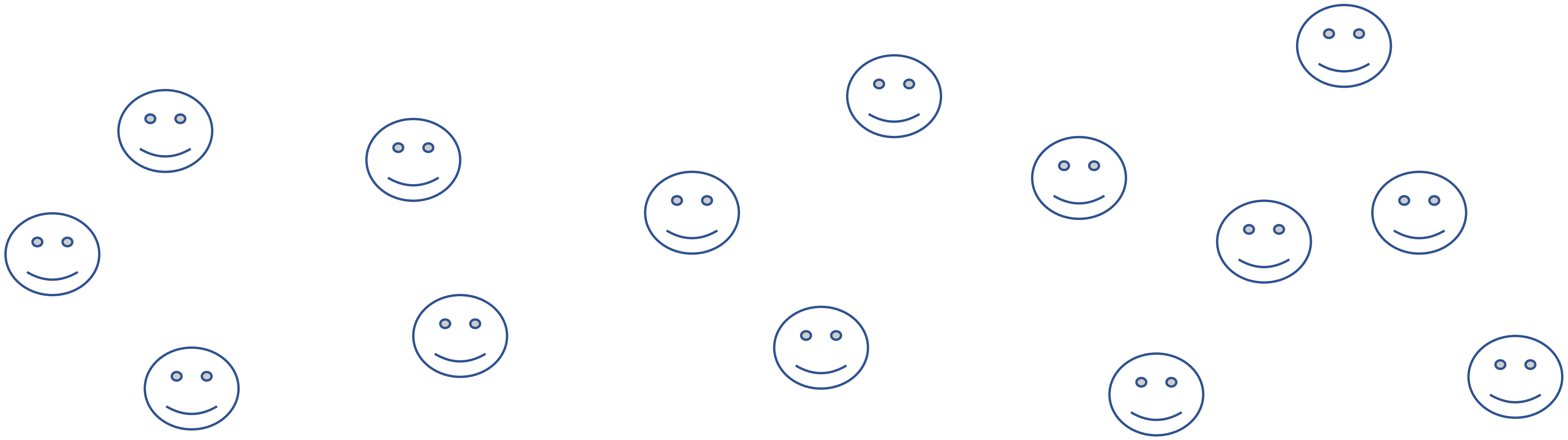
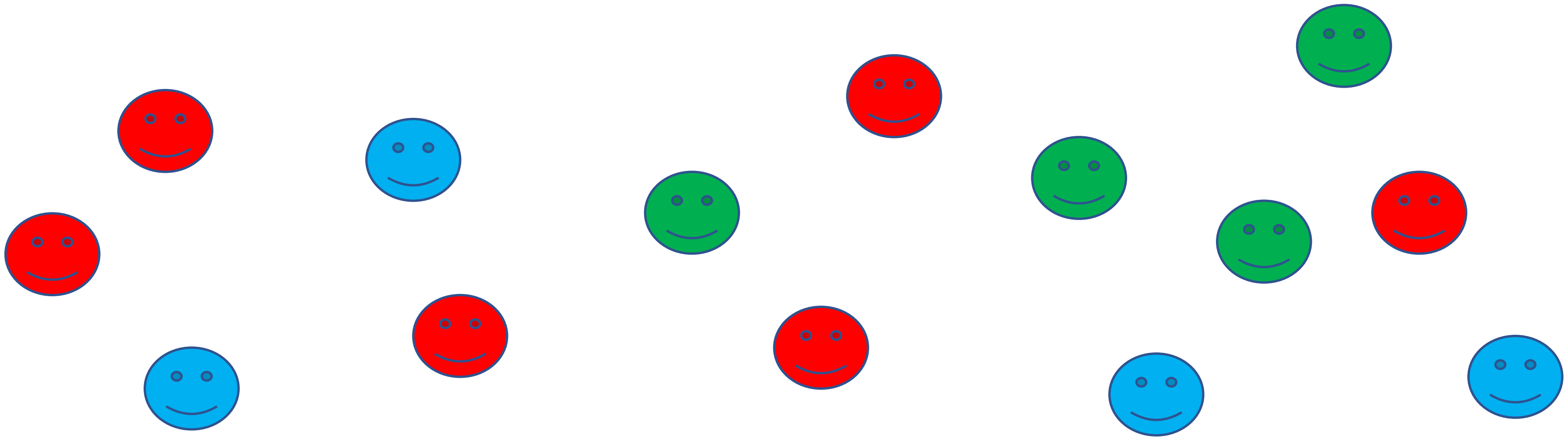
Sequential Community Mode Estimation

Shubham Jain, Shreyas Goenka, Divyam Bapna, Nikhil Karamchandani & **Jayakrishnan Nair**

(Department of Electrical Engineering, IIT Bombay)

