UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Foundations of IR Evaluation

Nicola Ferro
@frrncl

Department of Information Engineering
University of Padua, Italy

12th European Summer School in Information Retrieval (ESSIR 2019)
15-19 July 2019, Milan, Italy

Outline

- Introduction to IR Evaluation
- Evaluation Measures
- Reproducibility
- [Statistical Hypothesis Testing]

Evaluation Basics

Deeply Rooted…

Experimentation

Why Evaluation?

"To measure is to know"

"If you cannot measure it, you cannot improve it"

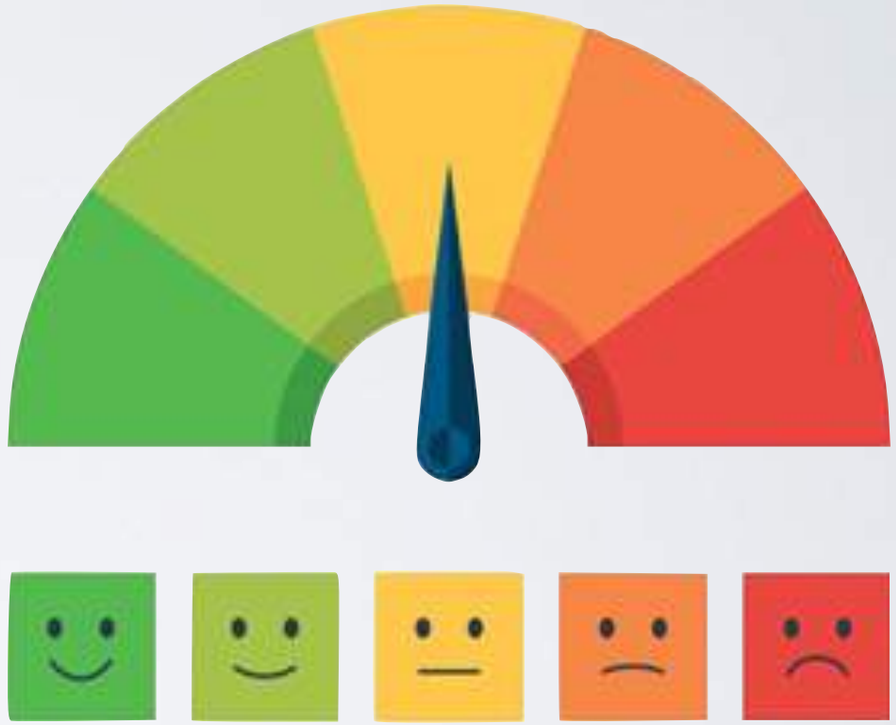
Lord William Thompson,
first Baron Kelvin (1824-1907)

What to Evaluate?

Efficiency

VS

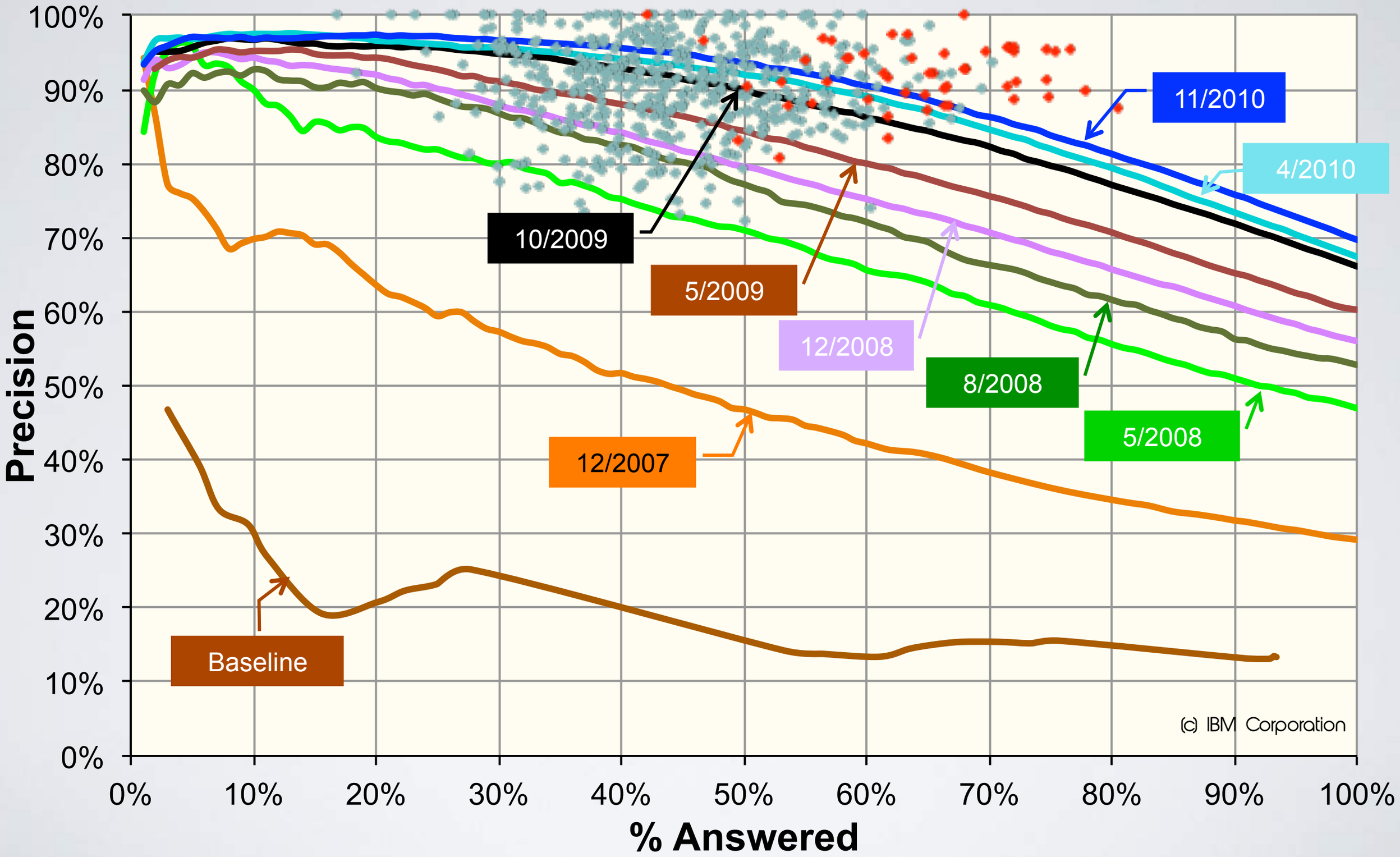
Effectiveness



Evaluation in Action

IBM Watson:
Deep QA Project

Evaluation in Action

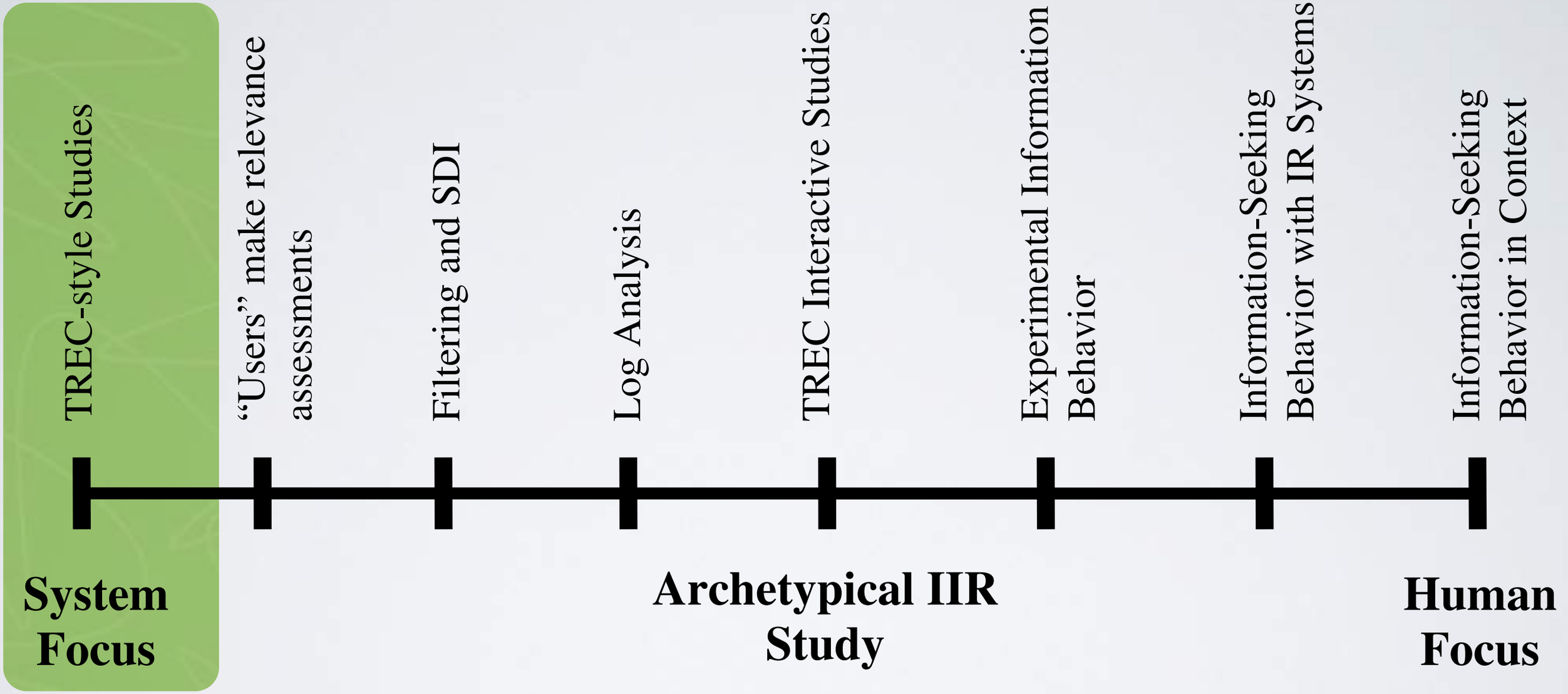


Critical Issues in Evaluation

- It must be scientifically **valid**
  - valid methodology, measures, and statistics
  - large-scale enough to be statistically valid
  - must be “repeatable” if possible
- It must be **realistic** for the applications that will be using the information retrieval systems
  - **task** and use cases
- It must be **understandable** to your audience/client

Harman, D. K. (2011). *Information Retrieval Evaluation*. Morgan & Claypool Publishers, USA.

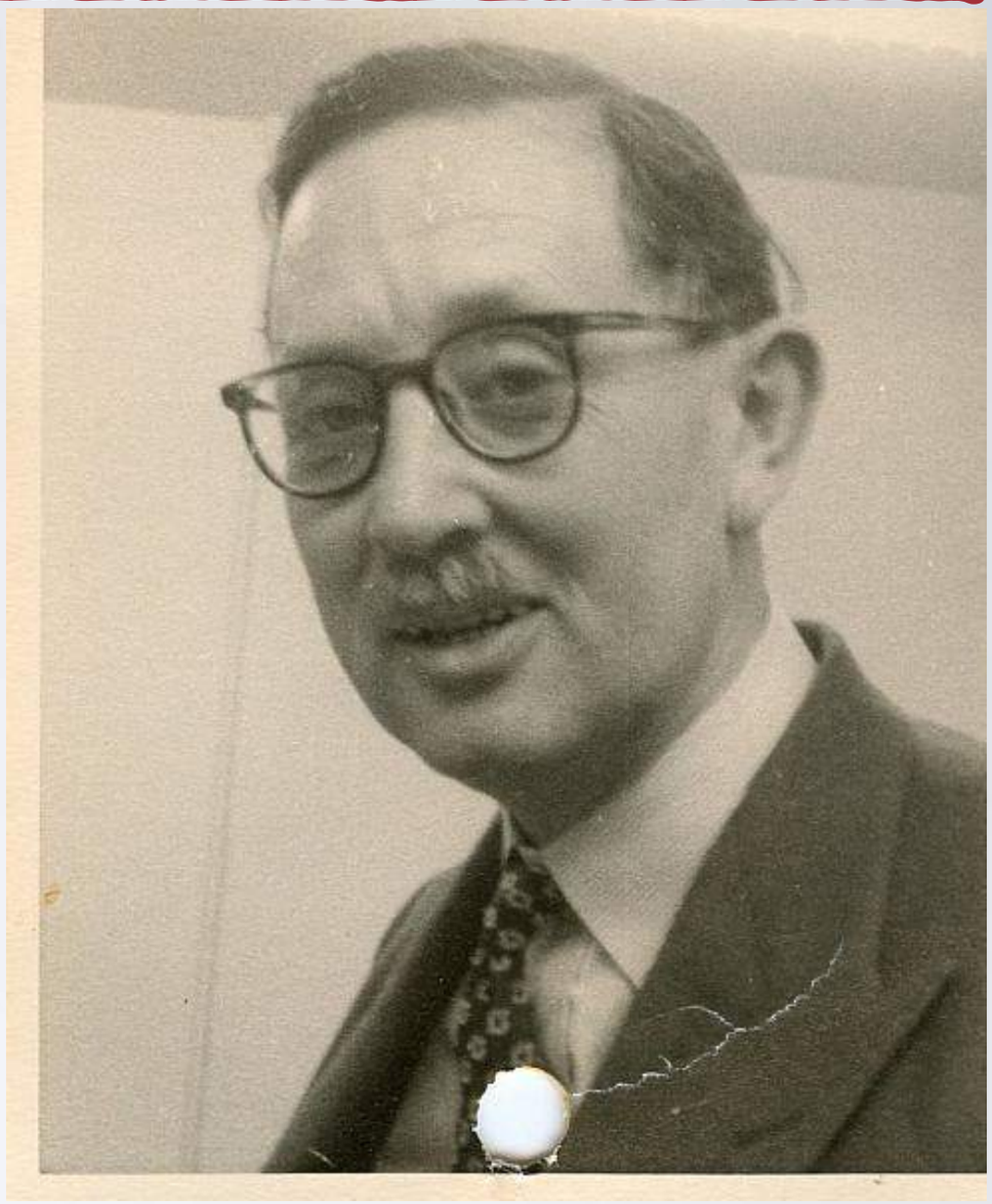
Evaluation Spectrum



Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval (FnTIR), 3(1-2), 1-224.

How Does Experimental Evaluation Work?

- **Cranfield Paradigm** by Cyril W. Cleverdon
  - Dates back to mid 1960s
- Makes use of **experimental collections**
  - documents (corpora)
  - topics
  - 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.

Some Document Collections

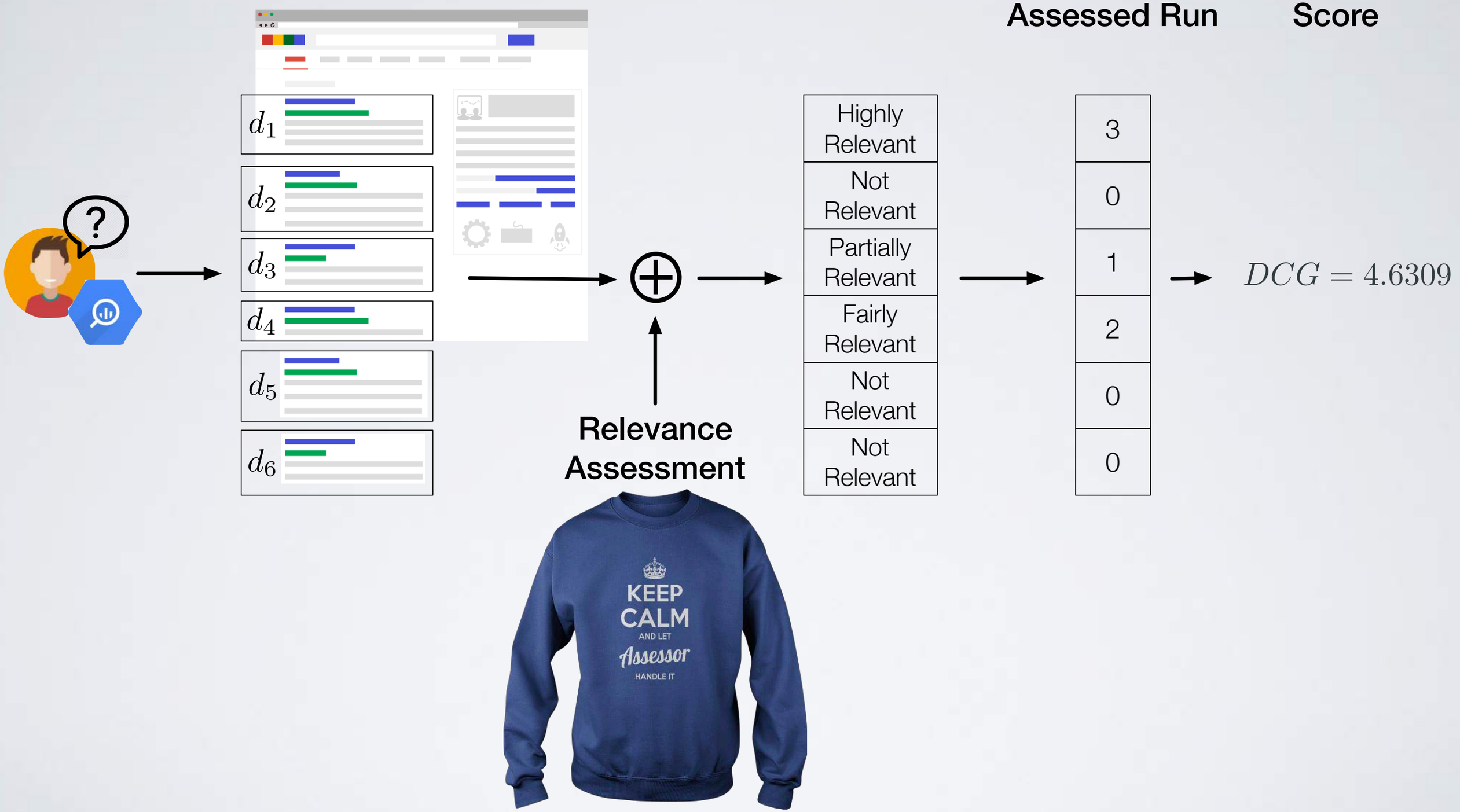
- Historical
  - **CACM**: 3,024 abstracts from the Communications of the ACM
    http://ir.dcs.gla.ac.uk/resources/test_collections/cacm/
- Mid-nineties
  - **TIPSTER**: 528,155 documents (news articles, US government reports, …), Disks 4 and 5 excluding Congressional Record subcollection
    https://catalog.ldc.upenn.edu/LDC93T3A
- Early 2000s
  - **WT10g**: 1,692,096 Web pages crawled in 2001
    http://ir.dcs.gla.ac.uk/test_collections/wt10g.html
  - **GOV2**: 25,205,179 Web pages crawled fro .gov sites in early 2004
    http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm
  - **CLEF Multilingual Corpus**: 4,883,227 multilingual news articles corpus in 13 languages (Bulgarian, Dutch, English, Farsi, Finnish, French, German, Hungarian, Italian, Portuguese, Spanish) gathered in 1994, 1995 and 2002. Topics in 28 different languages (Bengali, Bulgarian, Chinese, Czech, Dutch, English, Farsi, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Italian, Japanese, Marathi, Norwegian, Oromo Polish, Portuguese, Russian, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai)
- Today
  - **ClueWeb 2009**: 1,040,809,705 Web pages in 10 languages crawled between January and February 2009
    https://lemurproject.org/clueweb09/
  - **ClueWeb 2012**: 733,019,372 English Web pages crawled between February 10, 2012 and May 10, 2012
    https://lemurproject.org/clueweb12/
  - **TREC Washington Post Corpus**: 608,180 news articles and blog posts from January 2012 through August 2017 from Washington Post
    https://trec.nist.gov/data/wapost/

