800 ANNI 1222 · 2022
Università degli Studi di Padova
Dipartimento di Ingegneria dell'Informazione

# Evaluation of Quantum Computing for IR and RS

**Nicola Ferro**
@frrncl

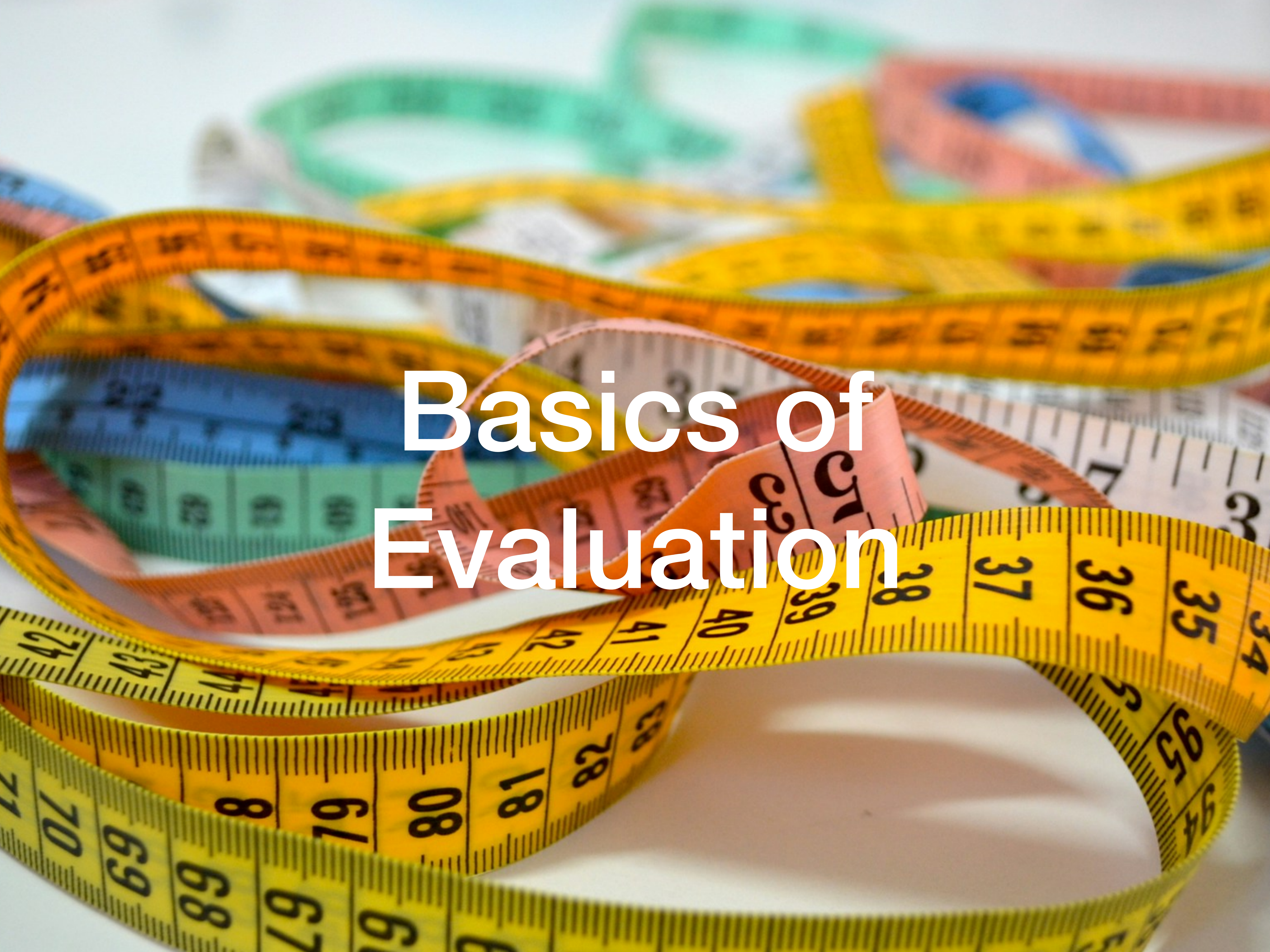
Intelligent Interactive Information Access (IIIA) Hub
Department of Information Engineering
University of Padua

SIGIR 2024 Washington, D.C.
Tutorial on Using and Evaluating Quantum Computing for Information Retrieval and Recommender Systems
SIGIR 2024, 14th July 2024, Washington, DC, USA

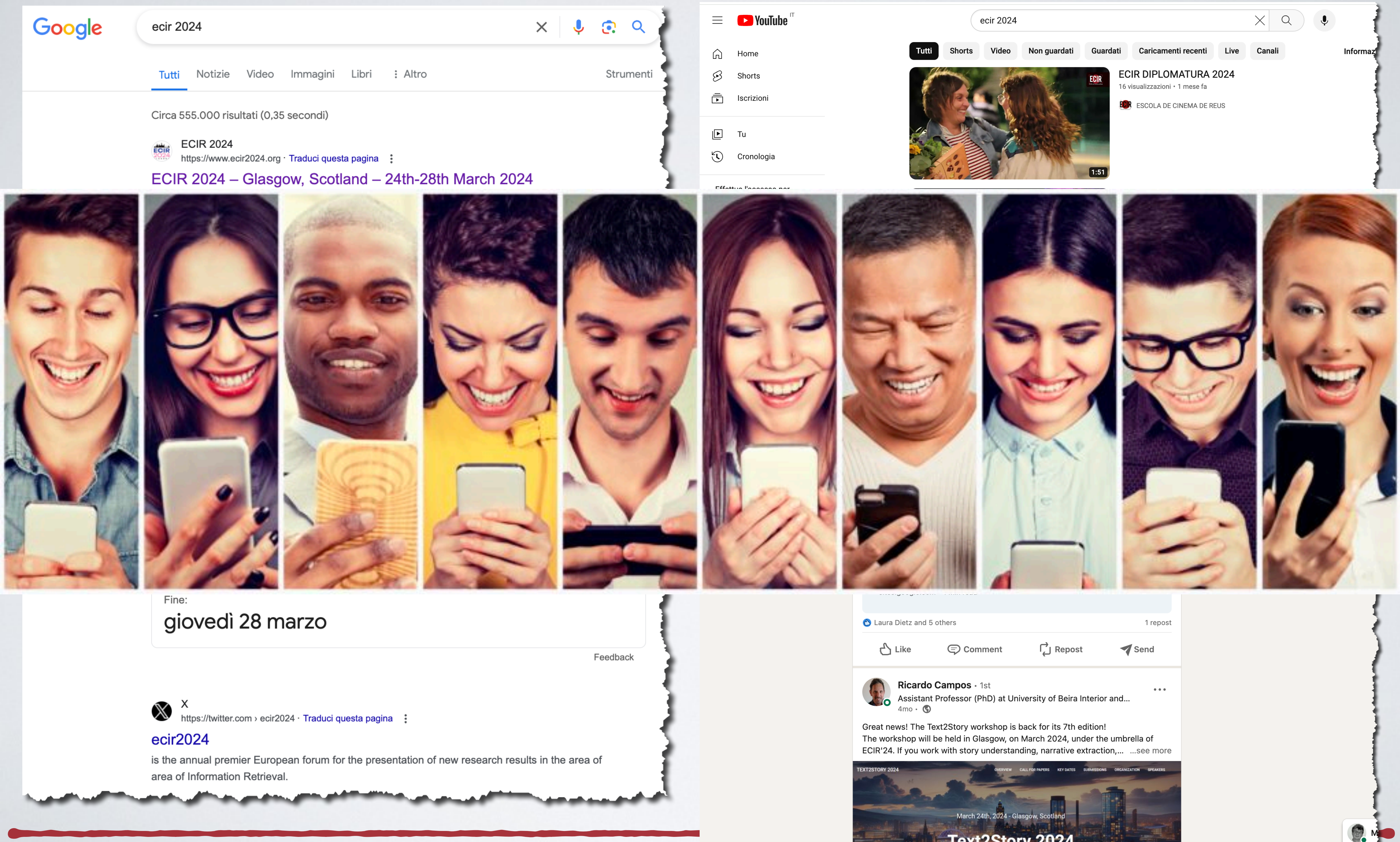
# Outline

- Basics of Evaluation
- QuantumCLEF

# Basics of Evaluation



# Our Goal



Google
ecir 2024
Tutti
Notizie
Video
Immagini
Libri
Altro
Strumenti
Circa 555.000 risultati (0,35 secondi)
ECIR 2024
https://www.ecir2024.org · Traduci questa pagina
ECIR 2024 – Glasgow, Scotland – 24th-28th March 2024
Fine:
giovedì 28 marzo
Feedback
X
https://twitter.com › ecir2024 · Traduci questa pagina
ecir2024
is the annual premier European forum for the presentation of new research results in the area of area of Information Retrieval.
YouTube
ecir 2024
Home
Shorts
Iscrizioni
Tu
Cronologia
Tutti
Shorts
Video
Non guardati
Guardati
Caricamenti recenti
Live
Canali
ECIR DIPLOMATURA 2024
16 visualizzazioni · 1 mese fa
ESCOLA DE CINEMA DE REUS
1:51
Laura Dietz and 5 others
1 repost
Like
Comment
Repost
Send
Ricardo Campos · 1st
Assistant Professor (PhD) at University of Beira Interior and...
4mo ·
Great news! The Text2Story workshop is back for its 7th edition!
The workshop will be held in Glasgow, on March 2024, under the umbrella of ECIR'24. If you work with story understanding, narrative extraction,... ...see more
TEXT2STORY 2024
March 24th, 2024 - Glasgow, Scotland
Text2Story 2024