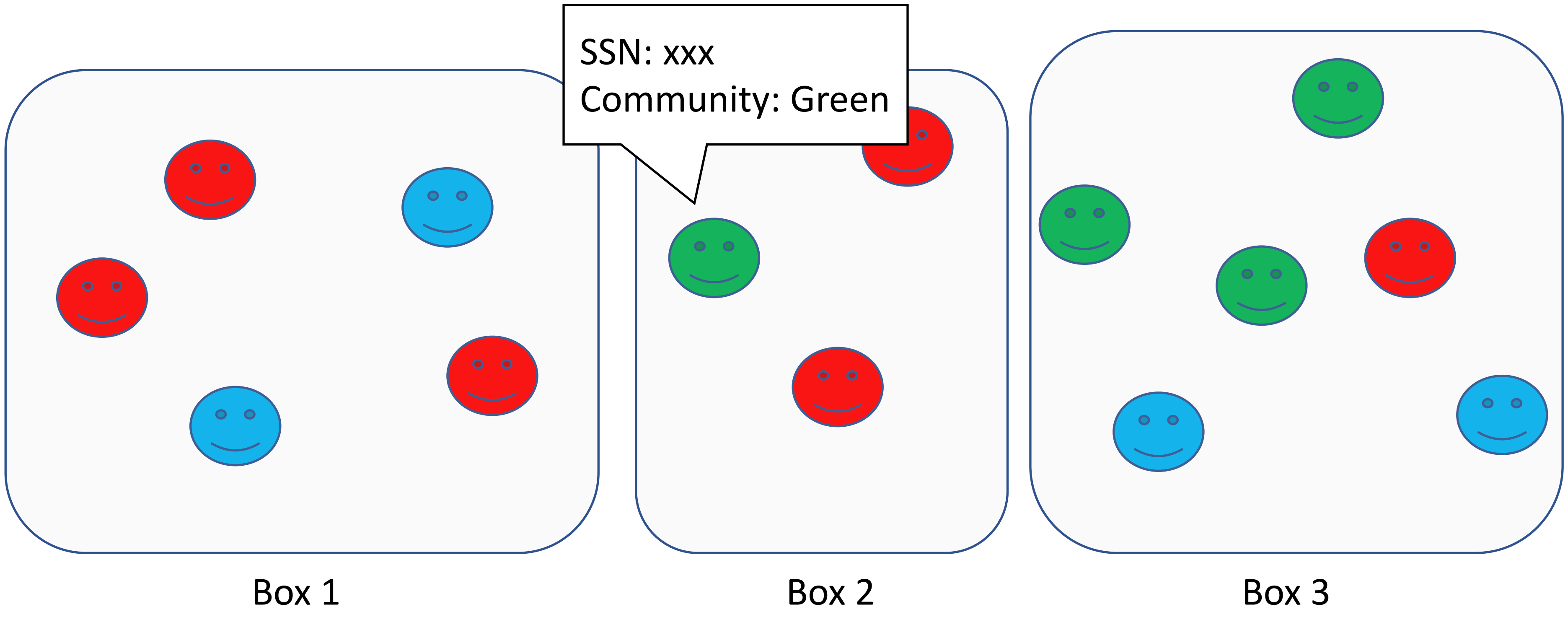


Given: Population consisting of N individuals

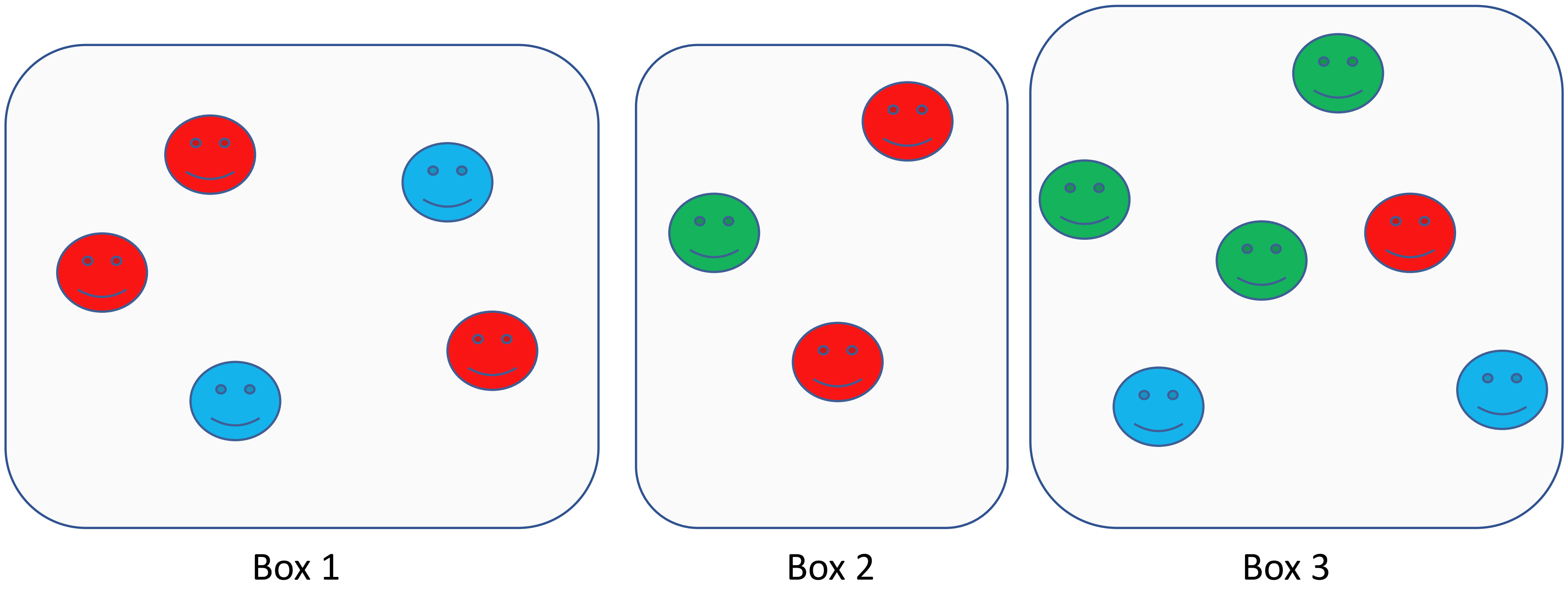


The population is partitioned into disjoint communities

Goal: Identify, via sequential sampling, the largest community



- Population also partitioned into sampling domains, a.k.a., *boxes*
- Learning agent can choose, at any time, which box to sample/query from
- Choosing Box i, a random individual gets sampled (with replacement) from that box; her community and identity gets revealed to agent
- Agent has *budget* of t queries *(fixed budget setting)*
- Goal: Minimize *probability of error*



