Clustering with Non-adaptive Subset Queries

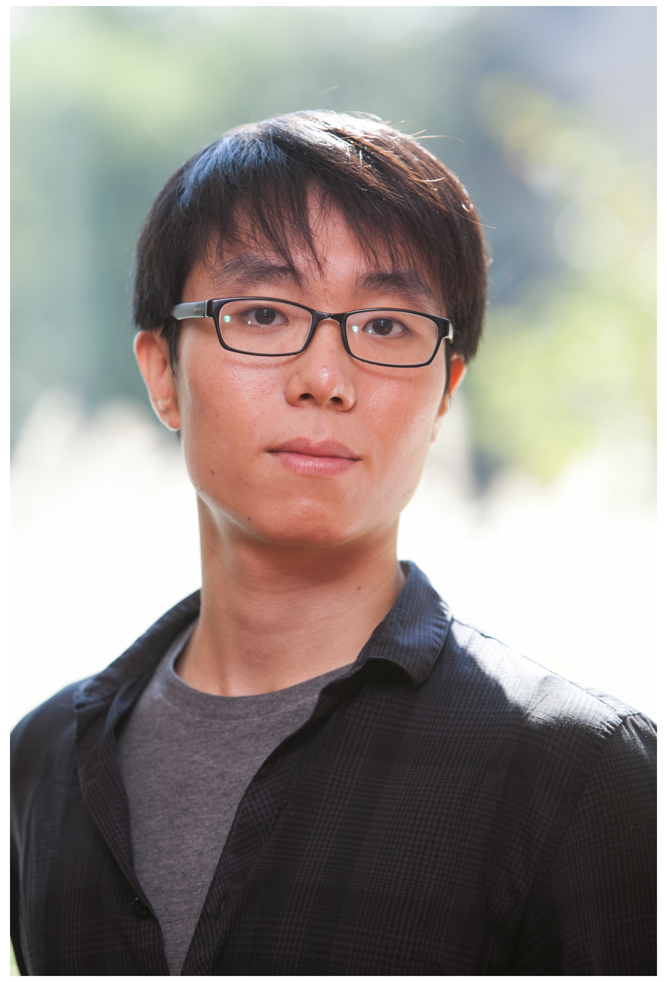
NeurIPS 2024

Hadley Black
UCSD

Euiwoong Lee
University of Michigan

Arya Mazumdar
UCSD

Barna Saha
UCSD



NSF UC San Diego Penn TEXAS UCLA ENCORE

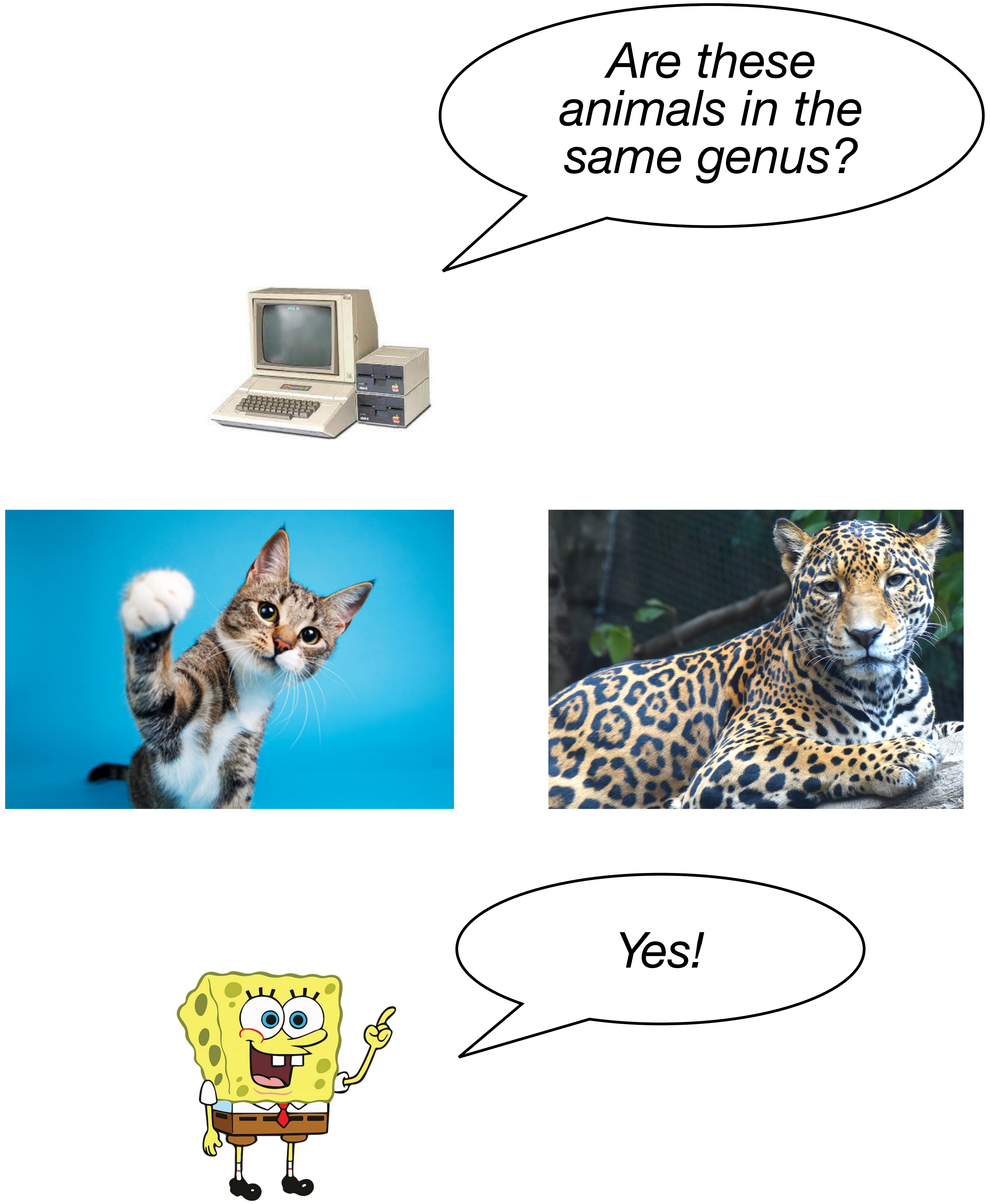
Clustering via Crowdsourcing

Clustering: group data based on similarity

- Fundamental task in data science with many instantiations

Clustering via crowdsourcing:

- Can we offload the work of computing a clustering by asking simple questions to external individuals?
- **Same-cluster queries:** Are these two points of the same type?



Are these animals in the same genus?

Yes!

Clustering via Crowdsourcing

Clustering: group data based on similarity

- Fundamental task in data science with many instantiations

Clustering via crowdsourcing:

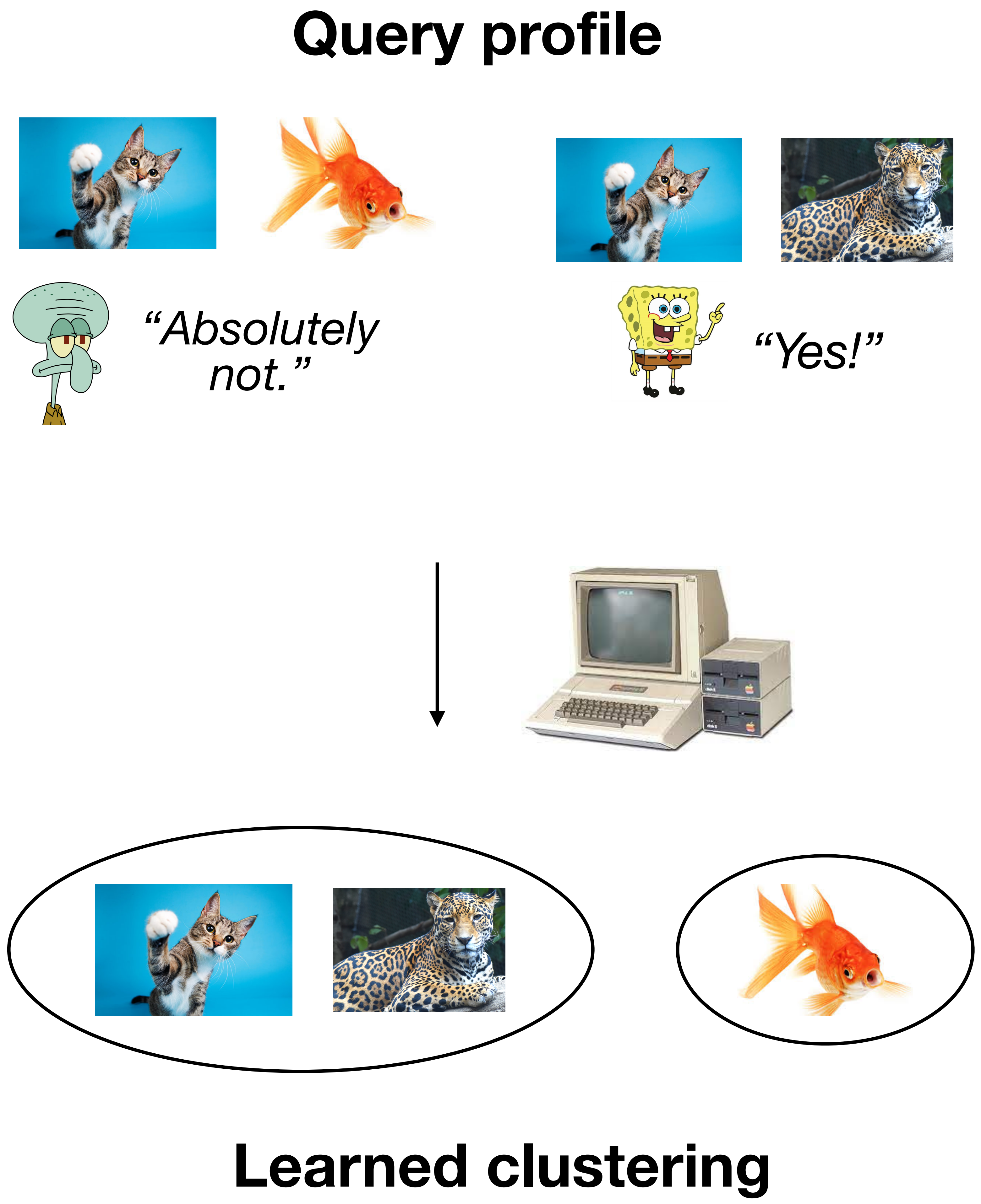
- Can we offload the work of computing a clustering by asking simple questions to external individuals?
- **Same-cluster queries:** Are these two points of the same type?

Wish list: (1) few queries, **(2)** queries specified in few rounds

- Spongebob & Squidward might be slow

  ⟹ Want to parallelize queries

- **Ideally**: non-adaptive (queries specified in one round)



Query profile

“Absolutely not.”

“Yes!”