Example of Topics

```
<?xml version="1.0" encoding="UTF-8"?>
<topic>
    <identifier>101</identifier>

    <title lang="en">Renewable Energy Sources</title>
    <title lang="it">Energie Rinnovabili</title>
    <title lang="bg">Възобновяеми източници на енергия</title>
    <title lang="zh">再生能源</title>
    <title lang="hi">नवीकरणीय ऊर्जा स्रोत</title>
    <title lang="om">Maddawwan Humnoota Haaromfamanii</title>

    <description lang="en">
        Find documents reporting on the exploitation of renewable resources for
        power/energy production.
    </description>
    <description lang="it">
        Si trovino documenti riguardanti lo sfruttamento di fonti rinnovabili per la
        produzione di energia.
    </description>
    <description lang="bg">
        Намерете документи, които дават информация за възобновяеми източници за
        производство на енергия.
    </description>
    <description lang="zh">尋找與再生能源有關的文件</description>
    <description lang="hi">
        वद्युत/ऊर्जा उत्पादन के लिए नवीकरणीय स्रोतों के दोहन की जानकारी देने वाले दस्तावेज खोजिए।
    </description>
    <description lang="om">
        Dokumantoota waa'ee qabeenyoota haaromfamanii irraa humnoota oomishuudhan itti
        fayyadamuu gabaasan barbaadi.
    </description>

    <narrative lang="en">
        Relevant documents will report on the use of power obtained from renewable
        resources, such as biomass, water, solar, geothermal or wind power. Information on
        energy-saving vehicles is not relevant.
    </narrative>
    <narrative lang="it">
        I documenti rilevanti riporteranno l'uso di energia ottenuta da fonti rinnovabili,
        come bio-masse, acqua, sole, energia geotermica o eolica. Informazioni su veicoli
        a risparmio energetico non sono da considerare rilevanti.
    </narrative>
    ...
</topic>
```

- Topics consists of:
  - **title**: a brief statement expressing the information need. It resembles the typical search engine query
  - **description**: more detailed formulation of the information need
  - **narrative**: instructions for assessors on when to consider a document relevant
- Typical experimental collections make use of **50 topics**

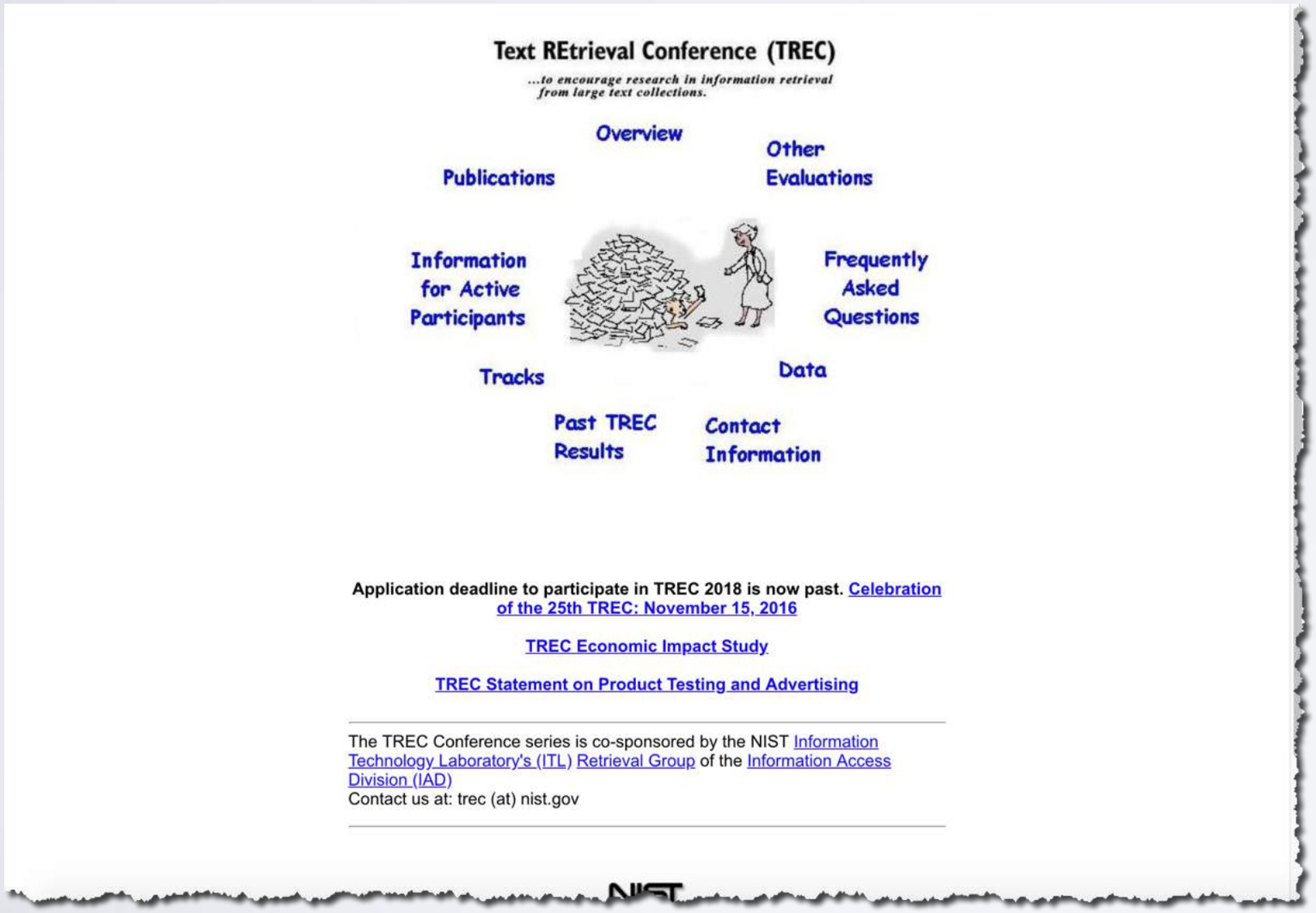
Evaluation with Test Collections in a Nutshell



Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR),* 4(4):247–375.

Large-scale Evaluation Initiatives: TREC

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



Donna Harman

Ellen M. Voorhees

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

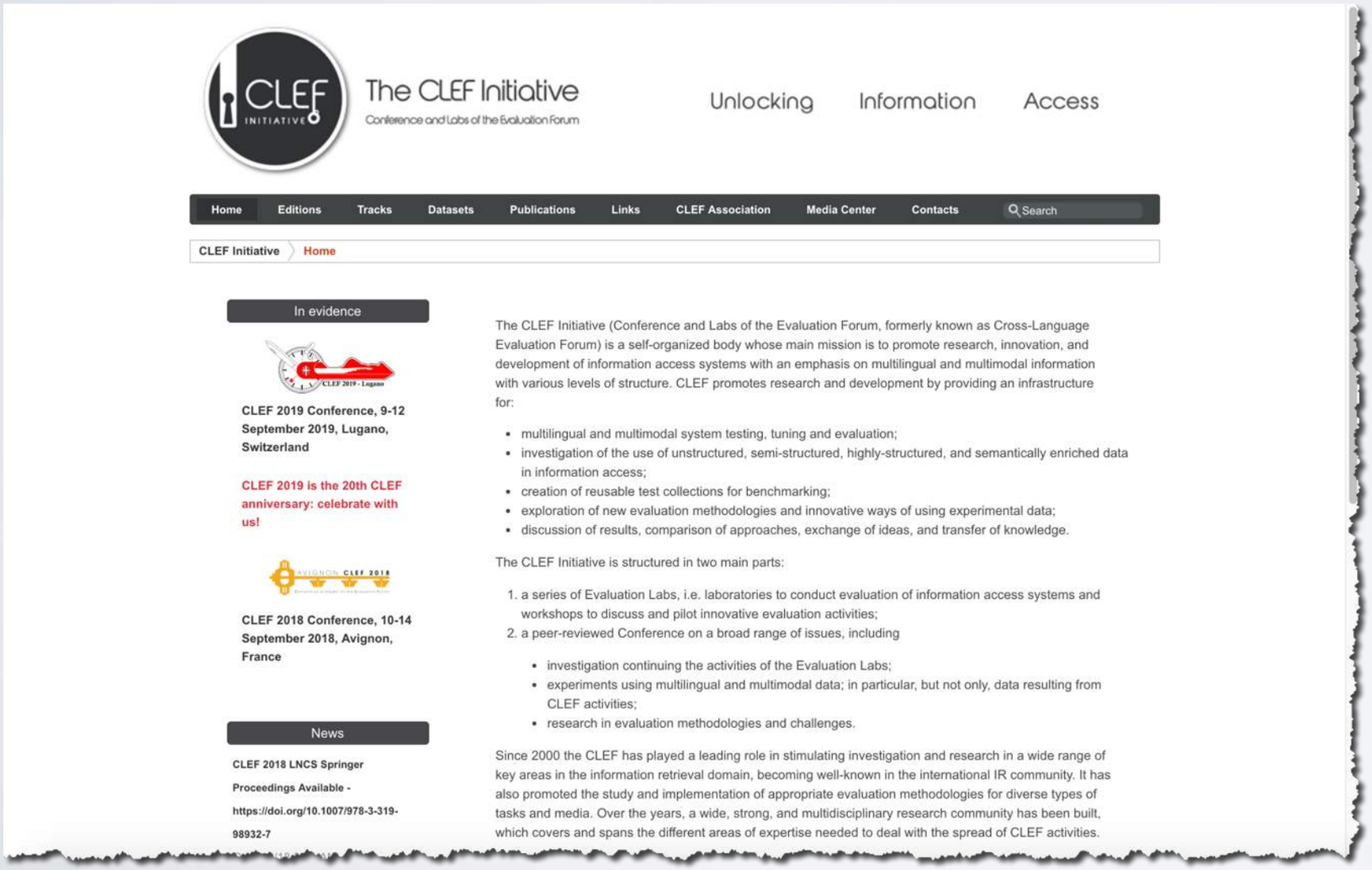
Large-scale Evaluation Initiatives: NTCIR

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

Noriko Kando

Large-scale Evaluation Initiatives: CLEF

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



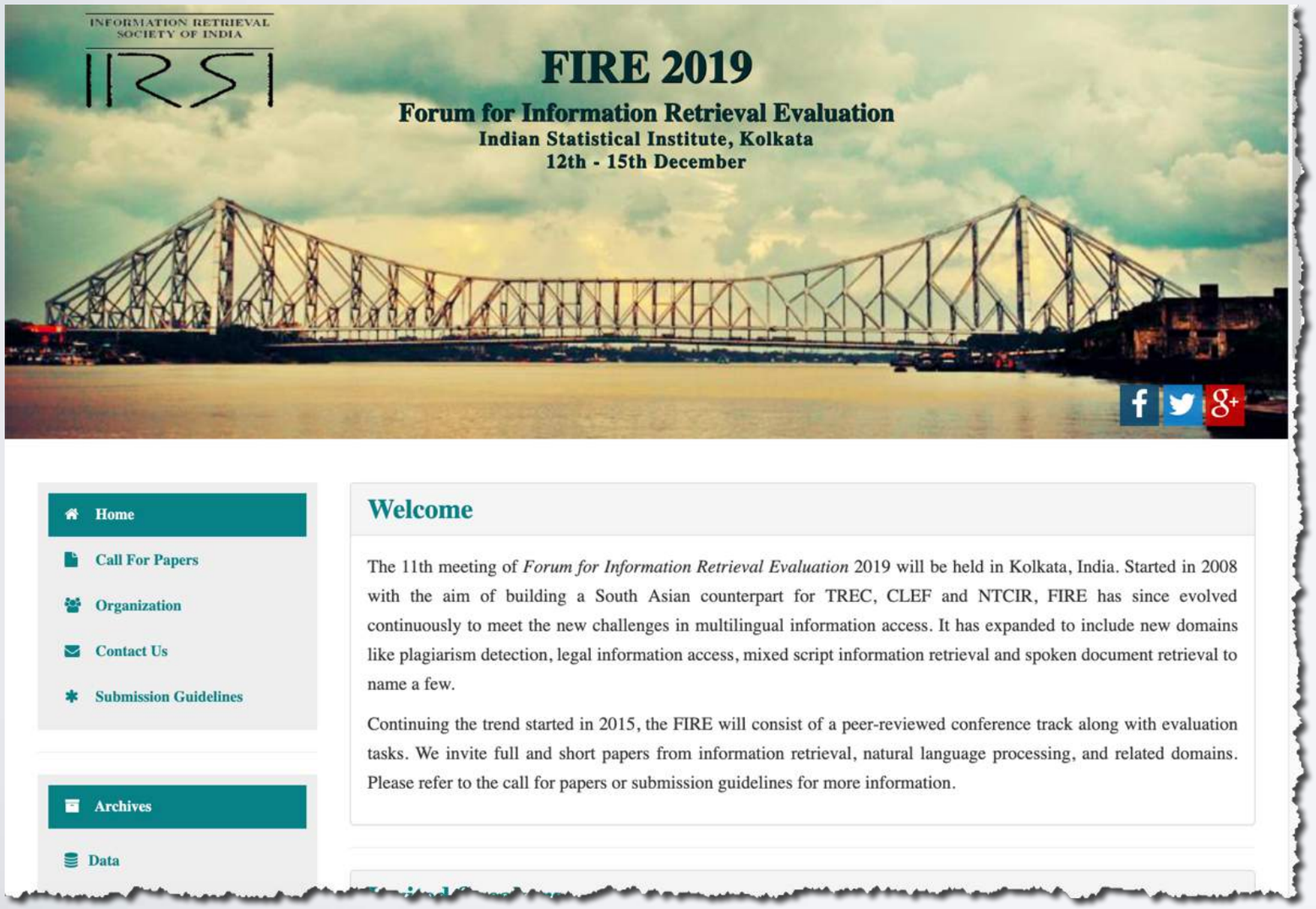
Carol Ann Peters

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

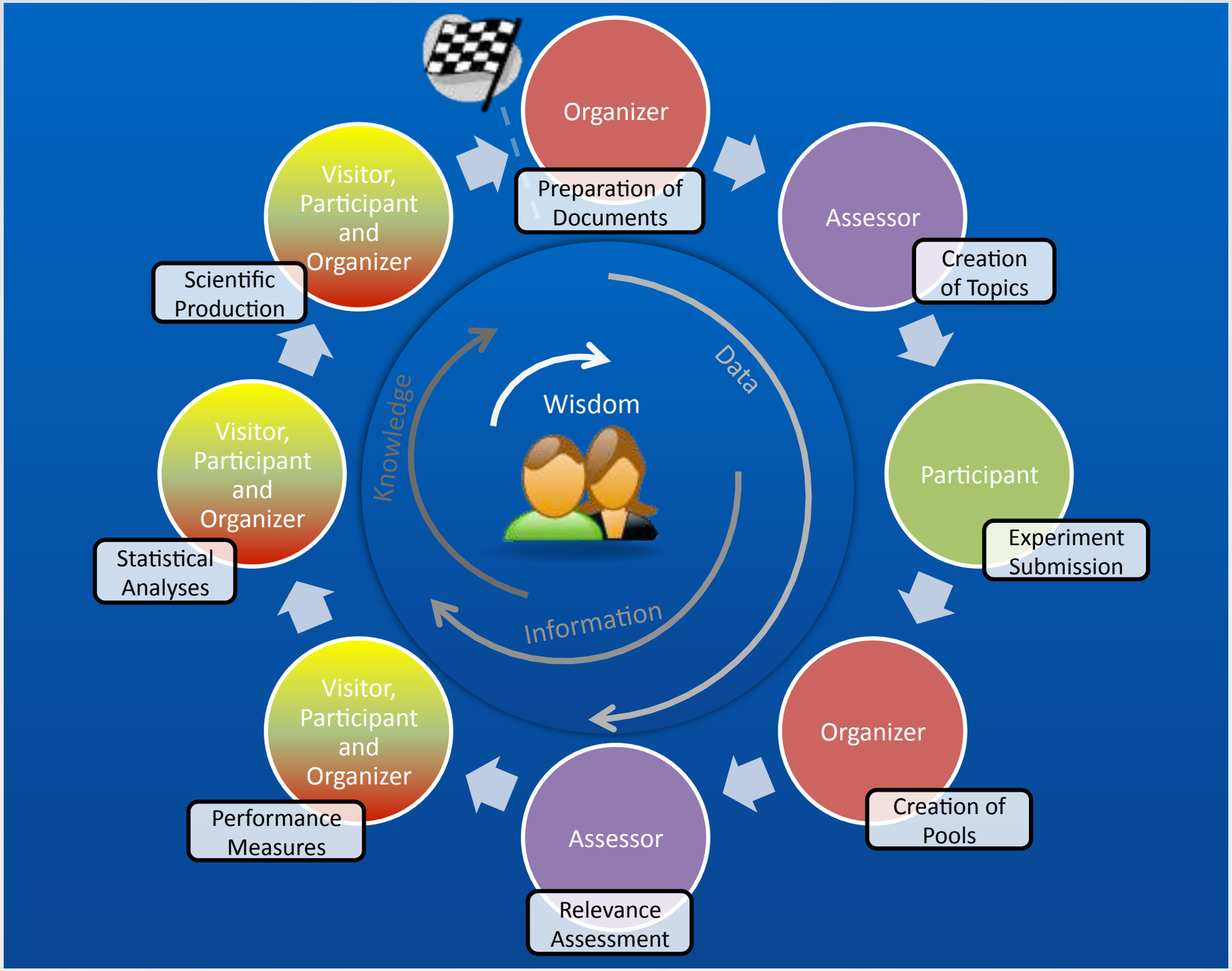
- **FIRE** (Forum for Information Retrieval Evaluation), India, since 2008
  - http://fire.irsi.res.in/



Mandar Mitra

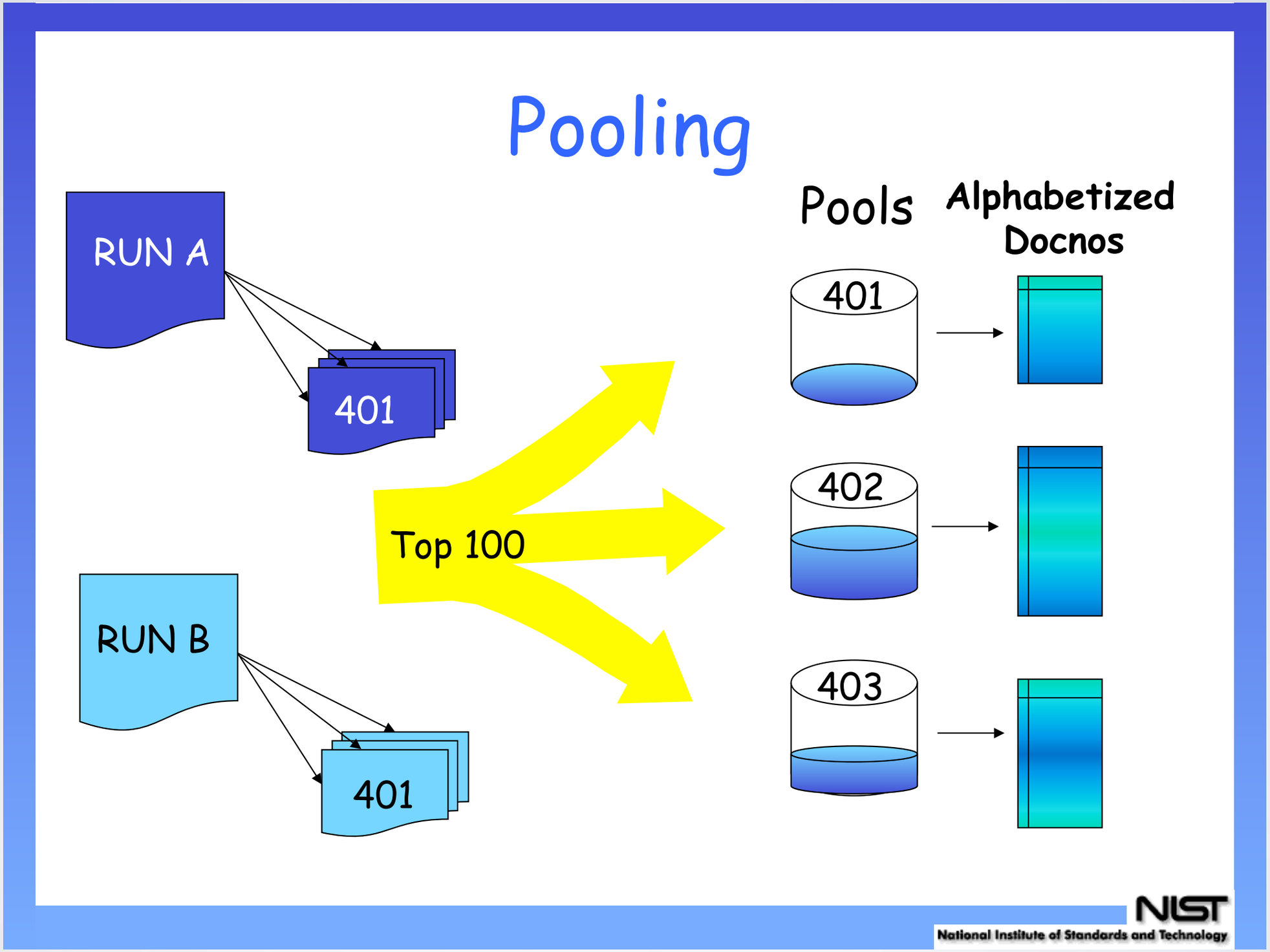
Prasenjit Majumder

Evaluation Initiatives: Typical Cycle



Dussin, M. and Ferro, N. (2009). Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., and Tsakonas, G., editors, Proc. *13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 63–74. LNCS 5714, Springer, Germany.

