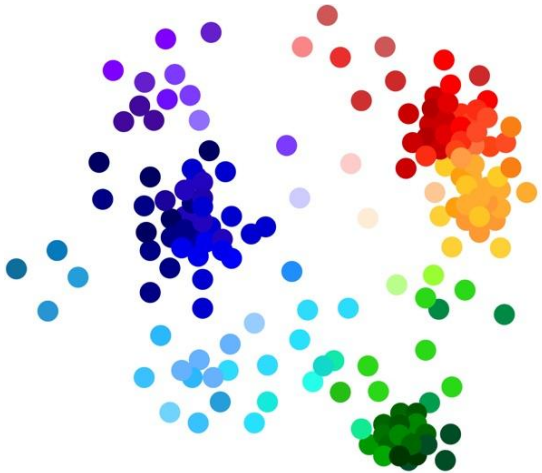
RICE

Diversified Streaming Sampling (for Terabyte-Scale Genomics)

Ben Coleman*, Benito Geordie*
Li Chou, R. A. Leo Elworth, Todd Treangen, Anshumali Shrivastava

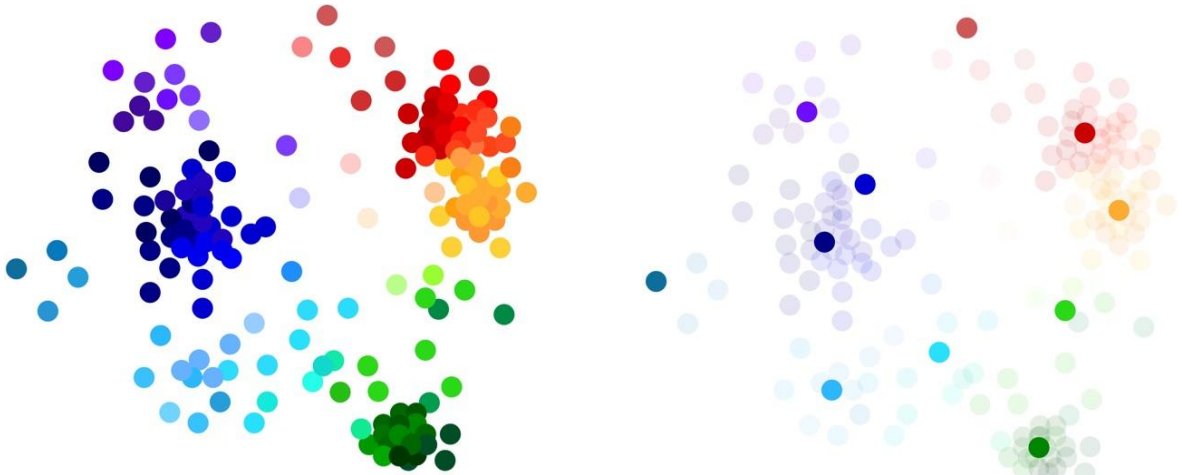
Diversified Sampling?

Choose a *representative subset*



Diversified Sampling?

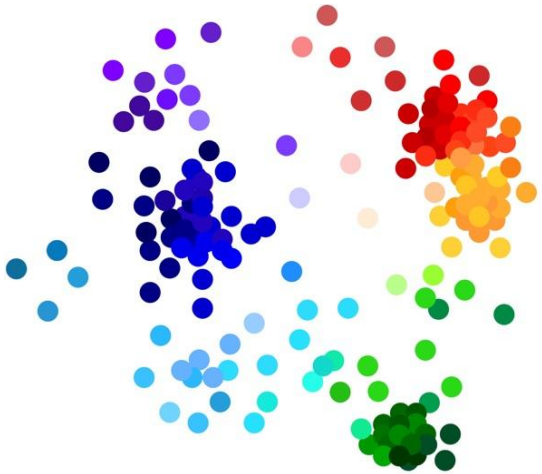
Choose a *representative subset*

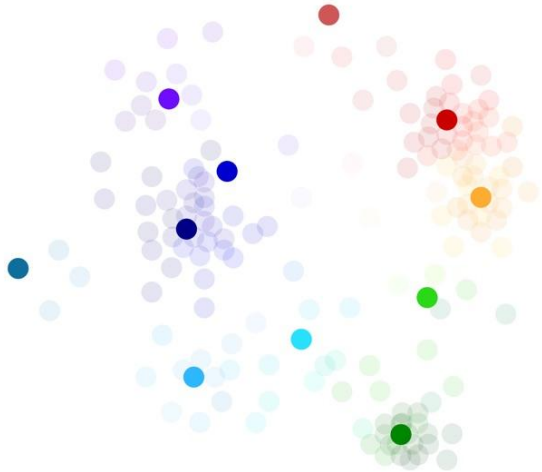


Diverse

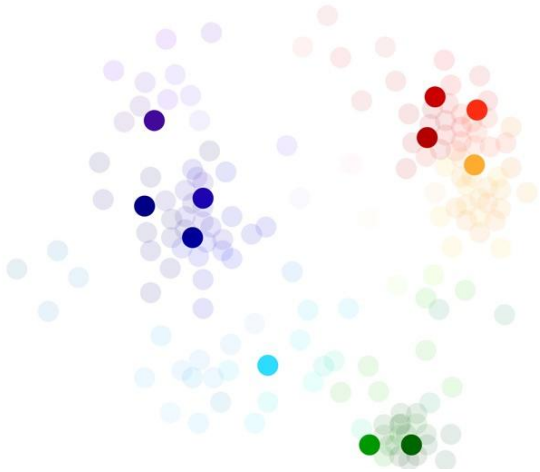
Diversified Sampling?

Choose a *representative subset*





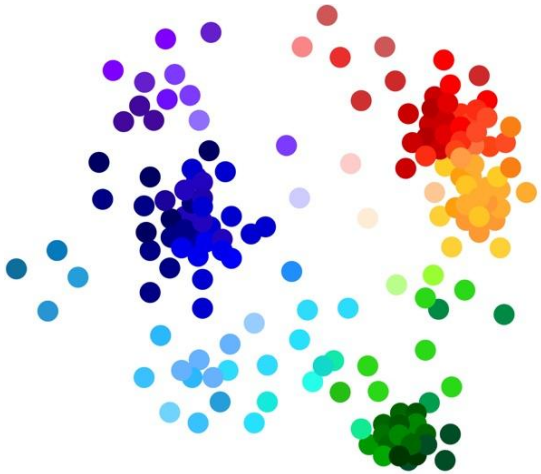
Diverse

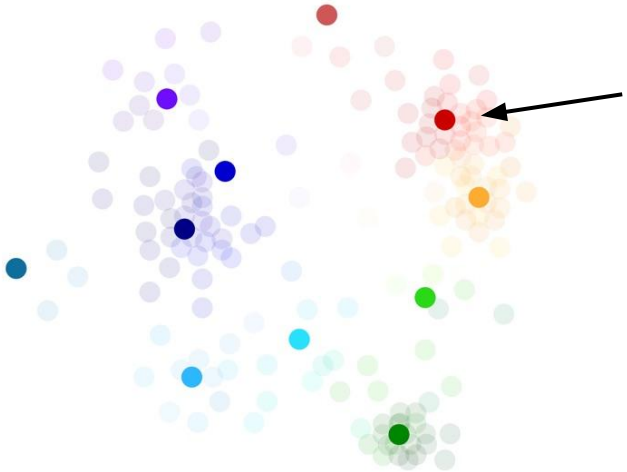


Non-diverse

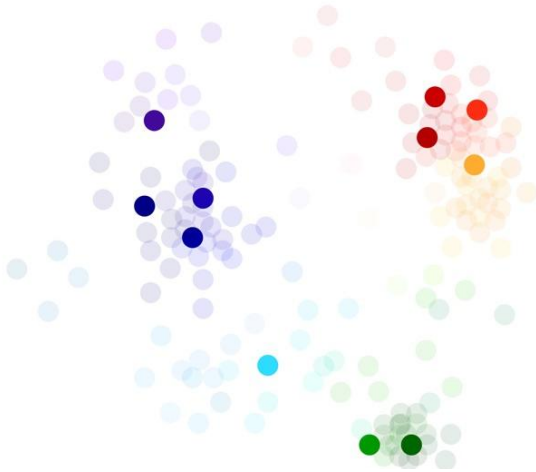
Diversified Sampling?

Choose a *representative subset*





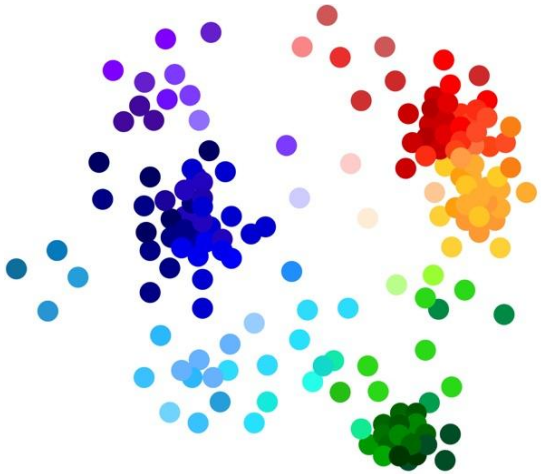
Diverse

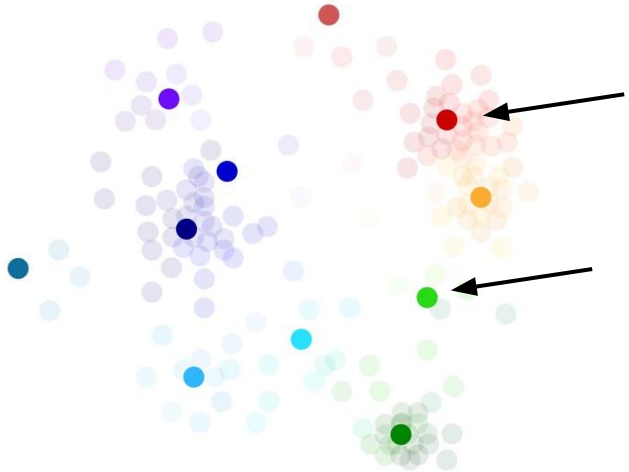


Non-diverse

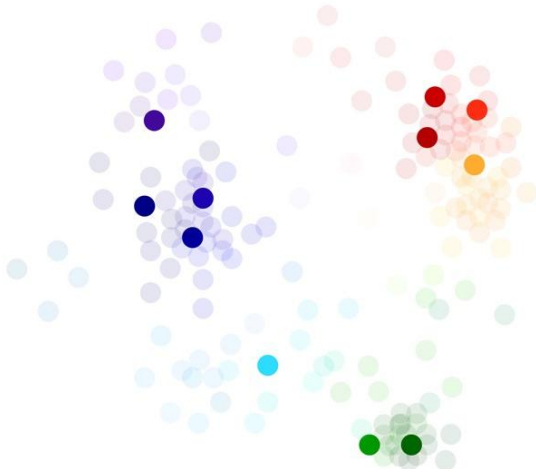
Diversified Sampling?