# 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

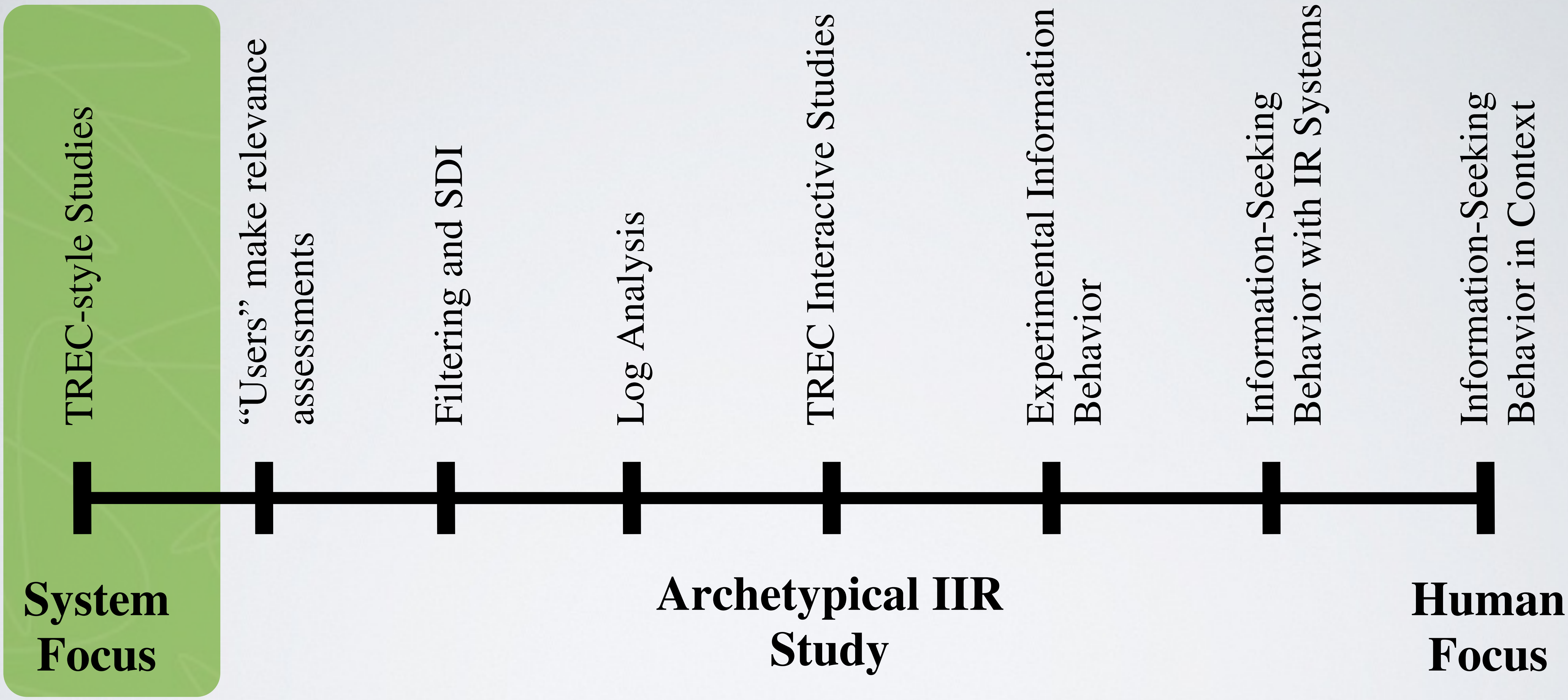
Effectiveness

# 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.

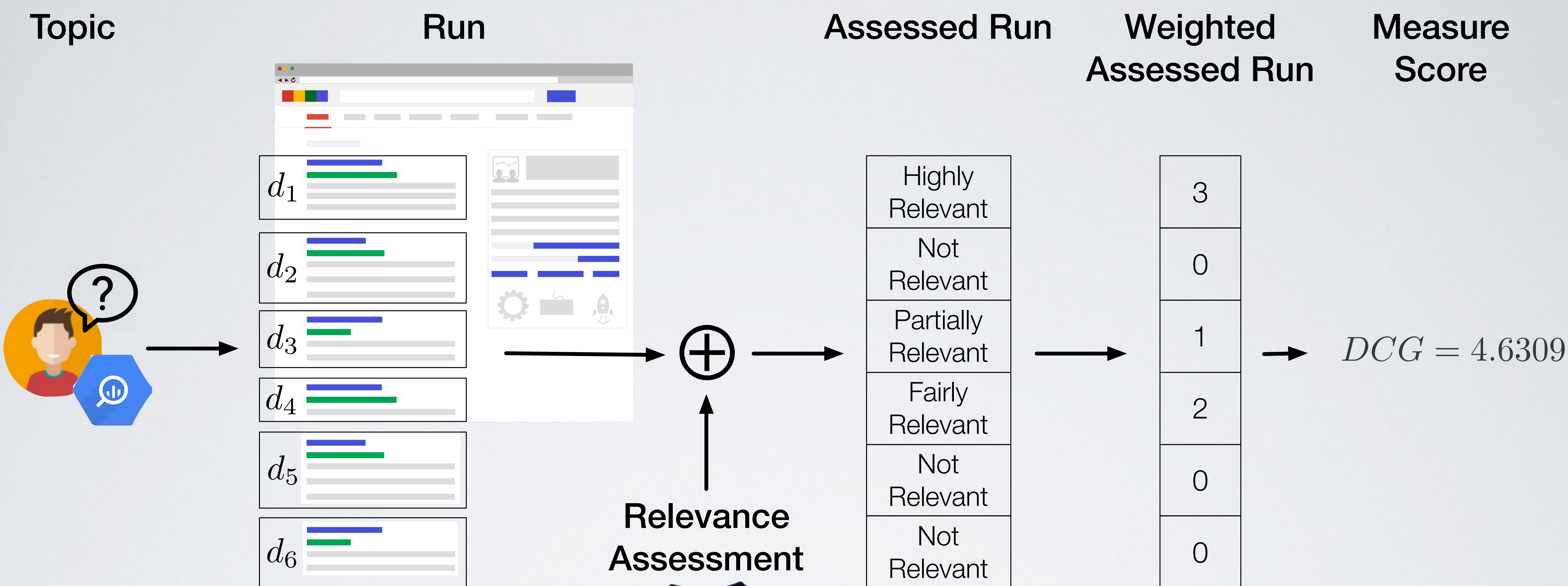
# How Does Experimental Evaluation Work

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

# Evaluation with Test Collections in a Nutshell





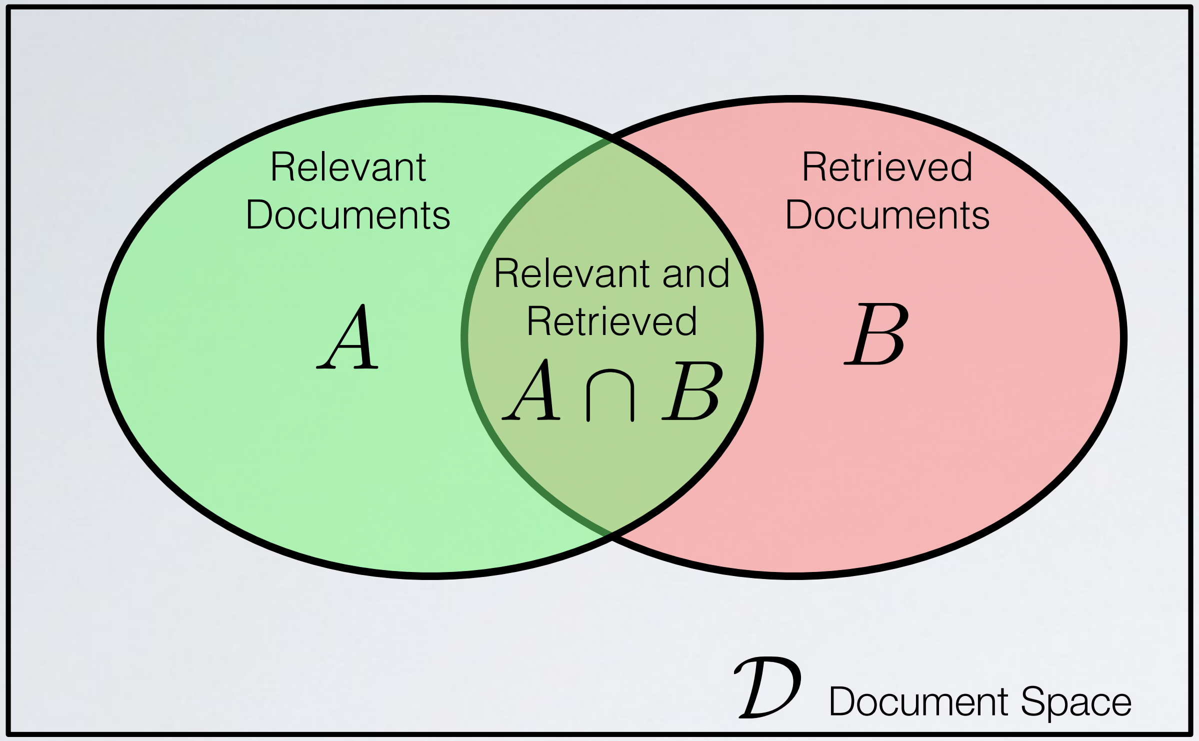
- Since we use set of topics, we can average the performance of a system over them
- We can compare two systems A and B run on the same test collection by comparing their average performance or,much better, by using statistical significance tests

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

# A Taxonomy of Evaluation Measures

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

# 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.

# Rank-based Measures: Discounted Cumulated Gain

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

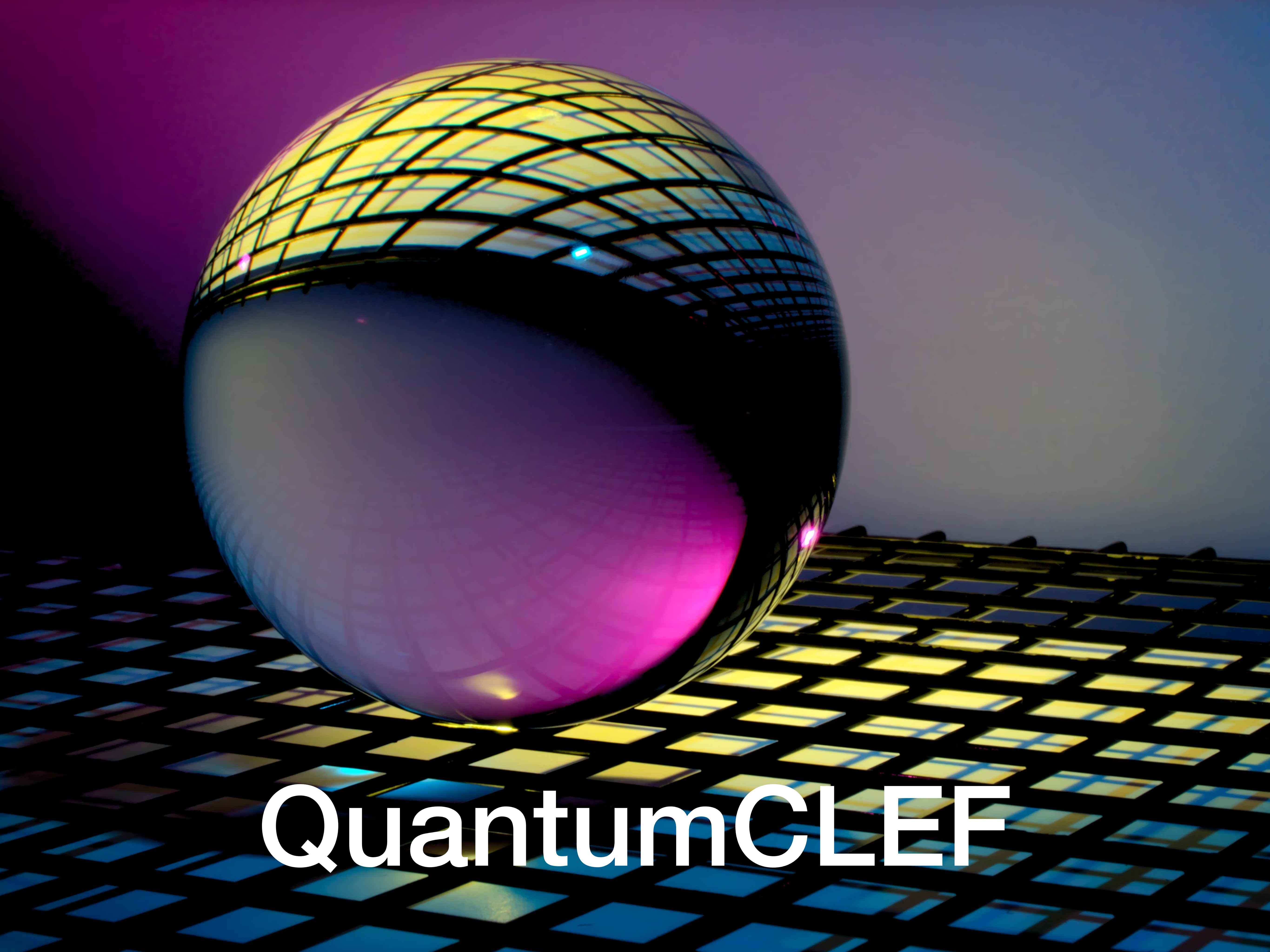
$$= \sum_{n=1}^{k} \frac{r_n}{\max(1, \log_b(n))}$$

- 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

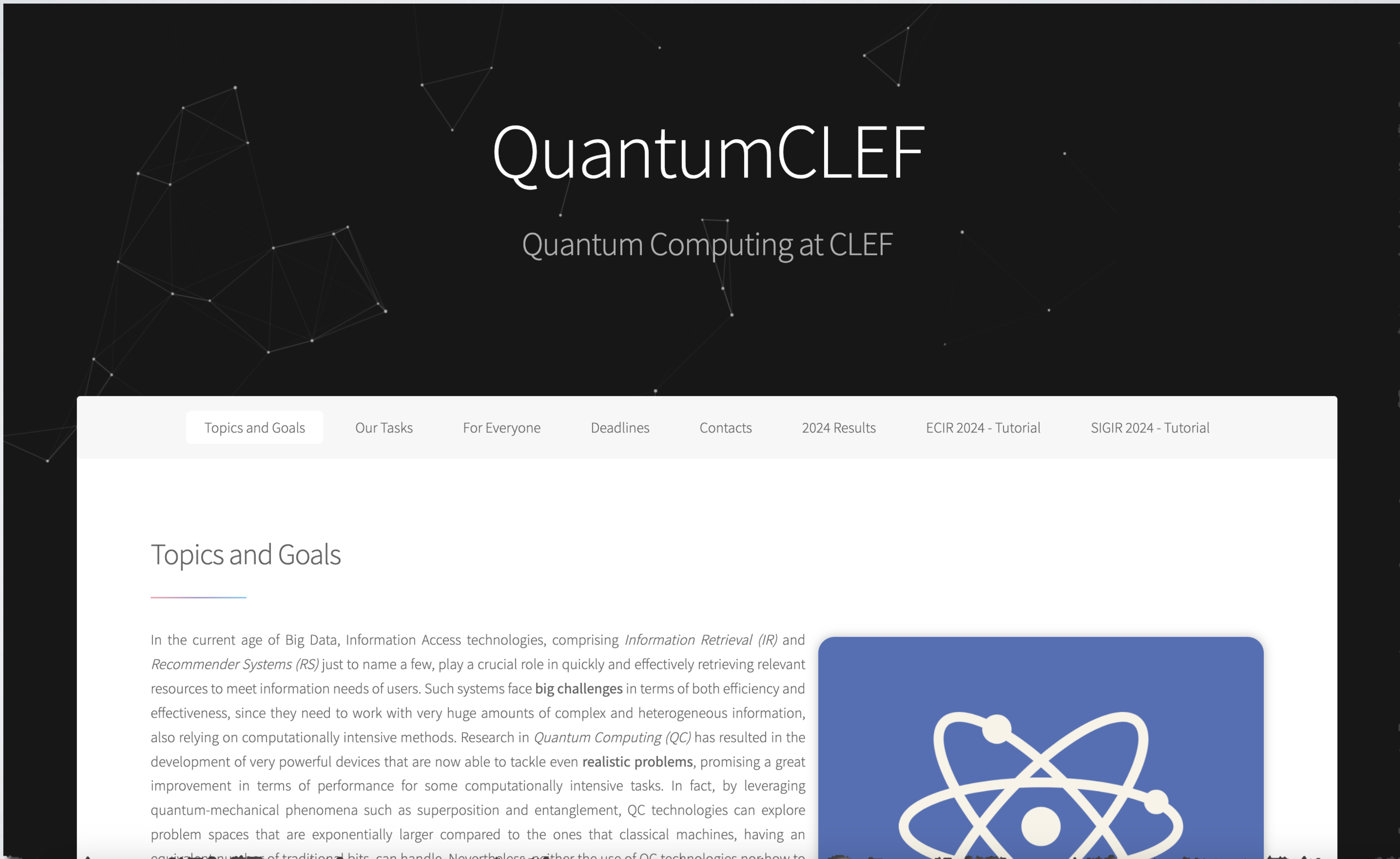
Jaana Kekäläinen

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


QuantumCLEF

# QuantumCLEF

https://qclef.dei.unipd.it/


QuantumCLEF

Quantum Computing at CLEF

Topics and Goals | Our Tasks | For Everyone | Deadlines | Contacts | 2024 Results | ECIR 2024 - Tutorial | SIGIR 2024 - Tutorial

Topics and Goals

In the current age of Big Data, Information Access technologies, comprising *Information Retrieval (IR)* and *Recommender Systems (RS)* just to name a few, play a crucial role in quickly and effectively retrieving relevant resources to meet information needs of users. Such systems face **big challenges** in terms of both efficiency and effectiveness, since they need to work with very huge amounts of complex and heterogeneous information, also relying on computationally intensive methods. Research in *Quantum Computing (QC)* has resulted in the development of very powerful devices that are now able to tackle even **realistic problems**, promising a great improvement in terms of performance for some computationally intensive tasks. In fact, by leveraging quantum-mechanical phenomena such as superposition and entanglement, QC technologies can explore problem spaces that are exponentially larger compared to the ones that classical machines, having an equivalent number of traditional bits, can handle. Nevertheless, neither the use of QC technologies nor how to

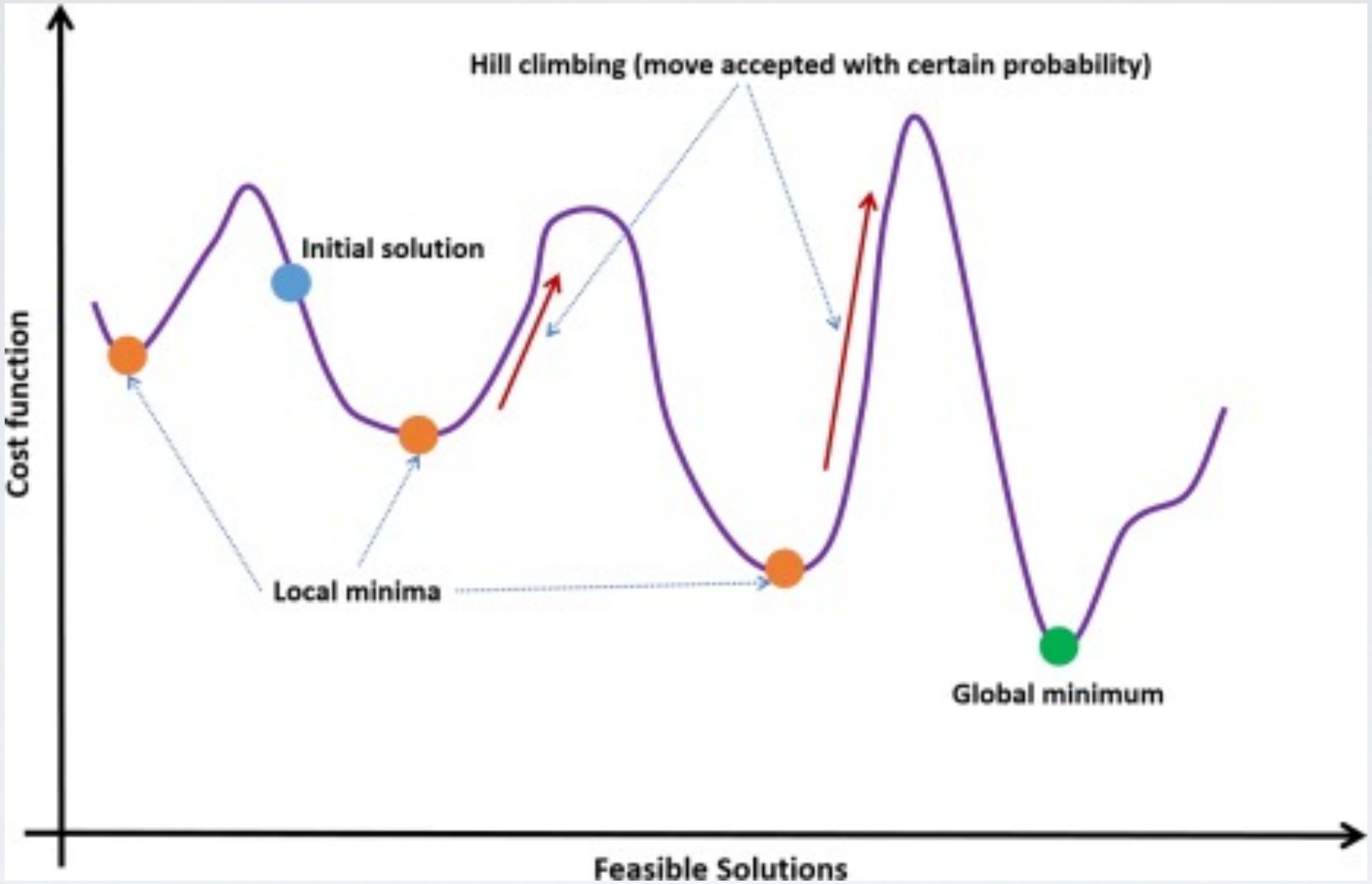
# Tasks 1 - Feature Selection

- **Quantum Feature Selection**: reducing the size of the input data to speed-up retrieval or enhancing effectiveness avoiding noisy features
- Task 1A - **Information Retrieval**
  - MQ2007 (one of the **LETOR** datasets), 46 features
  - **ISTELLA**, 220 features
  - Training of **LambdaMART** with the selected features to measure **nDCG@10**
- Task 1B - **Recommender Systems**
  - **150_ICM** (music recommendation): contains 150 features for each item
  - **500_ICM**: contains 500 features for each item
  - Training of an **Item-Based KNN** recommendation model to measure **nDCG@10**
- For each task submit runs using both Quantum Annealing (QA) and Simulated Annealing (SA)