Traditional Depth-k Pools



Harman, D. K. (2013). TREC-Style Evaluations. In Agosti, M., Ferro, N., Forner, P., Müller, H., and Santucci, G., editors, *Information Retrieval Meets Information Visualization – PROMISE Winter School 2012, Revised Tutorial Lectures*, pages 97–115. Lecture Notes in Computer Science (LNCS) 7757, Springer, Heidelberg, Germany.

Relevance Assessment

2 judgments per minute

75 person/days per pool

35,000-75,000 documents per pool

NIST
National Institute of Standards and Technology

Harman, D. K. (2013). TREC-Style Evaluations. In Agosti, M., Ferro, N., Forner, P., Müller, H., and Santucci, G., editors, *Information Retrieval Meets Information Visualization – PROMISE Winter School 2012, Revised Tutorial Lectures*, pages 97–115. Lecture Notes in Computer Science (LNCS) 7757, Springer, Heidelberg, Germany.

Voorhees, E. M. and Harman, D. K. (2001). Overview of TREC 2001. In Voorhees, E. M. and Harman, D. K., editors, *The Tenth Text REtrieval Conference (TREC 2001)*, pages 1–15. NIST, Special Publication 500-250, Washington, USA.

What Makes a Good Pool?

- Depth-*k* pools require large enough *k* and different enough pooled systems in order to produce "**complete**" judgements
  - Not pooled/not assessed documents are typically assumed to be not relevant
  - This is a motivation for organizing large-scale evaluation initiatives
- The objective is not to allow for computing the "exact" value of an evaluation measure but rather to **comparatively assess systems** and detect significant differences in a robust way
- **Leave-one-out tests**: are used to assess the **reusability** of a pool
  - one system/group of systems is removed from the pool
  - all the systems are evaluated using both the original pool and the newly created one
  - the two sets of results are compared by computing the Kendall's τ correlation among the ranking of systems on the original and the new pool and/or the maximum drop in ranking

Justin Zobel

Zobel, J. (1998). How Reliable are the Results of Large-Scale Information Retrieval Experiments. In Croft, W. B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R., and Zobel, J., editors, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, pages 307–314. ACM Press, New York, USA.

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

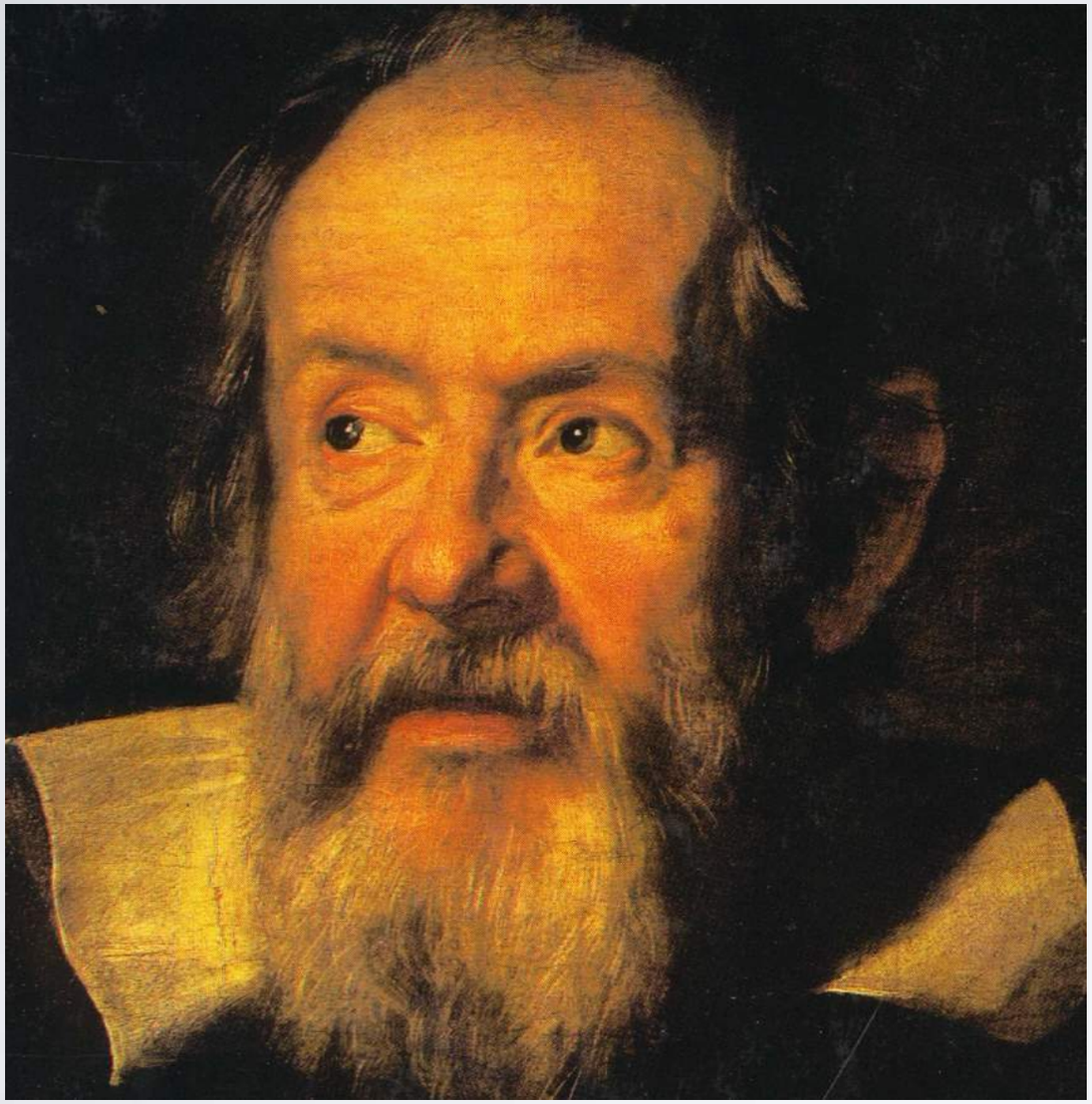
- They are the **pillars** for all the subsequent **scientific research** and **technology development**
  - TREC estimated the **return on investment** in the range of **3-5** for each invested dollar

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.

Questions?

Evaluation Measures

Evaluation Measures



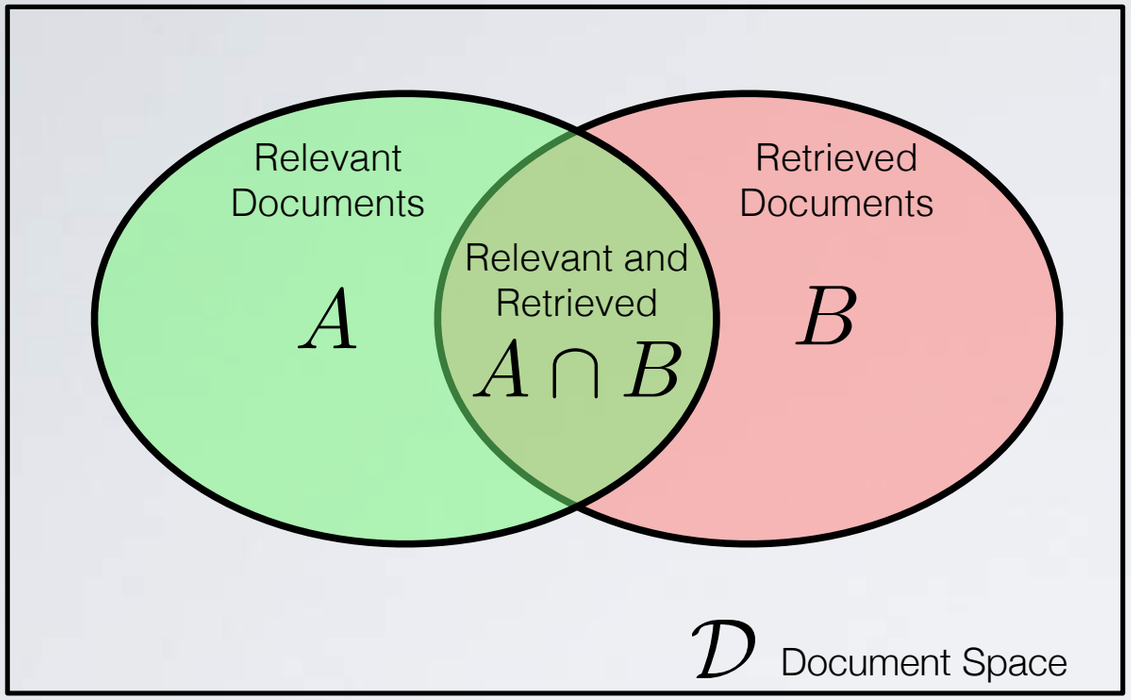
"Measure what is measurable and make measurable what is not"

Galileo Galilei (1564-1642)

A Taxonomy of Evaluation Measures

	Set-Based Retrieval	Rank-Based Retrieval
Binary Relevance	Precision (P) Recall (R) F-measure (F)	Precision at Document Cut-off (P@k) Recall at Document Cut-off (R@k) R-Precision (Rprec) Average Precision (AP) …
Multi-graded Relevance	Not widely agreed generalizations of Precision and Recall	Discounted Cumulated Gain (DCG) …

Set-based Measures: Precision, Recall and F-measure



$$P = \frac{|A \cap B|}{|B|} \qquad R = \frac{|A \cap B|}{|A|}$$

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = 2\frac{P \cdot R}{P + R}$$

- **Precision** is the the proportion of retrieved documents that are actually relevant
- **Recall** is the the proportion of relevant documents actually retrieved
- Together, Precision and Recall measure **retrieval effectiveness**, meant as the ability of a system to retrieve relevant documents while at the same time holding back non-relevant ones
  - maximizing Precision and Recall corresponds to optimal retrieval in the sense of the **Probability Ranking Principle**, i.e. ordering documents by their decreasing probability of being relevant, and creates a tight link between retrieval models and evaluation
- **F-measure** is the harmonic mean of Precision and Recall, summarising them into a single score

van Rijsbergen, C. J. (1974). Foundations of Evaluation. *Journal of Documentation*, 30(4):365–373.
van Rijsbergen, C. J. (1981). Retrieval effectiveness. In Spärck Jones, K., editor, *Information Retrieval Experiment*, pages 32–43. Butterworths, London, United Kingdom.

Set-based Measures: Example

$$P = \frac{4}{10} = 0.40$$

$$R = \frac{4}{8} = 0.50$$

$$F = 2 \cdot \frac{\frac{4}{10} \cdot \frac{4}{8}}{\frac{4}{10} + \frac{4}{8}} = \frac{4}{9} = 0.44$$

Assume

- $|A| = 8$ relevant documents in total
- Lenient mapping to binary relevance degrees

Rank-based Measures: Precision and Recall

- **Precision at Document Cut-off**:

$$P(k) = \frac{1}{k} \sum_{n=1}^{k} r_n$$

where $r_k \in \{0, 1\}$ is the relevance degree of the n-th document

- **Recall at Document Cut-off**:

$$R(k) = \frac{1}{RB} \sum_{n=1}^{k} r_n$$

where $RB = |A|$ is the **recall base**, i.e. the total number of relevant documents

- **Rprec** is Precision computed at the recall base

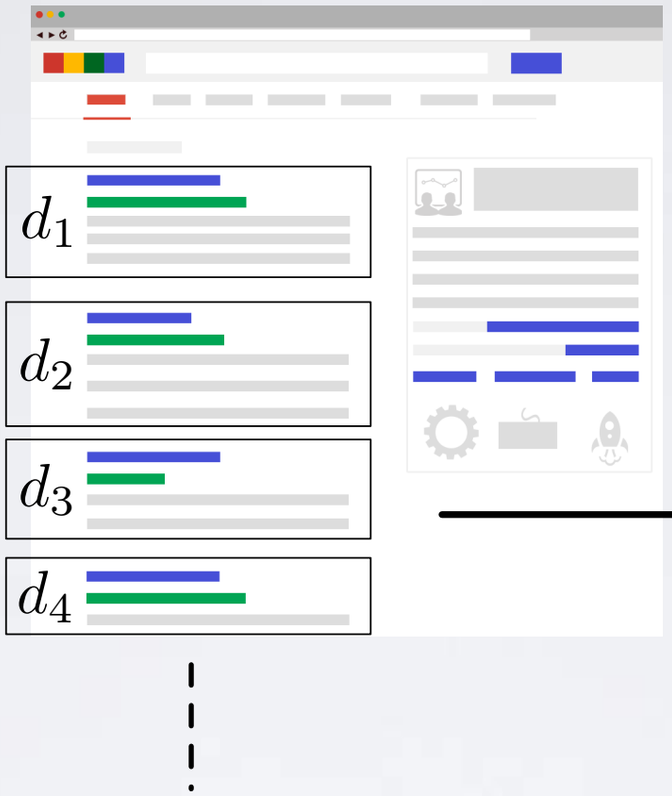
$$Rprec = P(RB)$$

Rank-based Measures: Example of Precision and Recall

Topic | Run | Assessed Run | Binary Weighted Assessed Run

Assume

- $RB = 8$ relevant documents in total
- Lenient mapping to binary relevance degrees



	Assessed Run	Binary Weighted Assessed Run
1	Highly Relevant	1
2	Not Relevant	0
3	Partially Relevant	1
4	Fairly Relevant	1
5	Not Relevant	0
6	Not Relevant	0
7	Not Relevant	0
8	Fairly Relevant	1
9	Not Relevant	0
10	Not Relevant	0

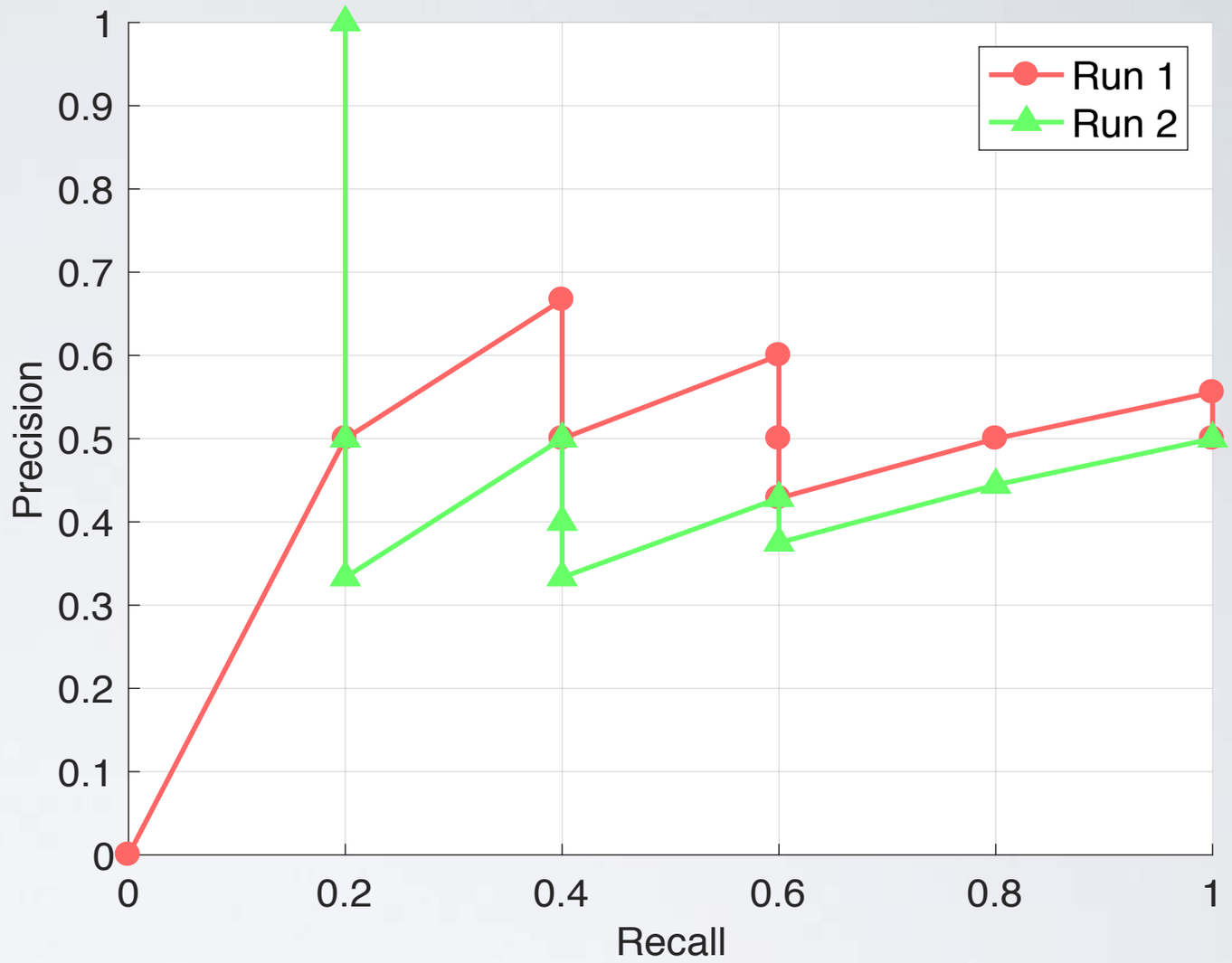
$$P(5) = \frac{3}{5} = 0.600$$

$$R(5) = \frac{3}{8} = 0.375$$

$$Rprec = P(8) = \frac{4}{8} = 0.500$$

Precision-Recall Curve

	Run1		Run2	
1	0	P = 0.00 R = 0.00	1	P = 1.00 R = 0.20
2	1	P = 0.50 R = 0.20	0	P = 0.50 R = 0.20
3	1	P = 0.66 R = 0.40	0	P = 0.33 R = 0.20
4	0	P = 0.50 R = 0.40	1	P = 0.50 R = 0.40
5	1	P = 0.60 R = 0.60	0	P = 0.40 R = 0.40
6	0	P = 0.50 R = 0.60	0	P = 0.33 R = 0.40
7	0	P = 0.42 R = 0.60	1	P = 0.42 R = 0.60
8	1	P = 0.50 R = 0.80	0	P = 0.37 R = 0.60
9	1	P = 0.55 R = 1.00	1	P = 0.44 R = 0.80
10	0	P = 0.50 R = 1.00	1	P = 0.50 R = 1.00