Choose a *representative subset*





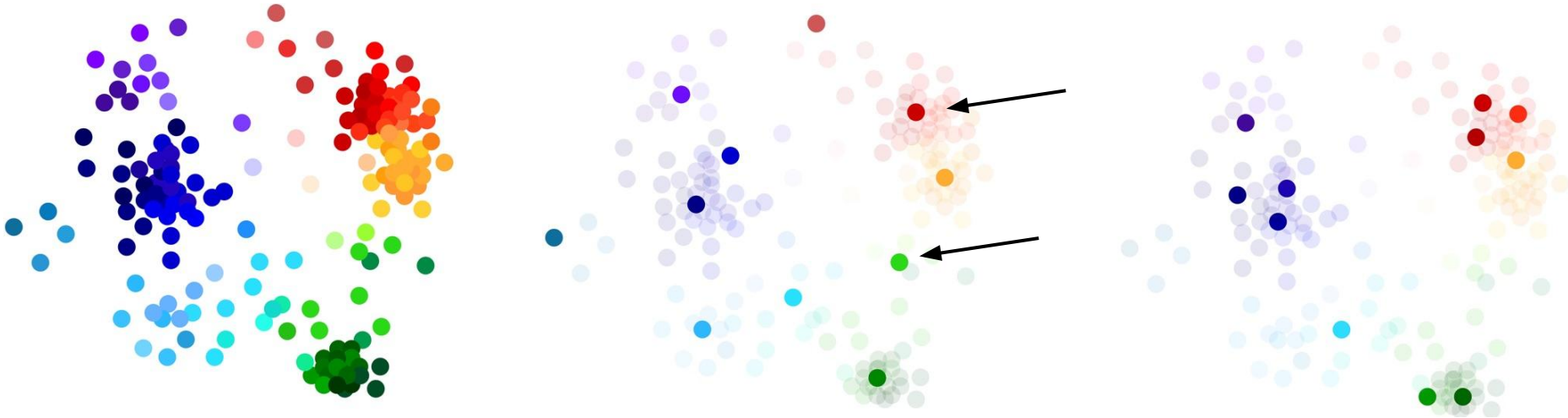
Diverse



Non-diverse

Diversified Sampling?

Choose a *representative subset*

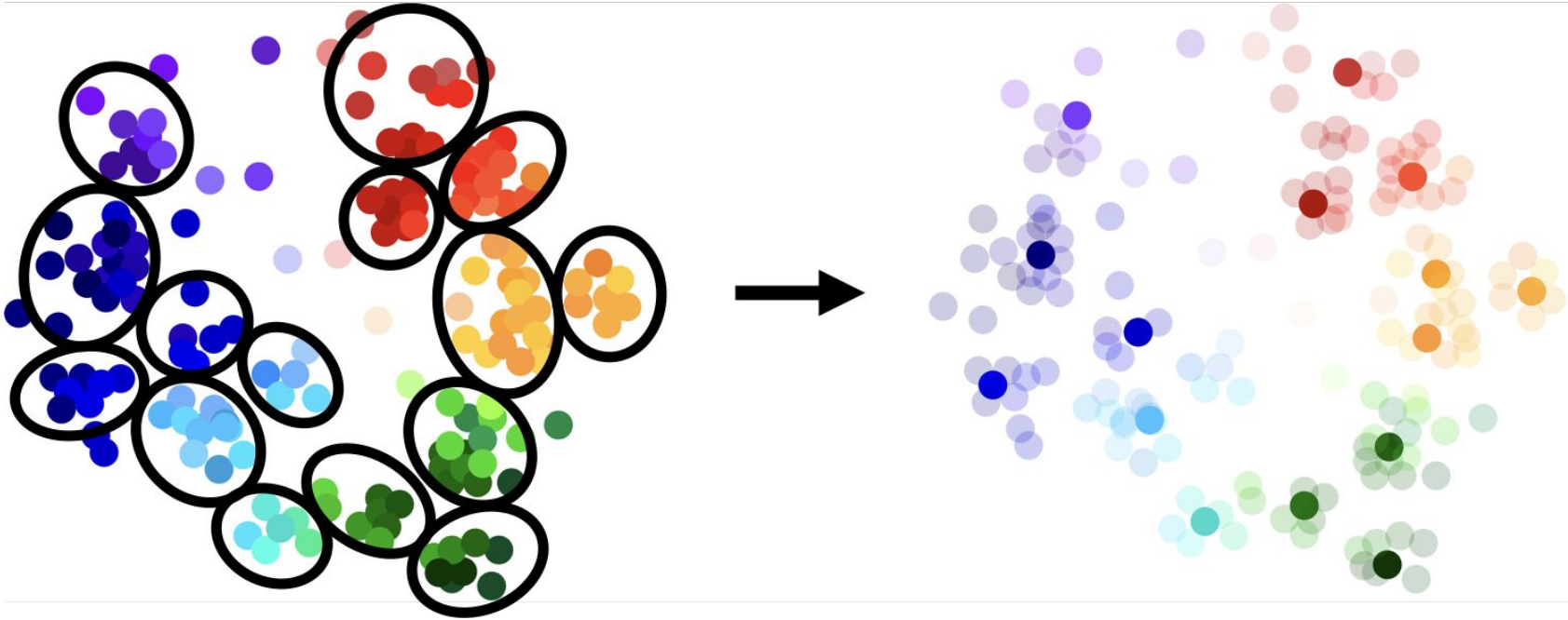


Diverse

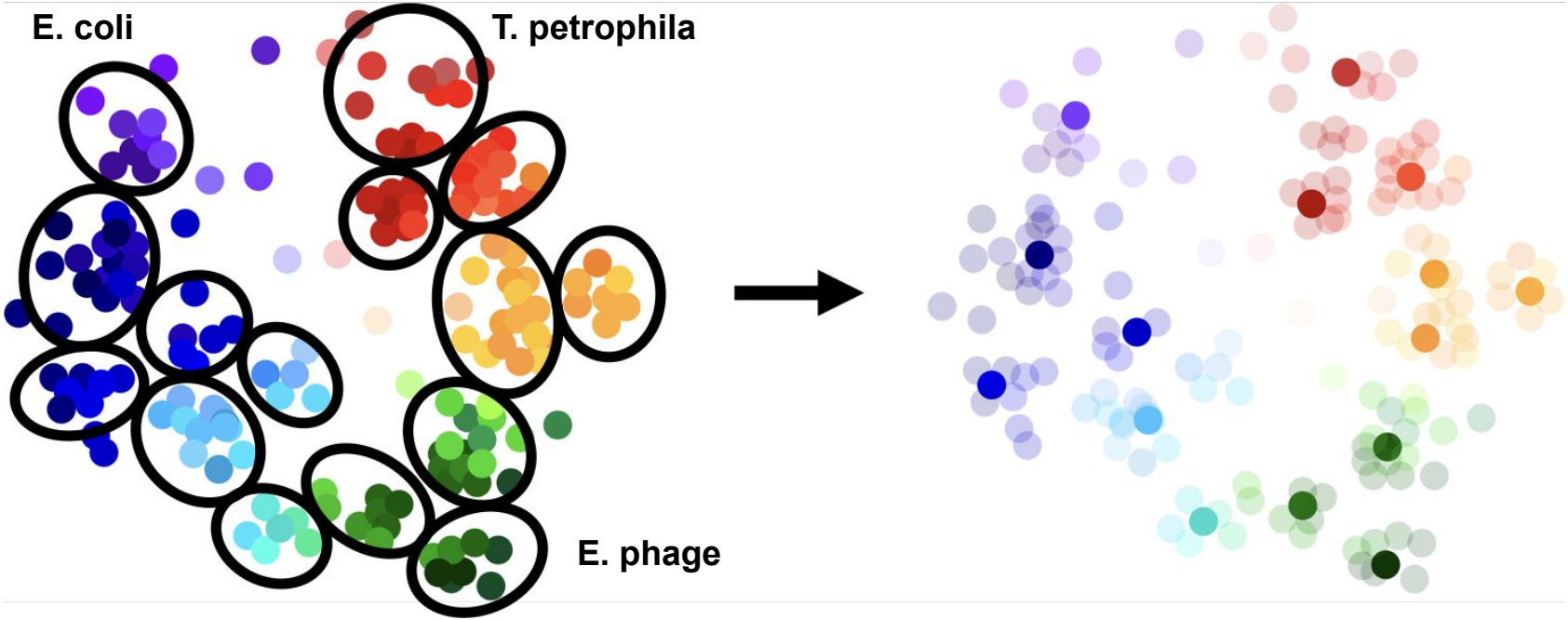
Non-diverse

Several **formal definitions** - see our paper

Roughly: Pick points from each cluster (class)