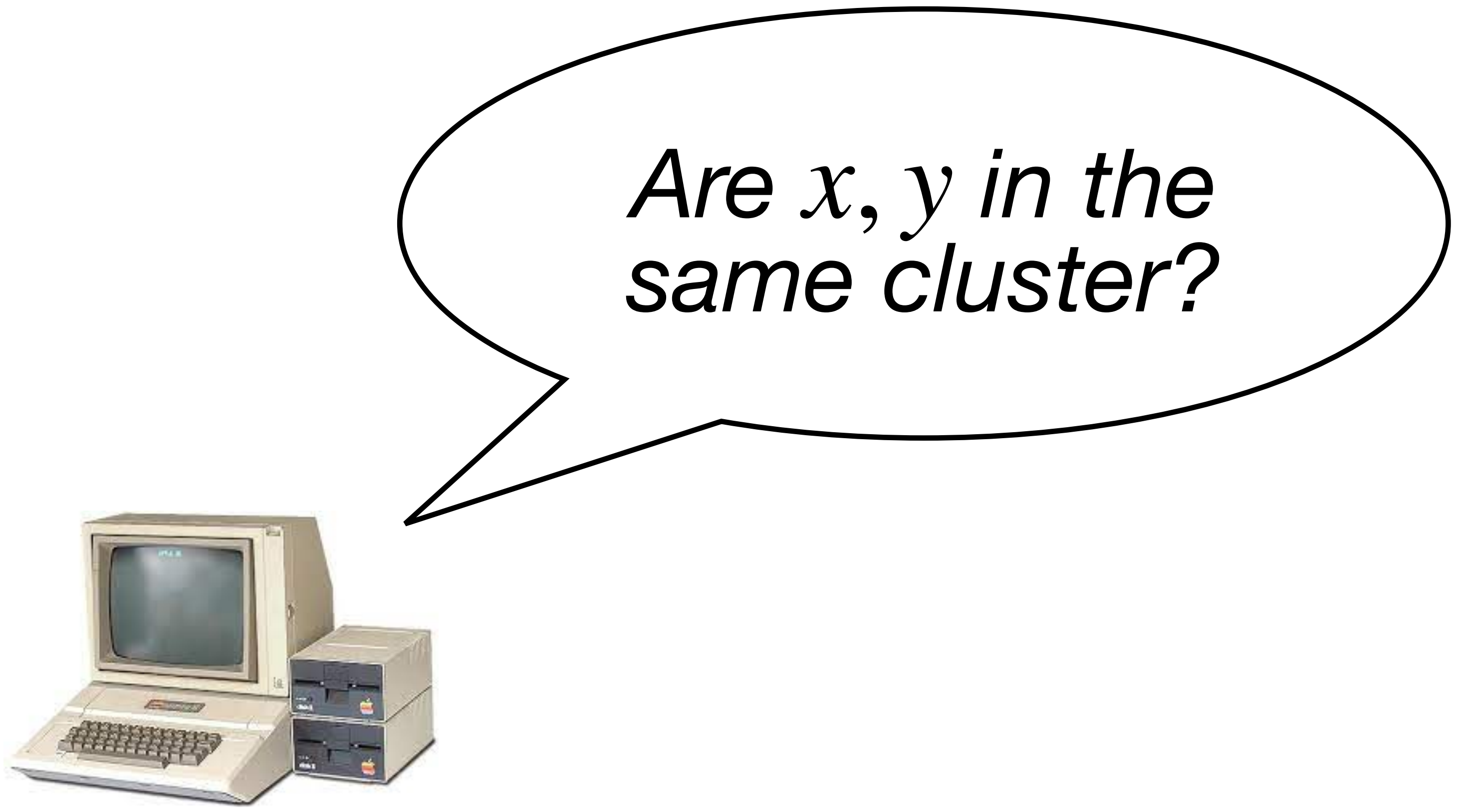
Learned clustering

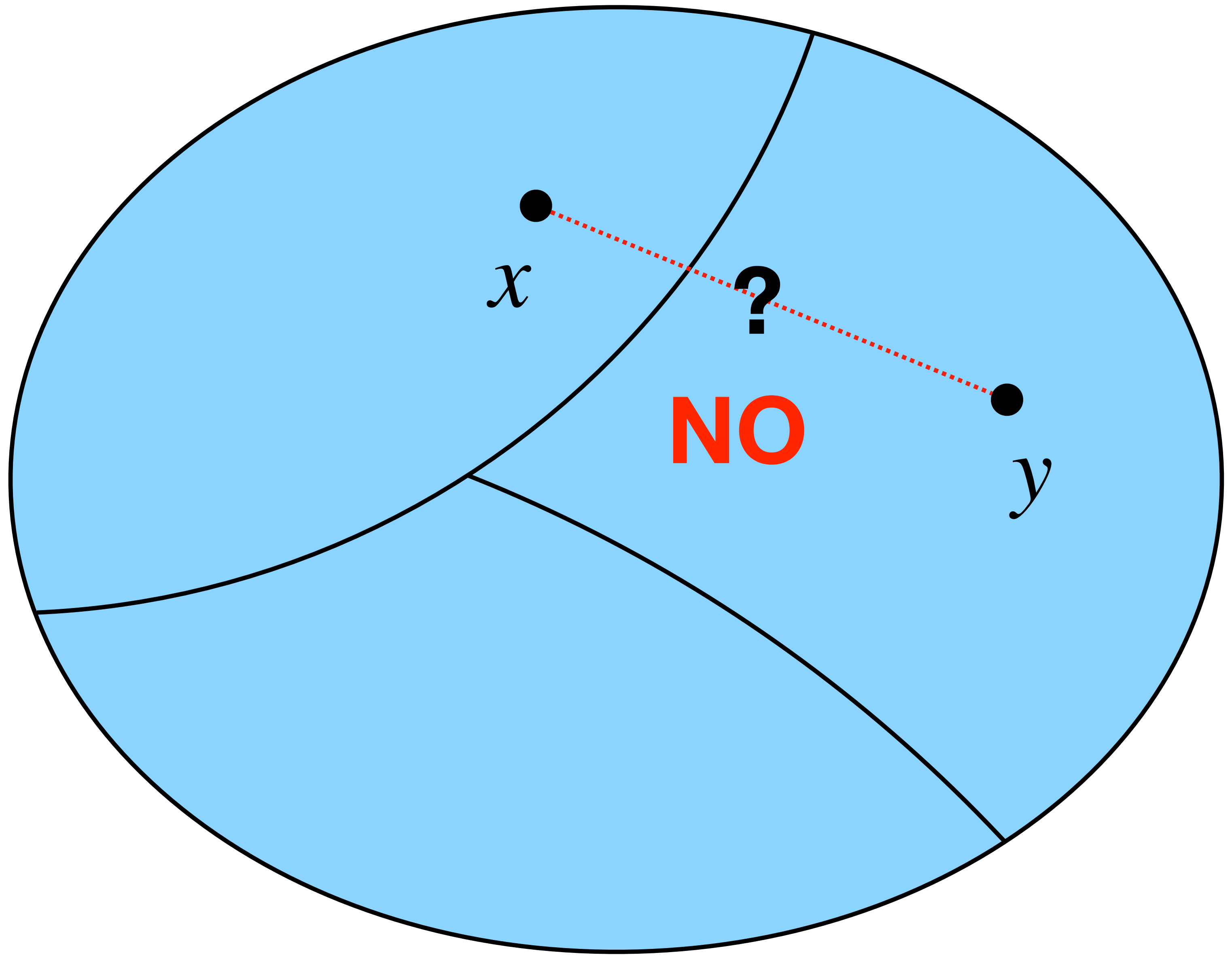
Clustering via Same-Cluster Queries

Mazumdar-Saha [Neurips 17], Mazumdar-Saha [AAAI 17], Mazumdar-Pal [Neurips 17], Mitzenmacher-Tsouraskis [16], Saha-Subramanian [ESA 19], Pia-Ma-Tzamos [COLT 22], Bressan-Cesa-Bianchi-Lattanzi-Paudice [Neurips 20], Huleihal-Mazumdar-Médard-Pal [Neurips 19]

- Set U of n points with hidden partition $C_1 \sqcup \cdots \sqcup C_k = U$
- Can **query** any $\{x, y\} \subset U$
  - Oracle says **YES** if x, y in same cluster and **NO** otherwise

Question: How many queries to learn $C_1, \ldots, C_k$ exactly?





Simple **adaptive** $O(nk)$ query algorithm ($k-1$ rounds) , **but…**

Theorem (MS 17, **BLMS 24)**

Non-adaptive algorithms require $\Omega(n^2)$ queries even for $k = 3$

$O(n^2)$ is trivial

Clustering via Subset Queries

Chakrabarty-Liao [FSTTCS 24], Vinayak-Hassibi [NeurIPS 16] (considered triangle queries)

- Set U of n points with hidden partition $C_1 \sqcup \cdots \sqcup C_k = U$
- Can **query** any $S \subseteq U$ and oracle returns $\# \, j: S \cap C_j \neq \emptyset$

Question: How many queries to learn $C_1, \ldots, C_k$ exactly?