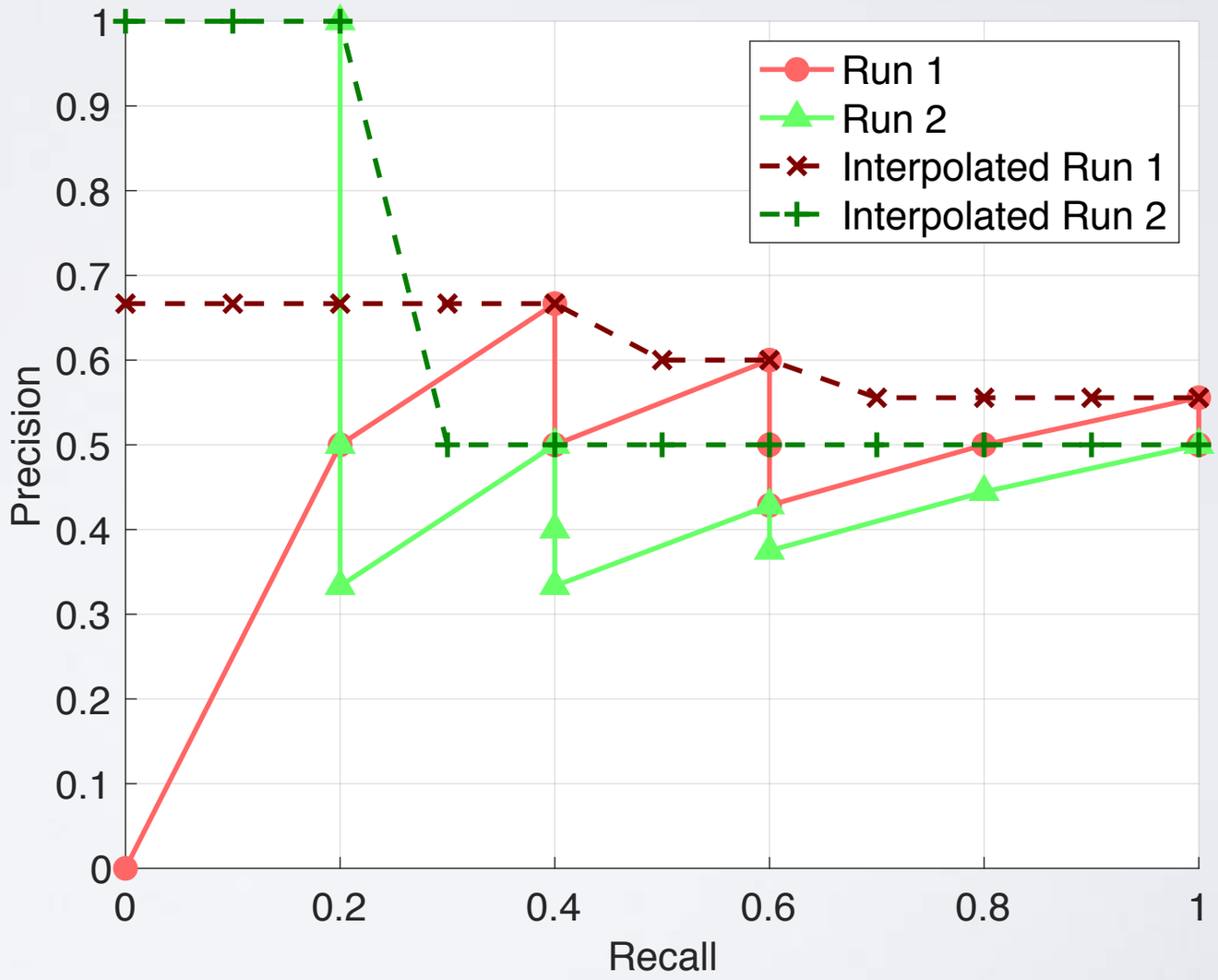
- Assume $RB = 5$ relevant documents in total
- The Precision-Recall curve has a typical saw-tooth shape
  - We may have multiple Precision values for the same Recall value
  - It is difficult to compare runs because they may not have the same Recall values

Interpolated Precision-Recall Curve

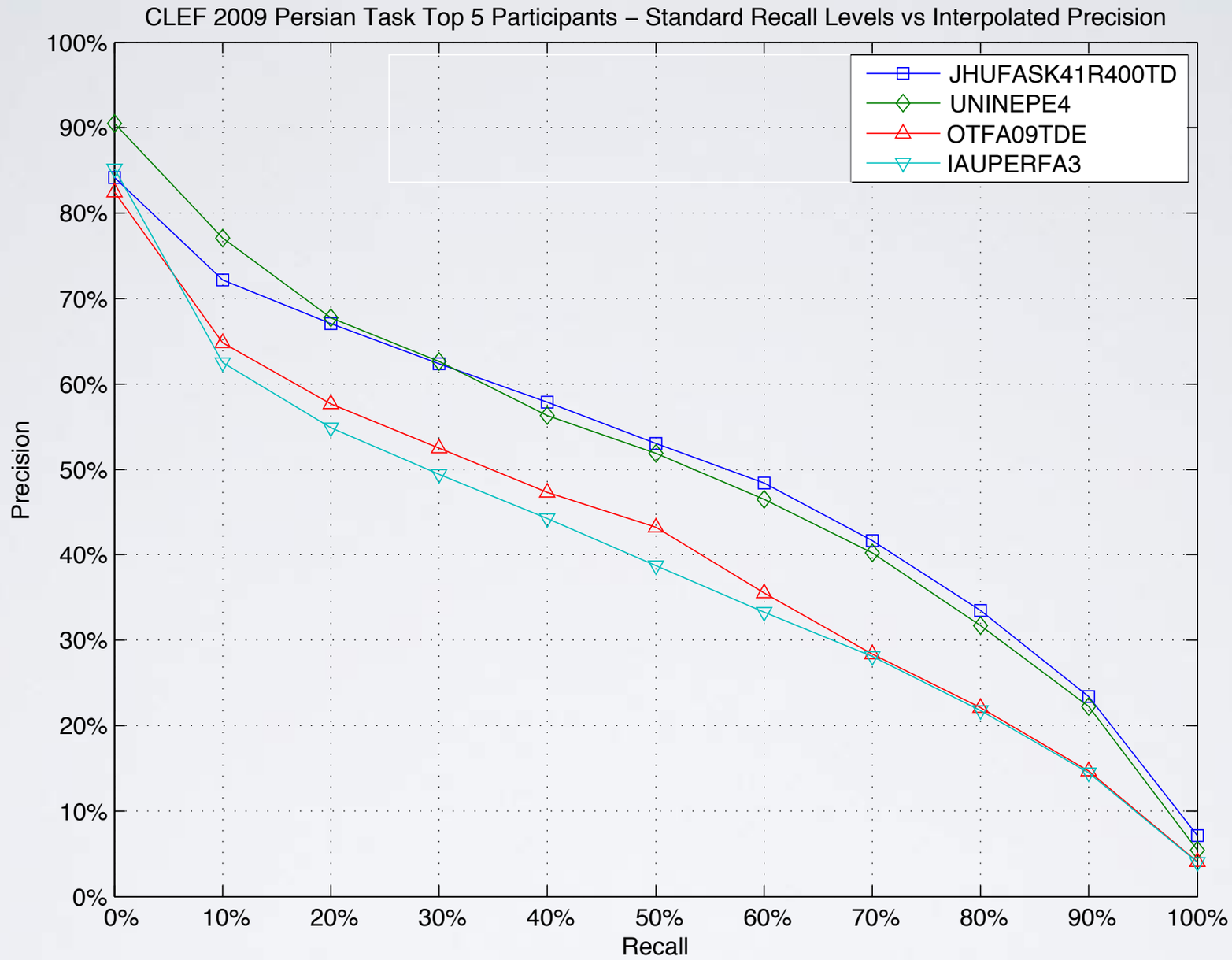
	Run1		Run2	
1	0	**iP = 0.66** P = 0.00 R = 0.00	1	**iP = 1.00** P = 1.00 R = 0.20
2	1	**iP = 0.66** P = 0.50 R = 0.20	0	**iP = 1.00** P = 0.50 R = 0.20
3	1	**iP = 0.66** P = 0.66 R = 0.40	0	**iP = 1.00** P = 0.33 R = 0.20
4	0	**iP = 0.66** P = 0.50 R = 0.40	1	**iP = 0.50** P = 0.50 R = 0.40
5	1	**iP = 0.60** P = 0.60 R = 0.60	0	**iP = 0.50** P = 0.40 R = 0.40
6	0	**iP = 0.60** P = 0.50 R = 0.60	0	**iP = 0.50** P = 0.33 R = 0.40
7	0	**iP = 0.60** P = 0.42 R = 0.60	1	**iP = 0.50** P = 0.42 R = 0.60
8	1	**iP = 0.55** P = 0.50 R = 0.80	0	**iP = 0.50** P = 0.37 R = 0.60
9	1	**iP = 0.55** P = 0.55 R = 1.00	1	**iP = 0.50** P = 0.44 R = 0.80
10	0	**iP = 0.55** P = 0.50 R = 1.00	1	**iP = 0.50** P = 0.50 R = 1.00

- To interpolate Precision at standard Recall value R_j we use the maximum Precision obtained for any actual Recall value R greater than or equal to R_j

$$iP@R_j = \max_{R \geq R_j} P@R$$



11 points Interpolated Precision-Recall Curve



- Standard Interpolated Precision-Recall curves exhibit a typical inverse relationship among Precision and Recall, indicating a trade-off between these two goals of effectiveness

Cleverdon, C. W. (1972). On the inverse relationship of recall and precision. *Journal of Documentation*, 28(3):195–201.

Buckland, M. and Gey, F. (1994). The relationship between Recall and Precision. *Journal of the American Society for Information Science and Technology (JASIST)*, 45(1):12–19.

Egghe, L. (2008). The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations. *Information Processing & Management*, 44(2):856–876.

Rank-based Measures: Average Precision

$$AP = \frac{1}{RB} \sum_{k \in \mathcal{R}} P(k) = \frac{1}{RB} \sum_{n=1}^{N} \underbrace{\left(\frac{1}{n} \sum_{m=1}^{n} r_m \right)}_{P(n)} r_n =$$

$$= \underbrace{\frac{rr}{RB}}_{\text{Recall}} \cdot \underbrace{\frac{1}{rr} \sum_{k \in \mathcal{R}} P(k)}_{\text{arithmetic mean of } P(k)}$$

where

- $\mathcal{R}$ is the set of the rank positions of the relevant retrieved documents
- $rr = |\mathcal{R}|$ is the total number of relevant retrieved documents
- N is the total number of retrieved documents, i.e. the length of the run

Chris Buckley

- The **Mean Average Precision (MAP)** is the mean of AP over a set of topics
  - Differently from the other measures, this mean has its own name since it is the most widely used single number to summarise the whole performance of a system

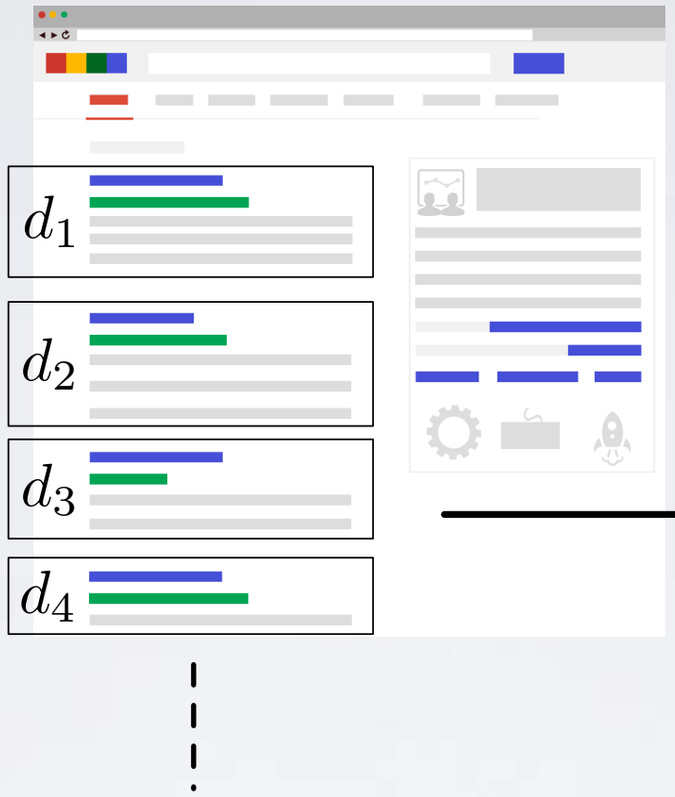
Buckley, C. and Voorhees, E. M. (2005). Retrieval System Evaluation. In Harman, D. K. and Voorhees, E. M., editors, *TREC. Experiment and Evaluation in Information Retrieval*, pages 53–78. MIT Press, Cambridge (MA), USA.

Rank-based Measures: Example of Average Precision

Topic | Run | Assessed Run | Binary Weighted Assessed Run

Assume

- $RB = 8$ relevant documents in total
- Lenient mapping to binary relevance degrees



	Assessed Run	Binary Weighted Assessed Run
1	Highly Relevant	1
2	Not Relevant	0
3	Partially Relevant	1
4	Fairly Relevant	1
5	Not Relevant	0
6	Not Relevant	0
7	Not Relevant	0
8	Fairly Relevant	1
9	Not Relevant	0
10	Not Relevant	0

$$AP = \frac{1}{RB}\Big(P(1) + P(3) + P(4) + P(8)\Big)$$

$$= \frac{1}{8}\left(1 + \frac{2}{3} + \frac{3}{4} + \frac{4}{8}\right) = \frac{35}{96} = 0.36$$

Rank-based Measures: Discounted Cumulated Gain

$$DCG(k) = \begin{cases} \sum_{n=1}^{k} r_k & \text{if } k < b \\ DCG(k-1) + \frac{r_k}{\log_b(k)} & \text{if } k >= b \end{cases} =$$

$$= \sum_{n=1}^{k} \frac{r_k}{\max(1, \log_b(k))}$$

- where the base of the logarithm b indicates the patience of the user in scanning the result list
  - $b = 2$ is an impatient user
  - $b = 10$ is a patient user
- DCG naturally handles multi-graded relevance
- DCG does not depend on the recall base
- DCG is not bounded in [0, 1]

Kalervo Järvelin

Jaana Kekäläinen

Järvelin, K. and Kekäläinen, J. (2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446

Rank-based Measures: Discounted Cumulated Gain

$$DCG(k) = \begin{cases} \sum_{n=1}^{k} r_k & \text{if } k < b \\ DCG(k-1) + \frac{r_k}{\log_b(k)} & \text{if } k >= b \end{cases} =$$

$$= \sum_{n=1}^{k} \frac{r_k}{\max(1, \log_b(k))}$$

Cumulated Gain (CG)

- where the base of the logarithm b indicates the patience of the user in scanning the result list
  - $b = 2$ is an impatient user
  - $b = 10$ is a patient user
- DCG naturally handles multi-graded relevance
- DCG does not depend on the recall base
- DCG is not bounded in [0, 1]

Kalervo Järvelin

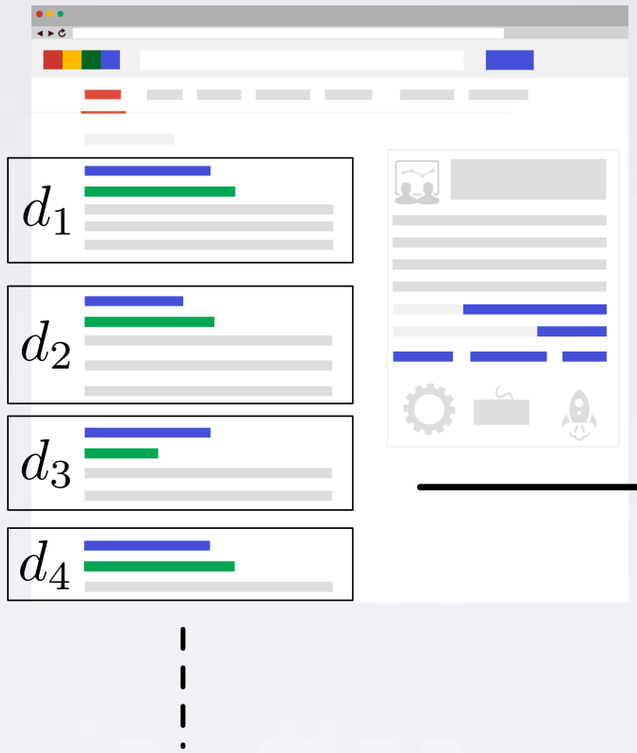
Jaana Kekäläinen

Järvelin, K. and Kekäläinen, J. (2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446

Rank-based Measures: Example of DCG

Assume

- $RB = 8$ relevant documents in total
- An impatient user



	Topic	Run	Assessed Run	Weighted Assessed Run
1			Highly Relevant	3
2			Not Relevant	0
3			Partially Relevant	1
4			Fairly Relevant	2
5			Not Relevant	0
6			Not Relevant	0
7			Not Relevant	0
8			Fairly Relevant	2
9			Not Relevant	0
10			Not Relevant	0

$$DCG = 3 + \frac{1}{\log_2(3)} + \frac{2}{\log_2(4)} + \frac{2}{\log_2(8)} = 5.2976$$

Rank-based Measures: Normalized DCG

- To normalize DCG in [0, 1], you need to compute the ideal run, i.e. the pool sorted in descending order of relevance, which represents the best retrieval possible and the maximum value of DCG

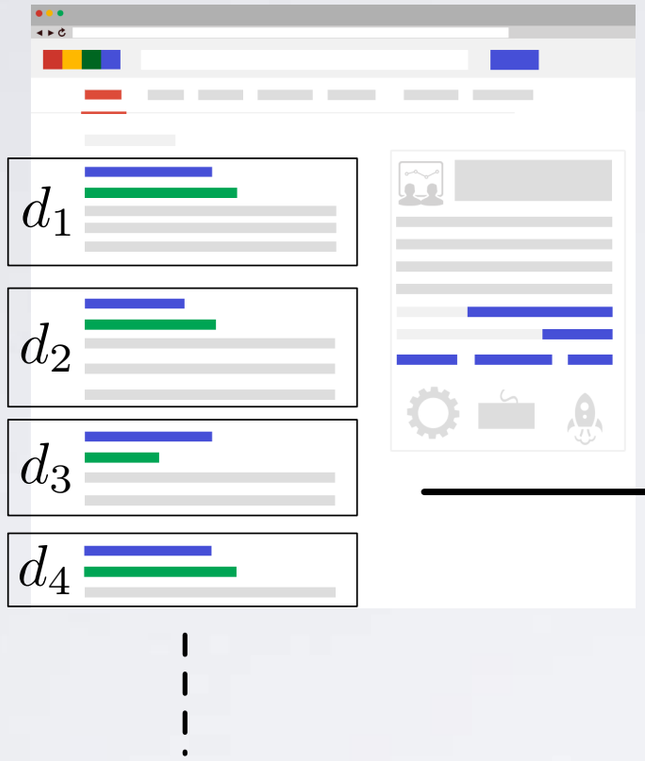
$$nDCG(k) = \frac{DCG(k)}{iDCG(k)}$$

- nDCG is given by the DCG of the run divided by the DCG of the ideal run

Rank-based Measures: Example of nDCG

Assume

- $RB = 8$ relevant documents in total
- An impatient user



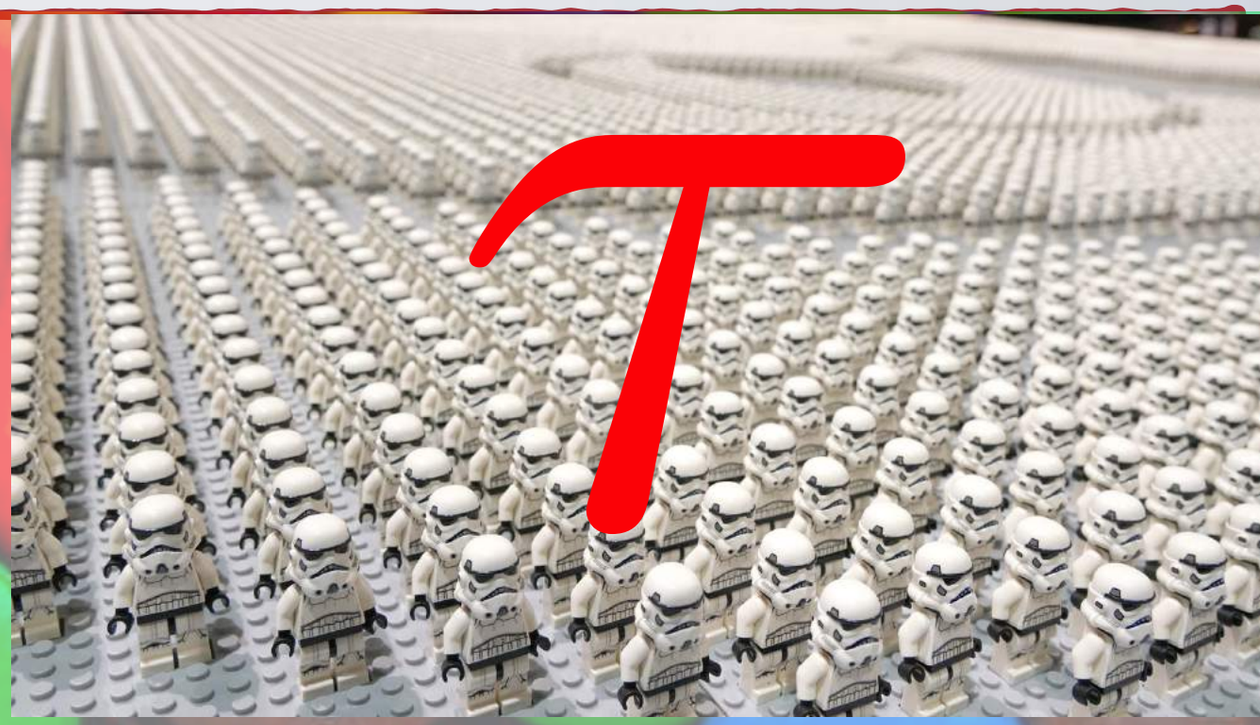
	Assessed Run	Weighted Assessed Run	Weighted Assessed Ideal Run
1	Highly Relevant	3	3
2	Not Relevant	0	3
3	Partially Relevant	1	2
4	Fairly Relevant	2	2
5	Not Relevant	0	2
6	Not Relevant	0	1
7	Not Relevant	0	1
8	Fairly Relevant	2	1
9	Not Relevant	0	0
10	Not Relevant	0	0

$$DCG = 5.2976$$

$$iDCG = 10.1996$$

$$nDCG = 0.5194$$

A Science of Measures...