# Task 2 - Clustering in IR

- Task: obtain a list of **representative centroids** of the given dataset of embeddings (10, 25 and 50 vectors that represent the final centroids)
- **ANTIQUE** dataset in which each sentence taken from Yahoo is turned into an embedding using a transformer.
  - 6,500 sentences for training
  - 2,200 sentences for testing
- Measures
  - the **Davies-Bouldin Index** is used to measure the overall cluster quality. The index is improved (lowered) by increased separation between clusters and decreased variation within clusters.
  - **nDCG@10** is used to measure the overall retrieval quality based on a set of 50 queries.
  - Each query is transformed into its corresponding embedding, then the Cosine Similarity is used to get the closest centroid and its corresponding cluster of documents, finally all the documents belonging to that cluster are retrieved and ranked using the Cosine Similarity between the documents and the query
- Submit runs using both Quantum Annealing (QA) and Simulated Annealing (SA)

# Simulated Annealing




- For both tasks (Feature Selection and Clustering), the QA approach will be compared against a SA approach, using the same QUBO formulation

# Effectiveness Challenges

Quantum Computing for Information Access - Feature Selection.ipynb

File Edit View Insert Runtime Tools Help Last edited on November 20

Comment Share

+ Code + Text Connect

Now we will use Quantum Annealing for real. This works as follows:

1. We need to have an API KEY which is required to have access to D-Wave's infrastructure.
2. Once we create our sampler and submit our problem, we will send it through internet. The problem will reach the D-Wave's infrastructure and will be enqued if there are currently other problems running.
3. Once it is our turn, our problem will be solved and the solution will be sent back to us.

```
%%time

from google.colab import userdata
token=userdata.get('API_TOKEN')

sampler_QPU = DWaveCliqueSampler(token=token)

response_QPU_4 = sampler_QPU.sample(kbqm,
                    label='Example - MI Feature Selection',
                    num_reads=num_reads)
```

```
CPU times: user 3.56 s, sys: 50.9 ms, total: 3.61 s
Wall time: 8.29 s
```

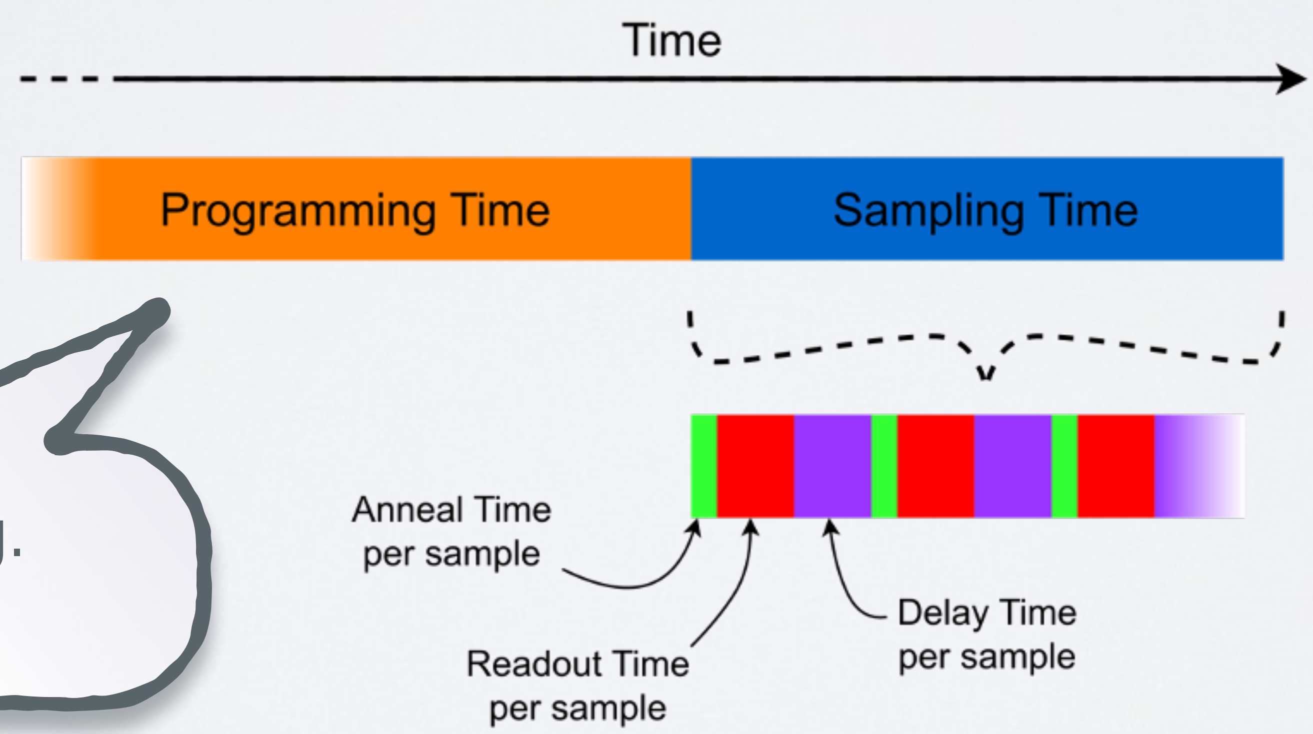
```
print_response_data(response_QPU_4.aggregate())
```

```
-----------------------------------------------------------------
          Set 0          Set 1    Energy          Count
-----------------------------------------------------------------
['age', 'alone', 'embarked port C', 'embarked port Q', 'embarked port S', 'fare', 'master', 'miss', 'mrs', 'rare']['cabin', 'mr', 'pclass'
['age', 'alone', 'cabin', 'embarked port C', 'embarked port Q', 'embarked port S', 'fare', 'master', 'miss', 'mrs']['mr', 'pclass', 'rare'
['age', 'alone', 'cabin', 'embarked port C', 'embarked port Q', 'embarked port S', 'fare', 'master', 'miss', 'rare']['mr', 'mrs', 'pclass'
['age', 'alone', 'cabin', 'embarked port C', 'embarked port Q', 'embarked port S', 'fare', 'master', 'mr', 'rare']['miss', 'mrs', 'pclass'
['age', 'alone', 'embarked port Q', 'embarked port S', 'fare', 'master', 'miss', 'mrs', 'pclass', 'rare']['cabin', 'embarked port C', 'mr'
['alone', 'embarked port C', 'embarked port Q', 'embarked port S', 'fare', 'master', 'miss', 'mrs', 'pclass', 'rare']['age', 'cabin', 'mr'
['age', 'alone', 'embarked port C', 'embarked port Q', 'embarked port S', 'fare', 'miss', 'mrs', 'pclass', 'rare']['cabin', 'master', 'mr'
['age', 'alone', 'embarked port C', 'embarked port Q', 'fare', 'master', 'miss', 'mrs', 'pclass', 'rare']['cabin', 'embarked port S', 'mr'
['age', 'alone', 'embarked port C', 'embarked port Q', 'embarked port S', 'fare', 'master', 'mrs', 'pclass', 'rare']['cabin', 'miss', 'mr'
['age', 'embarked port C', 'embarked port Q', 'embarked port S', 'fare', 'master', 'miss', 'mrs', 'pclass', 'rare']['alone', 'cabin', 'mr'
['age', 'alone', 'embarked port C', 'embarked port Q', 'embarked port S', 'fare', 'master', 'mr', 'pclass', 'rare']['cabin', 'miss', 'mrs'
['age', 'alone', 'embarked port C', 'embarked port Q', 'embarked port S', 'fare', 'master', 'miss', 'rare', 'sex']['cabin', 'mr', 'mrs', '
```

Can we leverage the stochastic nature of the QA process to perform a deeper statistical analysis?

# Efficiency Challenges

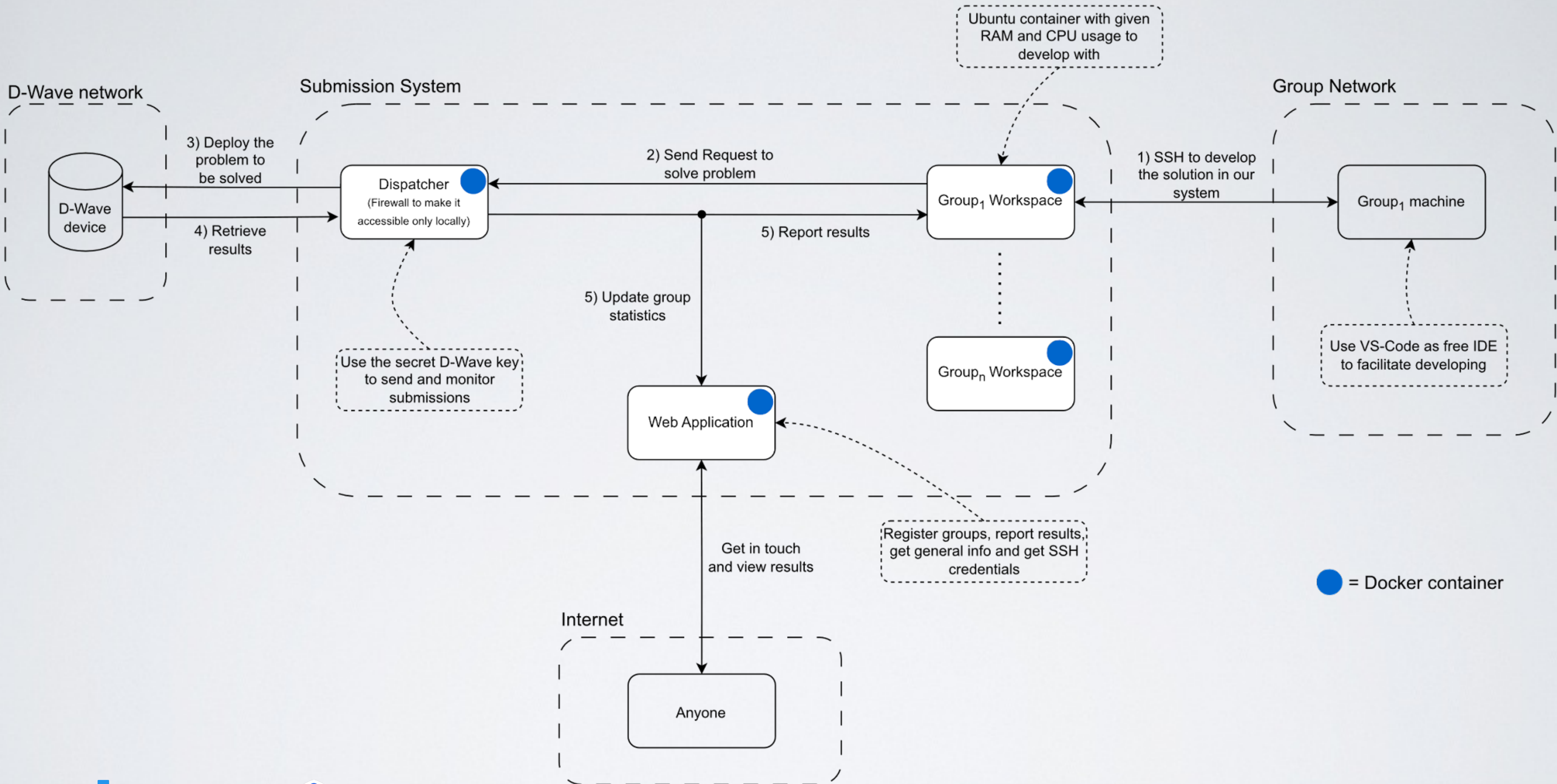
- There is not a standard way to measure the efficiency of Quantum Annealers
- There are several steps in the Annealing phase, each requiring a different amount of time based also on the used quantum annealer




Minor embedding.
Can we cache?

Resch Resch, S. and Karpuzcu, U. R. (2022). Benchmarking Quantum Computers and the Impact of Quantum Noise. *ACM Computing Surveys (CSUR)*, 54(7):142:1–142:35.