<u>Applications</u>

1. Election polling

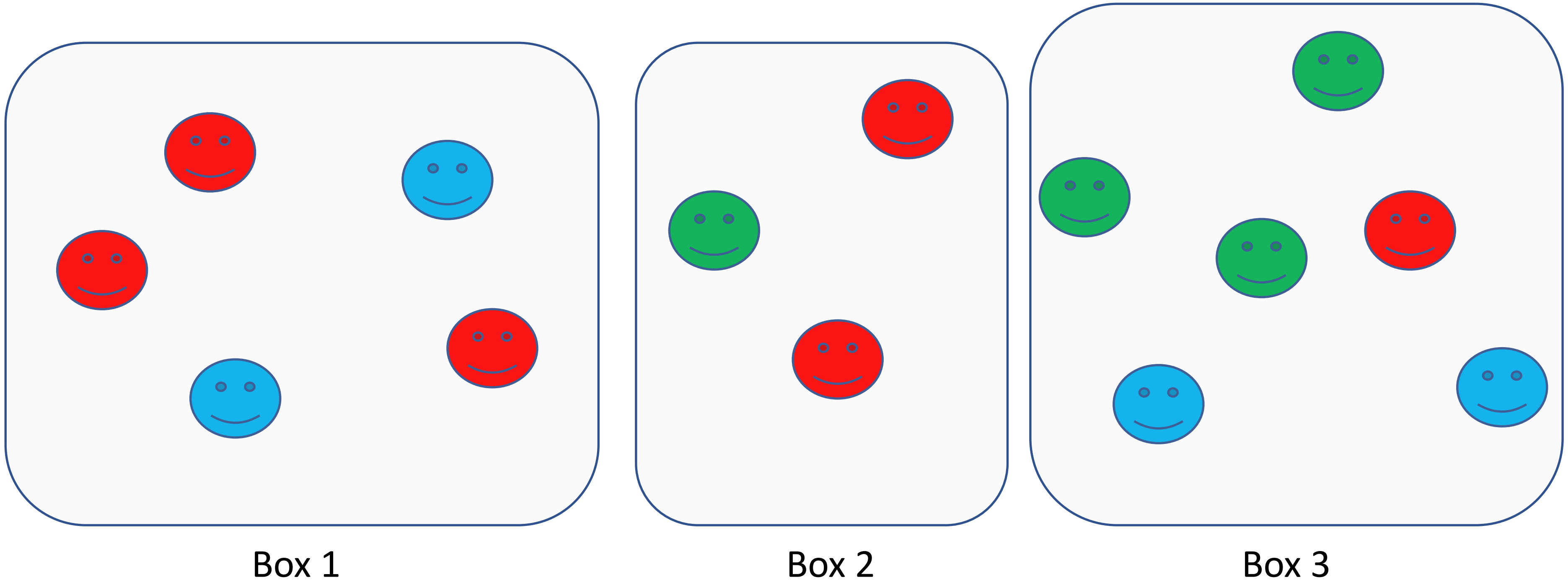
   communities → political parties

   boxes → states

2. Estimating dominant strain of a virus

   communities → virus strains

   boxes → localities

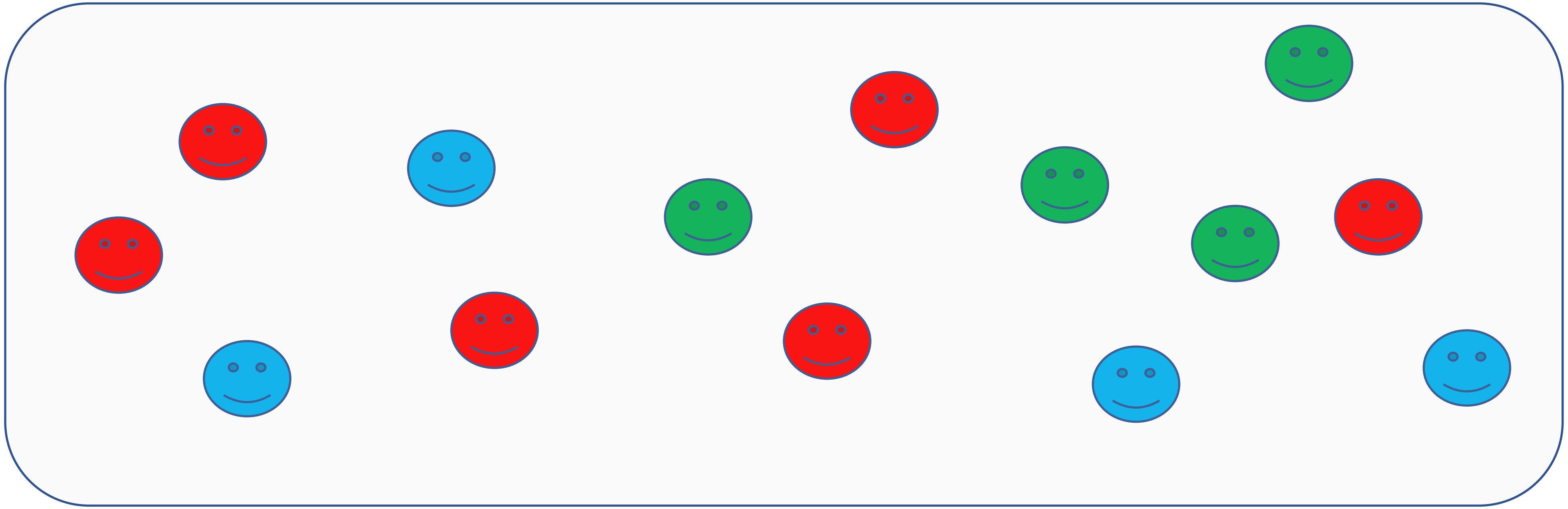


This is <u>not</u> an MAB problem

- Have only partial control on which community to sample from
- Observations are not *i.i.d.*