User Models

Top-heaviness

Sensitivity

Incomplete Information

… But, Wait, Assumptions?

- The **operations** you are **allowed** to perform with the values of a measure depend on the notion of **measurement scale**
  - mean
  - variance
  - correlation
  - statistical tests
  - ….

- ✦ Do IR measures **comply with** those assumptions?
- ✦ How much are (statistical) analyses impacted by **departures** from those assumptions?
- ✦ What is the **validity** of our experiments?

Ferrante, M., Ferro, N., and Pontarollo, S. (2019). A General Theory of IR Evaluation Measures. *IEEE TKDE*, 31(3):409–422.
Ferrante, M., Ferro, N., and Losiouk, E. (2019). How do interval scales help us with better understanding IR evaluation measures? *Information Retrieval Journal*.

… But, Wait, Assumptions?

Ferrante, M., Ferro, N., and Pontarollo, S. (2019). A General Theory of IR Evaluation Measures. *IEEE TKDE*, 31(3):409–422.
Ferrante, M., Ferro, N., and Losiouk, E. (2019). How do interval scales help us with better understanding IR evaluation measures? *Information Retrieval Journal*.

Questions?


It works...It doesn't....It works.....
Just a hunch....maybe we do need a better way to measure results.....

Reproducibility

Reproducibility: Why?

What we find reported in papers…

… what happens to us

https://twitter.com/DevilleSy/status/958761021421903872

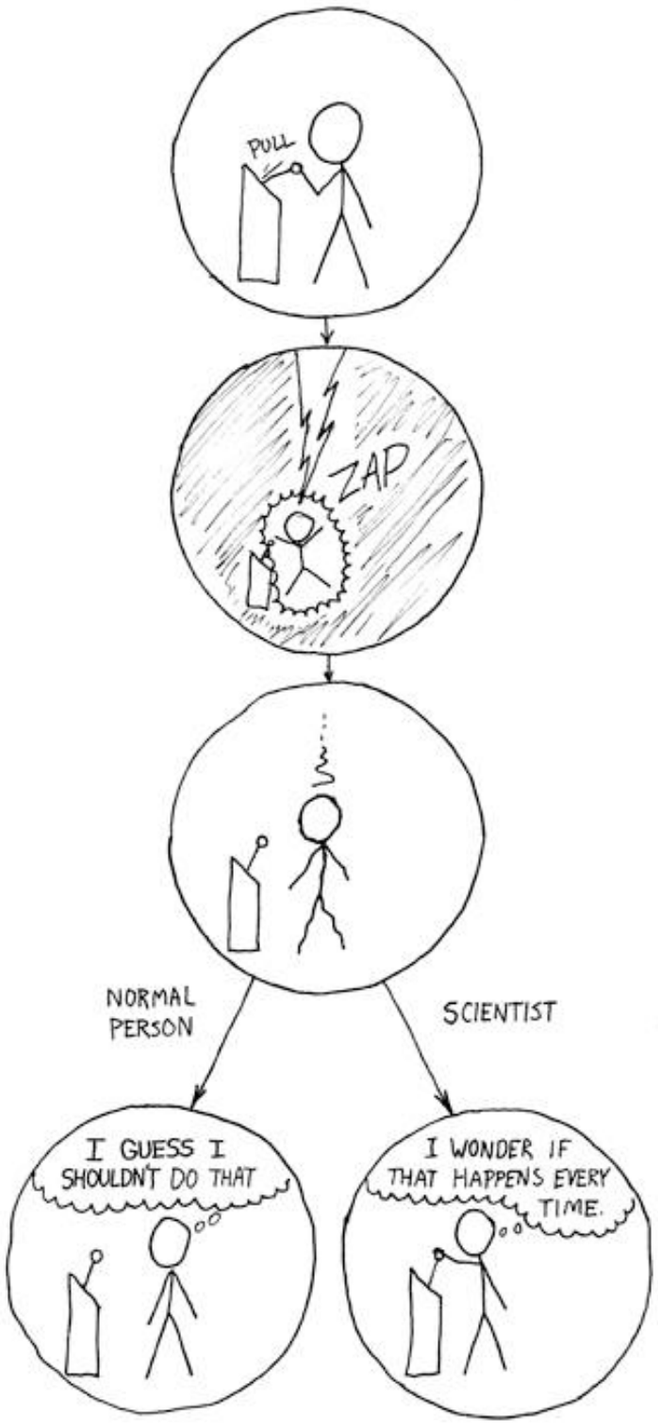
Reproducibility: How?

Everybody likes reproducibility…

…as soon as someone else does it

Reproducibility: What?

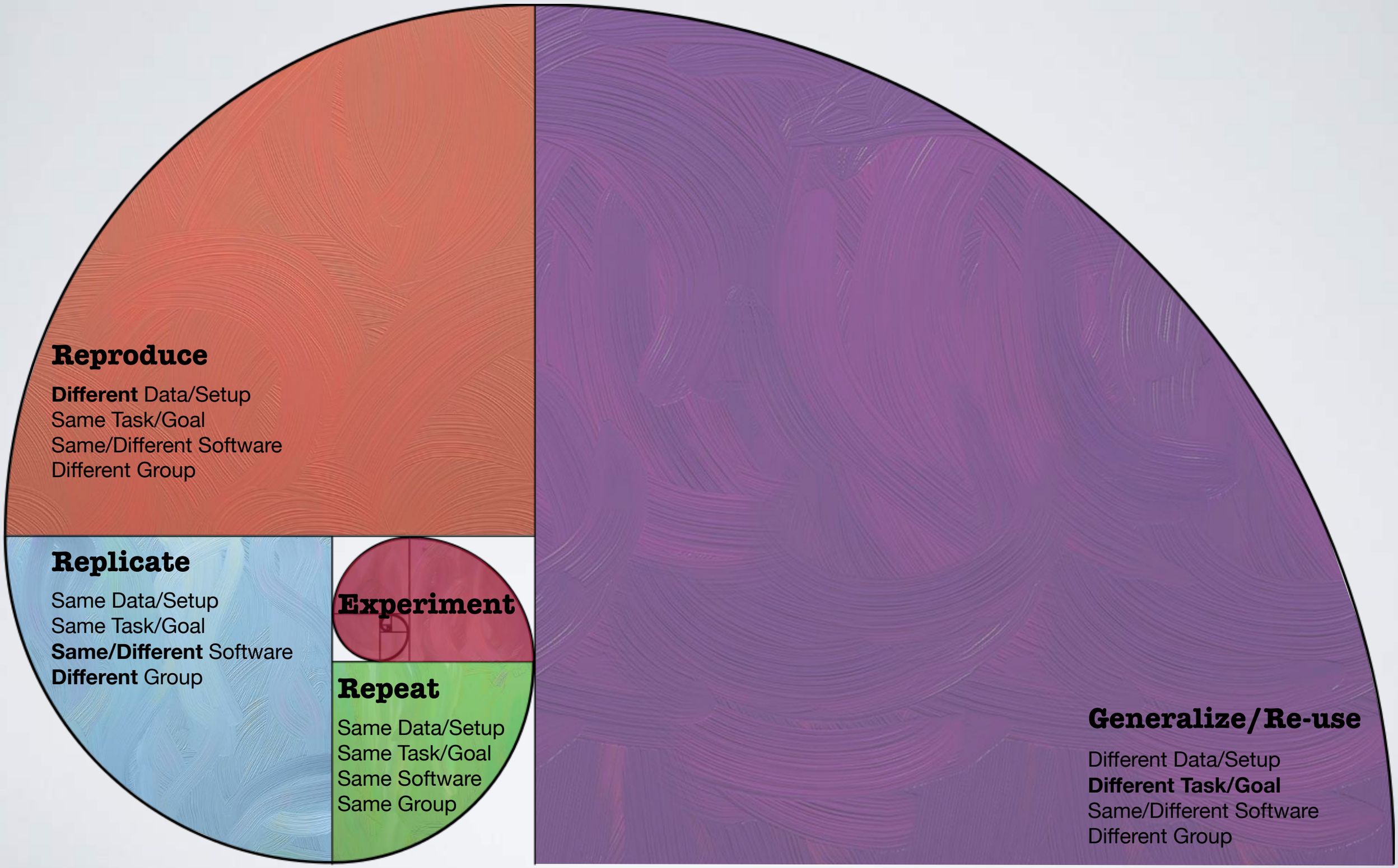
We know, reproducibly is at the core of science…



https://xkcd.com/242/

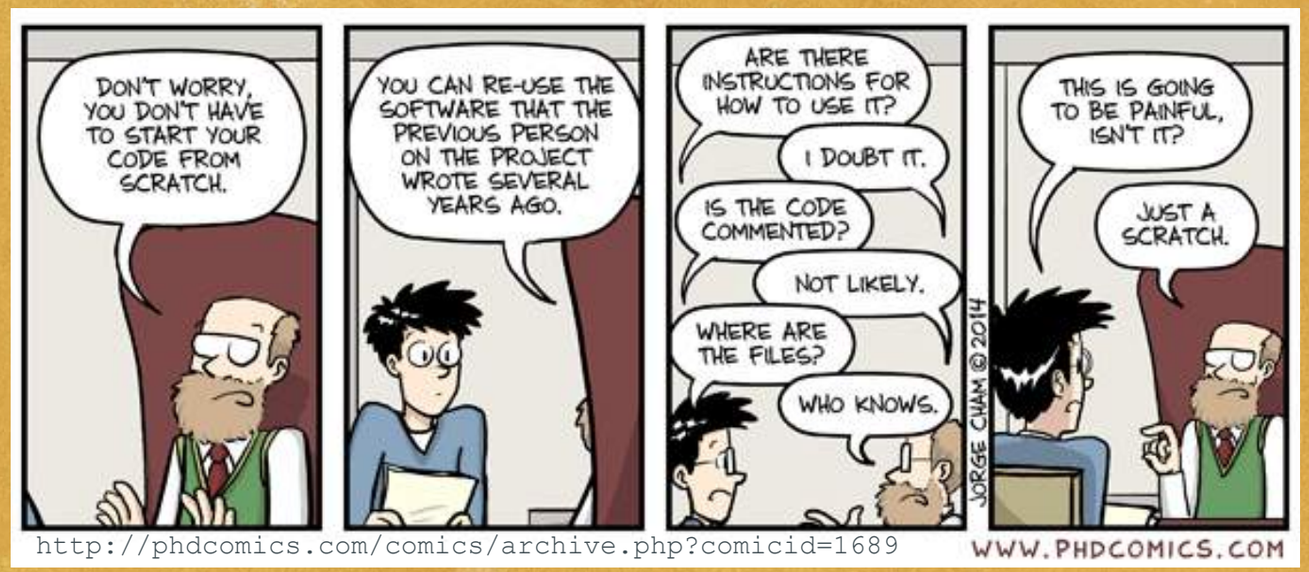
…but reproducing research is not new research

The “Reproducibility” Nautilus



The “Reproducibility” Nautilus

Is it really that easy?



Different Group

Replicate

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

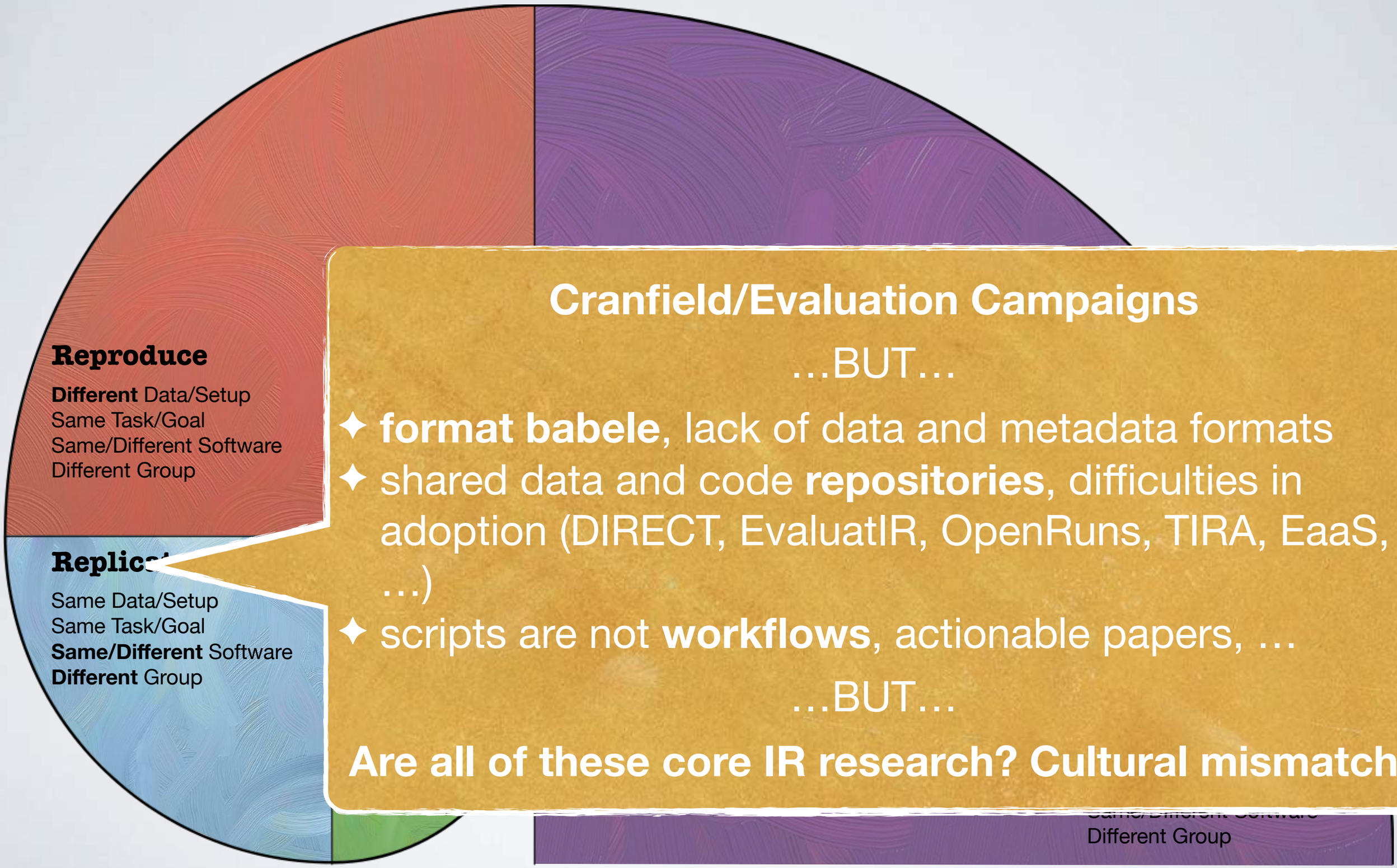
Repeat

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

Generalize/Re-use

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

The “Reproducibility” Nautilus



Cranfield/Evaluation Campaigns

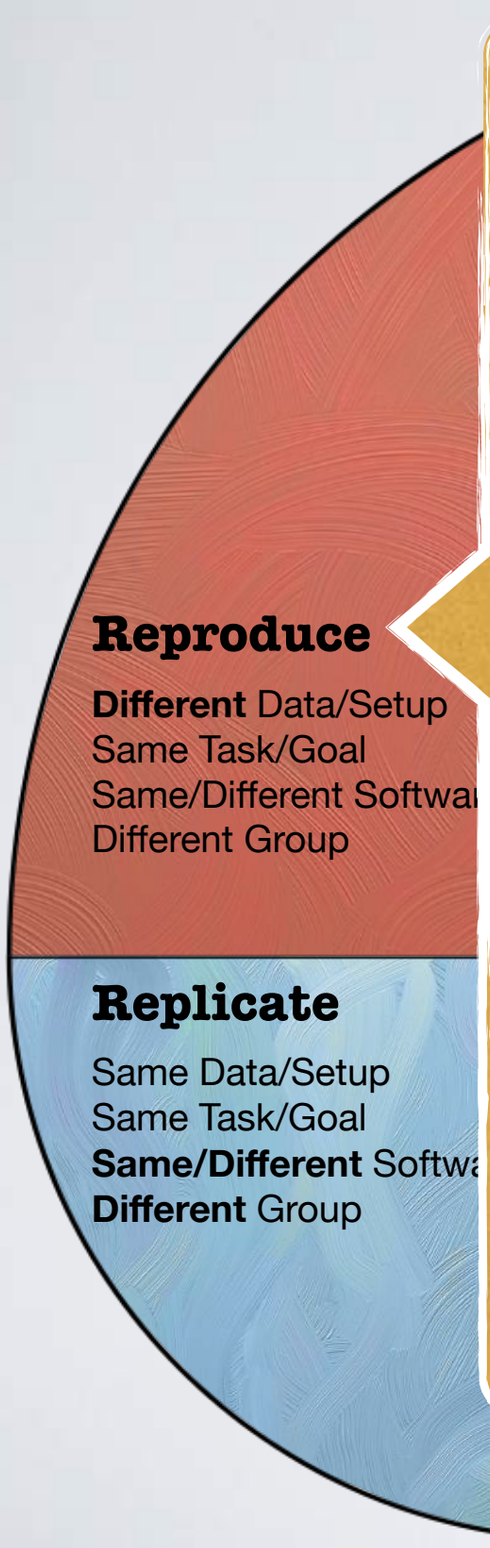
…BUT…

- **format babele**, lack of data and metadata formats
- shared data and code **repositories**, difficulties in adoption (DIRECT, EvaluatIR, OpenRuns, TIRA, EaaS, …)
- scripts are not **workflows**, actionable papers, …

…BUT…

Are all of these core IR research? Cultural mismatch

The “Reproducibility” Nautilus



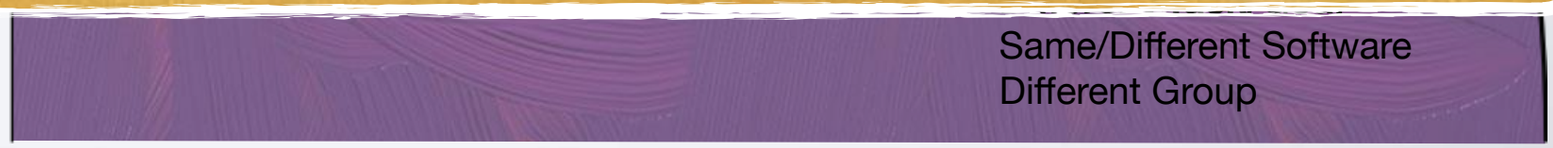
Somehow standard approach in IR evaluation

…BUT…

- typically done **with-in group**, in a repeatability-like fashion
- **how to quantify when “reproduced”** - same ranked list, correlation among ranked lists, same effectiveness score, confidence bounds on effectiveness score, close distributions of effectiveness score, similar statistical significance (p-values, effect sizes, …), …
- what about the **user-side** 😱?