# QuantumCLEF Infrastructure




docker kubernetes

# QuantumCLEF Participation

- 26 groups registered for participating
  - **7 groups** submitted runs
- Submissions
  - Simulated Annealing (SA): 32
  - Quantum Annealing/Hybrid (QA): 34
- Break-down of submitted runs
  - Task 1A - Feature Selection for IR: 5 groups; 20 runs QA, 19 runs SA
  - Task 1B - Feature Selection for RecSys: 2 groups; 7 runs QA, 8 runs SA
  - Task 2 - Clustering for IR: 1 group; 7 runs QA, 5 runs SA
- Computing time
  - Simulated Annealing: ~9 hours (1.2 cores of AMD EPYC 3,6 GHz, 10 GByte RAM)
  - Quantum Annealing/Hybrid: ~5 minutes

# Results are Available Online

https://qclef.dei.unipd.it/

# QuantumCLEF 2024 - Results

Task 1A | Task 1B | Task 2 | Home

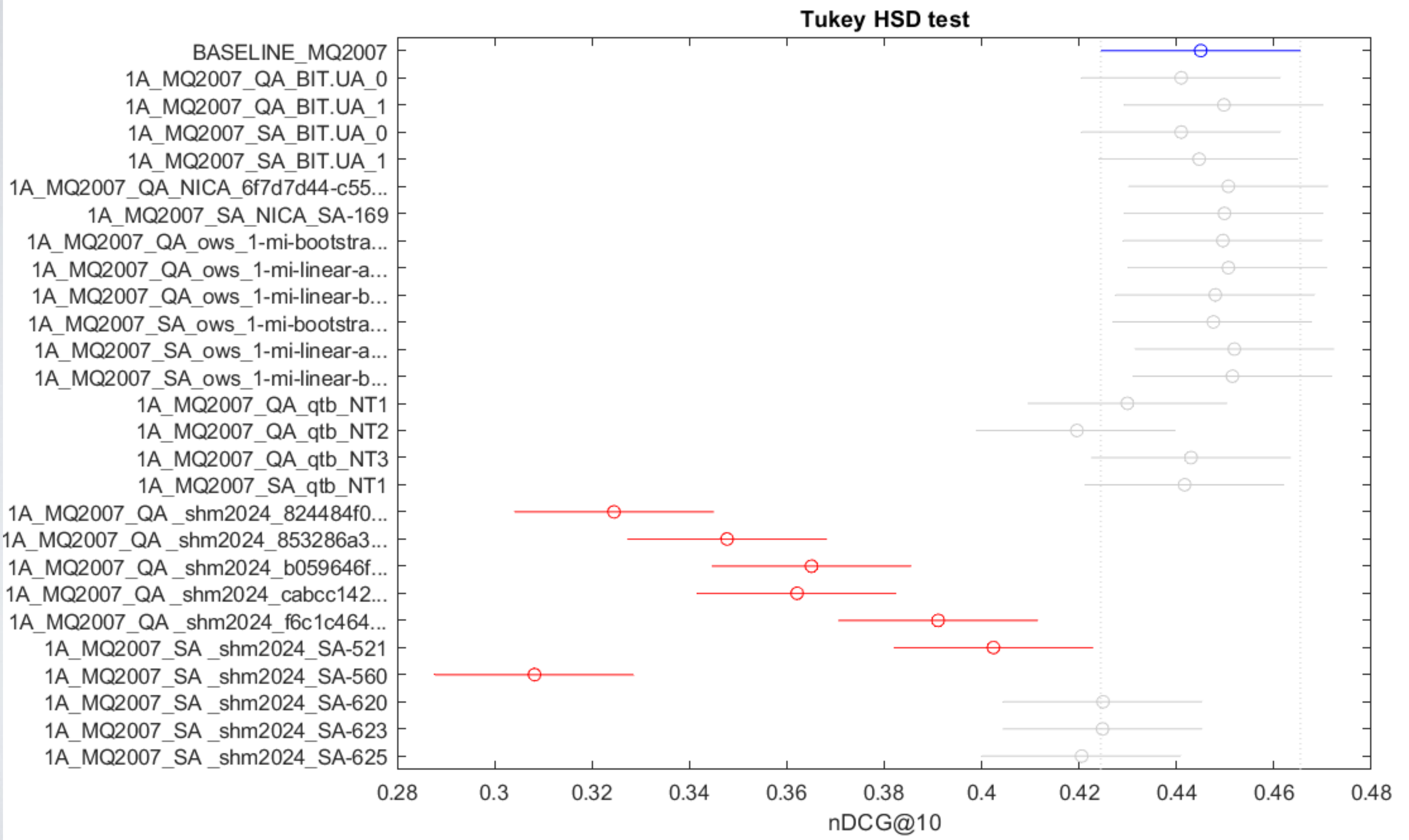
## Task 1A

The tables report the results for task 1A considering the results achieved by each team. The Annealing time measures the execution time of the approach. In the case of Quantum Annealing this consists in the programming time, sampling time and post-processing time. We will conduct a deeper analysis comparing the SA vs QA/Hybrid approaches used and we will run our baseline according to the number of features chosen by the participants' approaches to consider a more fair comparison.

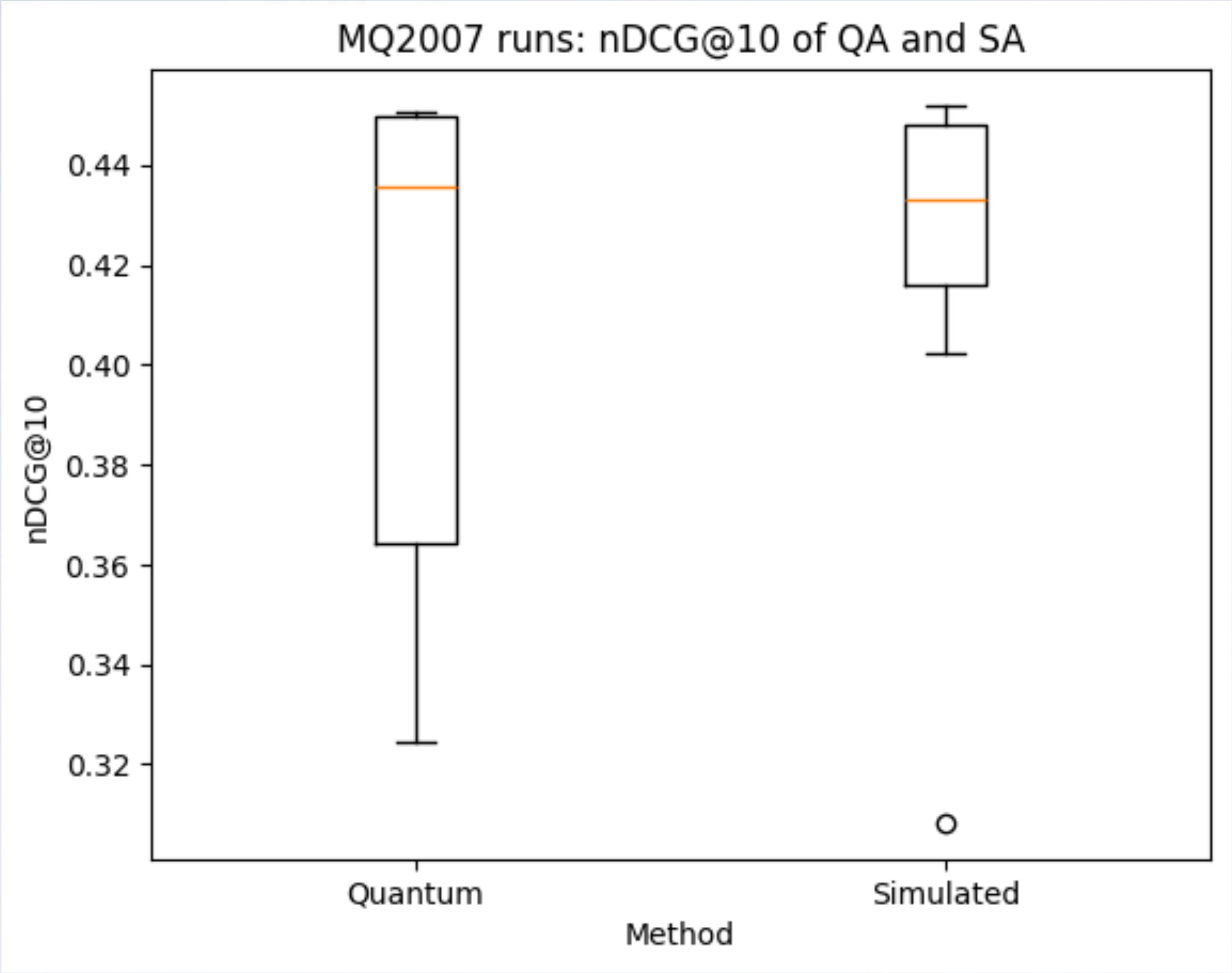
### MQ2007

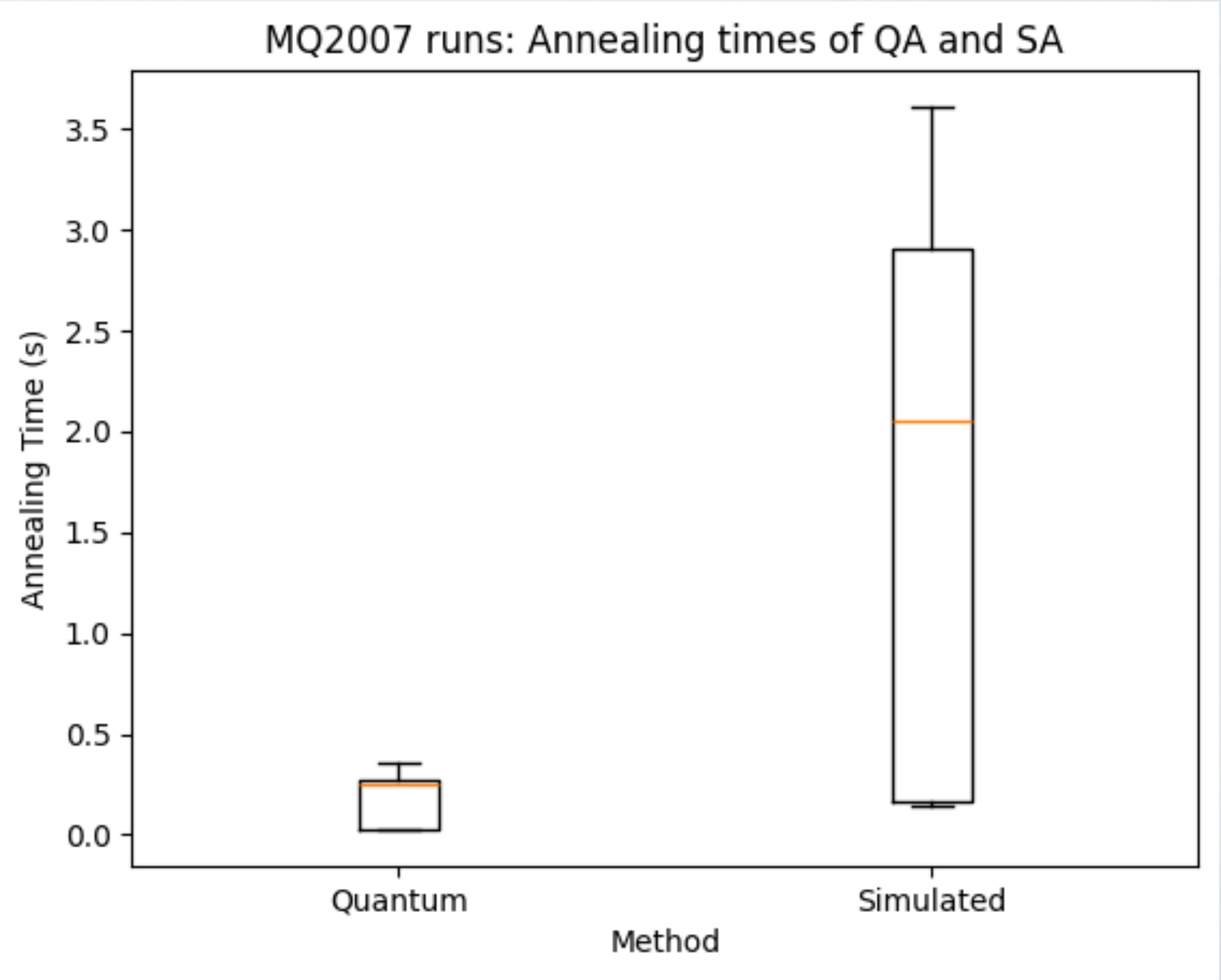
| Team | Submission id | ndcg@10 | Annealing Time (us) | Type | n° features |
|---|---|---|---|---|---|
| BIT.UA | 1A_MQ2007_QA_BIT.UA_0 | 0.441 | 273682 | Q | 18 |
| BIT.UA | 1A_MQ2007_QA_BIT.UA_1 | 0.4497 | 269805 | Q | 20 |
| BIT.UA | 1A_MQ2007_SA_BIT.UA_0 | 0.441 | 1351082 | S | 16 |
| BIT.UA | 1A_MQ2007_SA_BIT.UA_1 | 0.4446 | 3606880 | S | 18 |
| NICA | 1A_MQ2007_QA_NICA_6f7d7d44-c559- | 0.4506 | 274119 | Q | 17 |