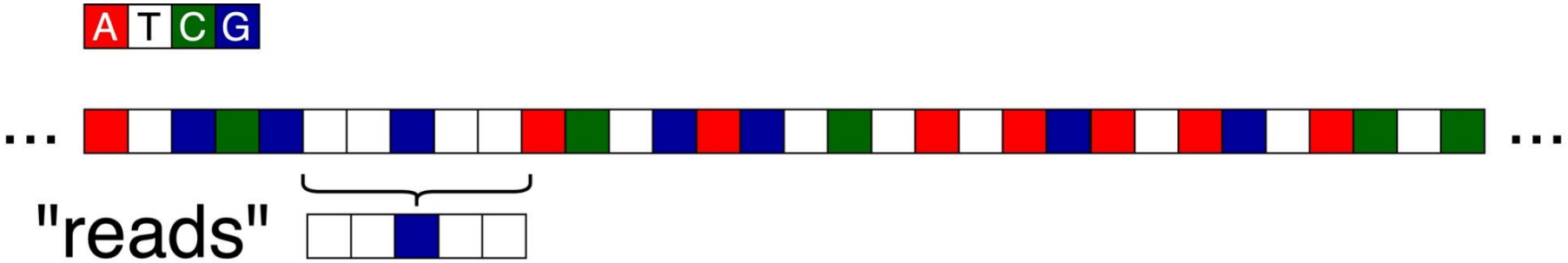


Metagenomic community profiling: *Cluster* = *Species*

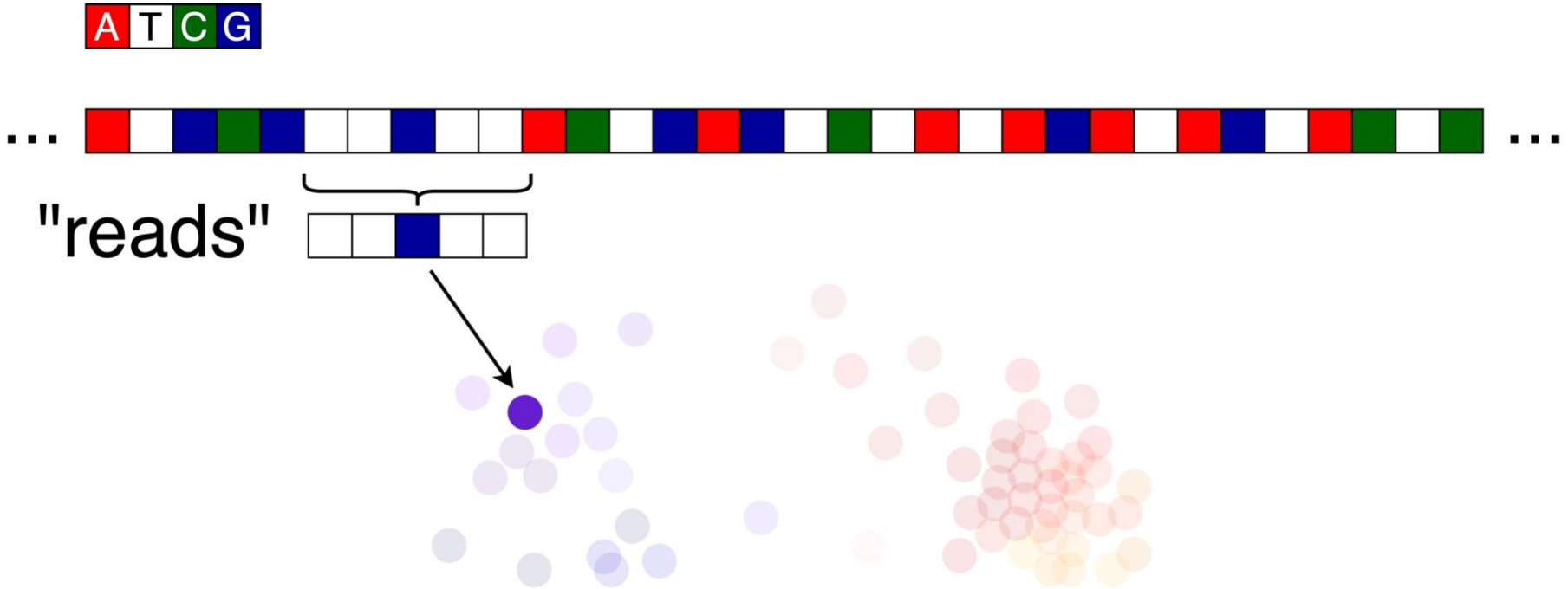


Disclaimer: Cluster model is for intuition only - real distributions are far more complex.

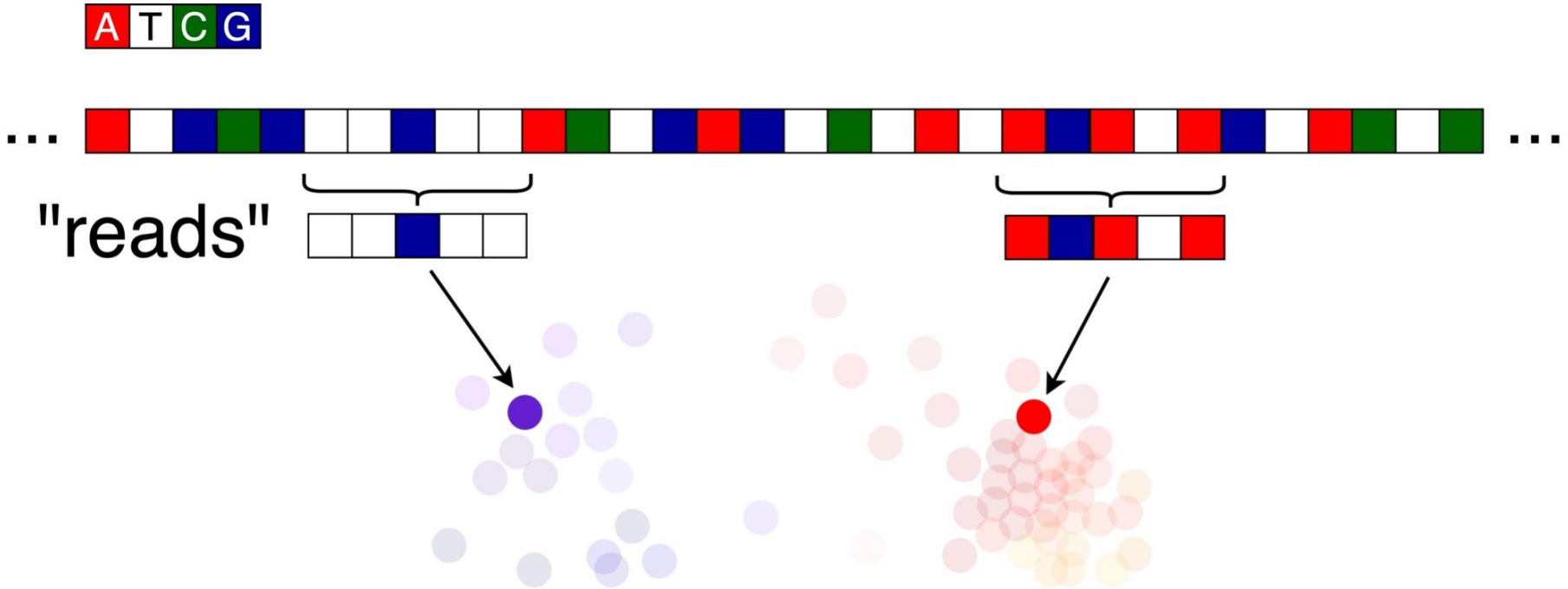
In bioinformatics, each point is a “read”



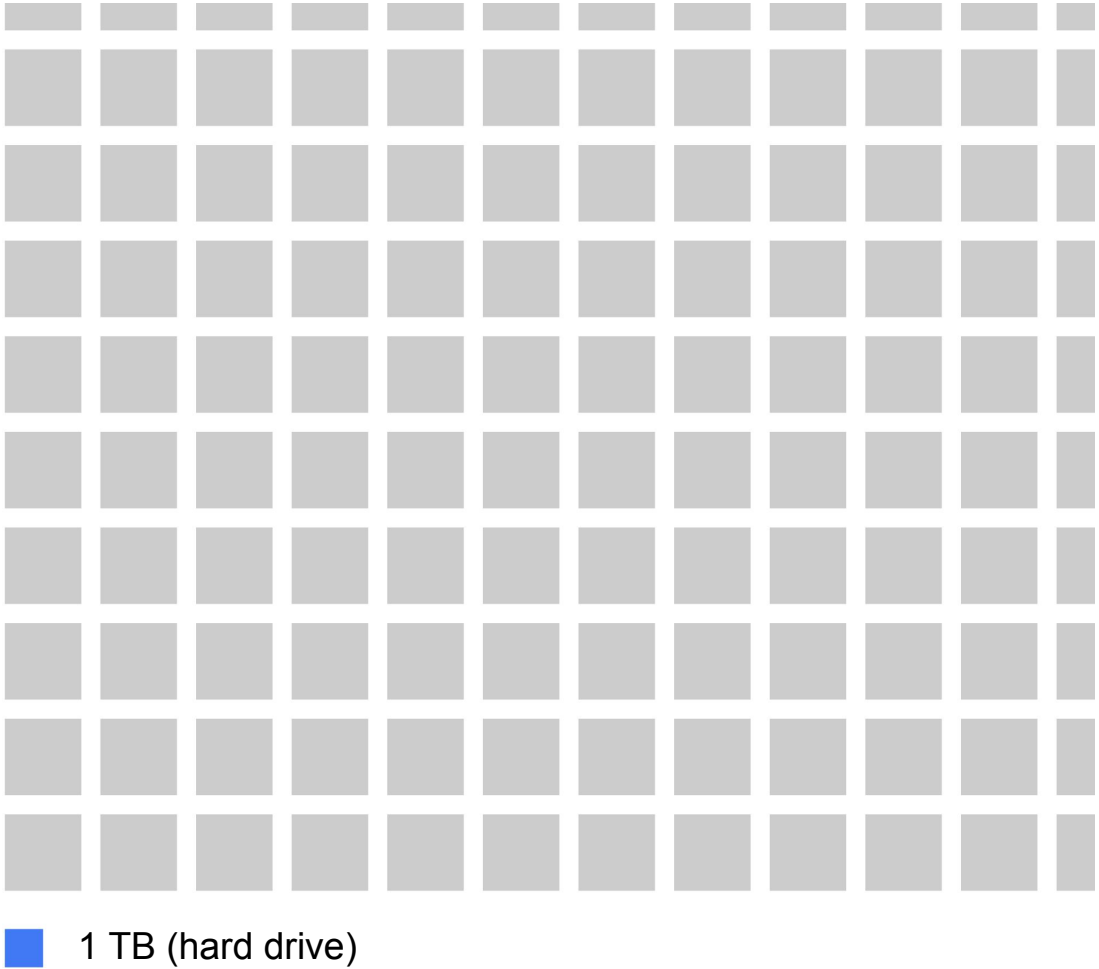
In bioinformatics, each point is a “read”



In bioinformatics, each point is a “read”

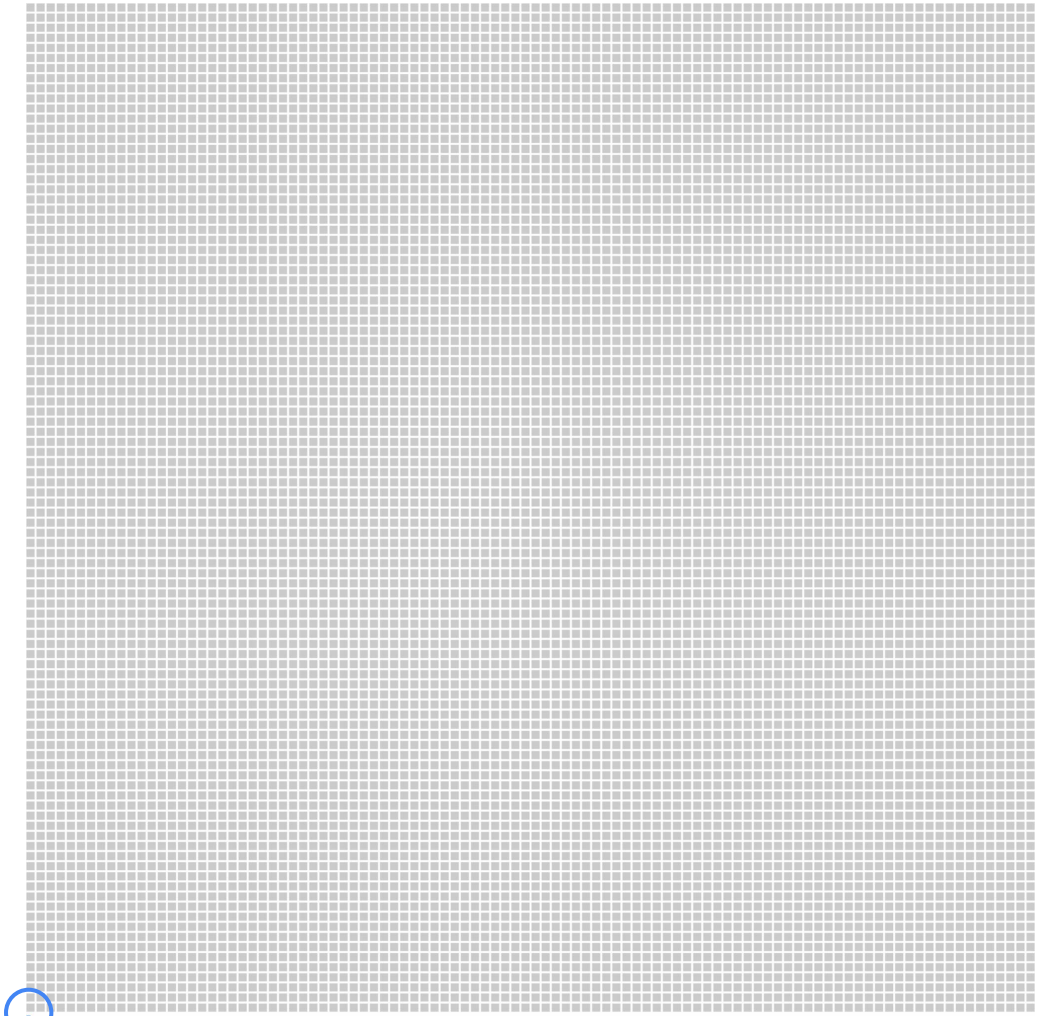


Why?