…Initial investigation in CENTRE@CLEF/NTCIR/TREC…

http://www.centre-eval.org/



The “Reproducibility” Nautilus

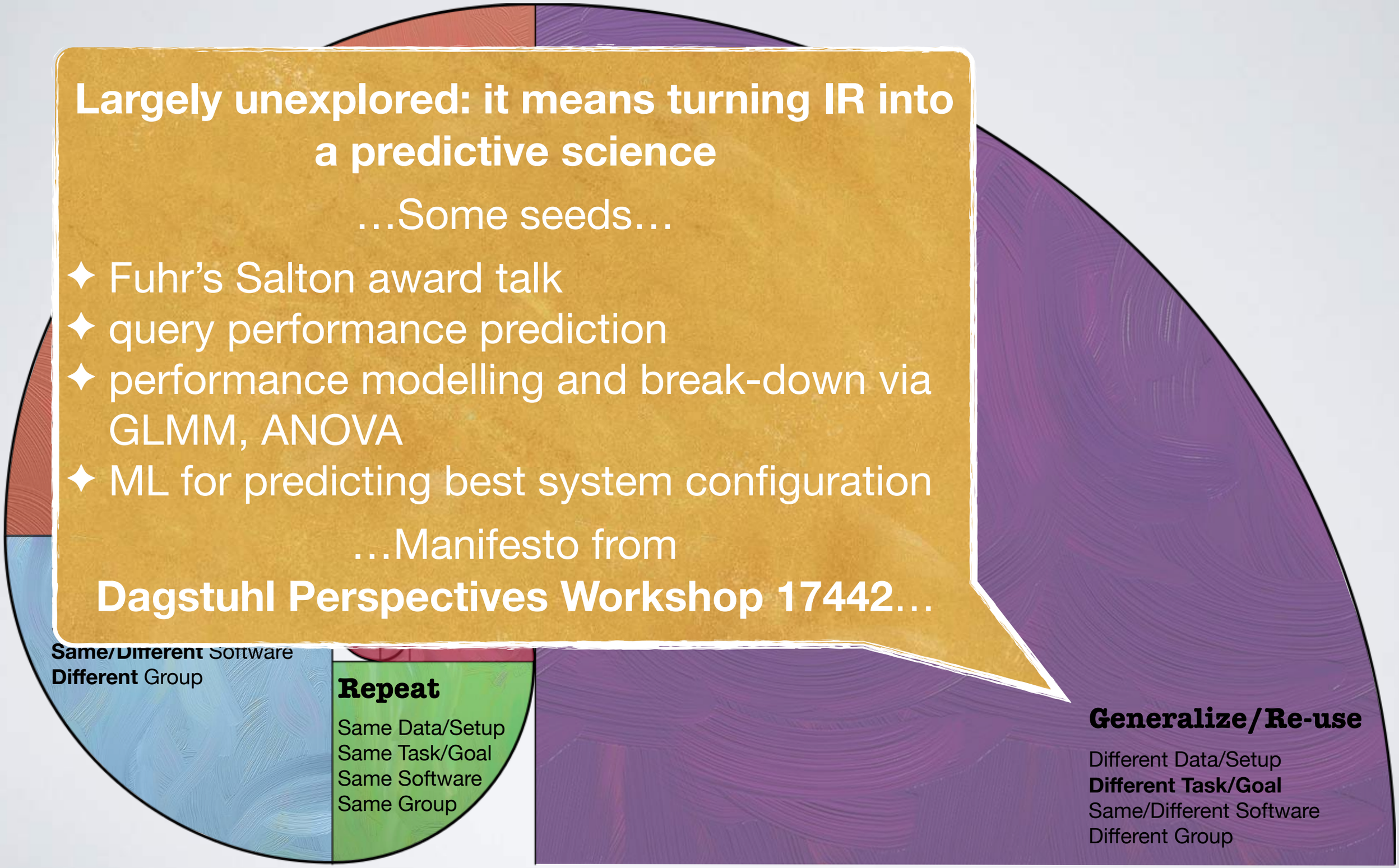
Largely unexplored: it means turning IR into a predictive science

…Some seeds…

- ✦ Fuhr’s Salton award talk
- ✦ query performance prediction
- ✦ performance modelling and break-down via GLMM, ANOVA
- ✦ ML for predicting best system configuration

…Manifesto from

Dagstuhl Perspectives Workshop 17442…

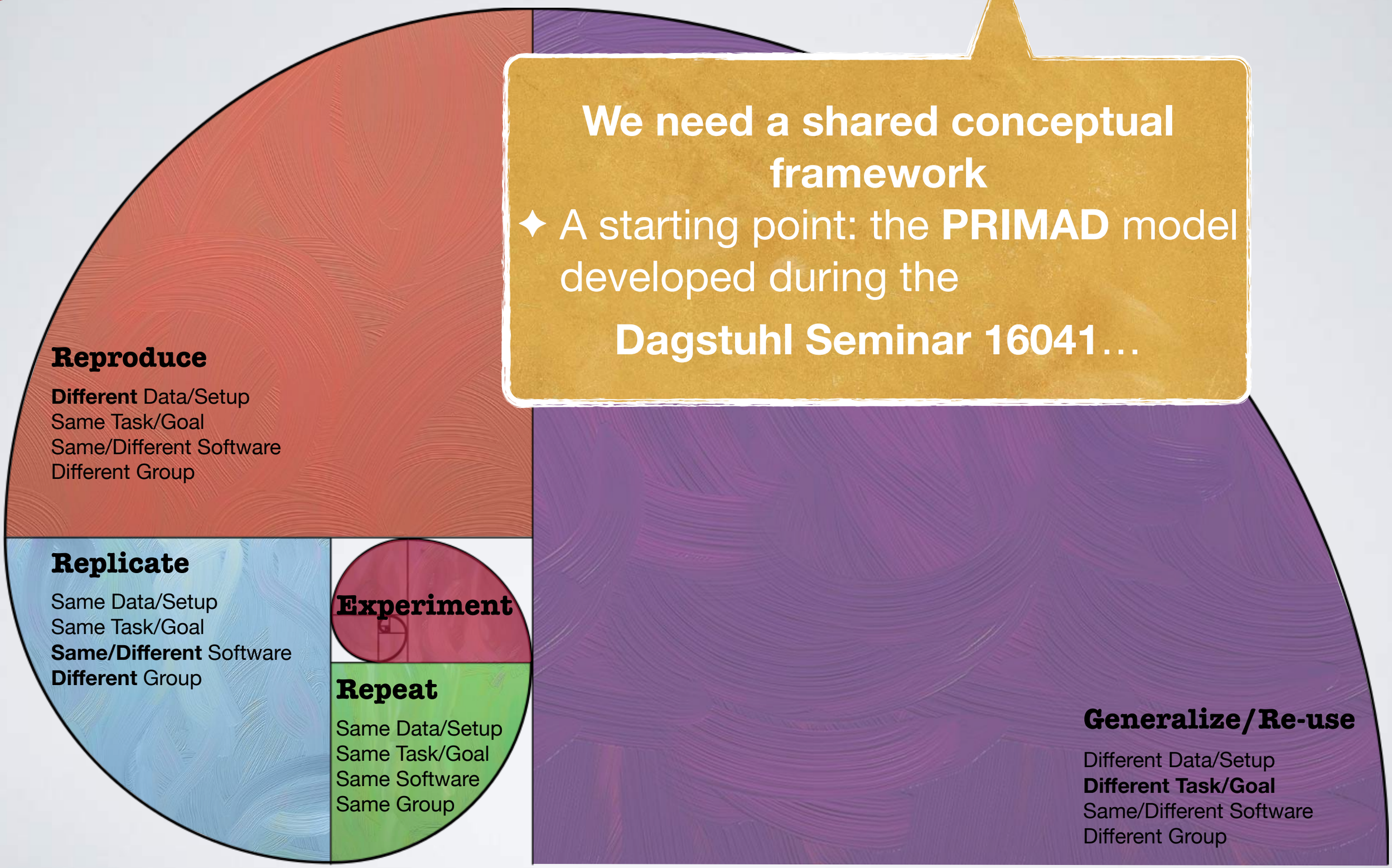


The “Reproducibility” Nautilus

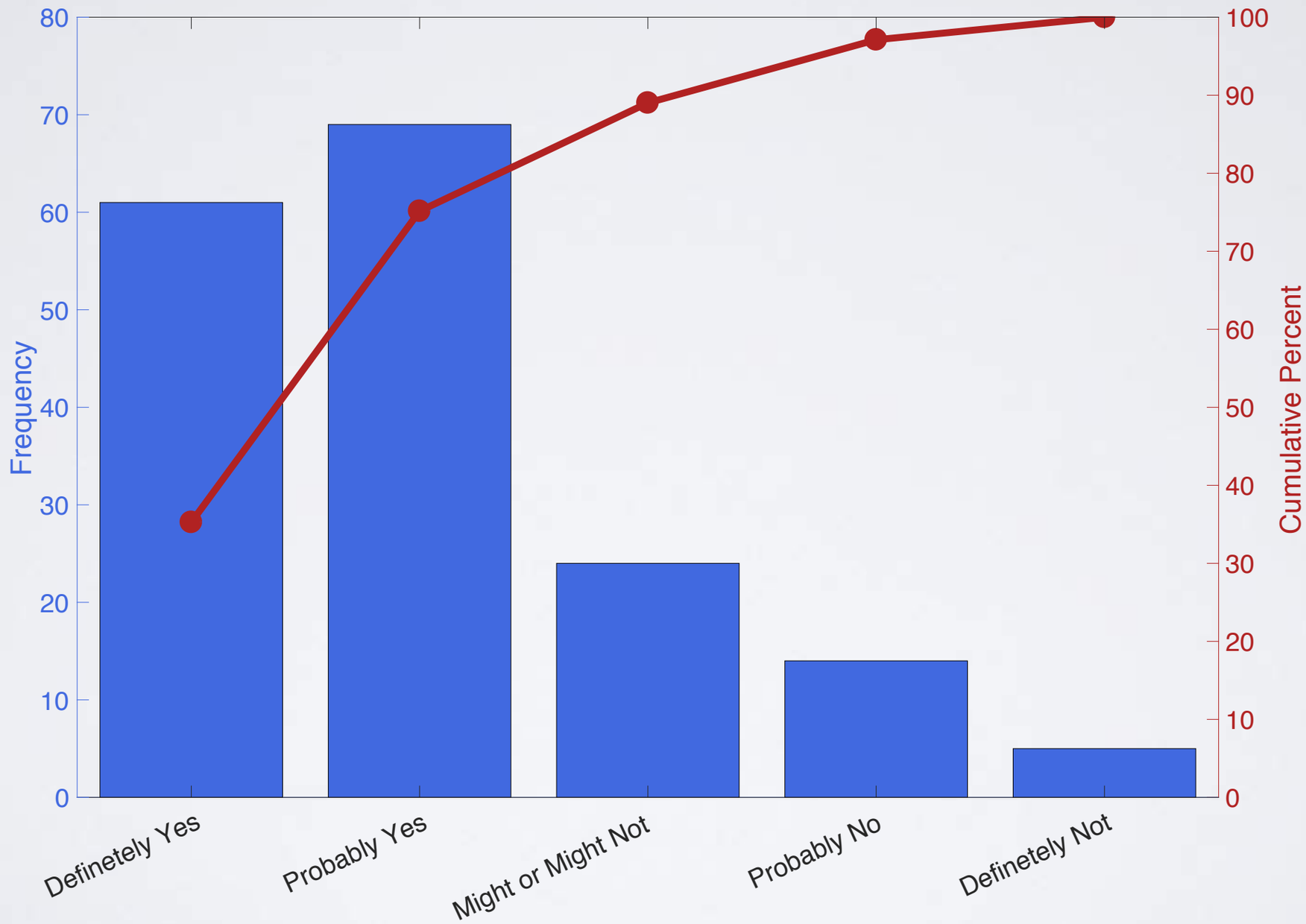
We need a shared conceptual framework

✦ A starting point: the **PRIMAD** model developed during the

Dagstuhl Seminar 16041…



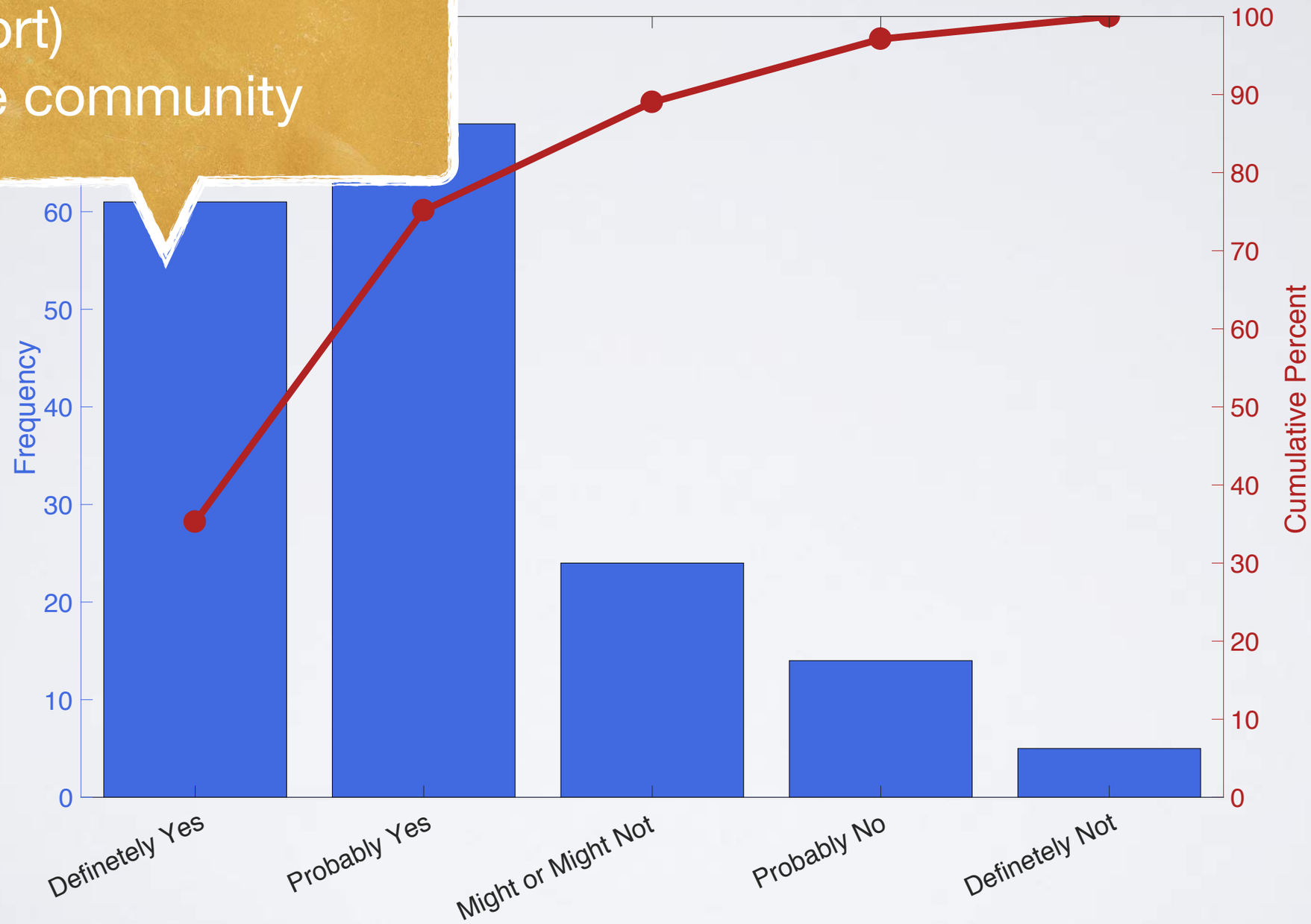
What about Introducing Badges?



Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

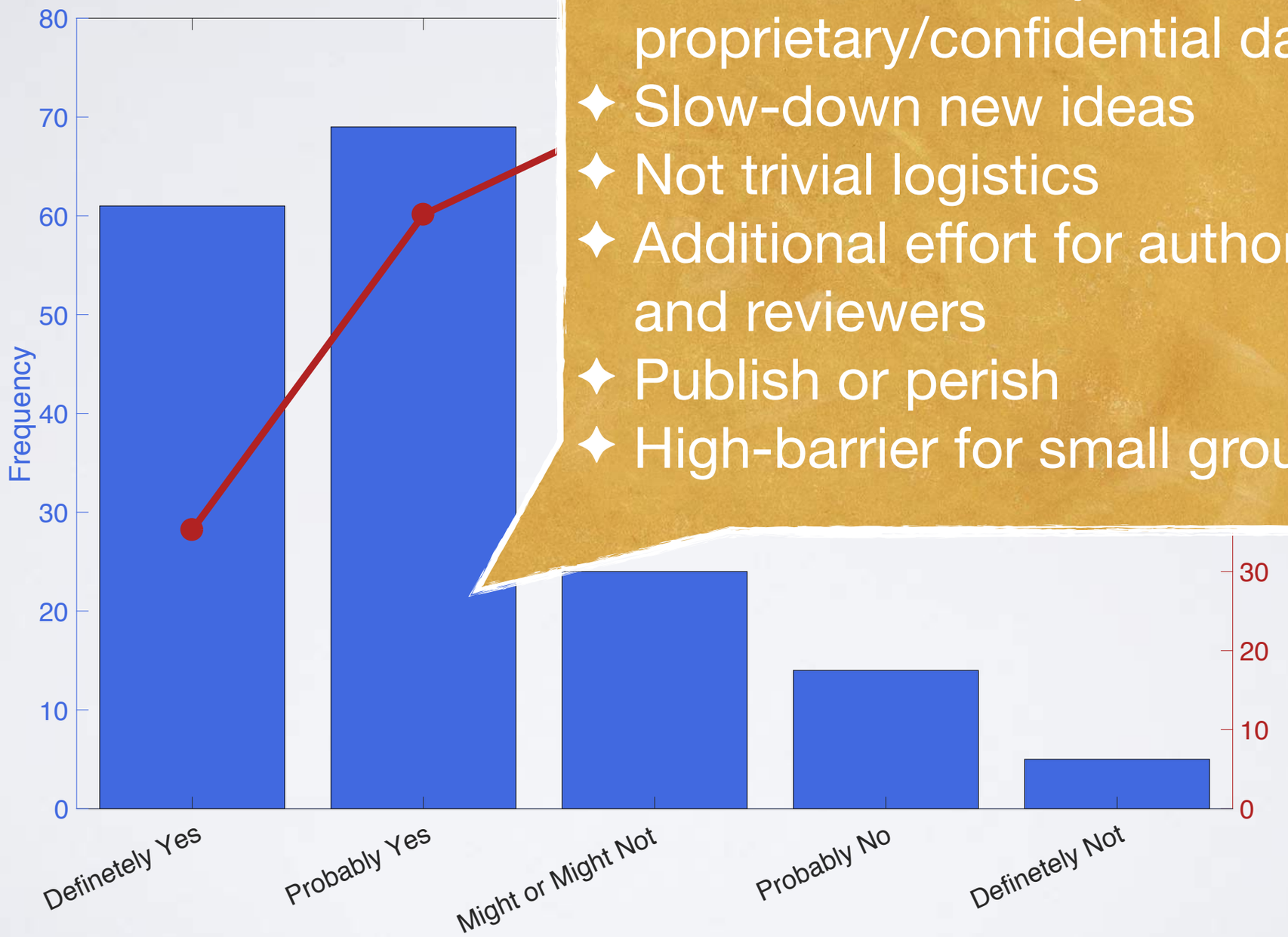
What about Introducing Badges?