# Task 1A IR - LETOR (46 features)

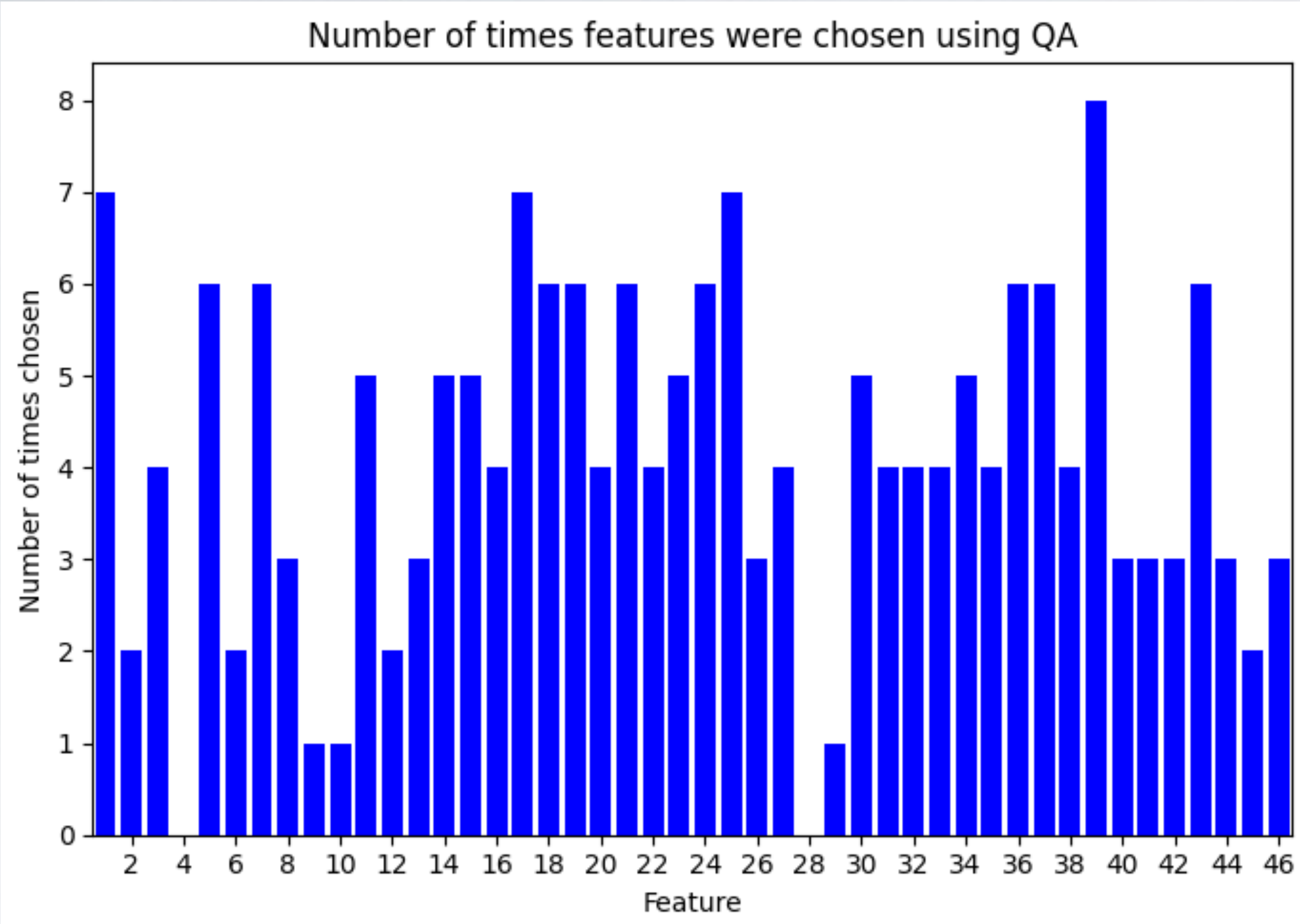


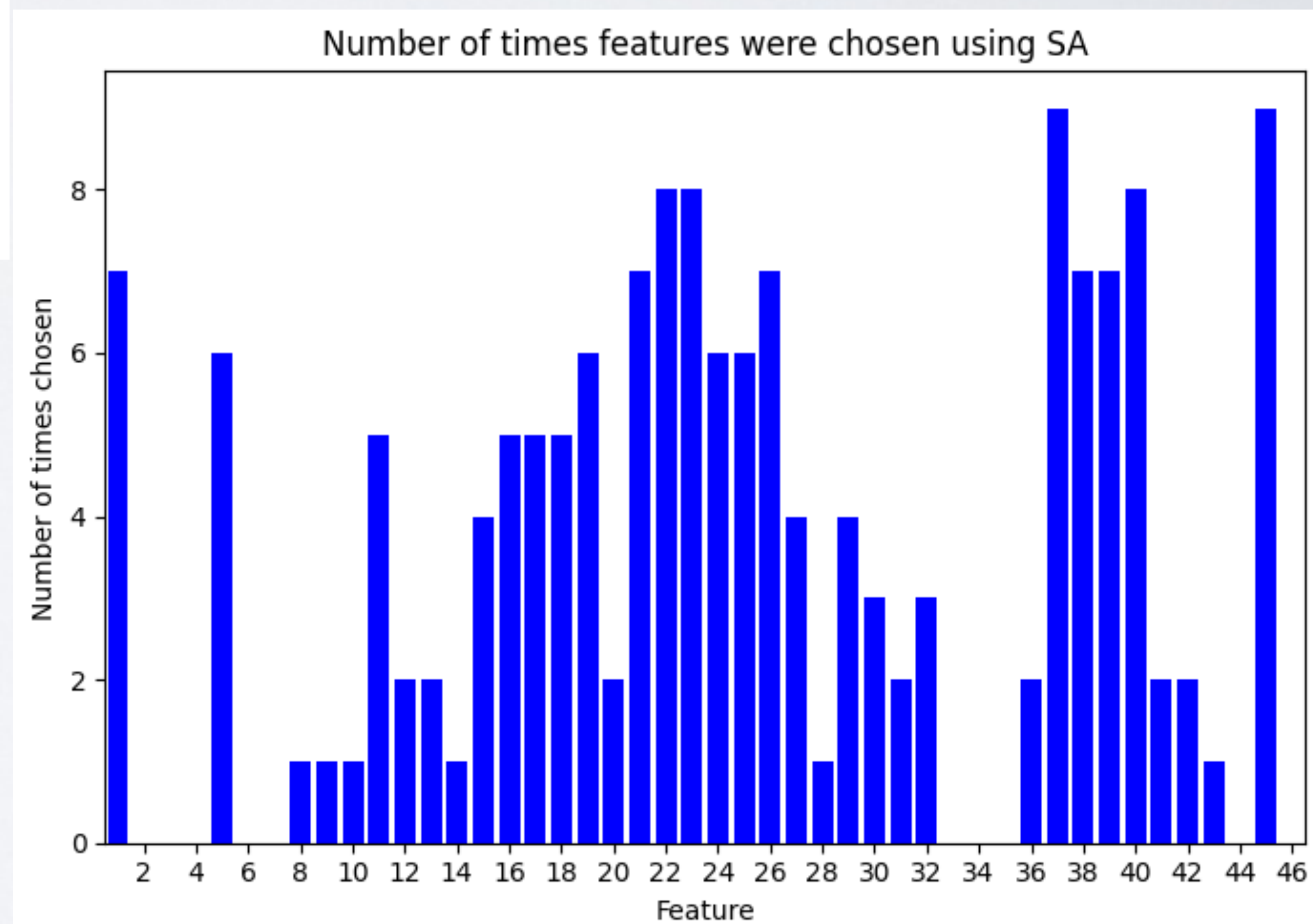
# Task 1A IR - LETOR (46 features)