Sequence Read Archive

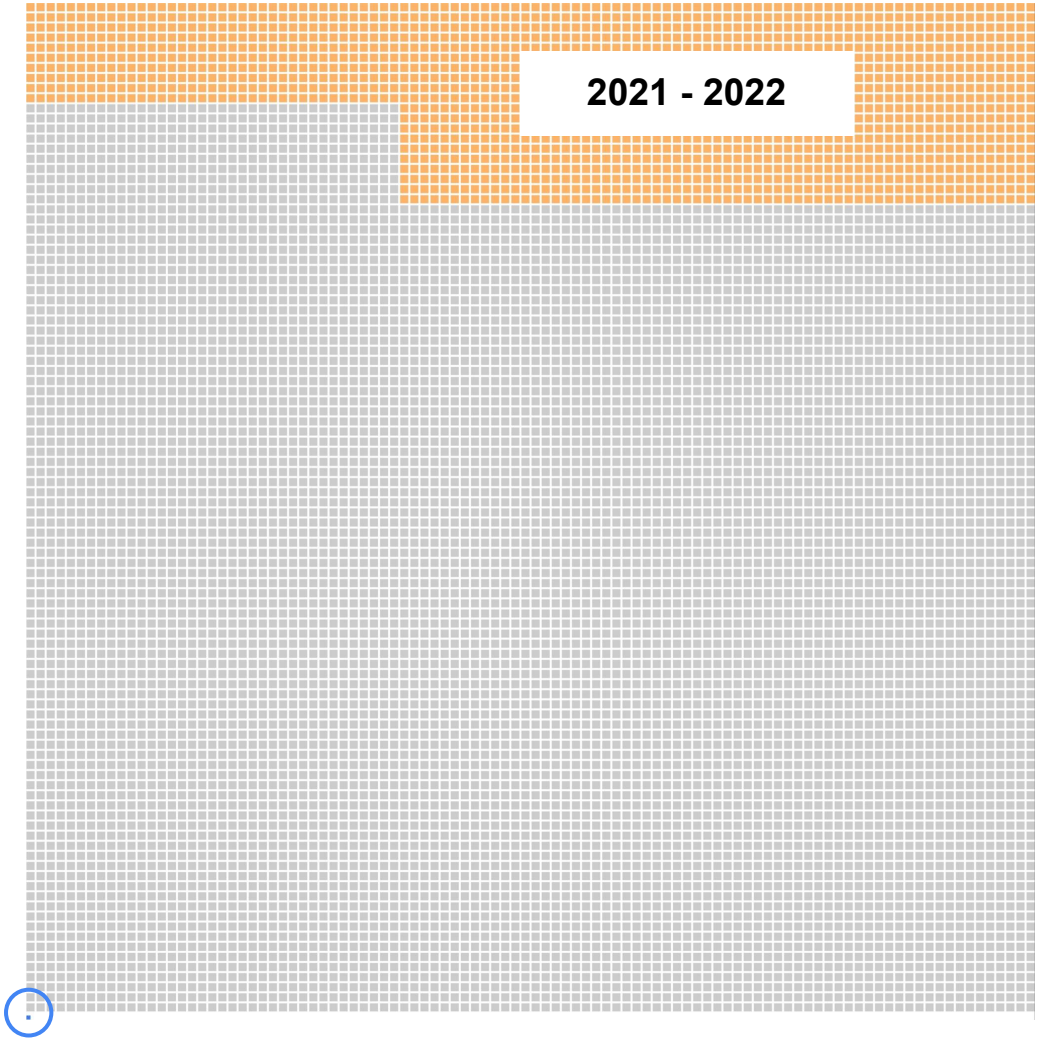
~ 43 petabytes



Hard drive

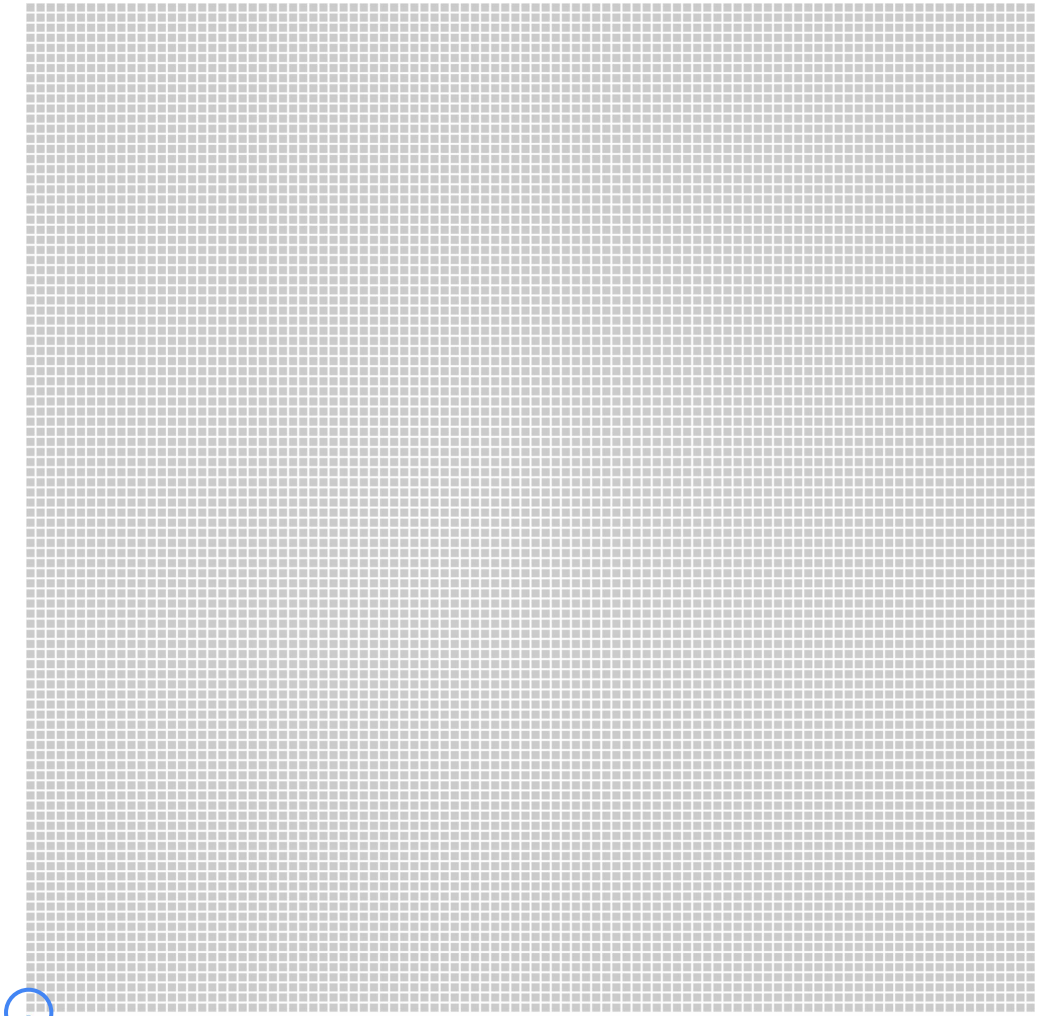
Sequence Read Archive

~ 43 petabytes



Diversified Sampling

Low memory
Fast - one pass!

