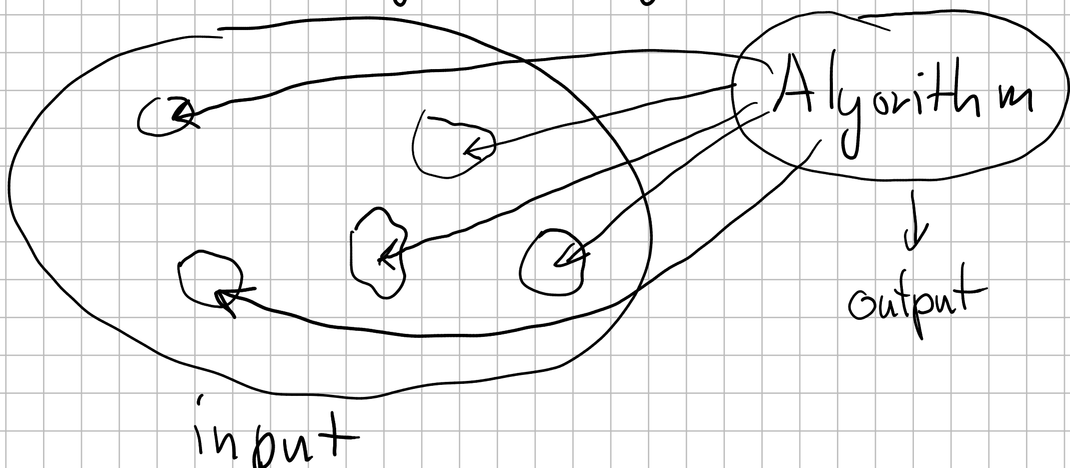
Today:

- overview of other types of distribution testing (see previous notes)
- new theme: "Querying Big Data Sets"
  example: estimating maximum matching or minimum vertex cover size

New theme: Querying Big Data Sets

- input: big data set
- don't want to read all of it
- what can we say about its properties after looking at tiny fraction?



Goal: sublinear-time algorithms

Note: results very dependent on type of queries/samples allowed

---

Simplest example: estimate fraction of elements with specific property, e.g., fraction of numbers that are prime

Input: sequence of numbers $s_1, s_2, \dots, s_n$

Queries: for any $i \in [n]$, can get s_i

How many queries needed to estimate fraction of prime numbers up to $\pm \varepsilon$?

Solution: - sample $O(1/\varepsilon^2)$ numbers
- return the fraction of primes in the sample

Now: graphs $G = (V, E)$

Allowed queries:
- get uniformly random $v \in V$
- for any $v \in V$, get degree of v
- for any $v \in V$ & $i \in \mathbb{Z}_+$, get i-th neighbor of v

Maximum matching & minimum vertex cover

Matching = subset of edges that share no endpoints

Maximum matching = highest cardinality matching

Maximal matching = no edge can be added or it won't be a matching

$\circledast$ = (# edges in maximal matching) $\geq \frac{1}{2}$ (# edges in maximum matching)

Vertex cover = subset of vertices such that at least one endpoint of each edge in the set

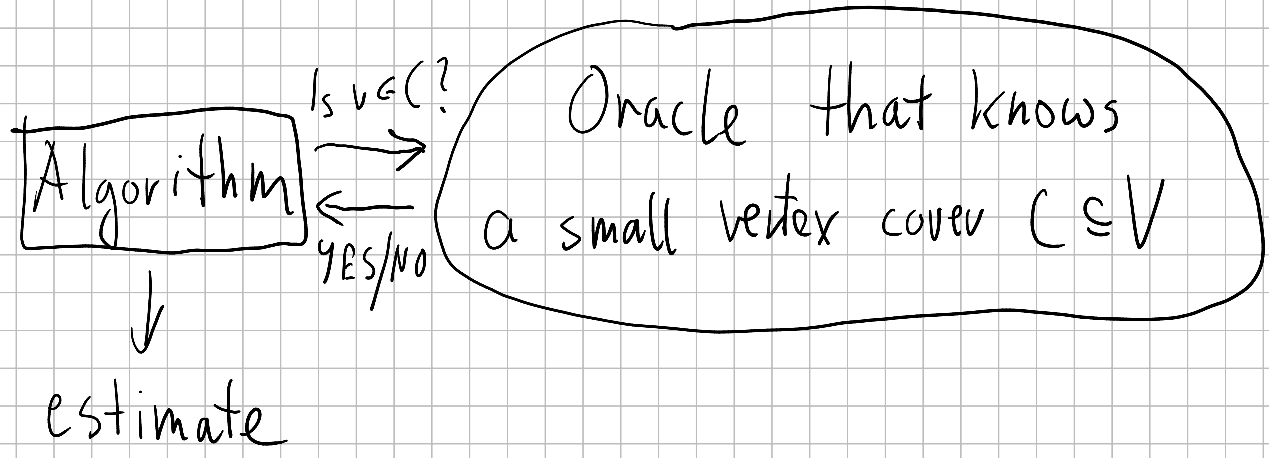
Minimum vertex cover = minimum cardinality vertex cover

$\circledast \leq$ (size of minimum vertex cover) $\leq 2 \circledast$

Our main goal: estimate this ↑ (size of minimum vertex cover)

Ideal situation: "Our lucky day"

Someone gave us an oracle that provides access to a small vertex cover



$O(1/\varepsilon^2)$ queries about randomly selected vertices suffice to estimate $|C|$ up to $\pm \varepsilon n$ with probability $99/100$

Bad news: today is not our lucky day
and Tuesday won't be either

We'll have to construct our oracle ourselves