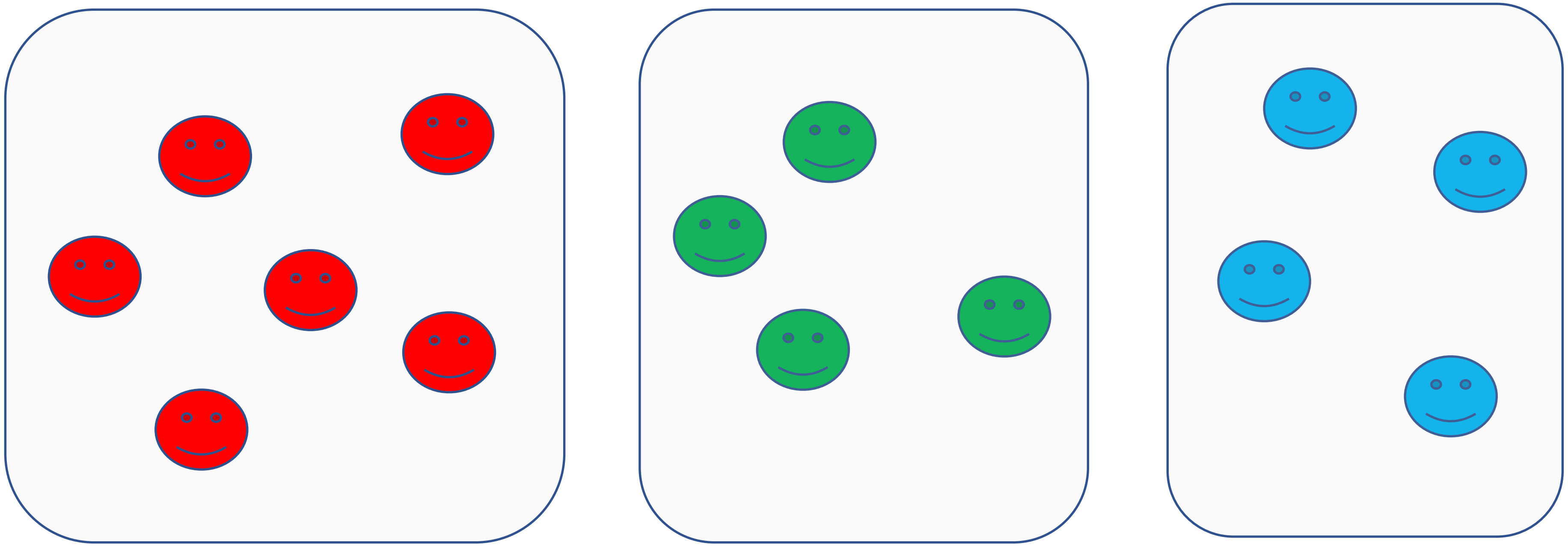
Our paper

- **Mixed community setting**



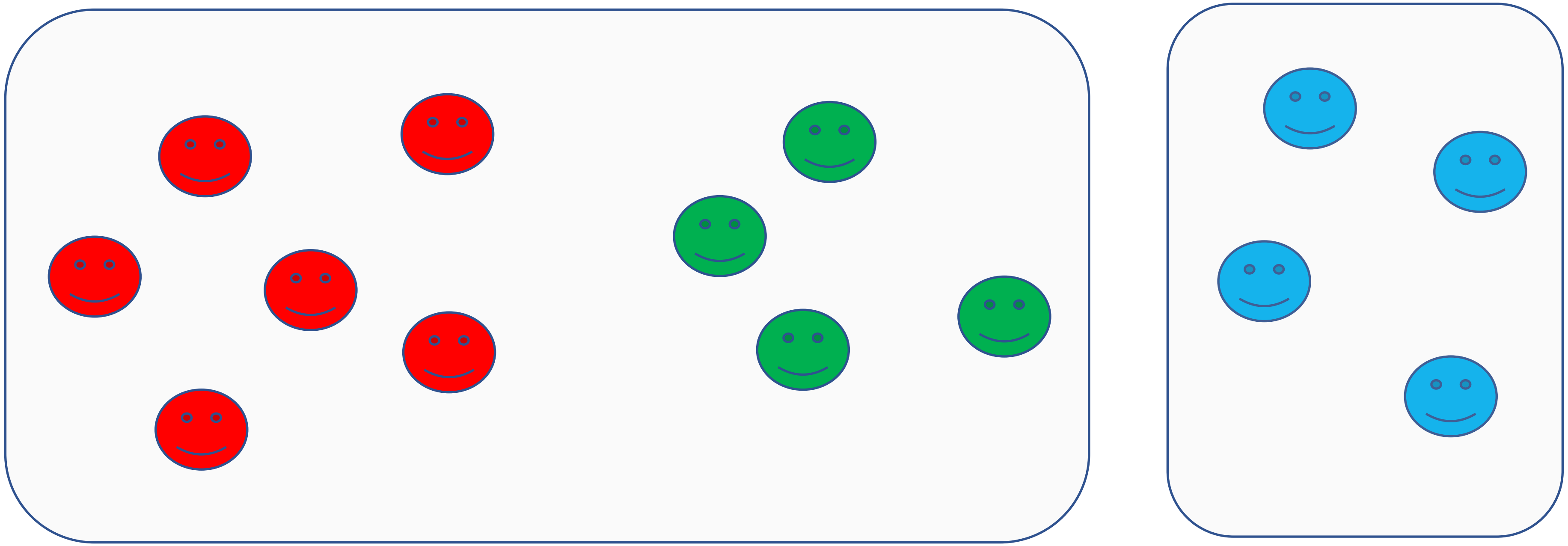
Our paper

- Mixed community setting
- **Separated community setting**



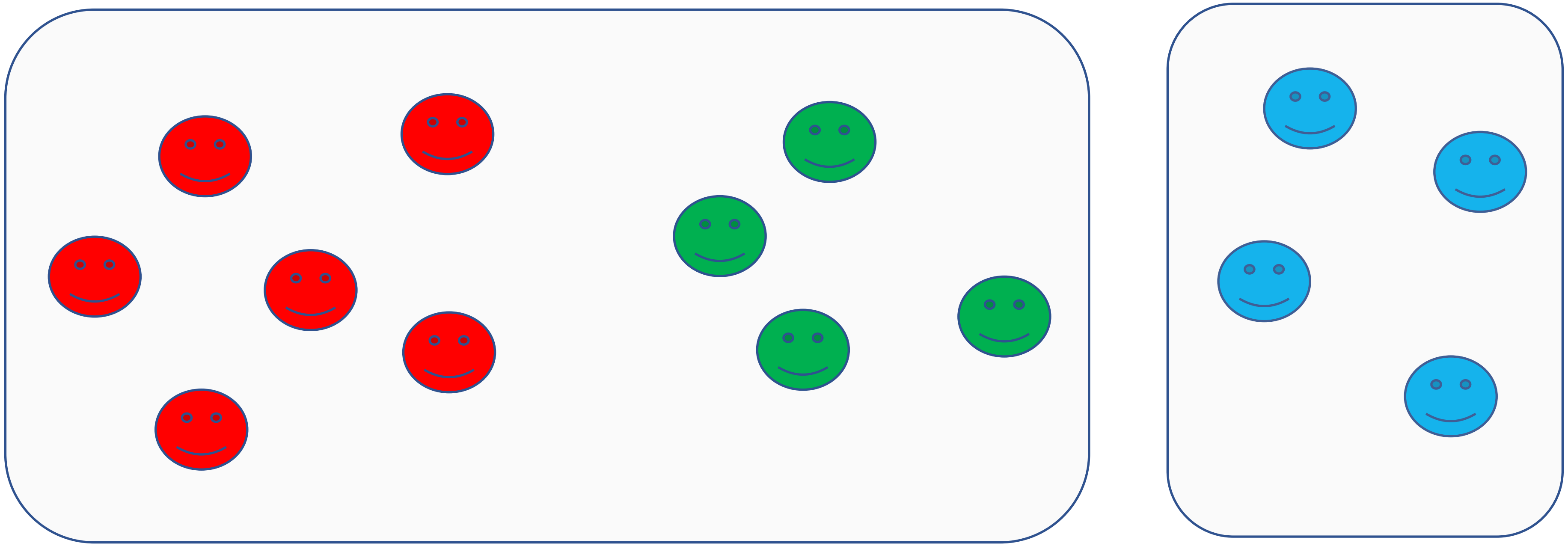
Our paper