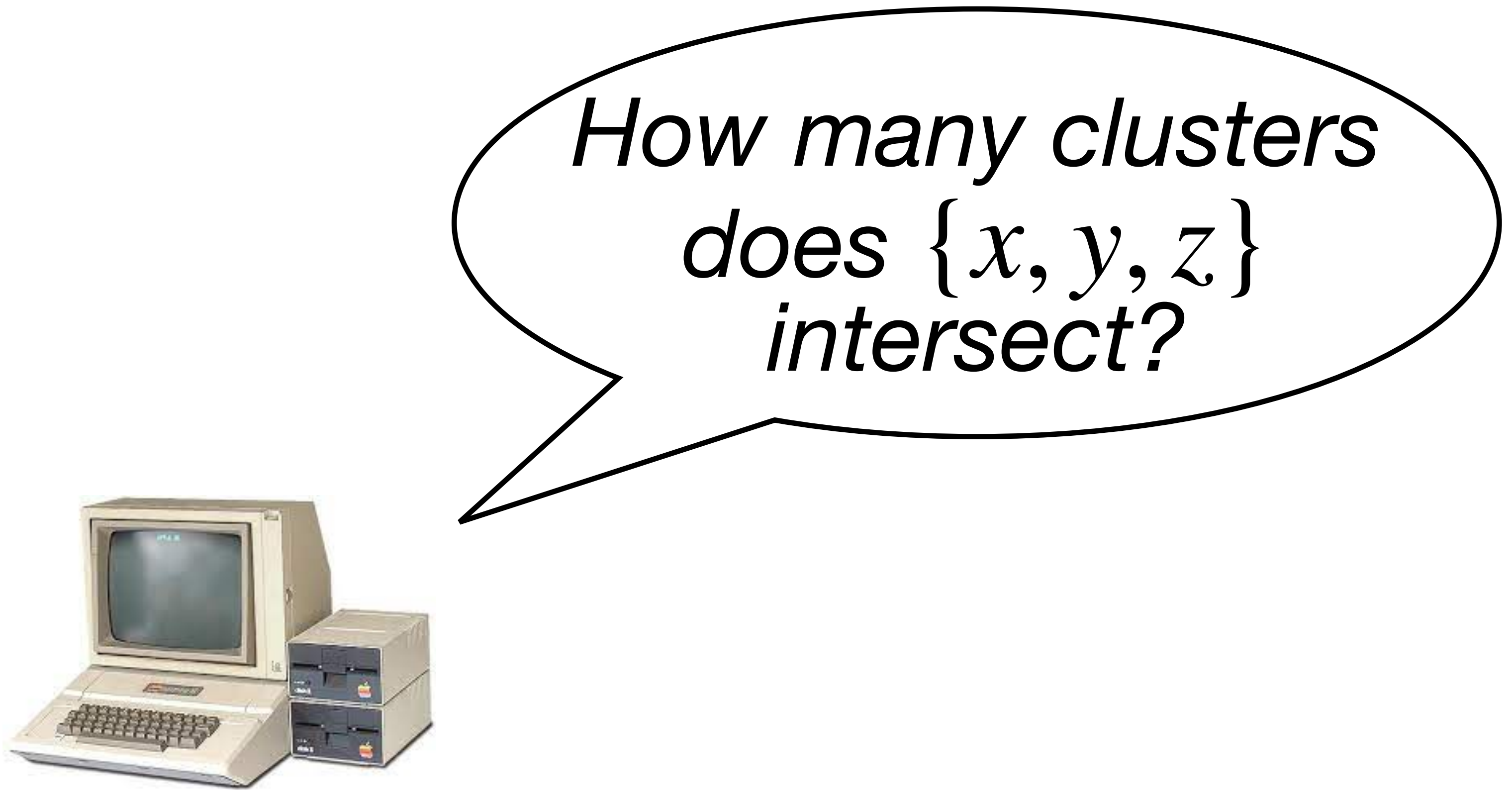
Information-theoretic lower bound: $\Omega(n)$ $\Longleftarrow$ # bits per query: $O(\log k)$ # partitions: k^n