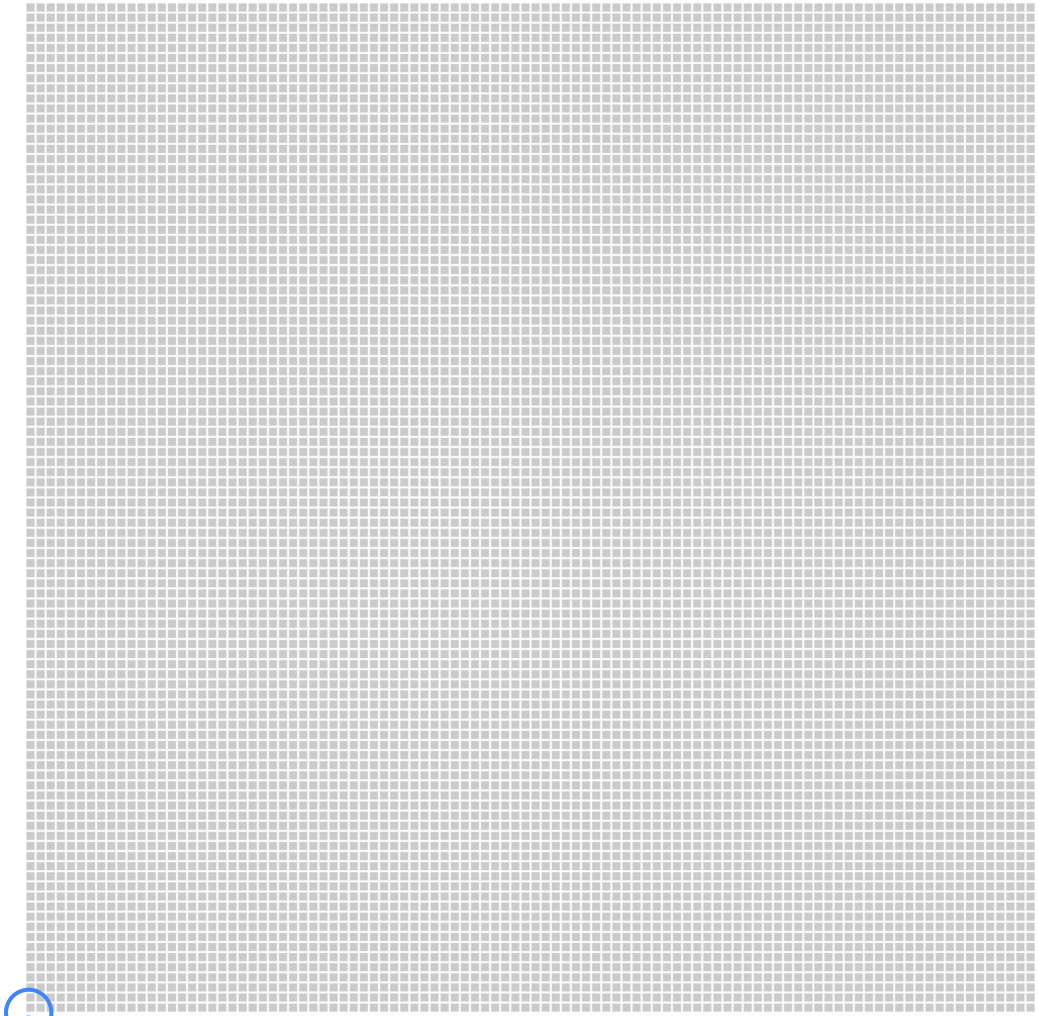


Hard drive

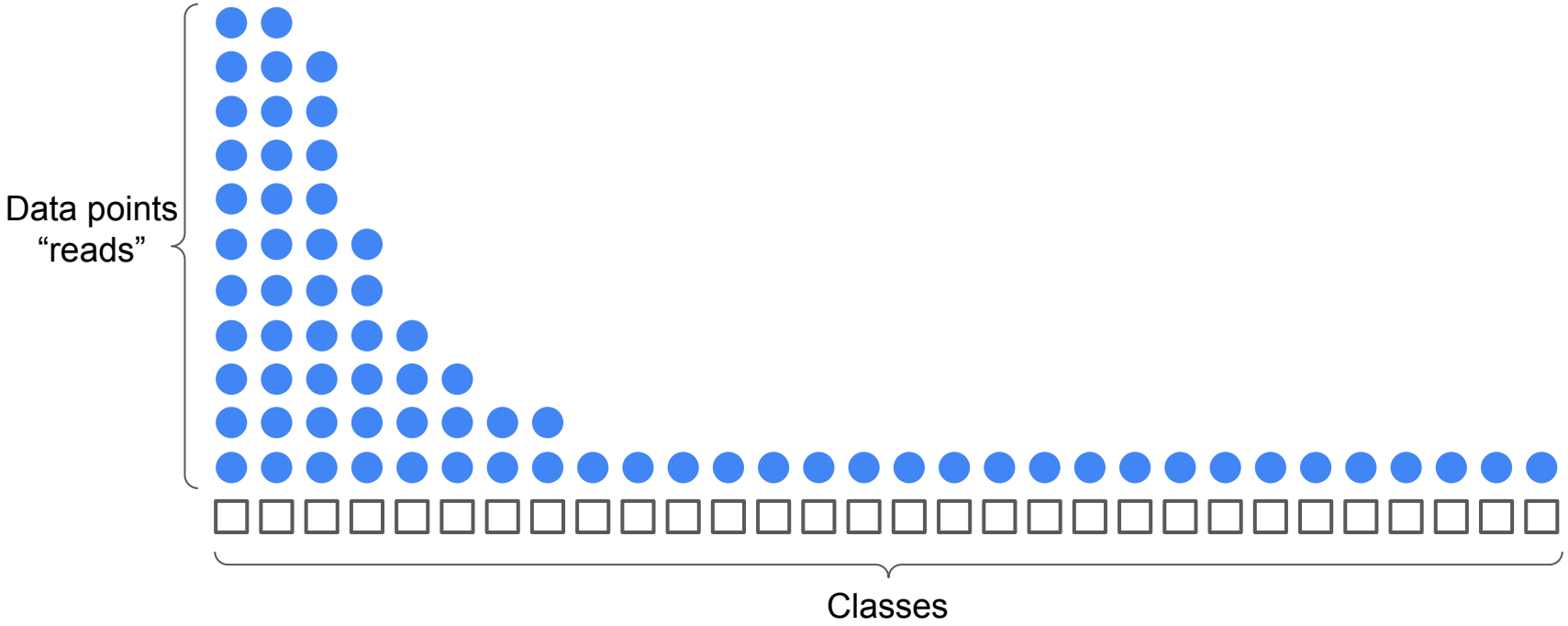
Diversified Sampling

Low memory
Fast - one pass!