- Mixed community setting
- Separated community setting
- **Community-disjoint box setting**

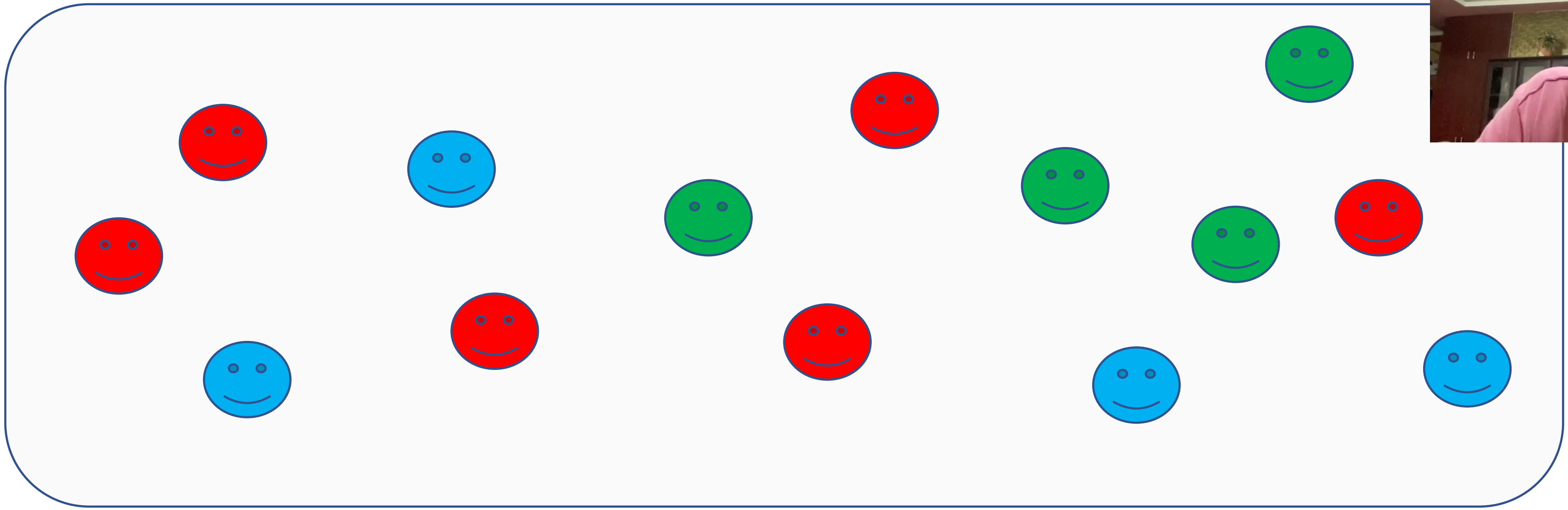


Our paper

- Mixed community setting
- Separated community setting
- **Community-disjoint box setting => general case**

