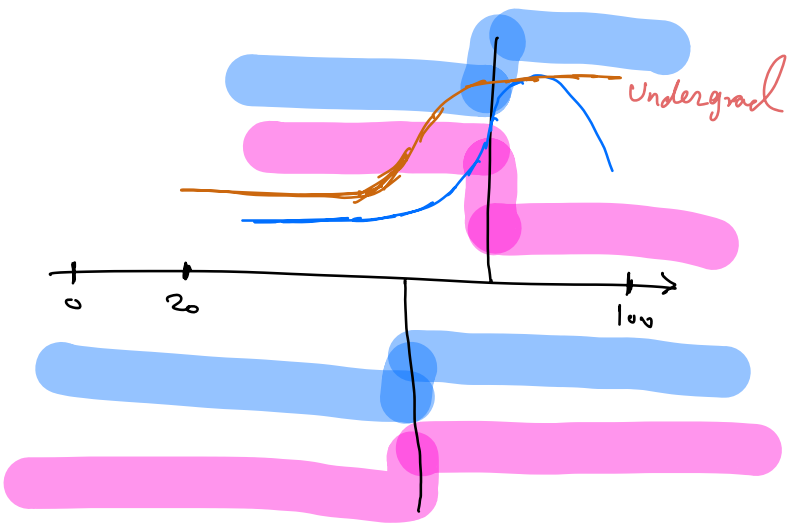
- Measurement Bias: when:
  - The measured attribute is not exactly what the goal is
    - ↳ Crime likelihood (Proxy) Arrest Likeli
  - Labeling Bias:
  - Attribute Definition Bias.



- Sampling Bias: when the Samples are not iid from the underlying distribution
  - ↳ the distribution of Samples is diff from the true one.
    - → The Survey Collection Example.
- Representation Bias: when data misses to represent some demographic groups.

↳ Rep. Bias can be due to Sampling Bias

↑ " " & Sampling Bias can be in Trade off.

↳ we should introduce an slightly amount of over sampling to resolve rep. Bias.

- Aggregation Bias: when Data is recorded at a specific hierarchy in the Data Cube

e.g., Location | Country / State / City / County / ⋮

↳ Berkely Admission Example. | University / College ...

- Spatio-Temporal Bias: when data from wrong time spans / geo. locations is recorded

Reasons of Bias in data

- Historical Bias (Discriminations)
- Internal human Bias
- Population Bias
- Behavioural Bias
- Social Bias: when people in Social Networks impact each other

---

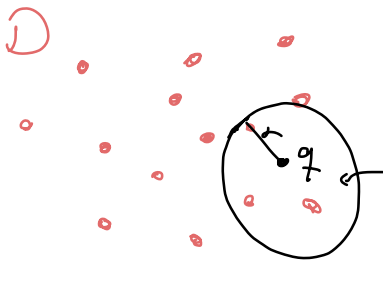
Representation Bias: When there is not **enough** Samples **Similar** to a query Point

enough:
- Almost equal from each group.
  - ↳ Equal Base Rate
  - → Considers equal Ratios: Representation Rate $R = |g_1| / |g_2|$
- An absolute number: e.g., ~~50~~ at least 50 from each group.

a query q is covered

if $|Sim(q) \cap D| \geq \tau \simeq 50$

$Sim(q)$: the Samples from the Same demo. group, $q \in$ Black Female



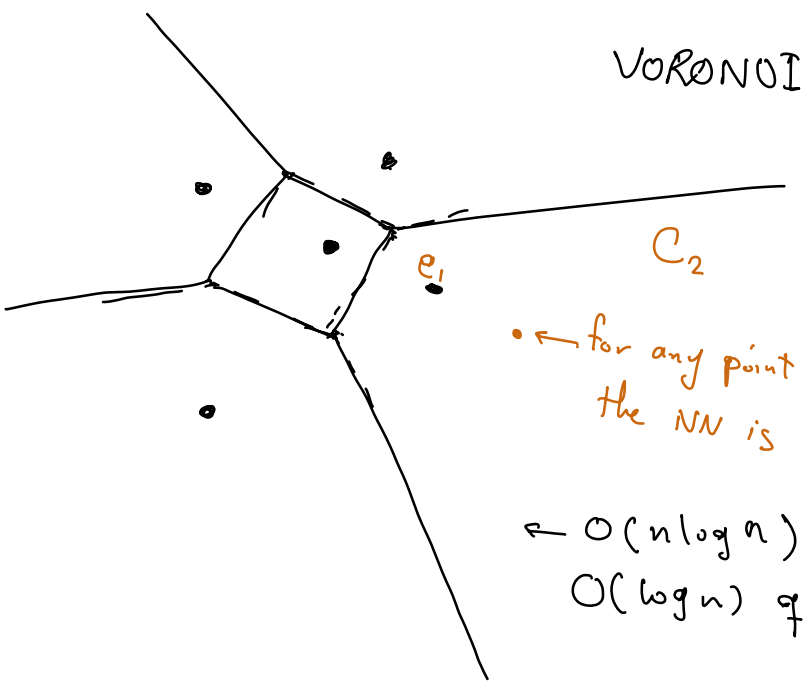
any thing inside the ball is similar to q

- q is covered if

$$\left| \underset{t_i \in D}{dist(q, t_i) \leq r} \right| \geq k$$

- The uncovered Region as the set of points that are not covered

$$\{ q \in \mathbb{R}^d \mid Cov(q, D) = False \}$$



← $O(n \log n)$ Construct

$O(\log n)$ query

k-voronoi diagrams : The K-NN for every cell is the same