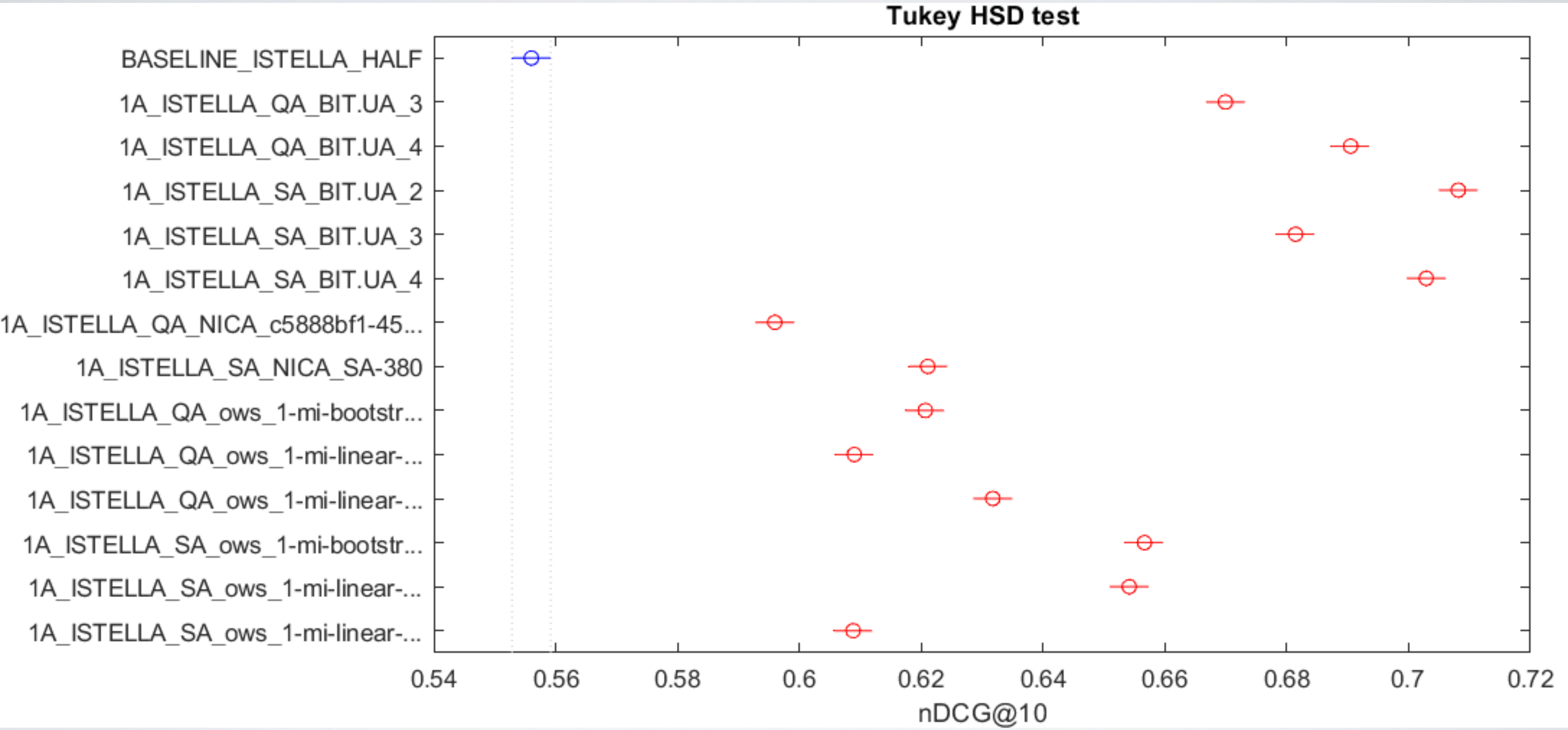



# Task 1A IR - LETOR (46 features)

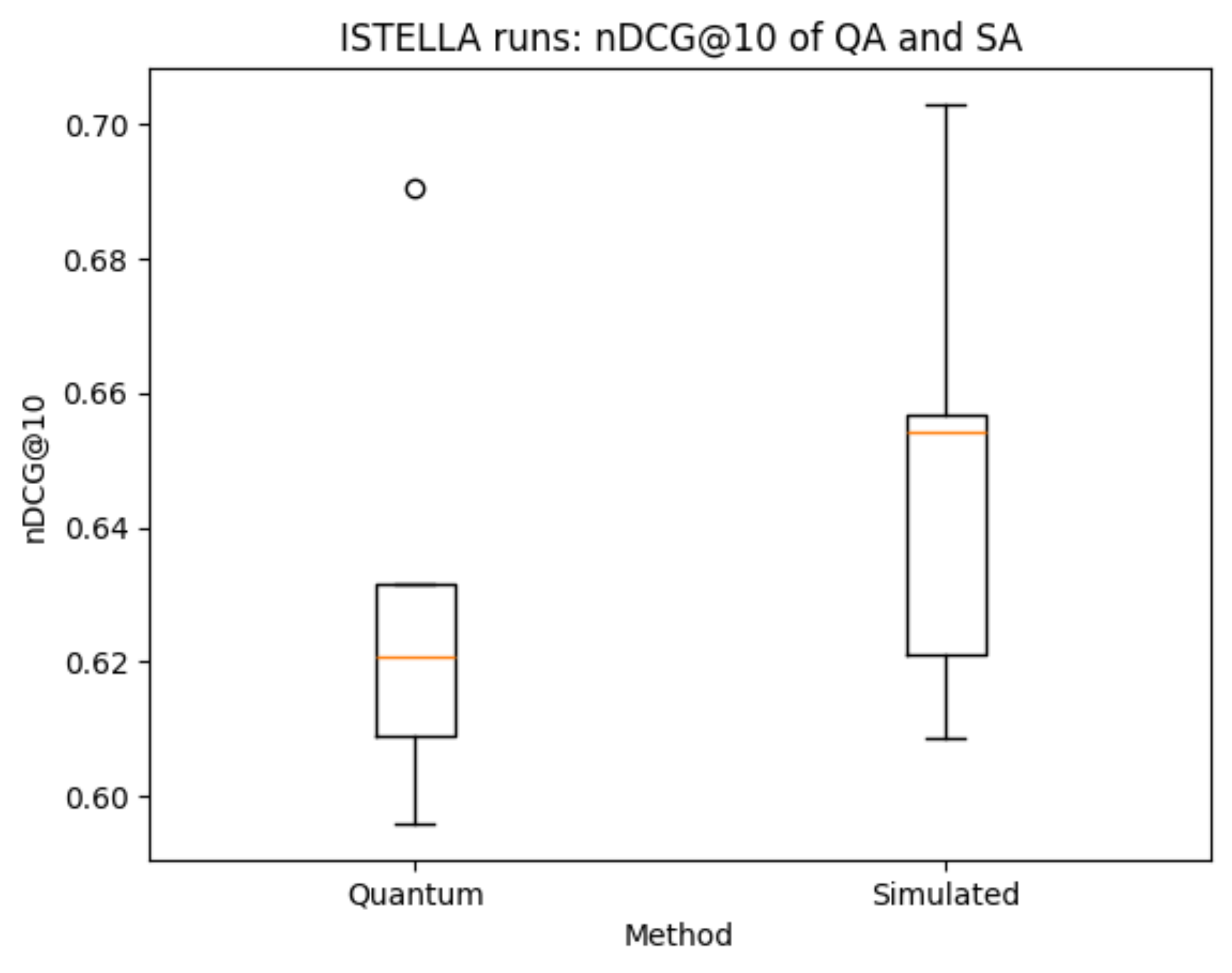


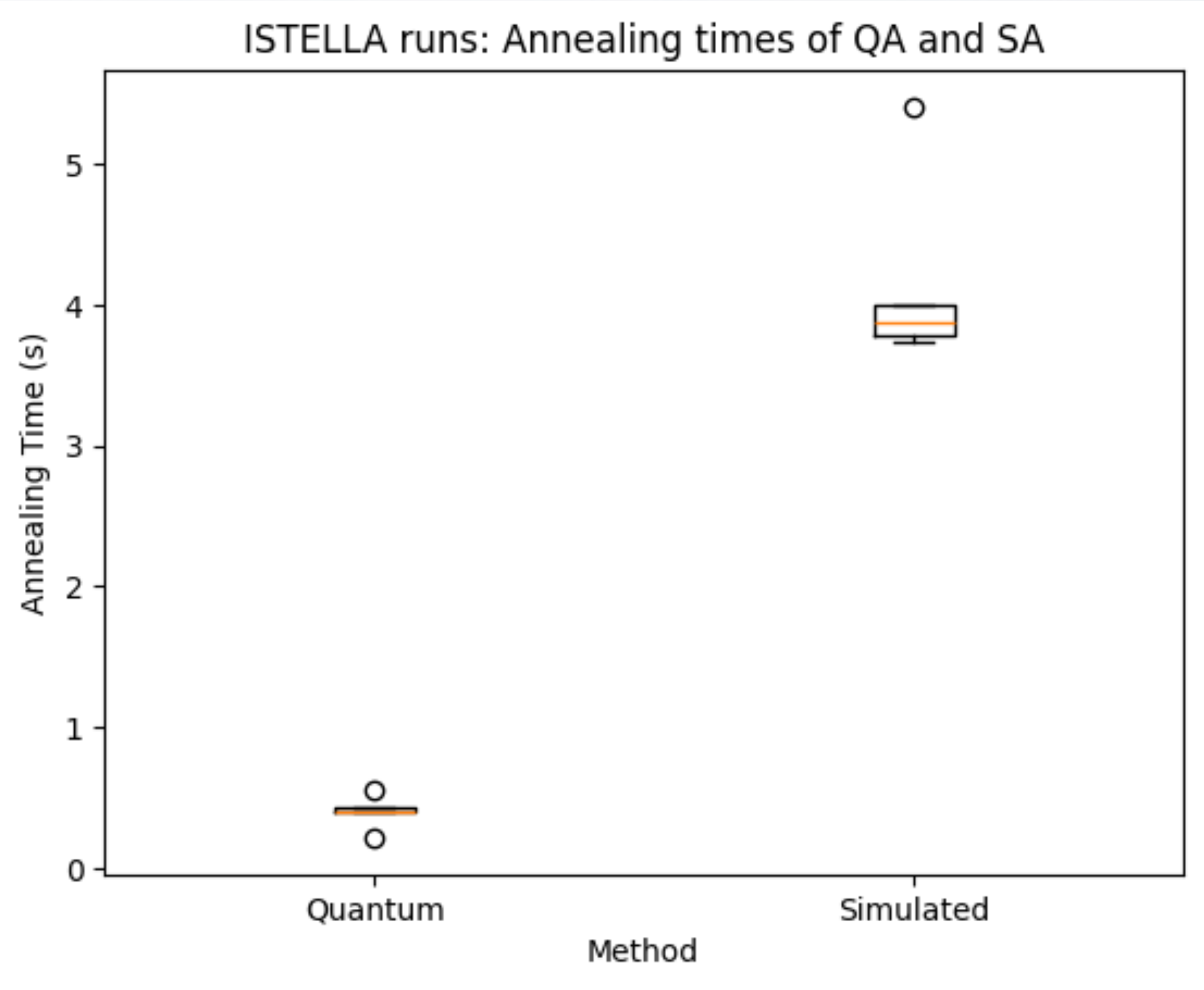



# Task 1A IR - iSTELLA (220 features)

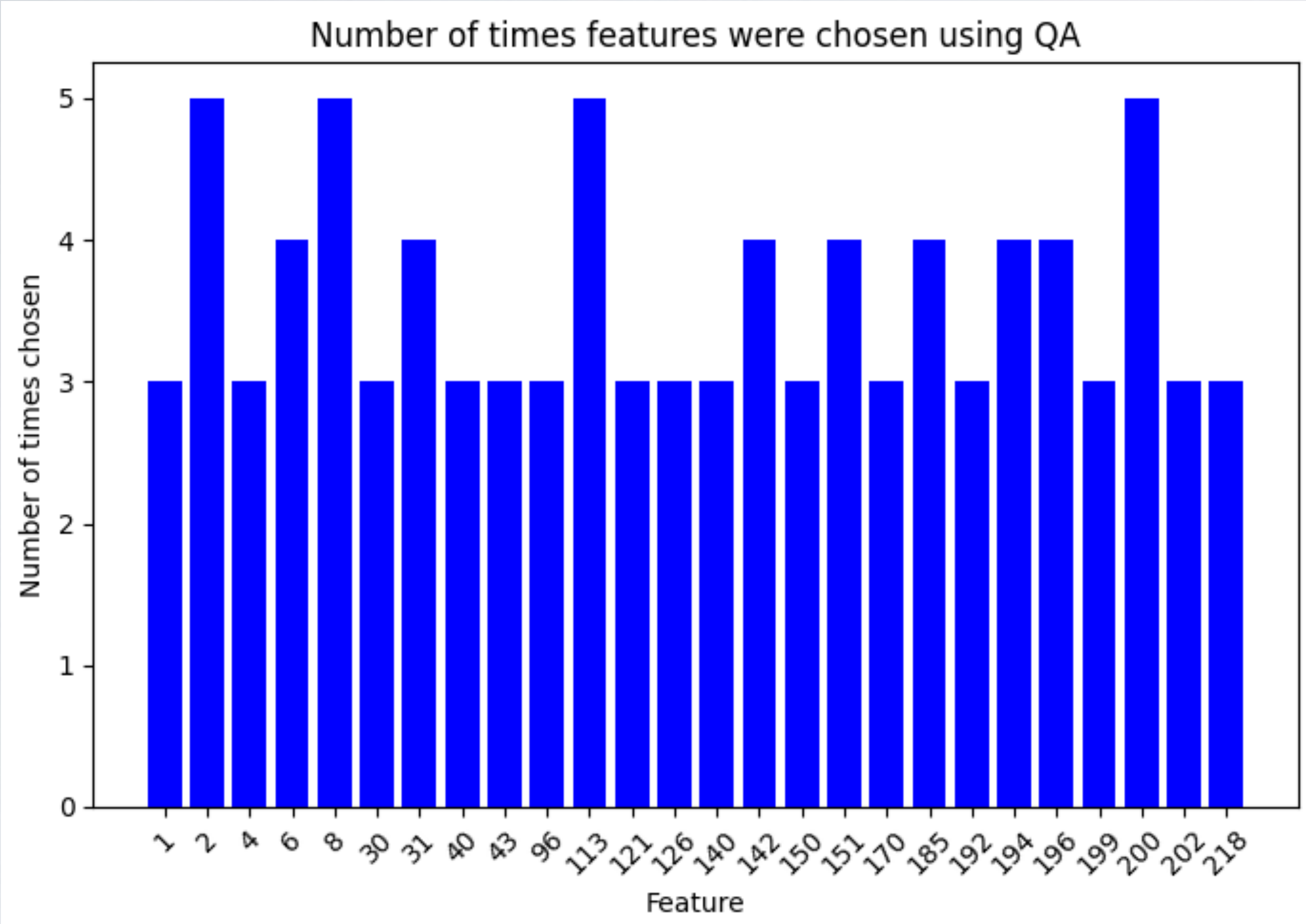


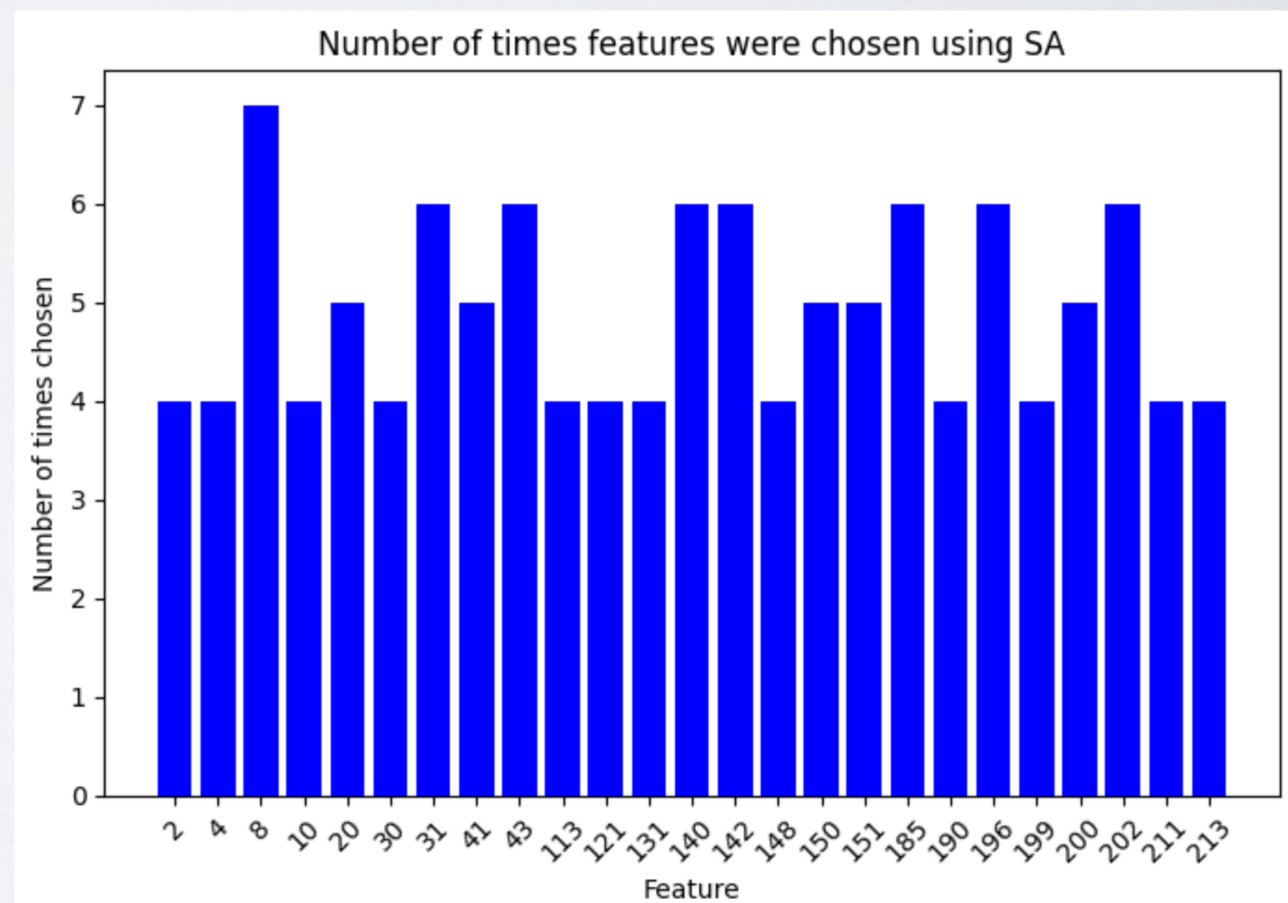
# Task 1A IR - iSTELLA (220 features)



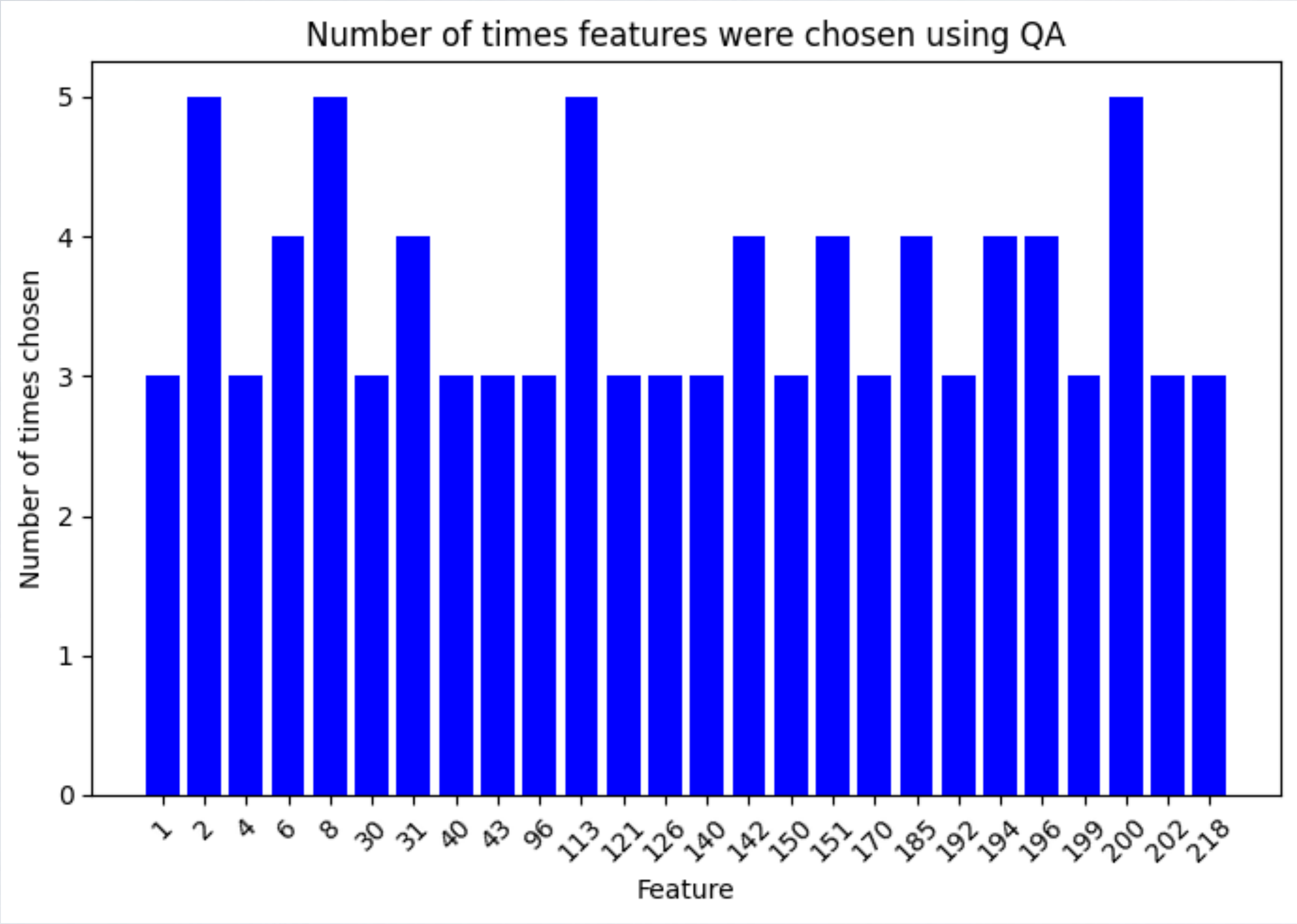



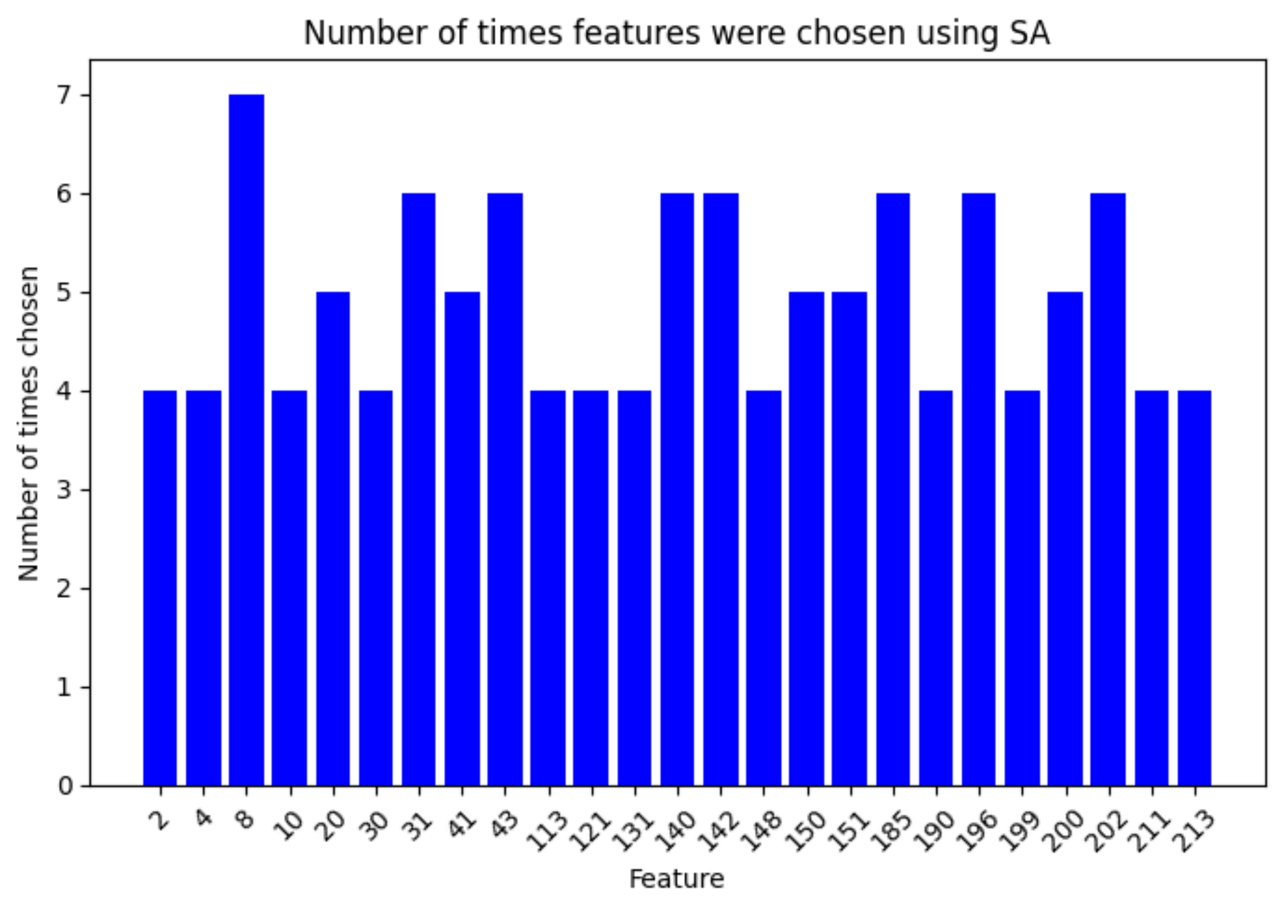
# Task 1A IR - iSTELLA (220 features)