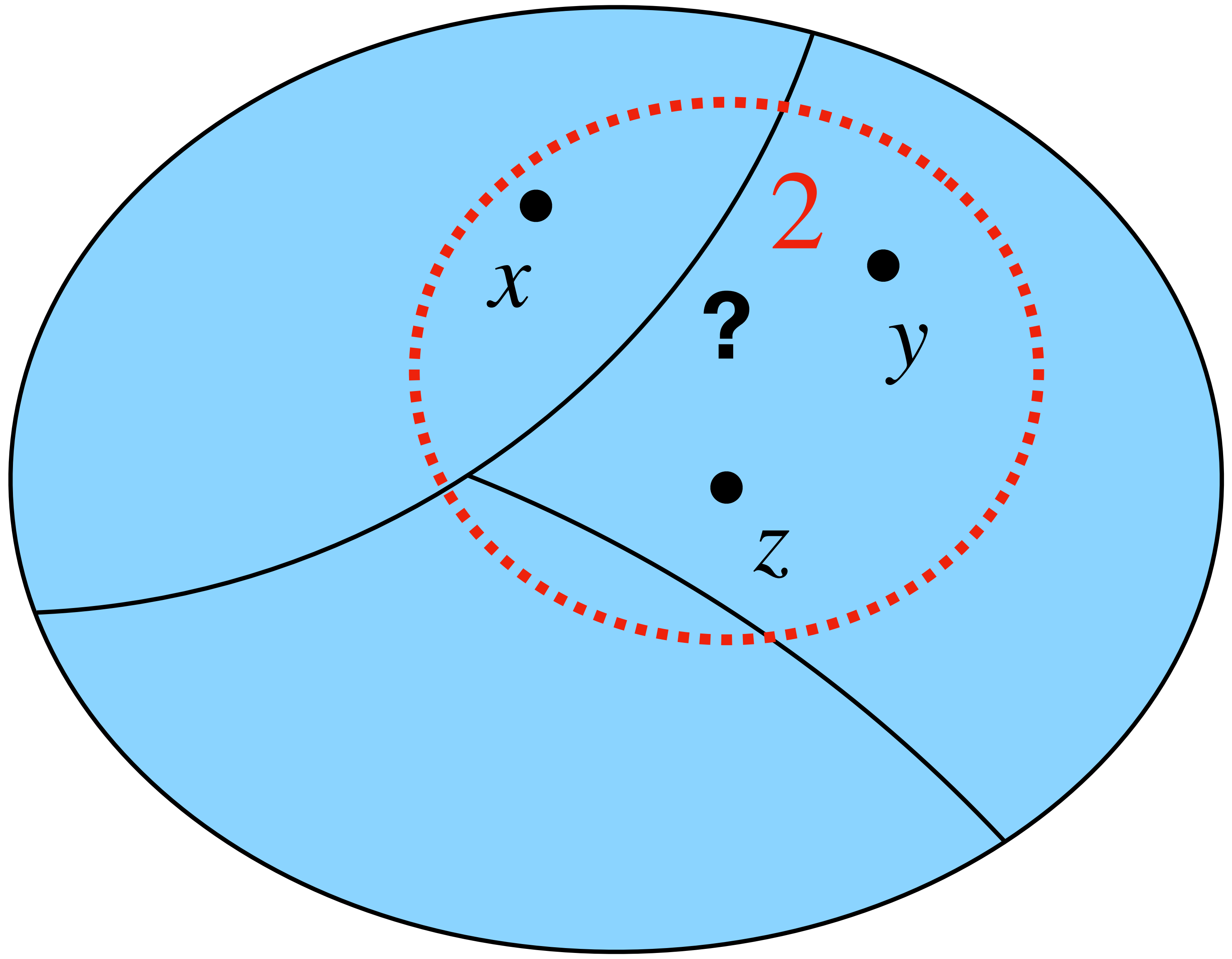
Theorem
(Chakrabarty-Liao 24)
$O(n)$ **adaptive algorithm**

Questions

How close to linear can we get non-adaptively?

How small of queries can we get away with?





Some of our Results

(all algorithms and lower bounds are non-adaptive)

Unbounded subset queries

Theorem

$O(n \log \log n)$ for $k = O(1)$

$O(n \log k) + \widetilde{O}(k)$ for **balanced** clustering

Question

Is $O(n)$ for $k = 3$ possible using non-adaptive algorithms?

Subset queries of size $|S| \leq s$

Theorem

$O(n \log n \log \log n)$ for $s = O(\sqrt{n})$, $k = O(1)$

Getting near-linear requires $s = \Omega(\sqrt{n})$

Question

Can we get near-linear with $s = O(\sqrt{n})$ for all k?

Thank you!