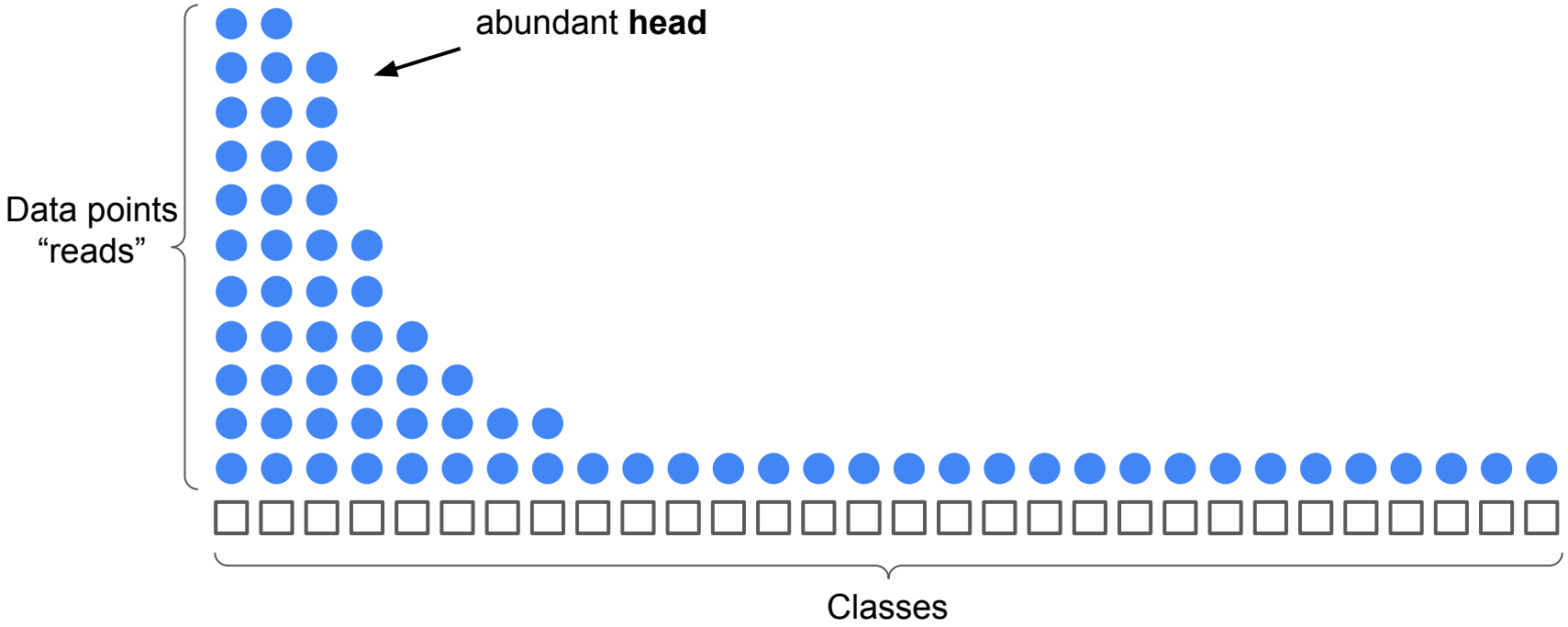
1000x less memory
2x faster



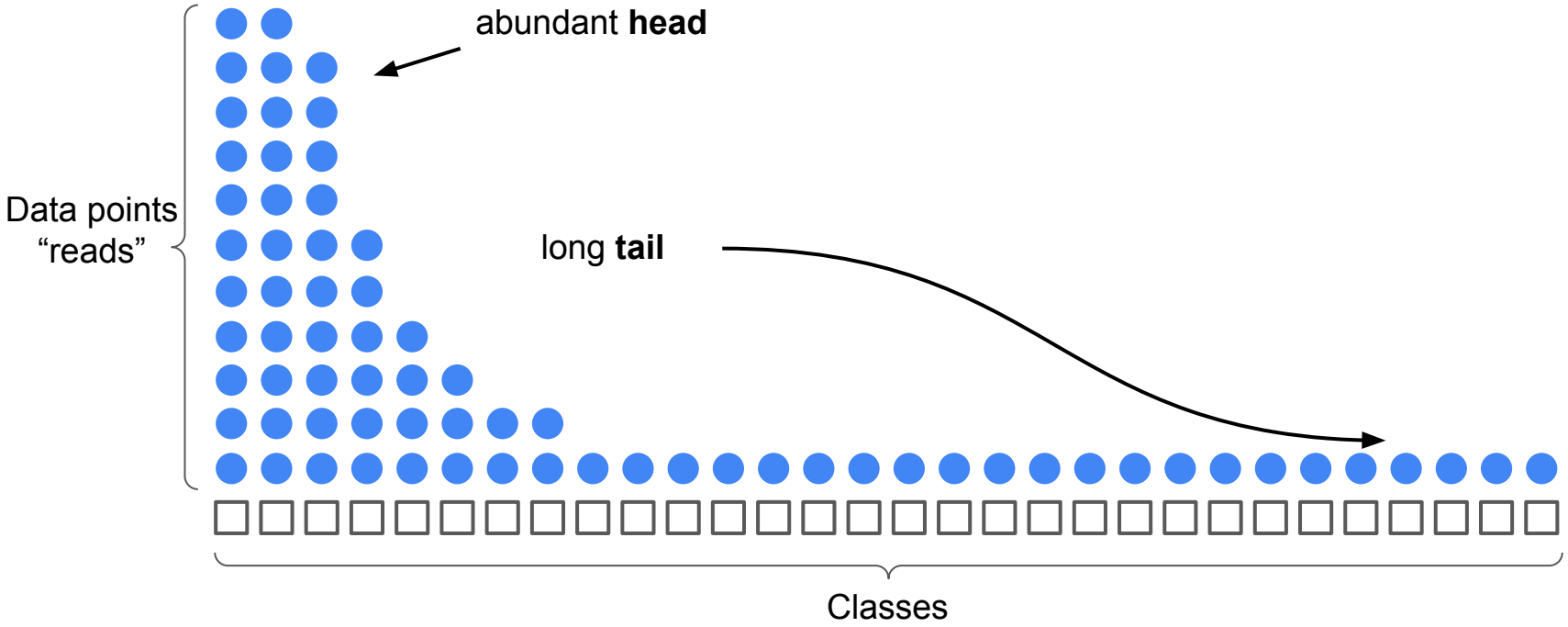
Our datasets



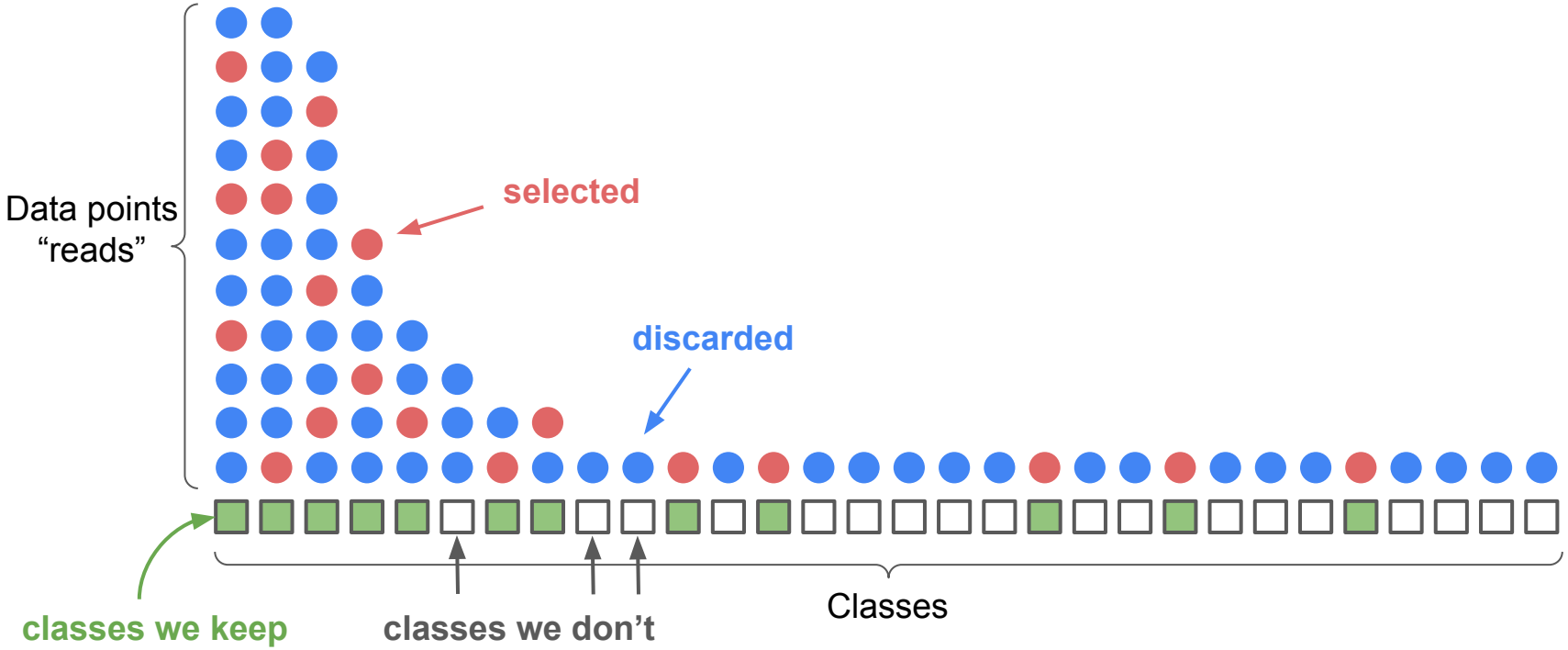
Our datasets



Our datasets

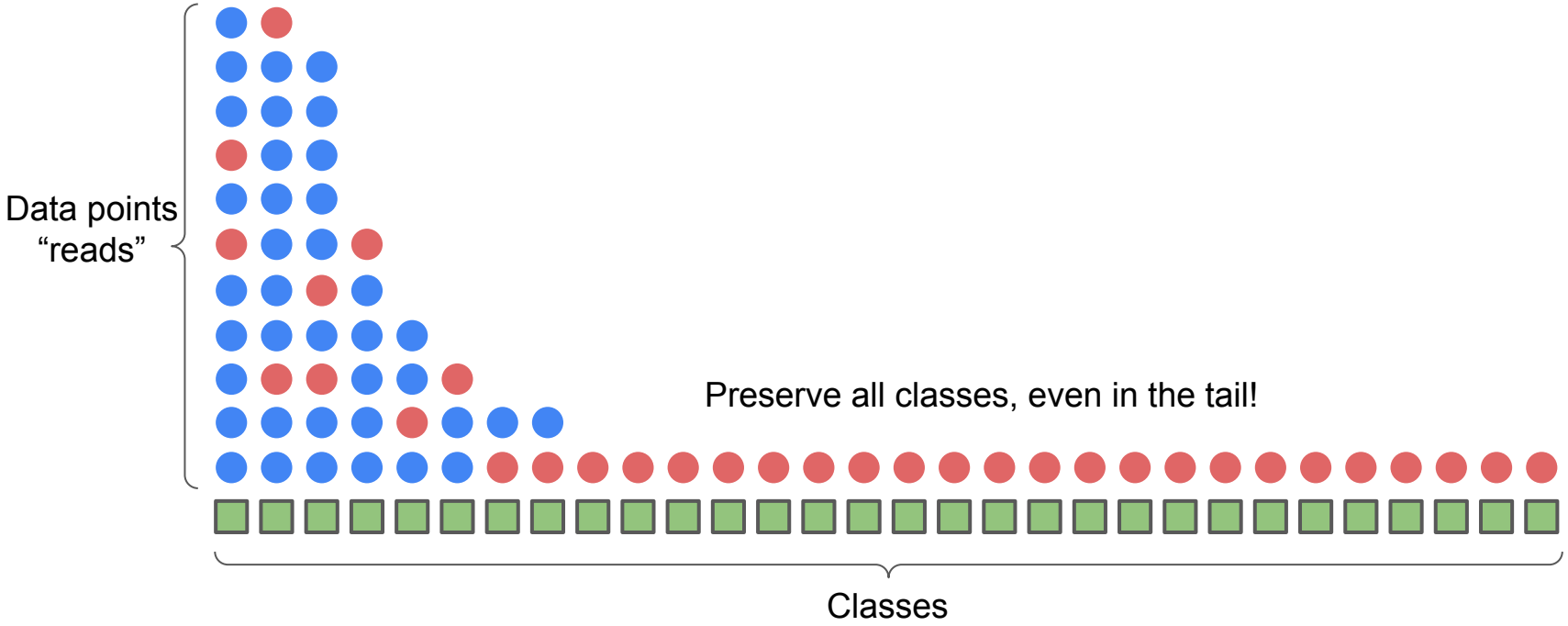


Random sampling

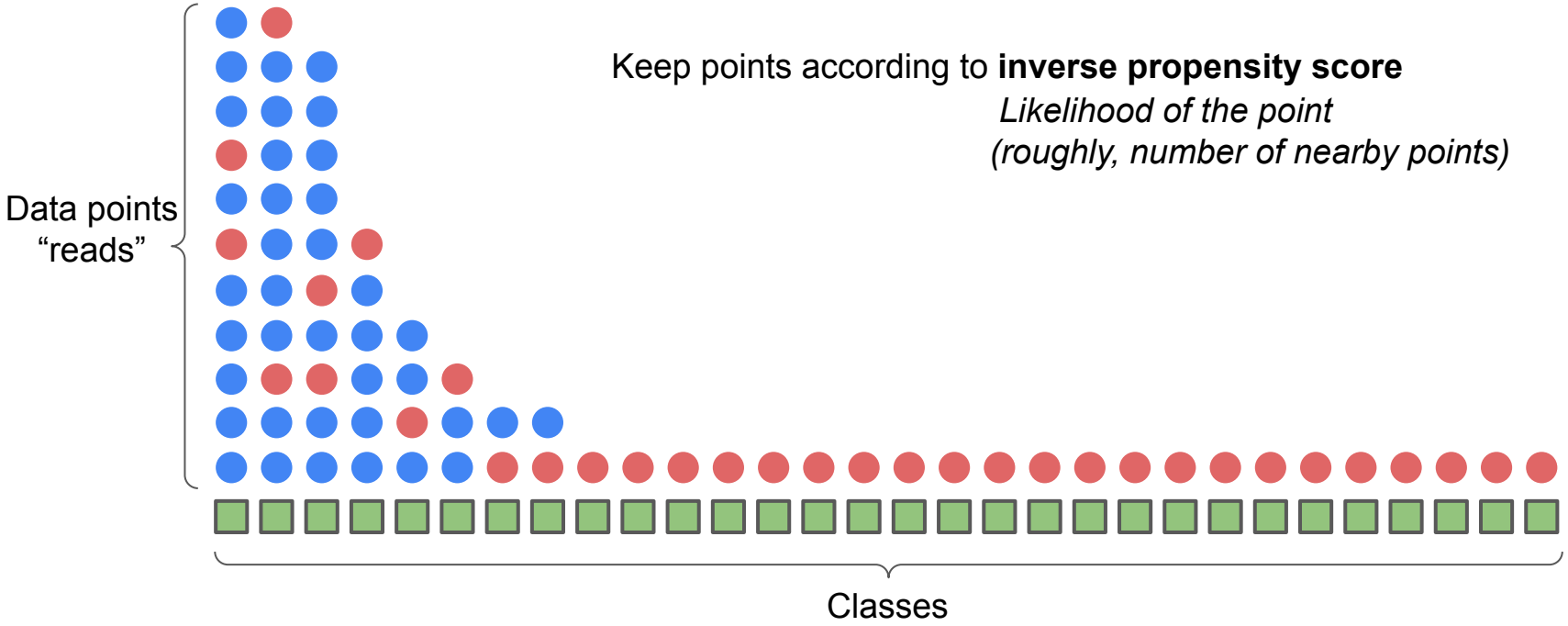


Random sampling

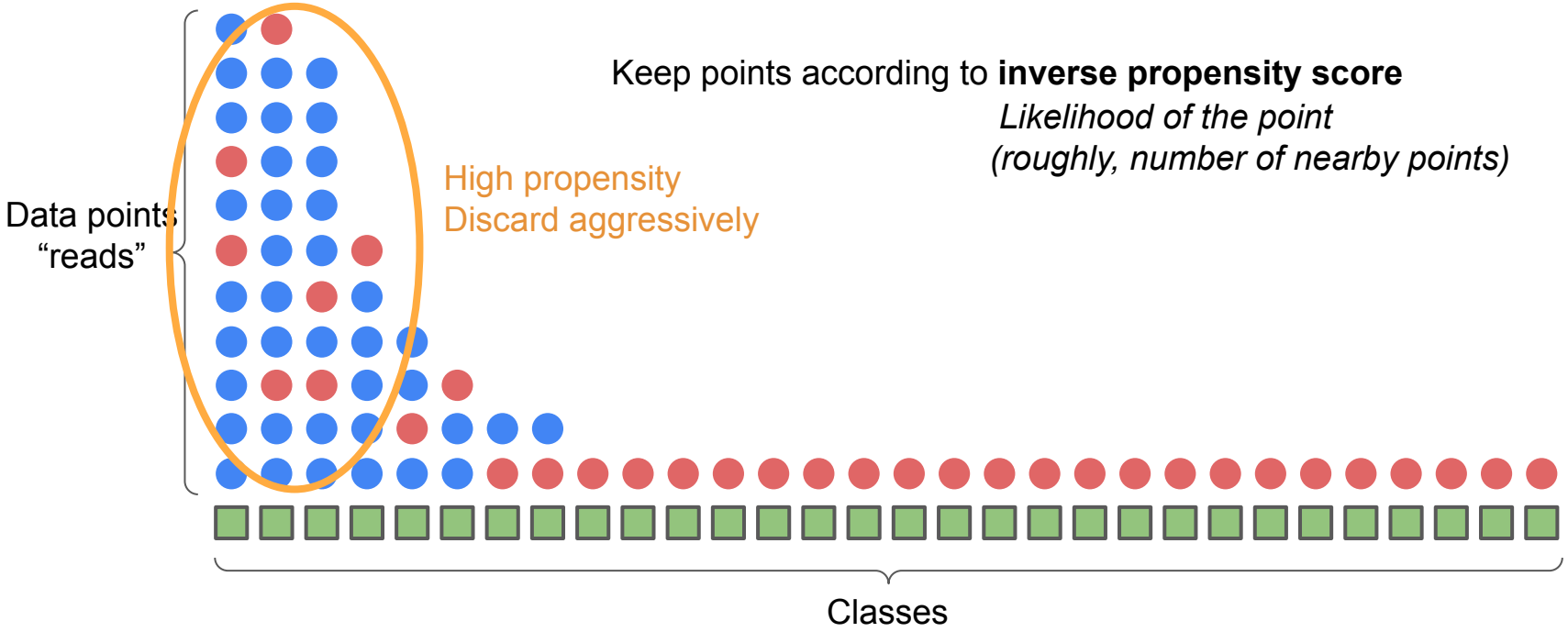
What we want


