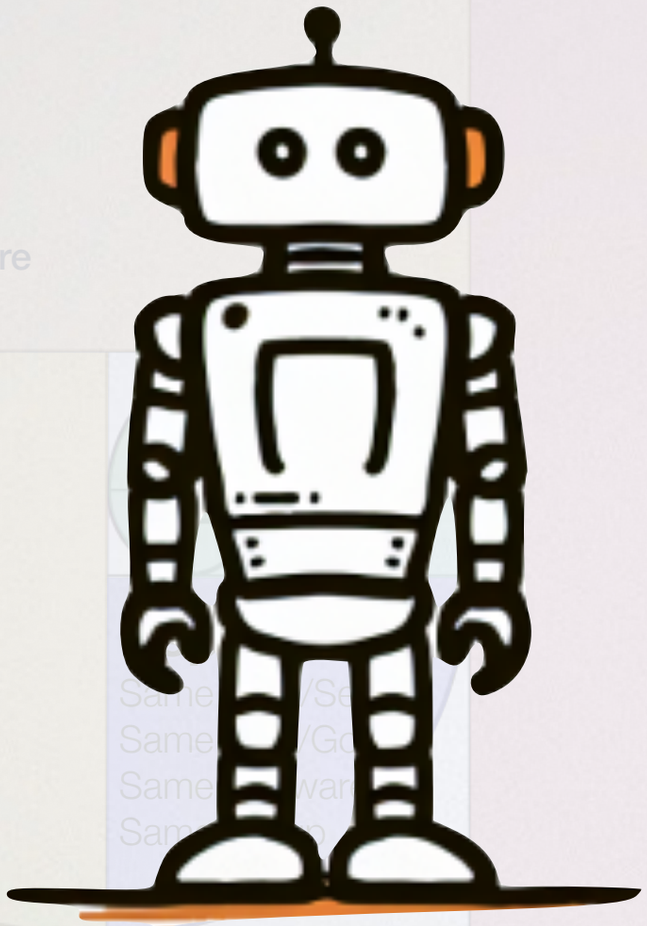
- **Human judgment**. On one extreme, humans make all relevance judgments manually without being influenced by an LLM
- **Model in the loop**. To make it easier for human assessors to decide on relevance in a consistent manner, an advanced level of automatic support could be provided, e.g. summarizing documents
- **Human in the loop**. Automated judgments could be produced by an LLM and then verified by humans
- **Fully automated**. If LLMs were able to reliably judge relevance, they could completely replace humans when judging relevance.

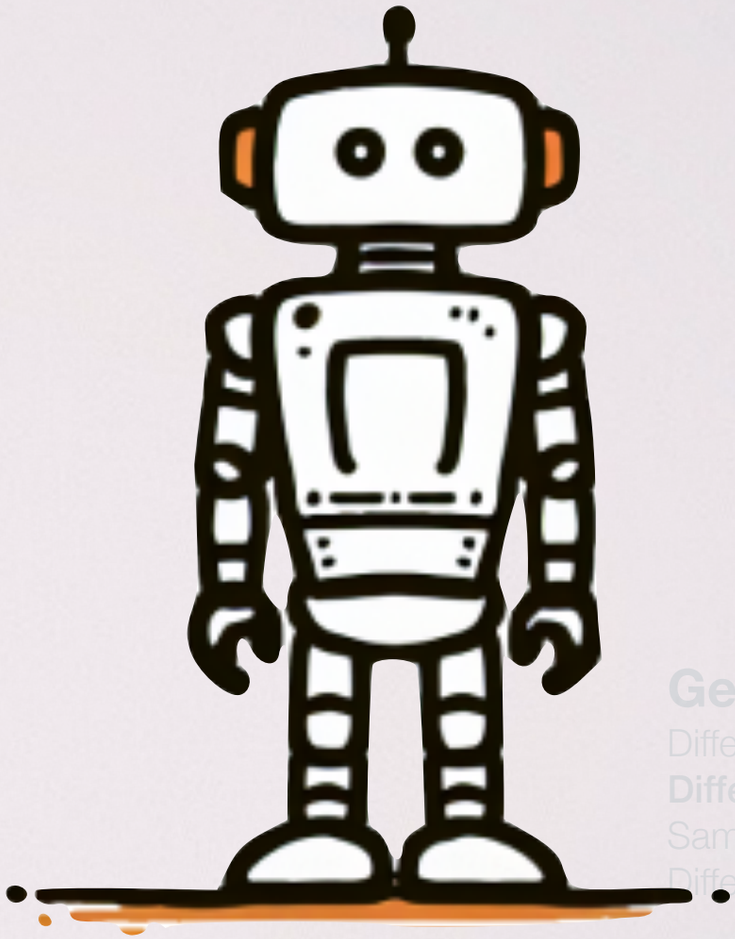
Faggioli, G., Dietz, L., Clarke, C. L. A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., and Wachsmuth, H. (2024). Who determines what is relevant? Humans or AI? Why not both! A Spectrum of Human-AI Collaboration in Assessing Relevance. *Communications of the ACM (CACM)*, 67(4):31–34.