- Better science
- Better baselines (and less effort)
- Improve community



Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

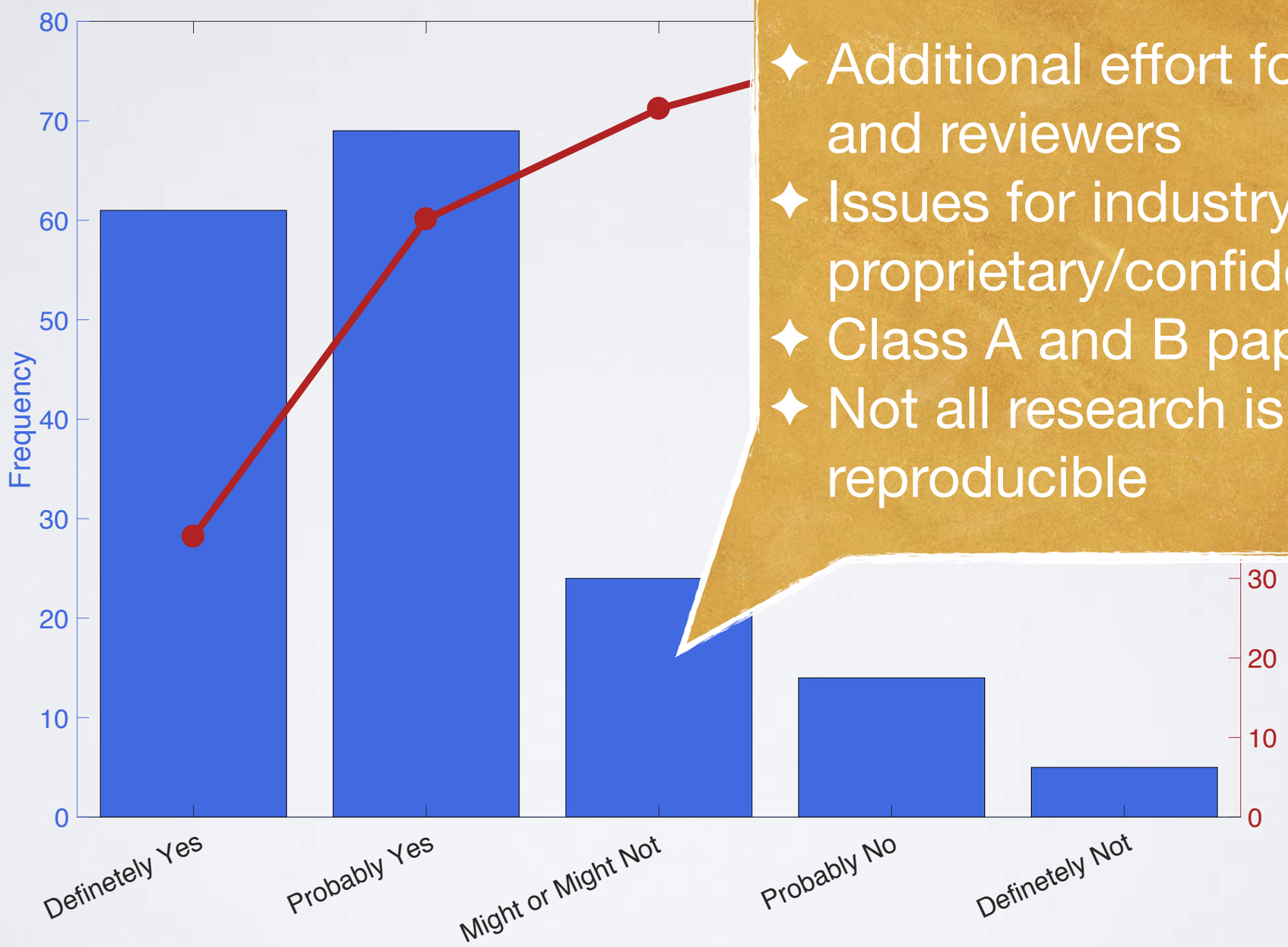
What about Introducing Badges?



- Better science
- Issues for industry, proprietary/confidential data
- Slow-down new ideas
- Not trivial logistics
- Additional effort for authors and reviewers
- Publish or perish
- High-barrier for small groups

Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

What about Introducing Badges?

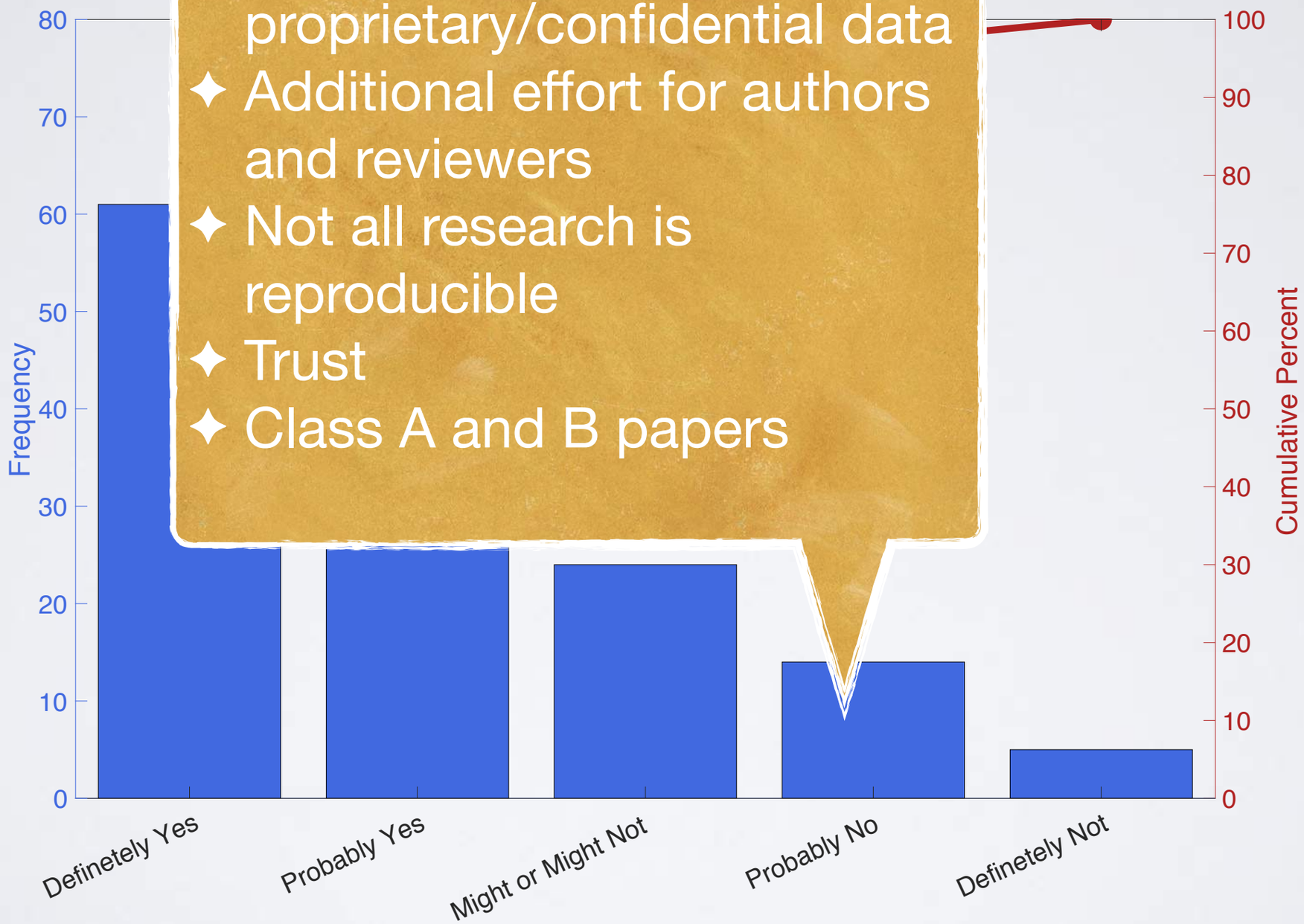


- Additional effort for authors and reviewers
- Issues for industry, proprietary/confidential data
- Class A and B papers
- Not all research is reproducible

Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

What about Introducing Badges?

- Issues for industry, proprietary/confidential data
- Additional effort for authors and reviewers
- Not all research is reproducible
- Trust
- Class A and B papers

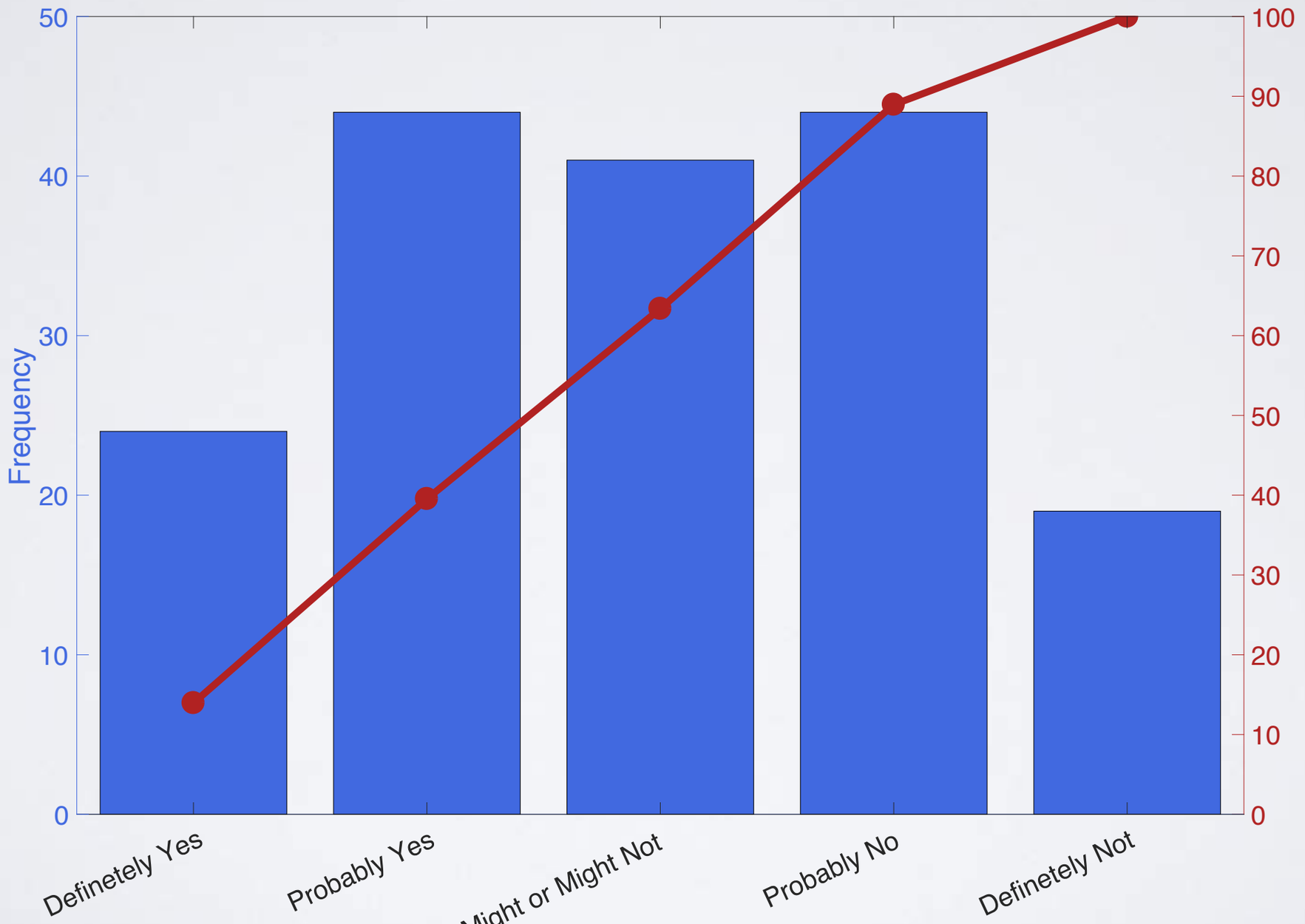


Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

What about Introducing Badges?

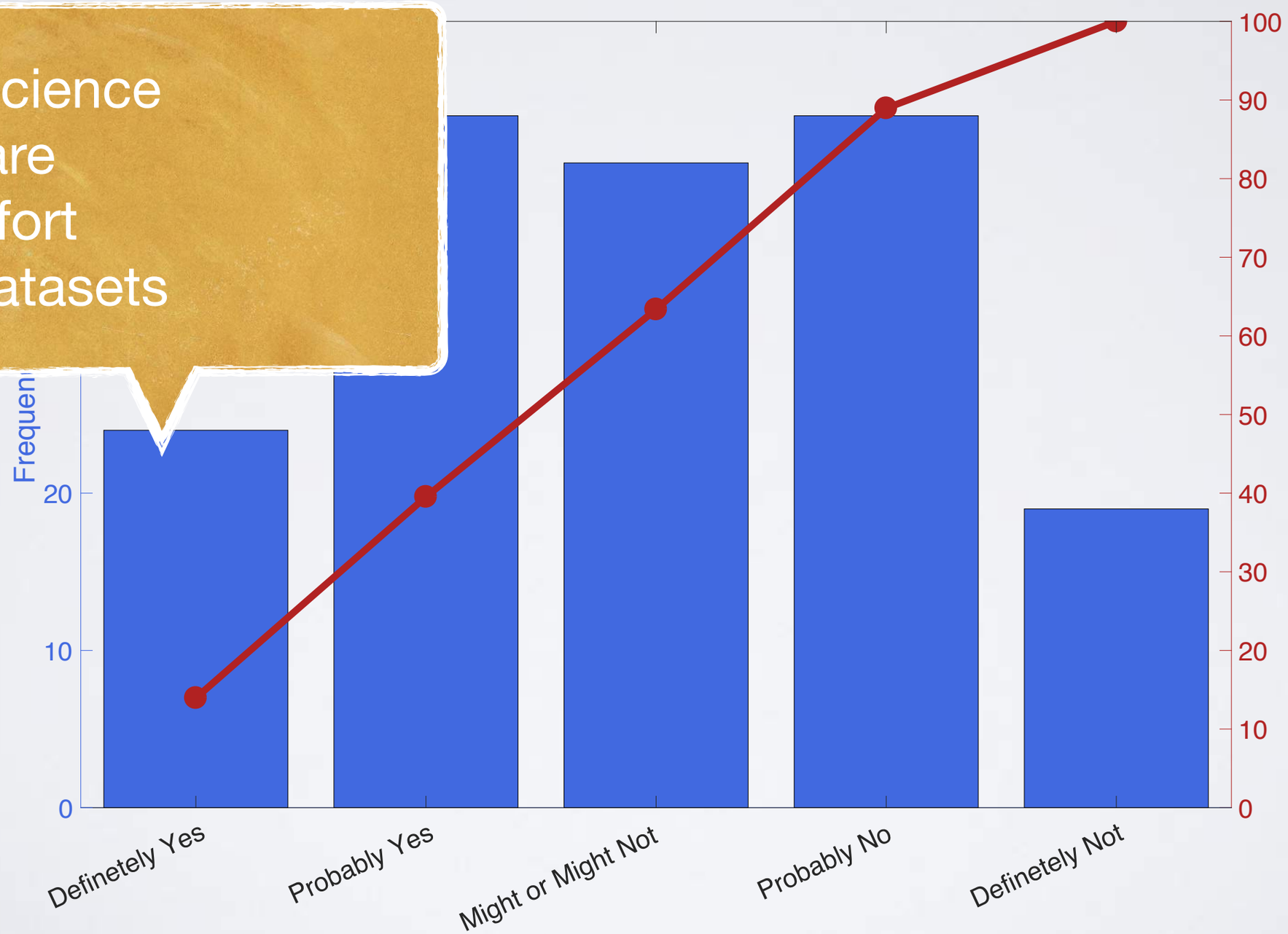
Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

Would Badges Change Your Research?



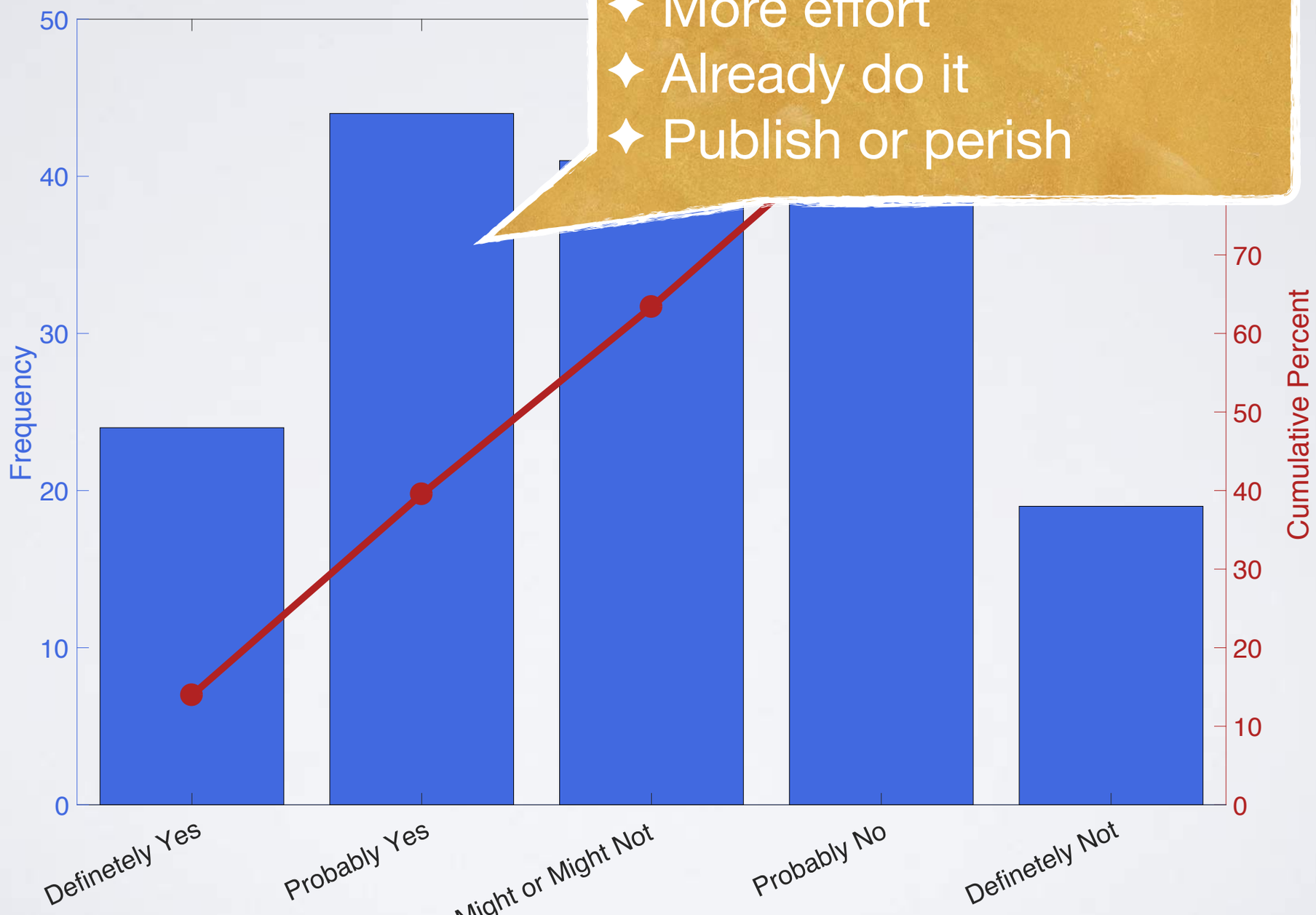
Would Badges Change Your Research?