Data points
“reads”
Preserve all classes, even in the tail!
Classes

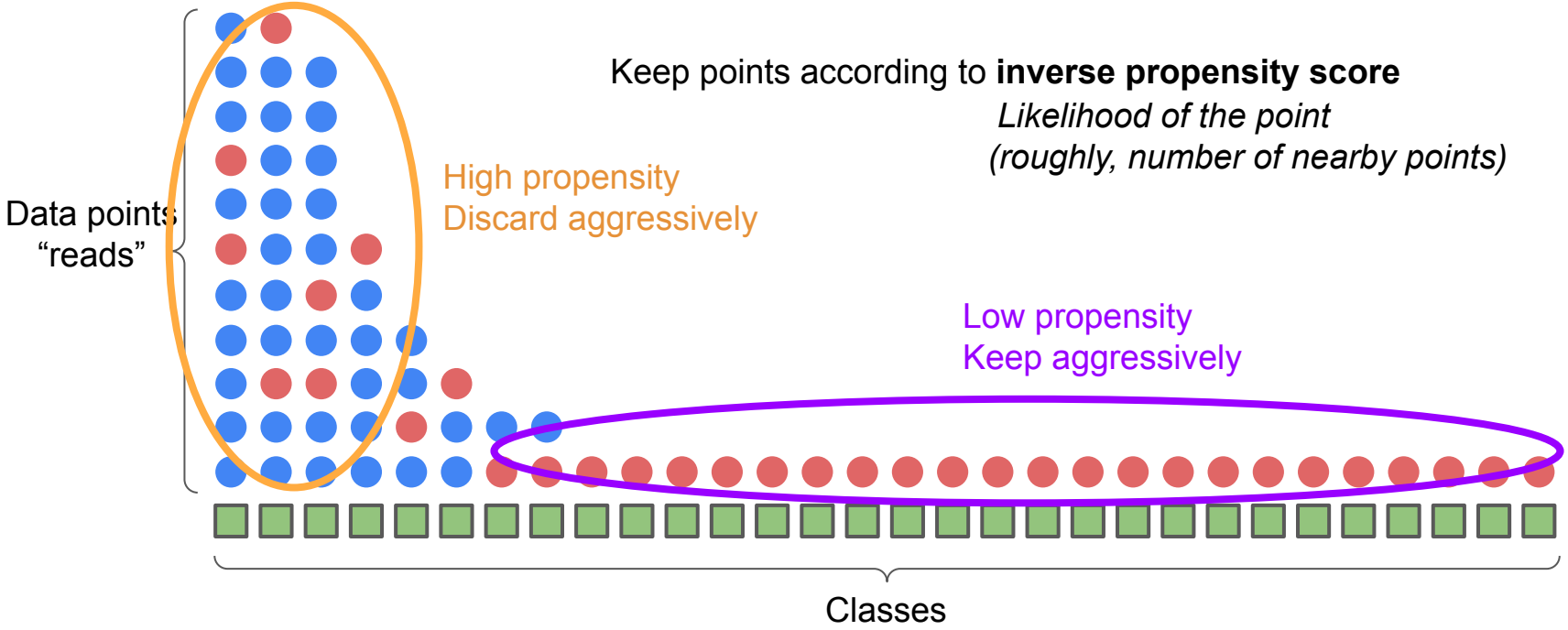
Inverse Propensity Sampling



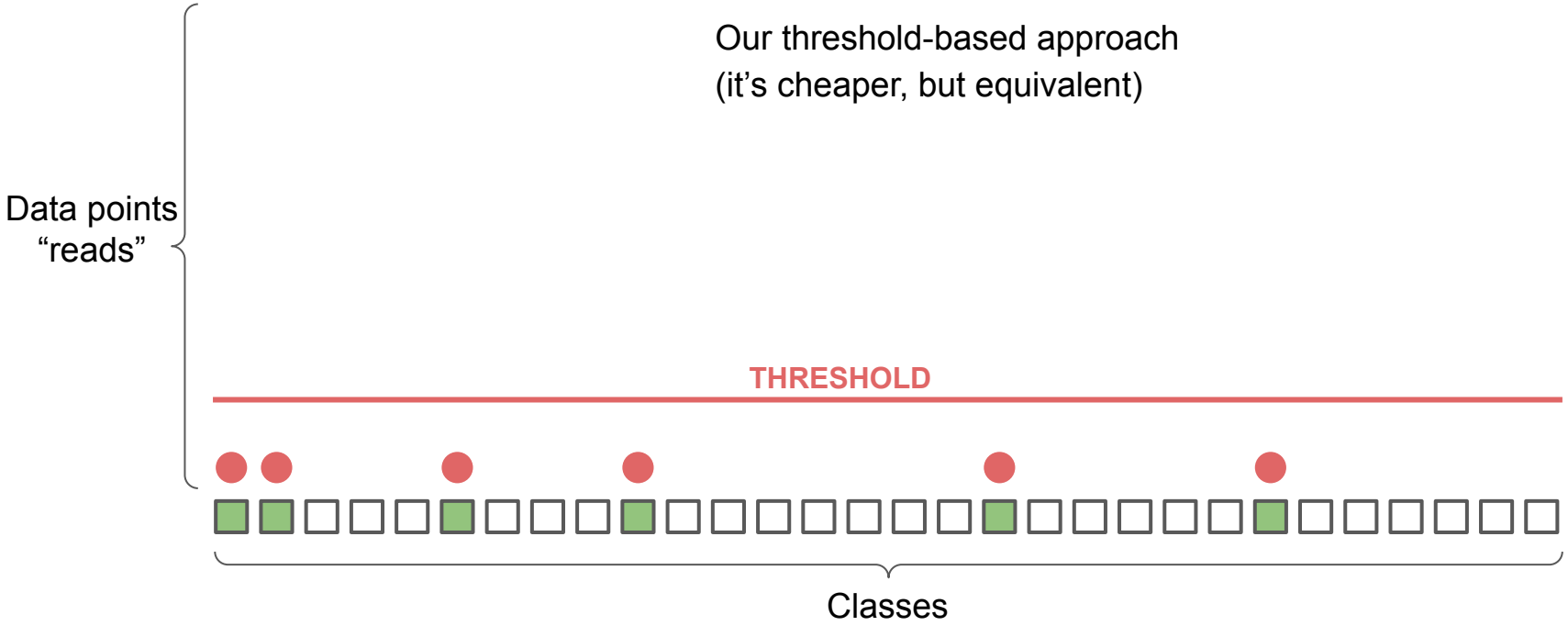
Inverse Propensity Sampling



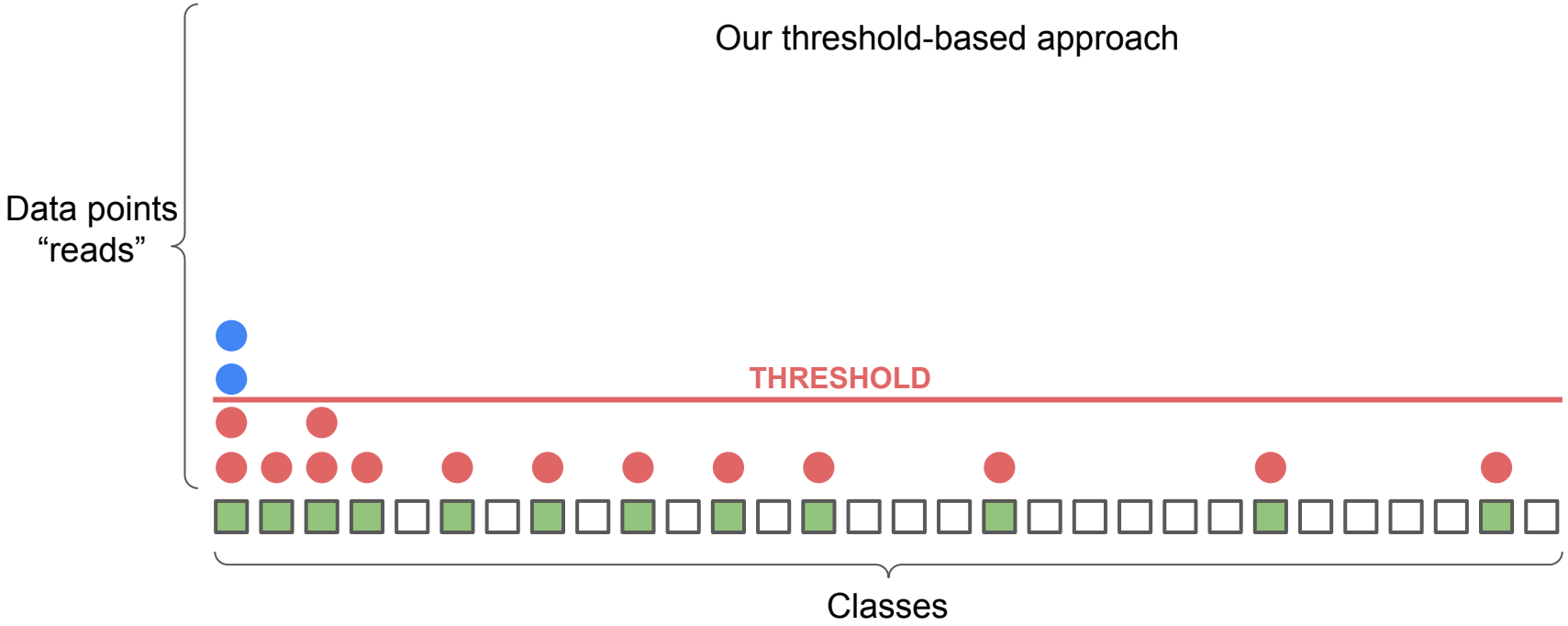
Inverse Propensity Sampling



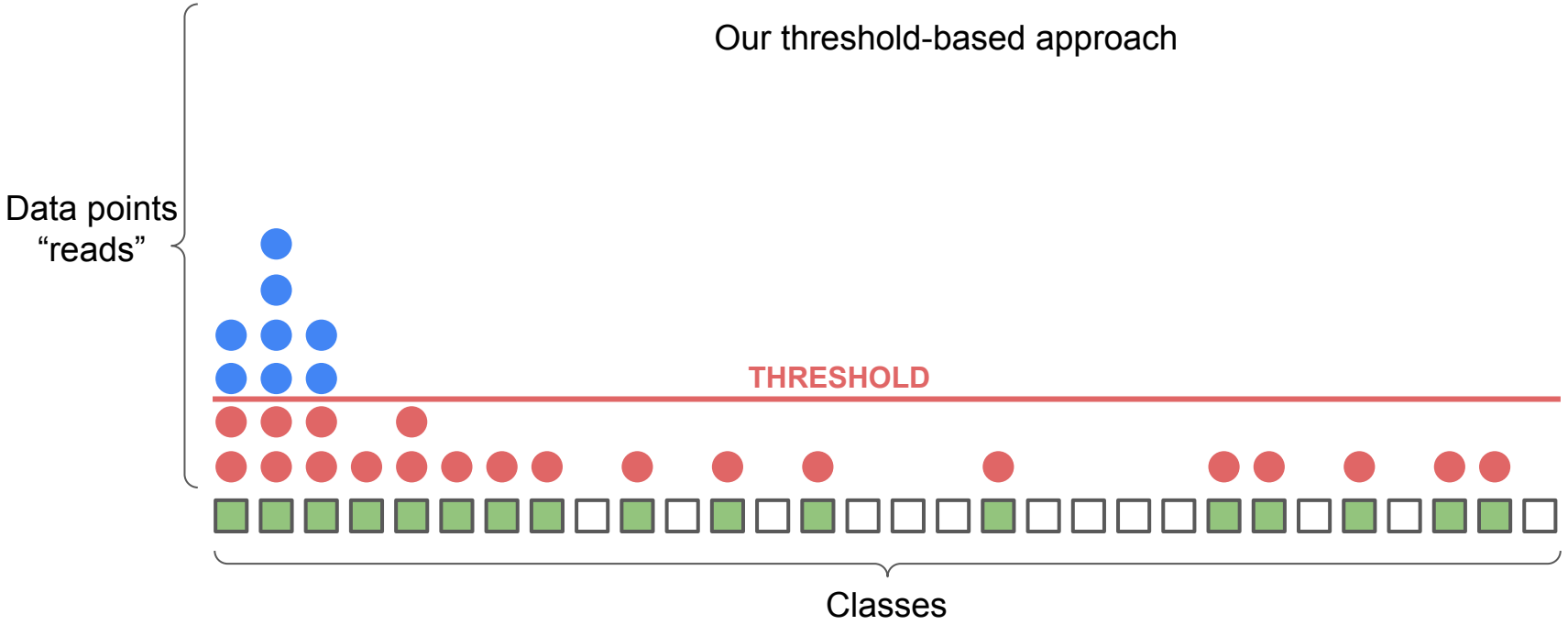
Inverse Propensity Sampling



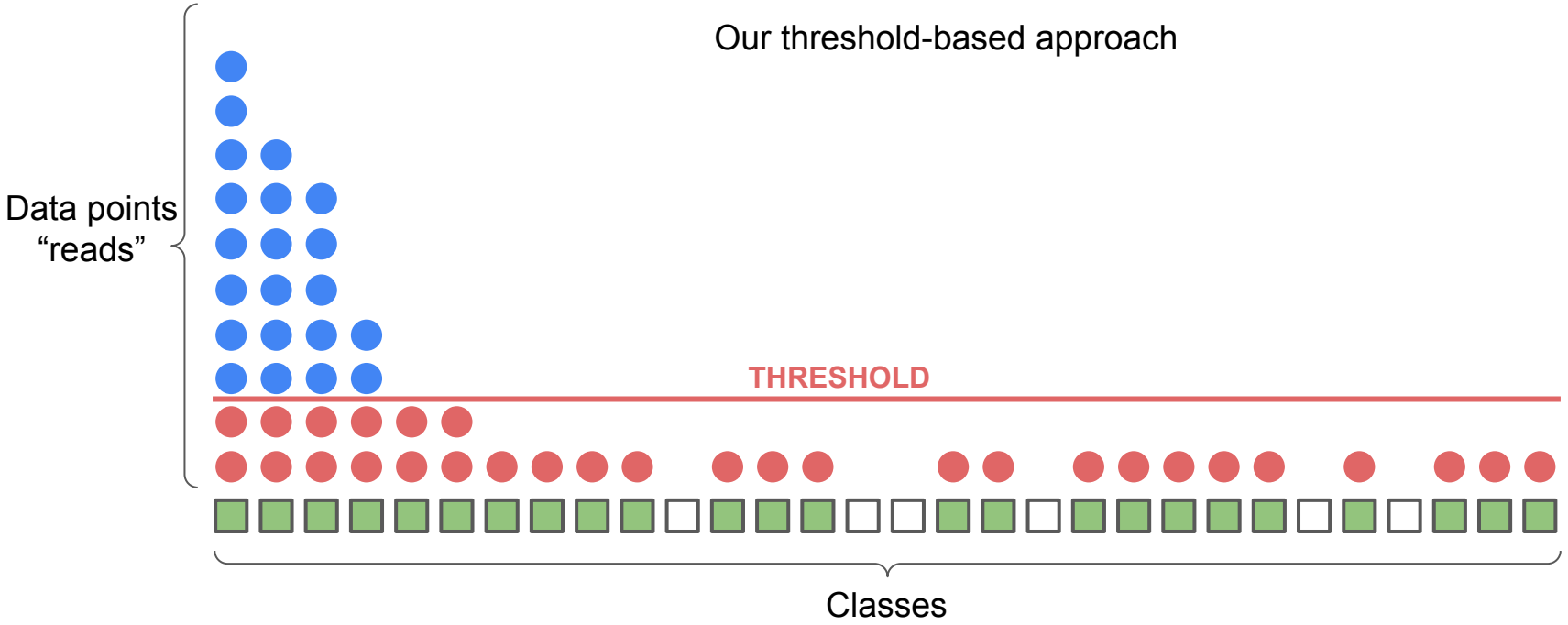