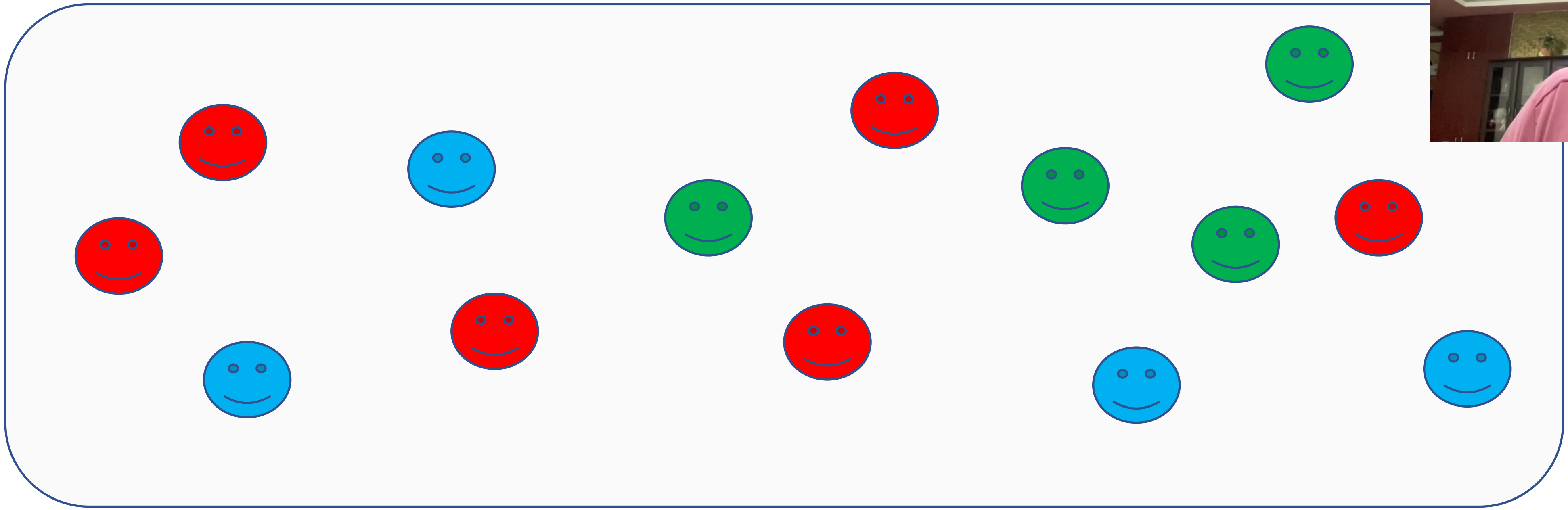


Mixed community setting

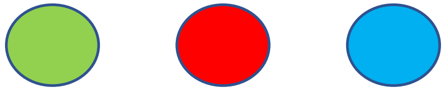


- No sampling control here; individuals sampled uniformly at random
- Baseline scenario: *identityless* sampling

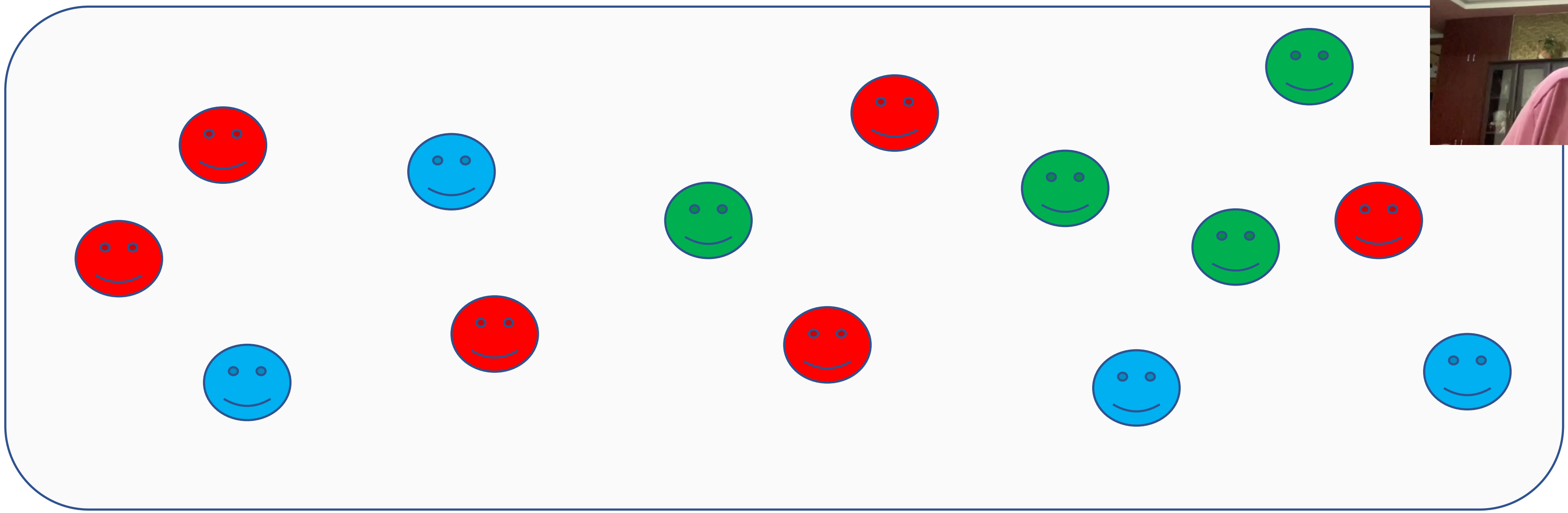
Mixed community setting



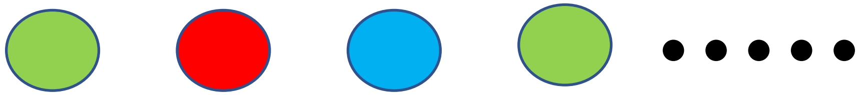
- No sampling control here; individuals sampled *i.i.d.*, uniformly at random
- Baseline scenario: *identityless* sampling