- Better science
- More care
- More effort
- Open datasets



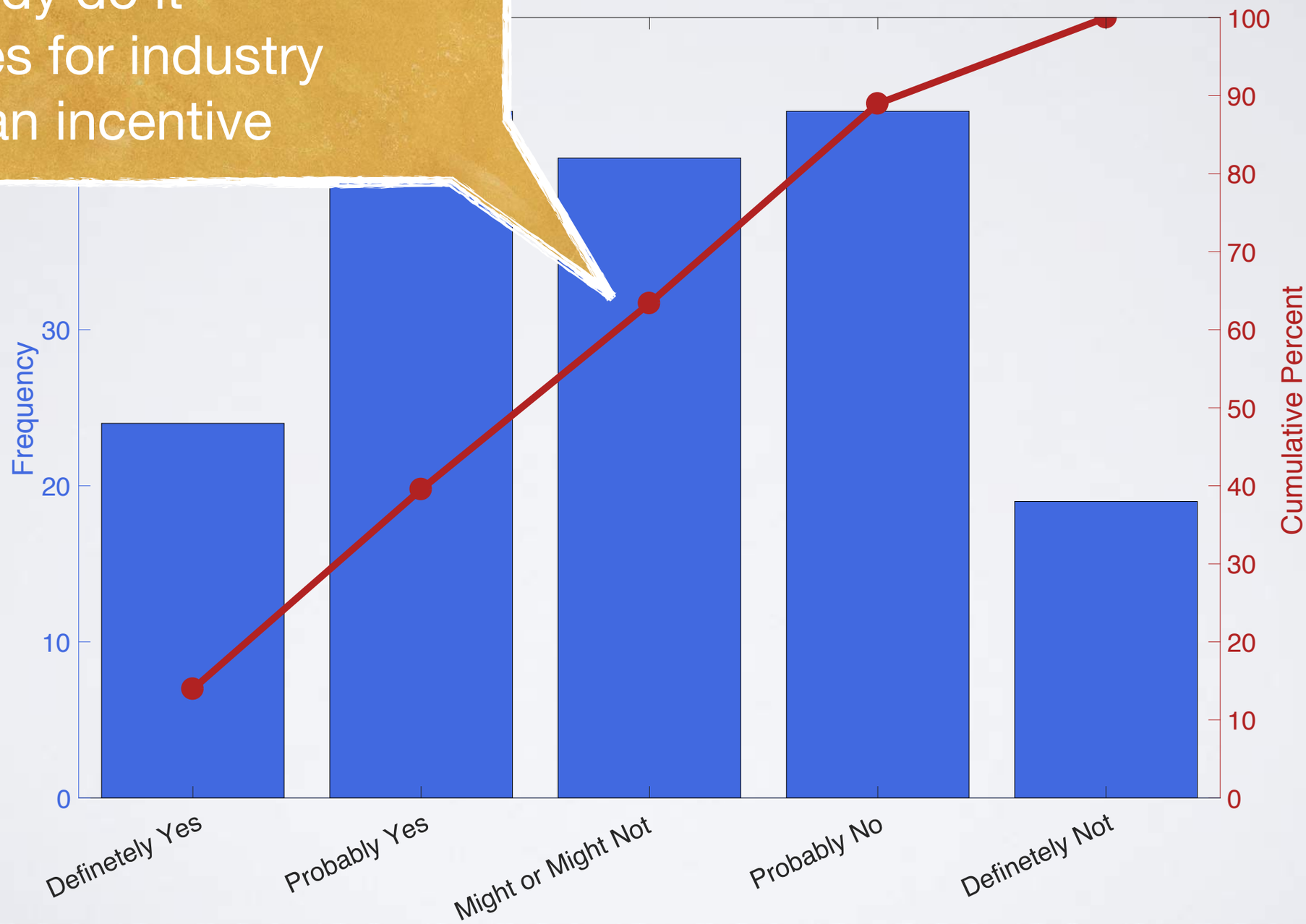
Would Badges Change Your Research?

- Better science
- More care
- More effort
- Already do it
- Publish or perish



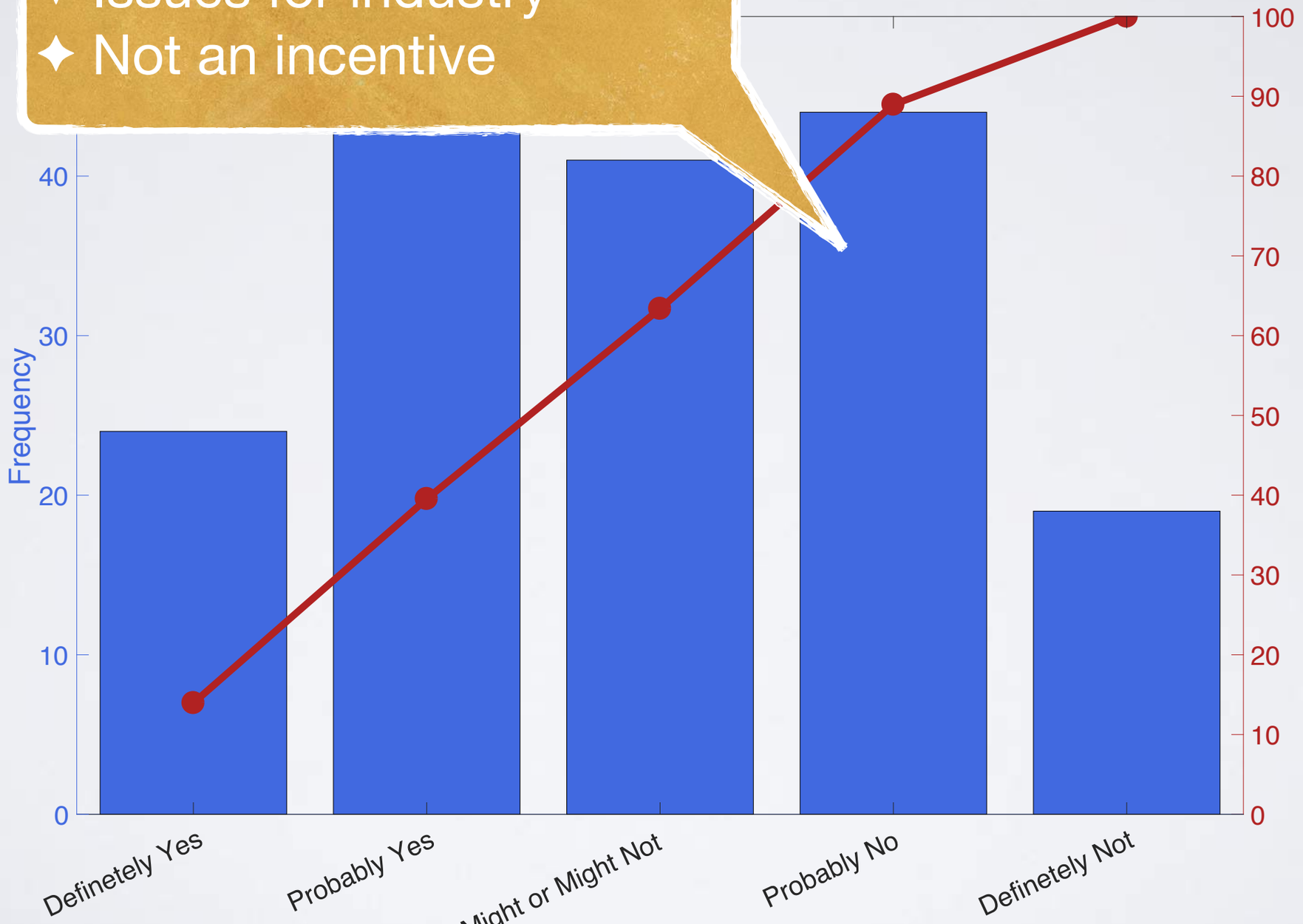
Would Badges Change Your Research?

- More care
- More effort
- Already do it
- Issues for industry
- Not an incentive

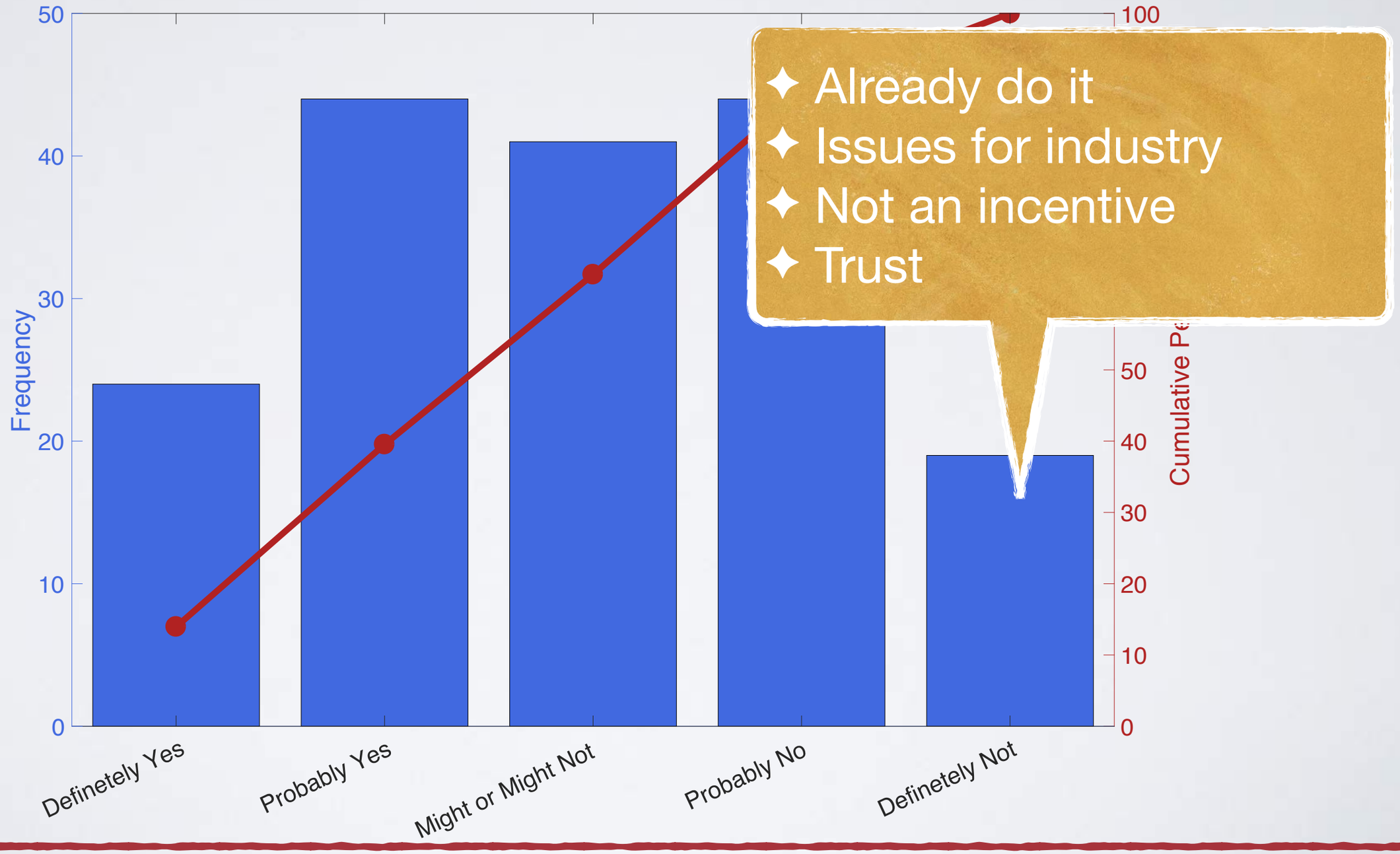


Would Badges Change Your Research?

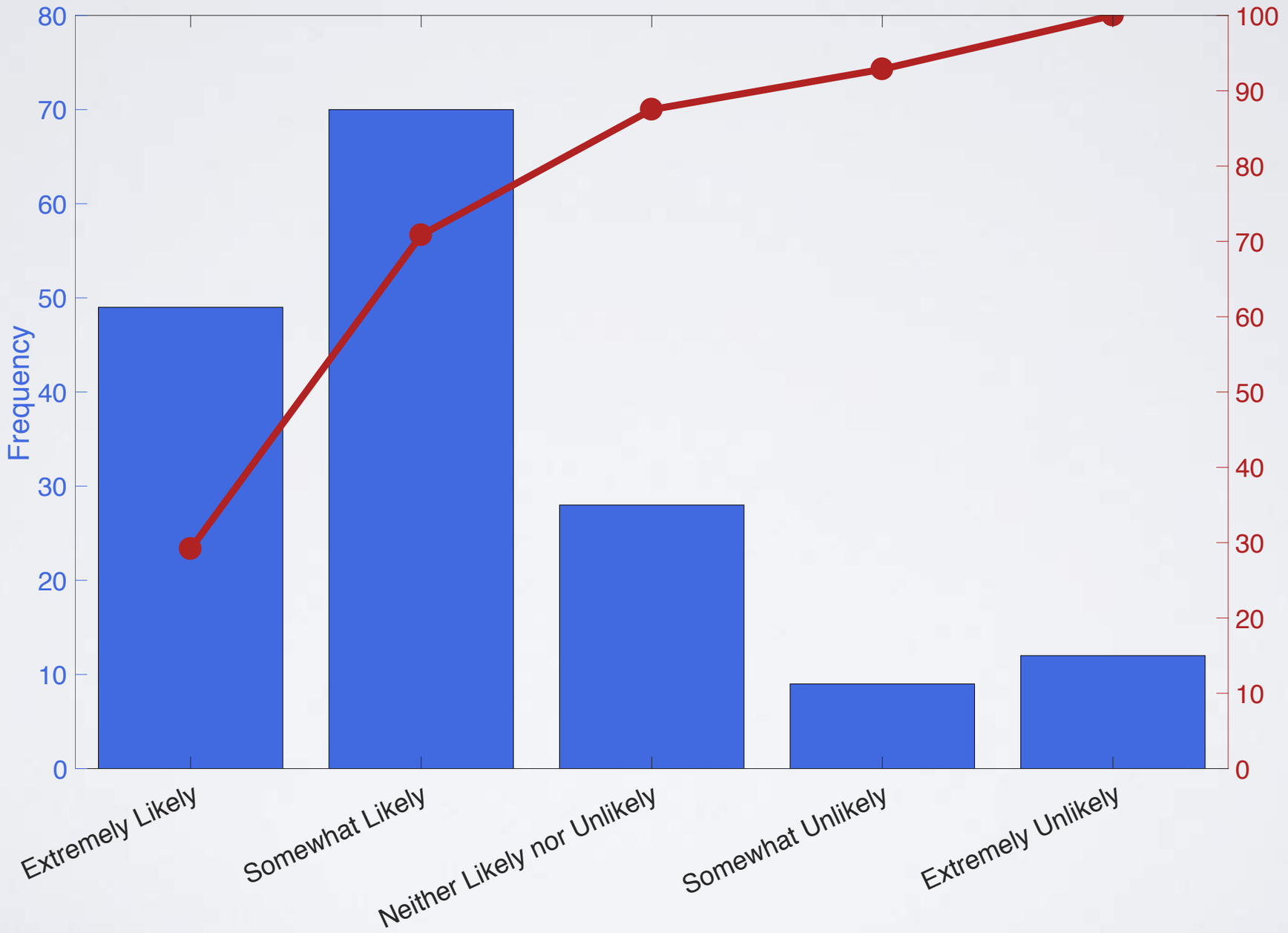
- ✦ Already do it
- ✦ Issues for industry
- ✦ Not an incentive



Would Badges Change Your Research?



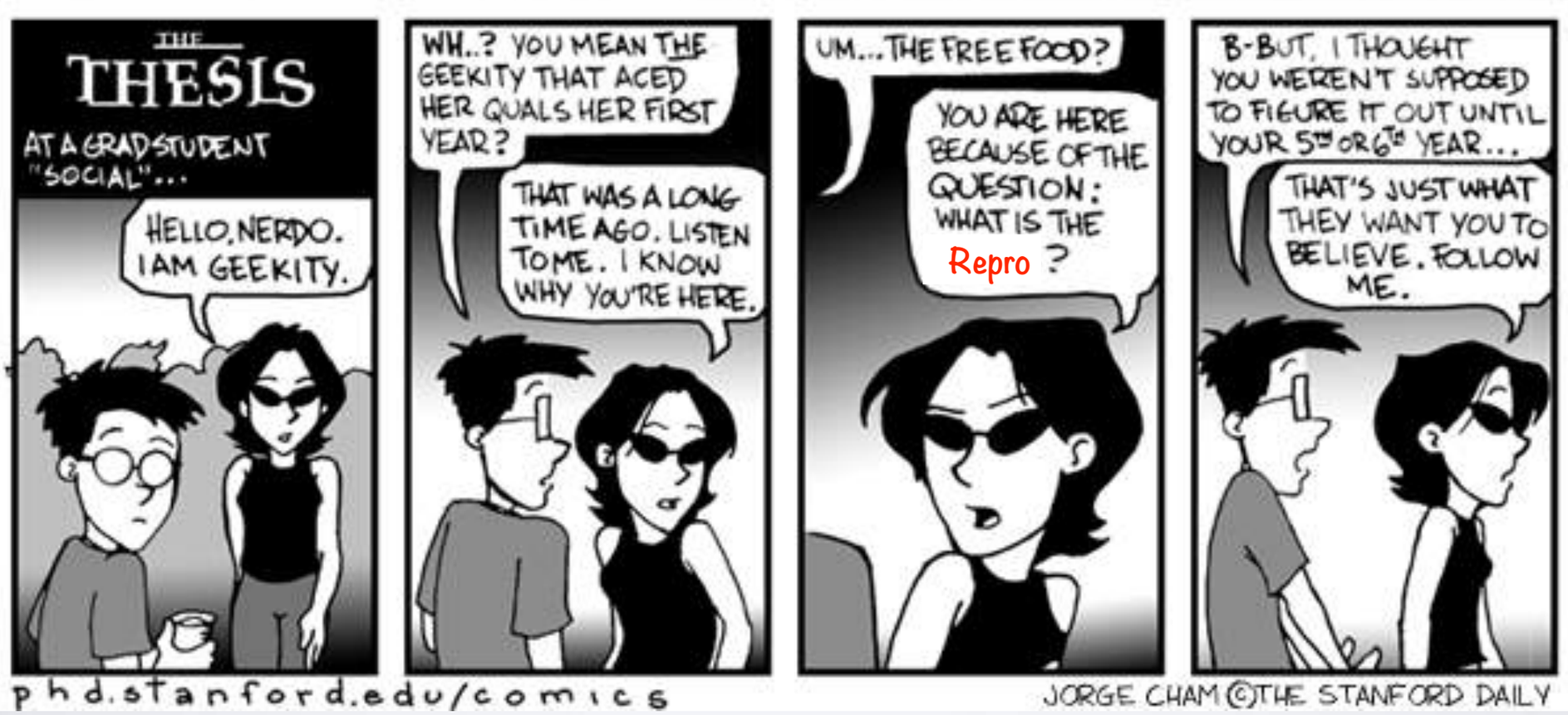
Request for Badges?



Reproducibility: Some Needs

- Shift in culture
  - more work needed to put reproducibility in action
  - what about the pressure to publish?
  - acknowledgment in careers
- Systematic but focused approach
  - how to choose what to reproduce?
- Quantitative assessment
  - when do we consider something as “reproduced”?
- Infrastructures (evaluation campaigns?)
  - lightweight tools and protocols… but they need adoption!

Questions?



http://phdcomics.com/comics/archive.php?comicid=139