Reproducibility



Reproducibility



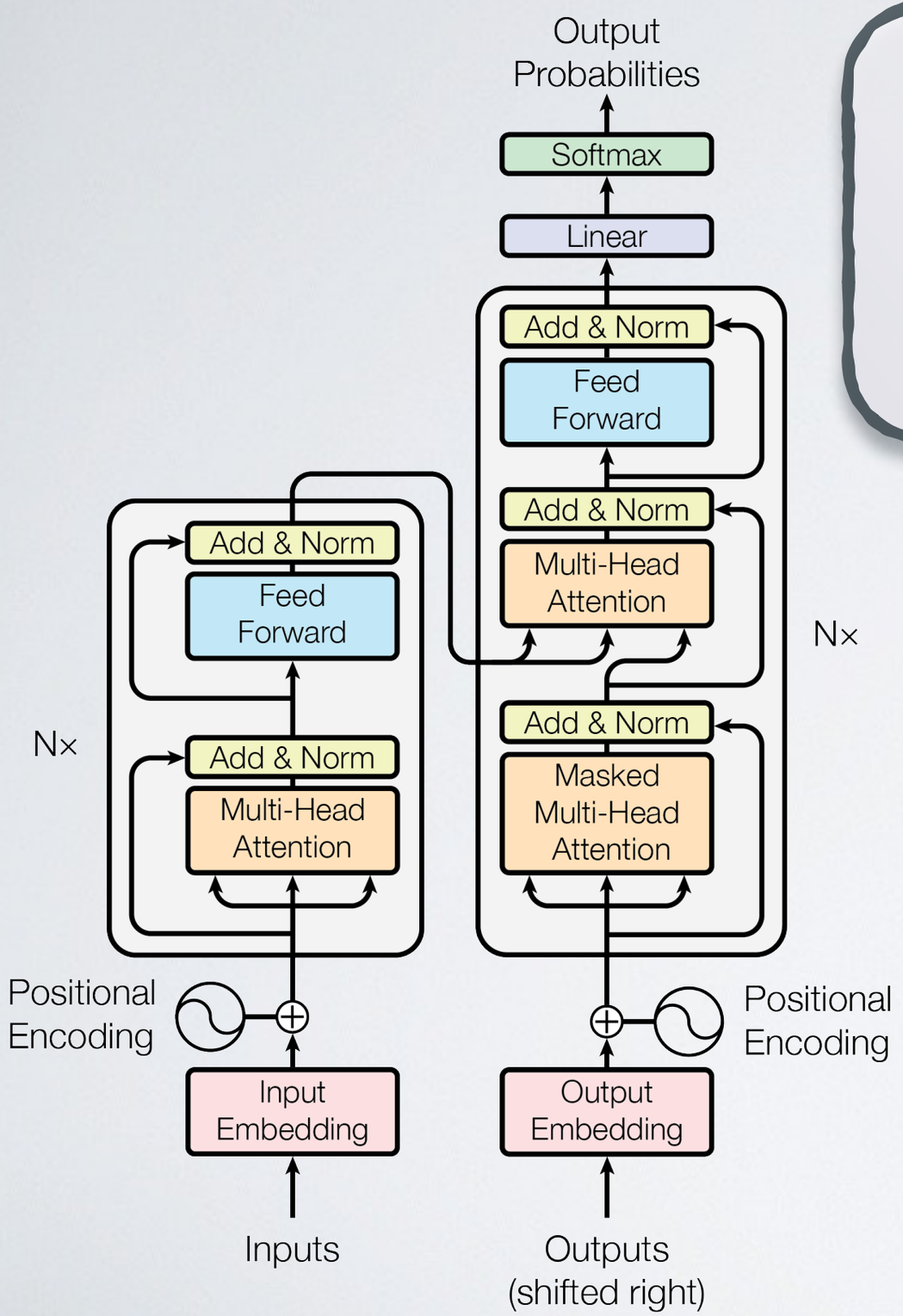


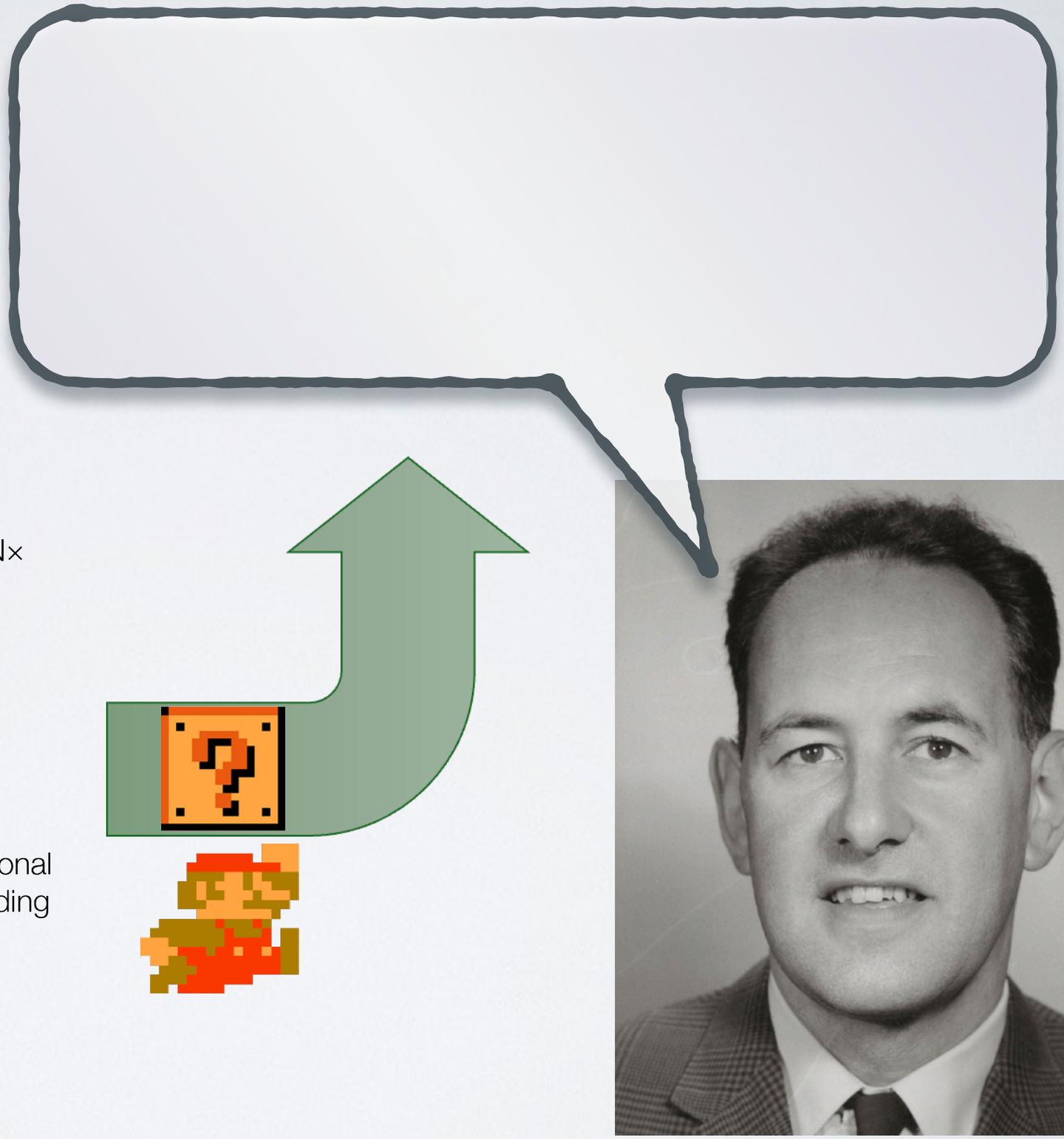
Replicate
Different Data/Setup
Same Task/Goal
Same/Different Software
Different Group

Reproduce
Same Data/Setup
Same Task/Goal
Same Software
Different Group

Generalize
Different Data/Setup
Different Task/Goal
Same/Different Software
Different Group

Back to the Roots: What is IR Today?





Back to the Roots: What is IR Today?

Back to the Roots: What is IR Today?

Evaluation Beyond the Surface

Thank You!