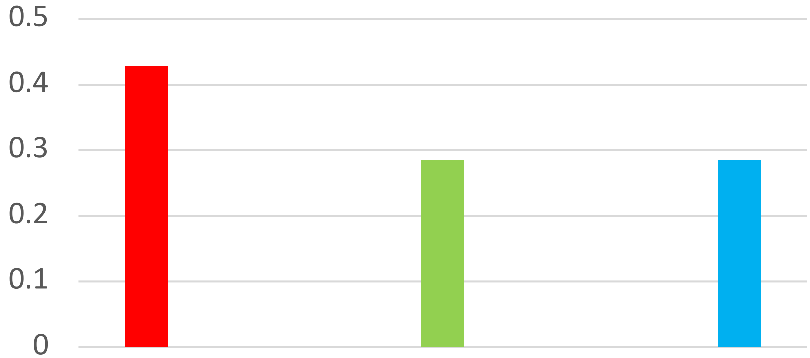
Mixed community setting



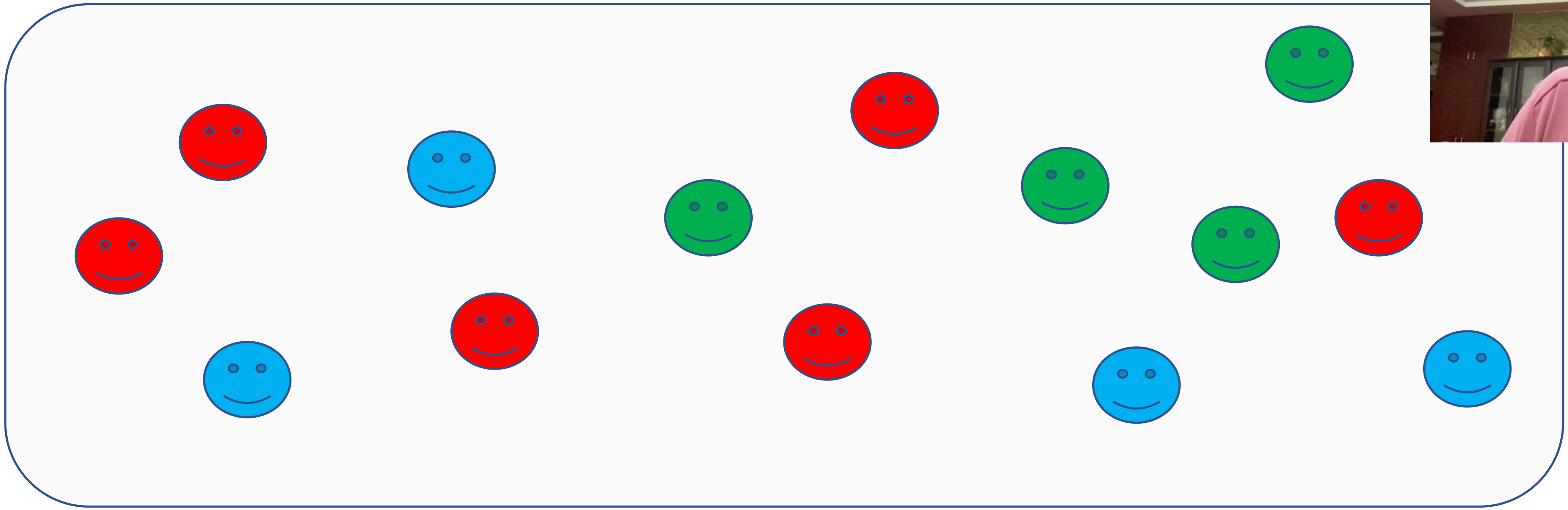
- No sampling control here; individuals sampled *i.i.d.*, uniformly at random
- Baseline scenario: *identityless* sampling



i.i.d. samples from distribution



Mixed community setting



- No sampling control here; individuals sampled *i.i.d.*, uniformly at random
- Baseline scenario: *identityless* sampling

  Natural algorithm: Output community with max # of samples

  Theorem: This algorithm satisfies

$$P_e \leq (m-1)\exp\left(-t\log\left(\frac{N}{N-(\sqrt{d_1}-\sqrt{d_2})^2}\right)\right)$$

Decay rate is focus here
d_1 - largest comm. size
d_2 - 2nd largest comm. size
Decay rate is optimal

Mixed community setting

<u>Identity-based sampling</u>

Algorithm: Output community with max # of *distinct individuals* seen

Theorem: This algorithm satisfies:

$$P_e \leq \binom{d_1}{d_2} \exp\left(-t \log\left(\frac{N}{N-(d_1-d_2)}\right)\right)$$

$$\log\left(\frac{N}{N-(d_1-d_2)}\right) \gg \log\left(\frac{N}{N-(\sqrt{d_1}-\sqrt{d_2})^2}\right)$$