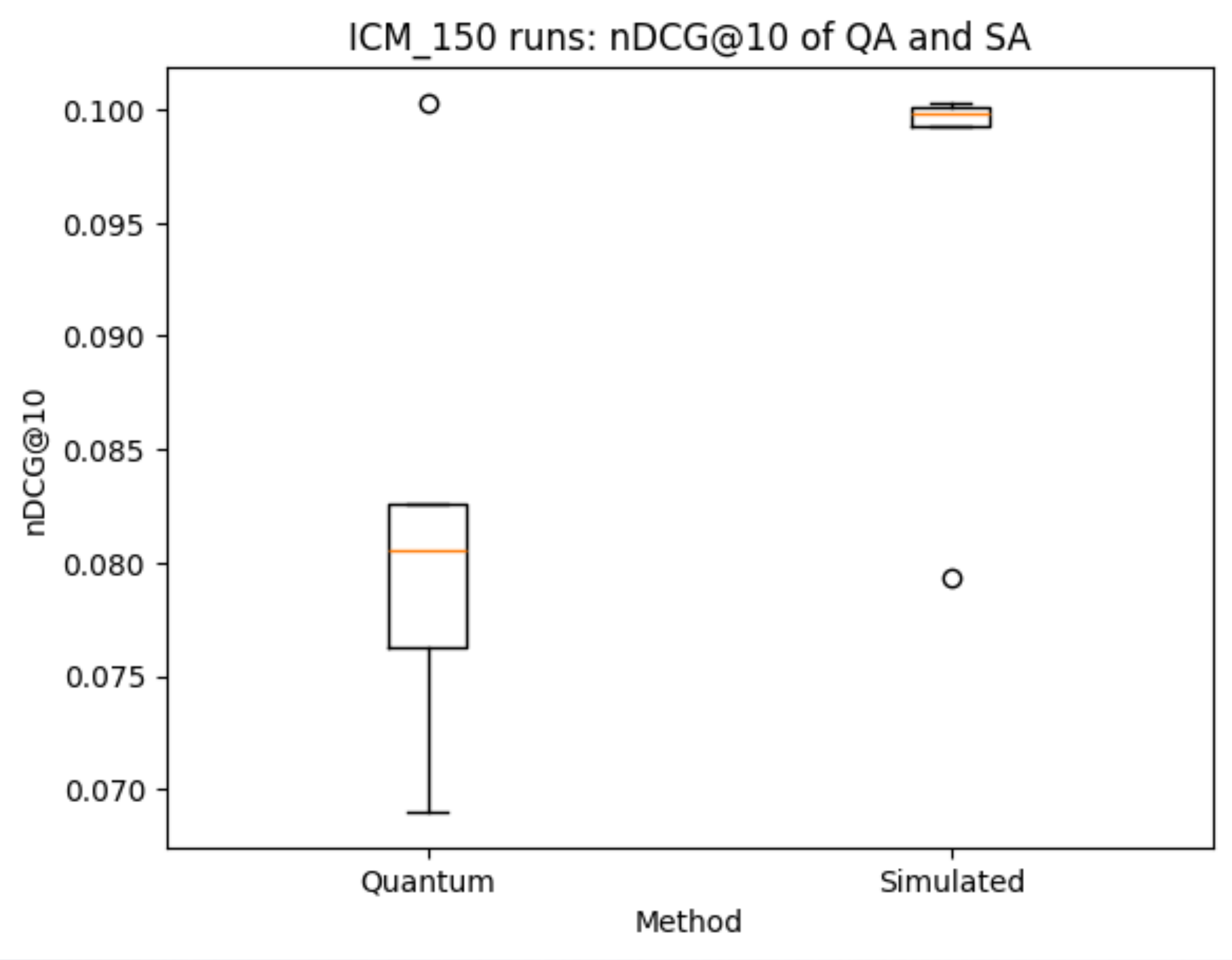



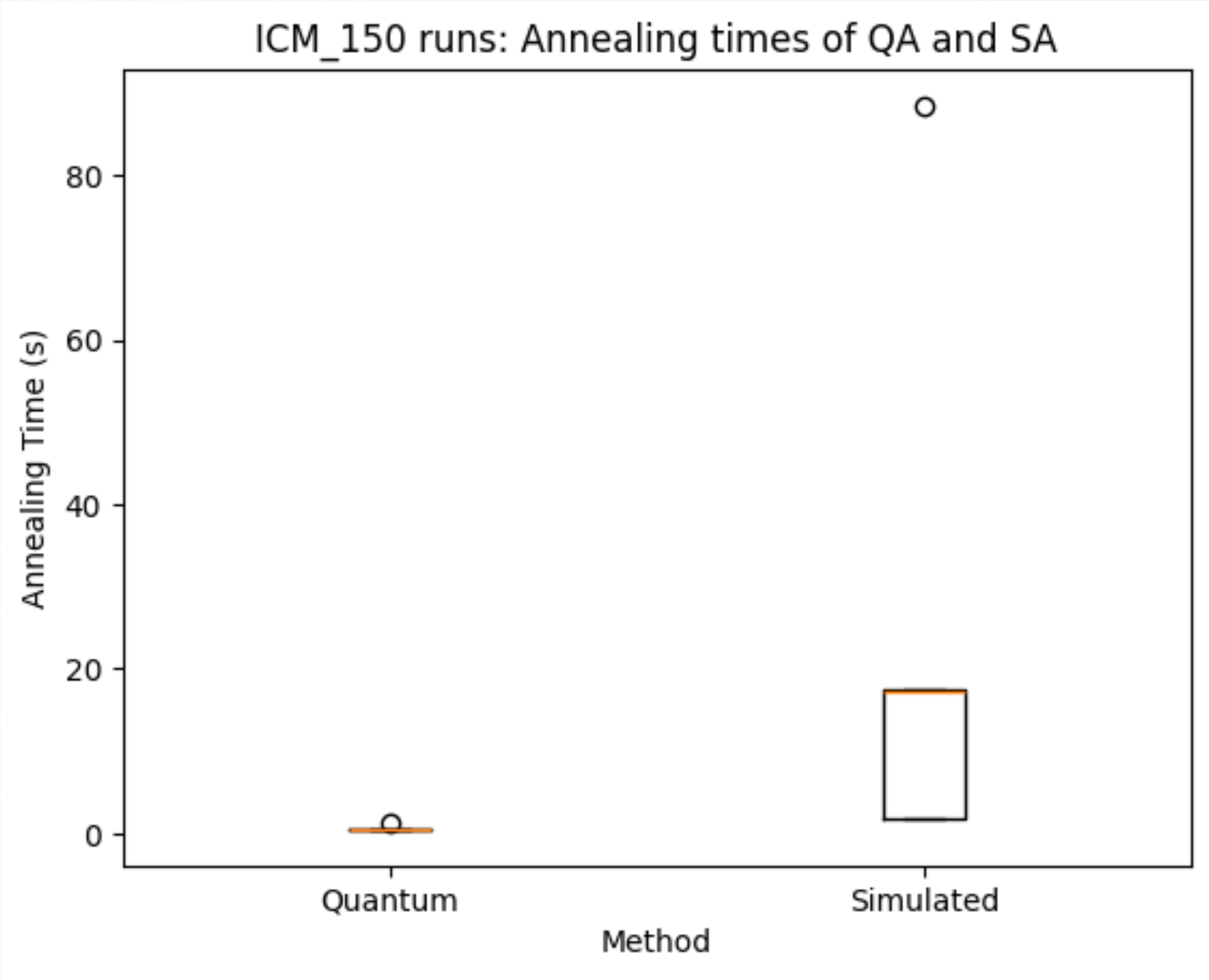
# Task 1A IR - iSTELLA (220 features)



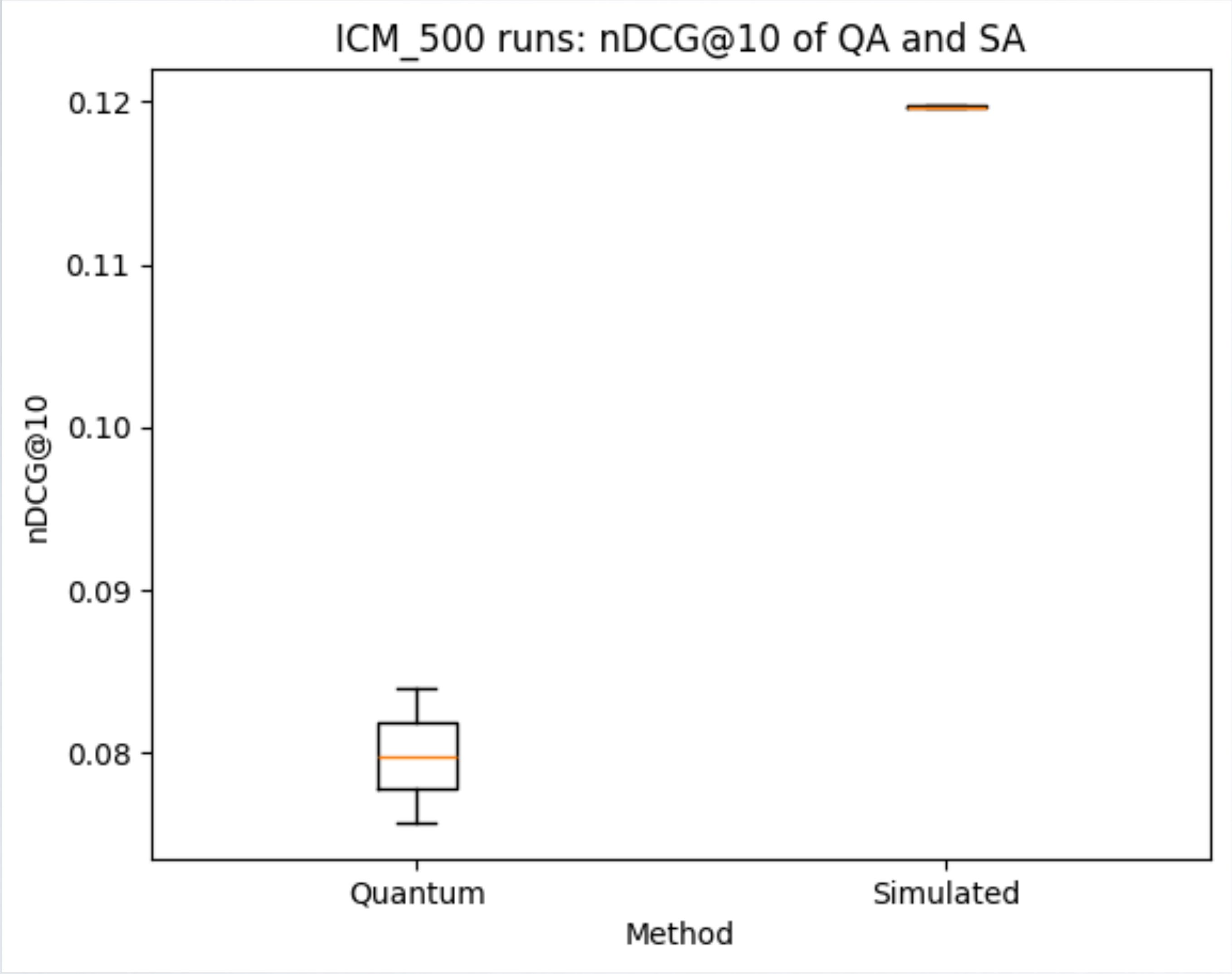



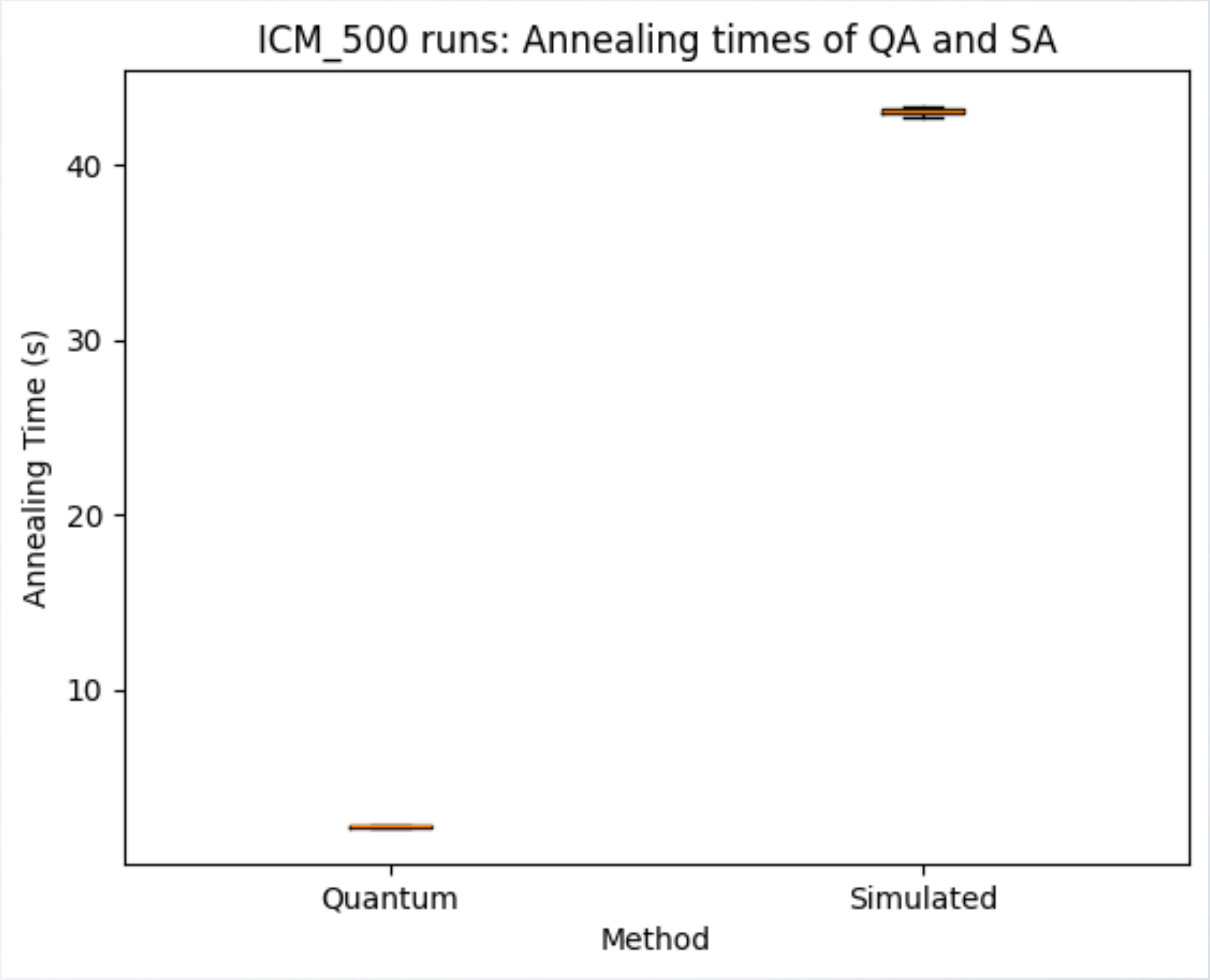
# Task 1B RS - ICM (150 features)