$\Rightarrow$ *identity information improves performance of mode estimation*

Mixed community setting

<u>Identity-based sampling</u>

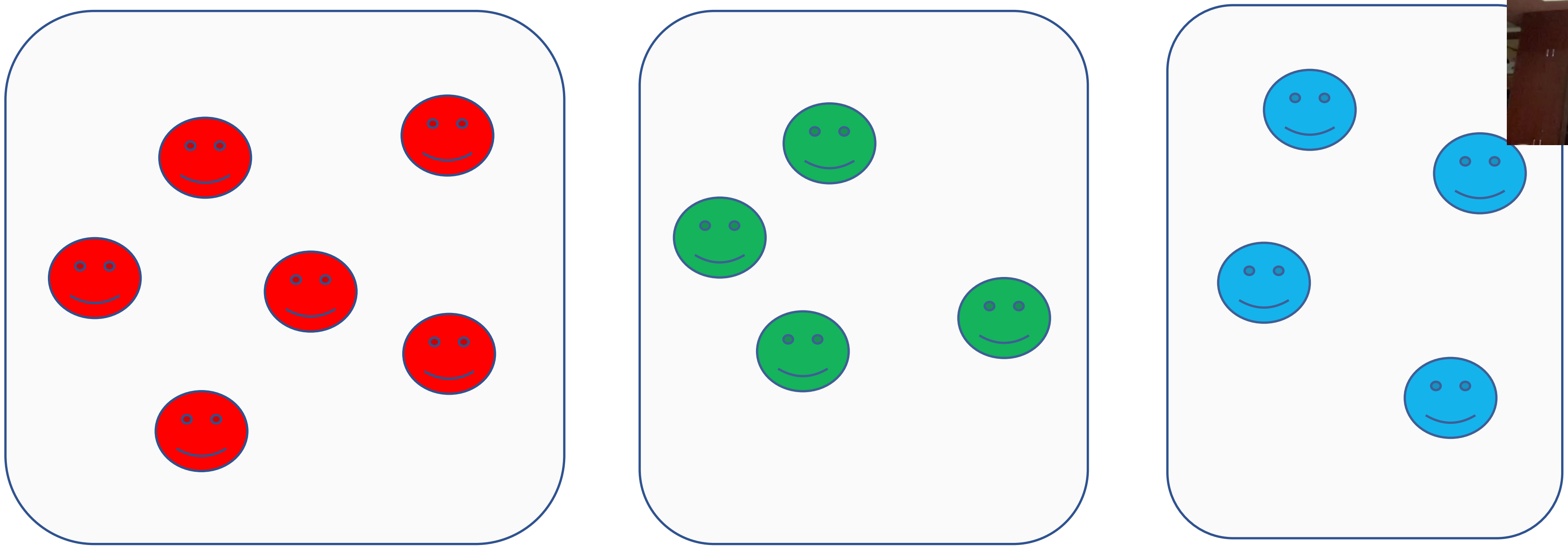
Algorithm: Output community with max # of *distinct individuals* seen

Theorem: This algorithm satisfies:

$$P_e \leq \binom{d_1}{d_2} \exp\left(-t \log\left(\frac{N}{N-(d_1-d_2)}\right)\right)$$

- This decay rate is optimal
- Result follows from a coupon collector style argument
- Error most likely caused by certain set of individuals of largest community never getting sampled

Separated community setting

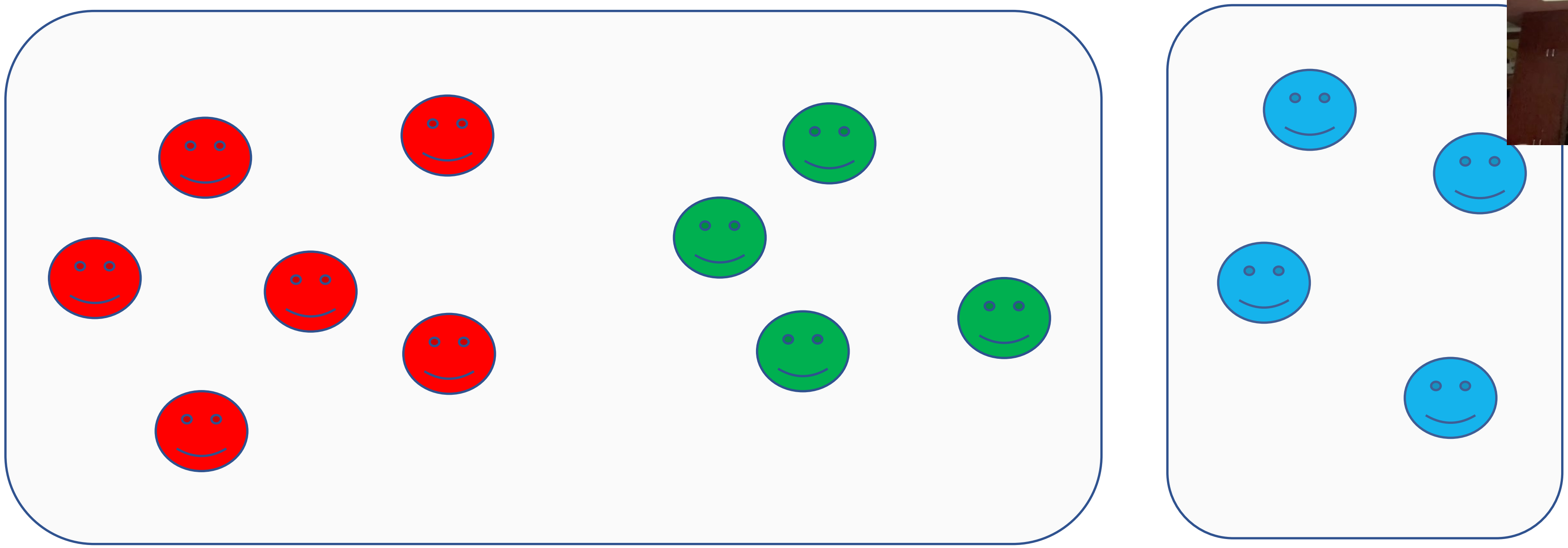


Successive elimination style algorithm:

- Partition the learning budget into $b-1$ phases
- Eliminate the `worst looking' box at the end of each phase *(metric: # of distinct individuals seen)*
- Uniform sampling among `surviving' boxes in each phase

Decay rate optimal up to a logarithmic (in b) factor *(lower bound uses MAB-style change of measure argument)*

Community-disjoint box setting



Combines elements from previous settings. Two sub-tasks:

a) Identify the box containing largest community (successive elimination style algorithm)
b) Identify largest community from said box (mixed community mode estimation)

Lower bound matches upper bound decay rate up to log. factors for a broad class of instances
Sub-task (a) is `harder’

Concluding remarks

- Online community mode estimation using semi-targeted querying
  - Algorithms & information theoretic lower bounds
- Several generalizations
  - Sampling without replacement
  - Fixed confidence variant
  - MAB problems with imprecise arm selection
  - *Trade-off between privacy and learning efficiency*

Sequential Community Mode Estimation

Shubham Jain, Shreyas Goenka, Divyam Bapna, Nikhil Karamchandani & **Jayakrishnan Nair**

(Department of Electrical Engineering, IIT Bombay)