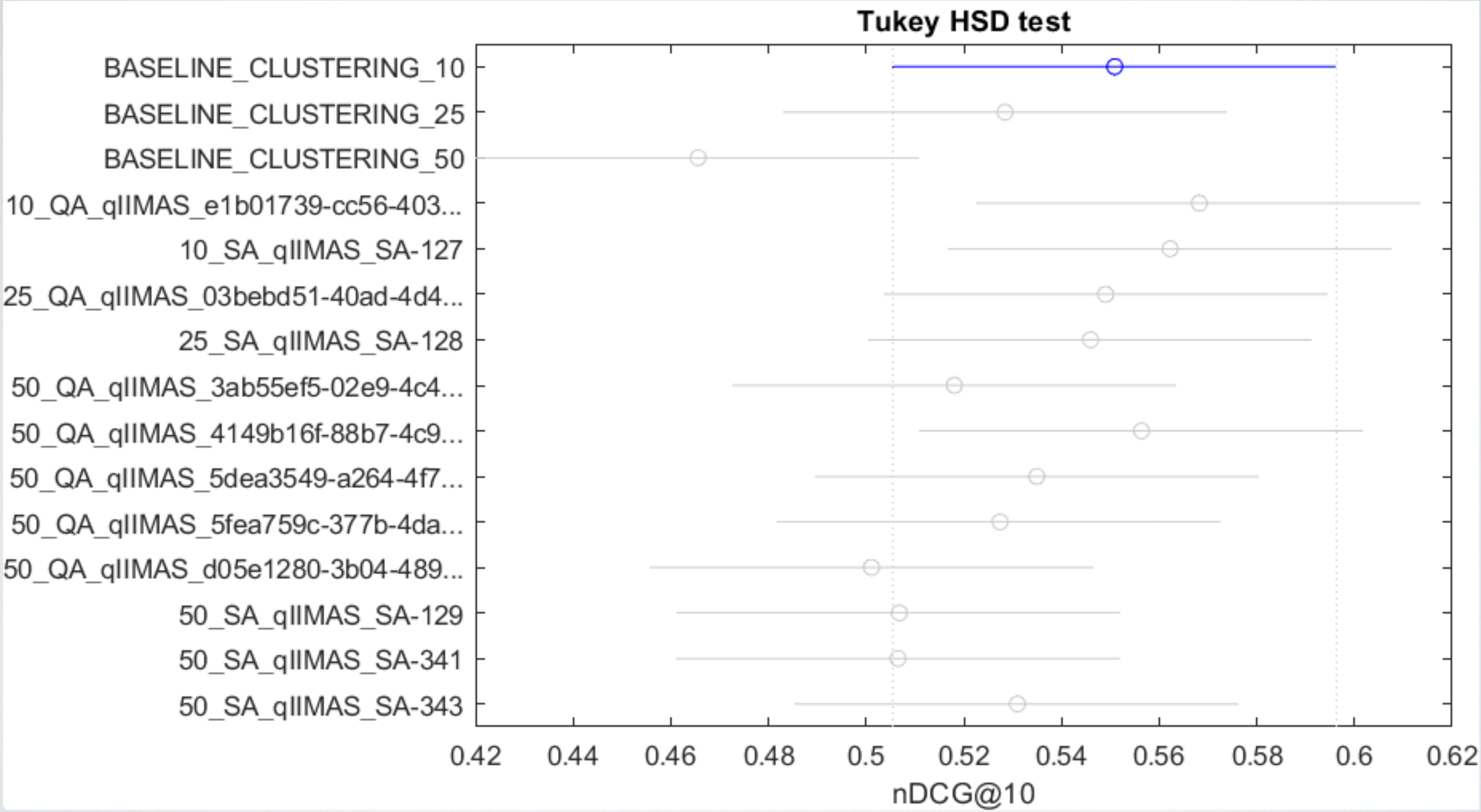



# Task 1B RS - ICM (500 features)

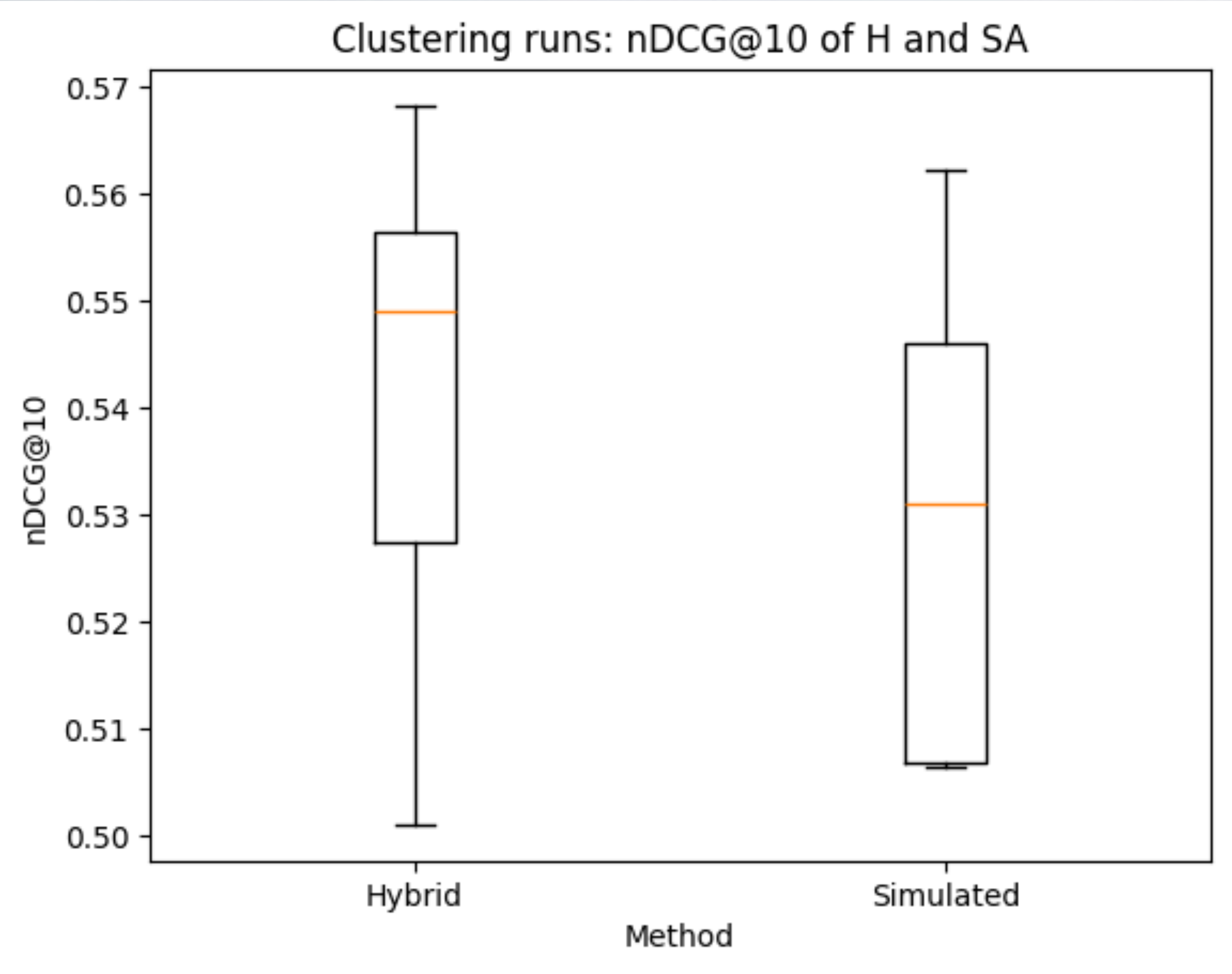


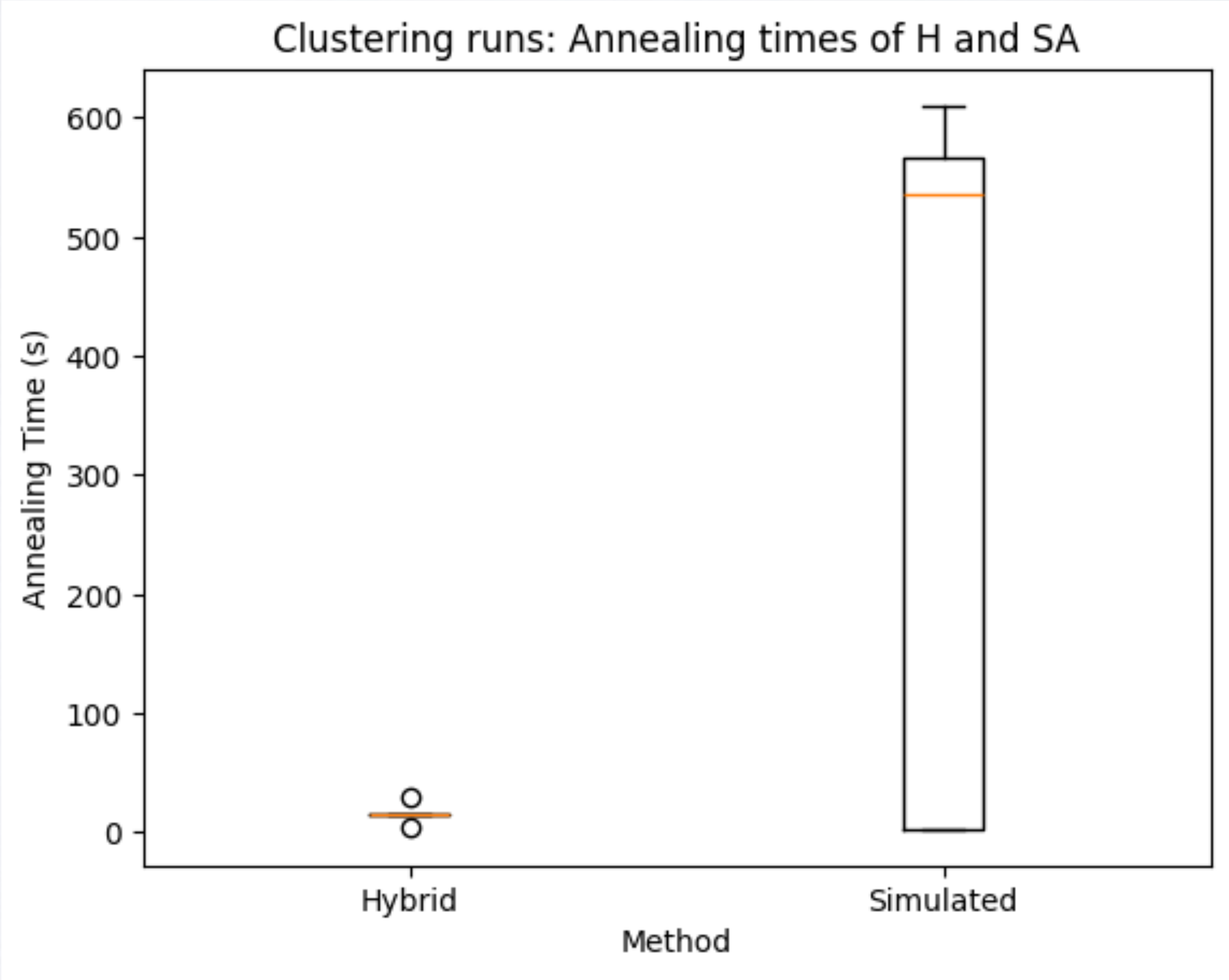



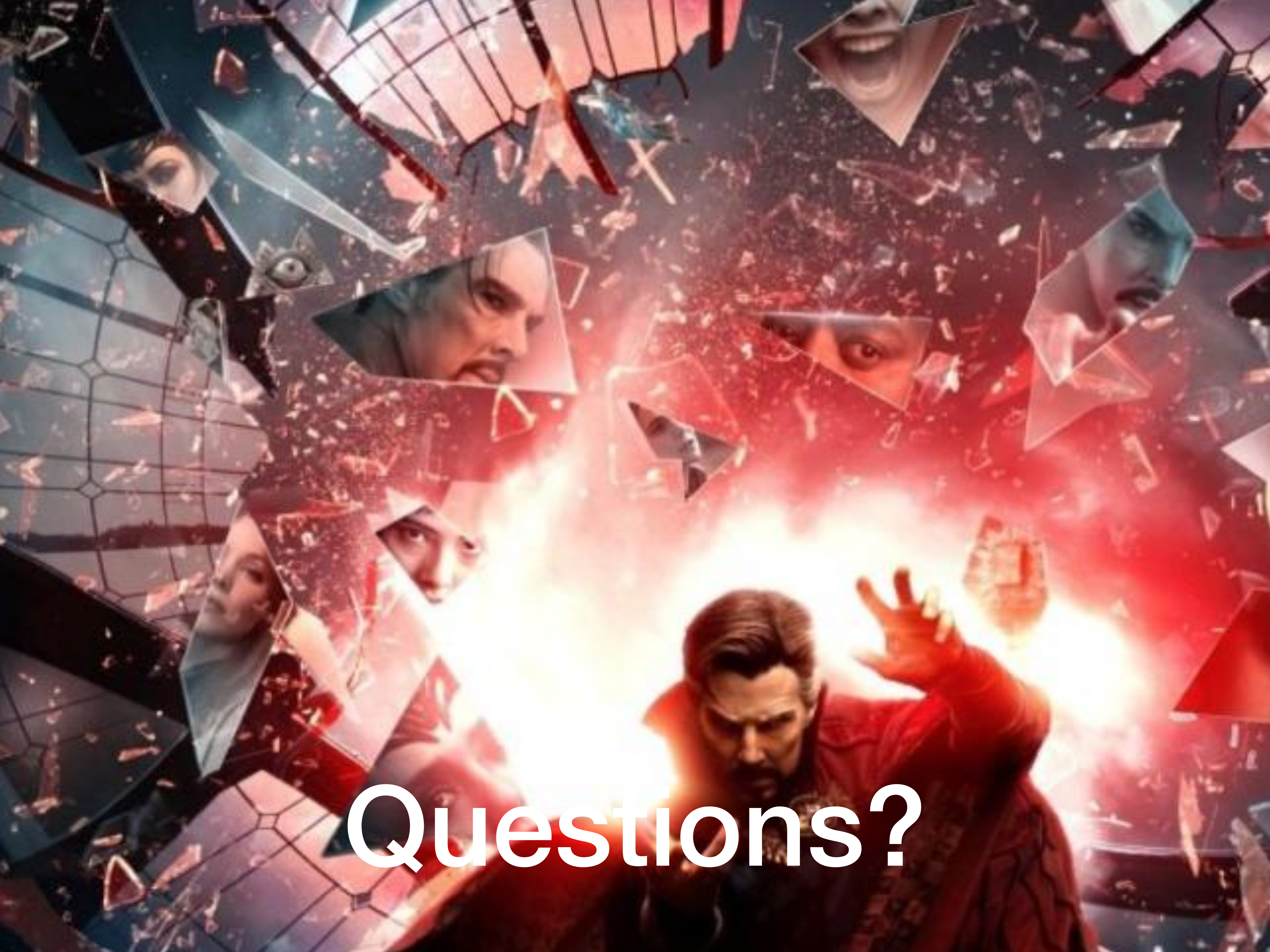
# Task 2 Clustering



# Task 2 Clustering




Questions?