Inverse Propensity Sampling



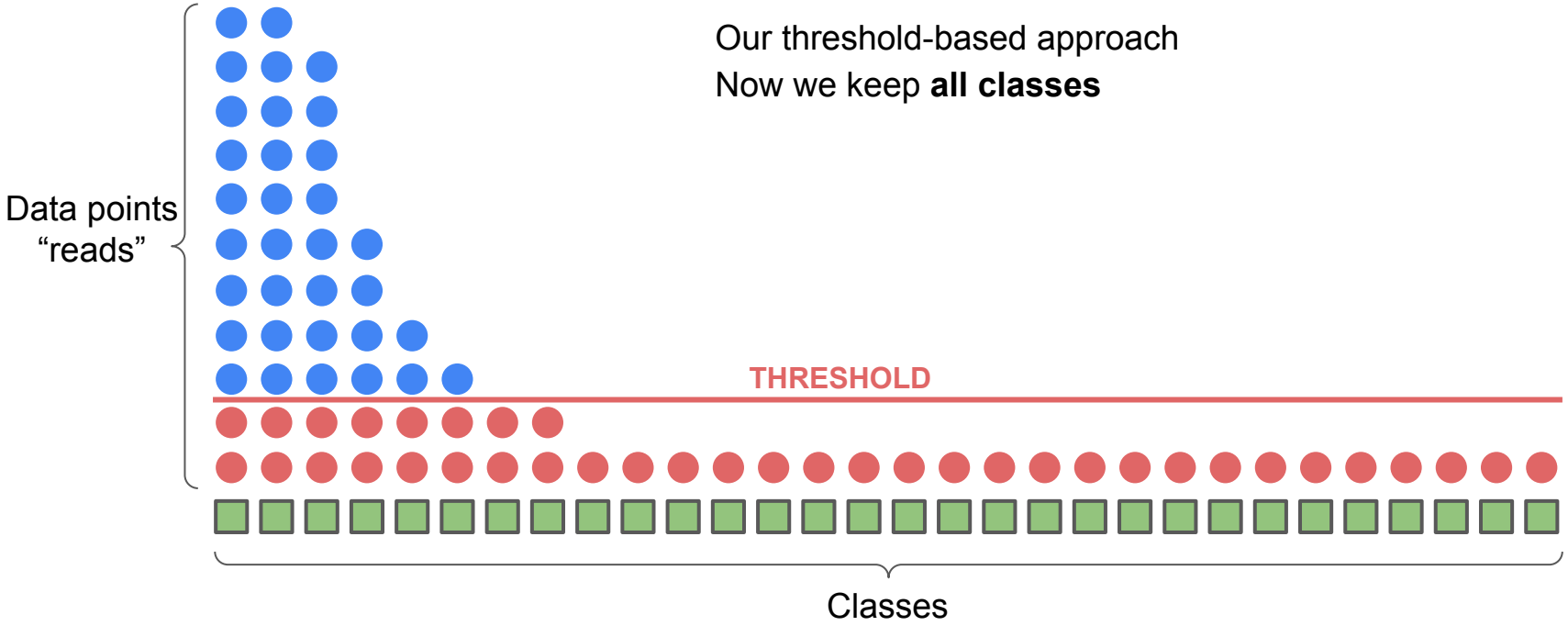
Inverse Propensity Sampling



Inverse Propensity Sampling

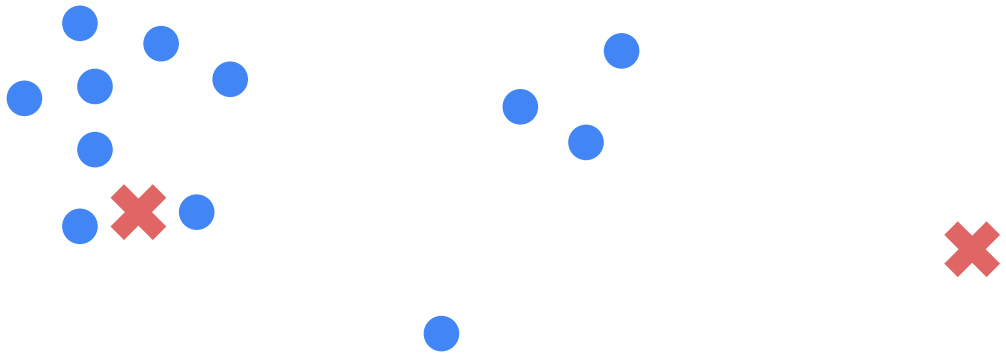


Inverse Propensity Sampling



How do we calculate propensity scores?

- We need a likelihood model
- We estimate this with Kernel Density Estimation
- Density = sum of similarity kernels

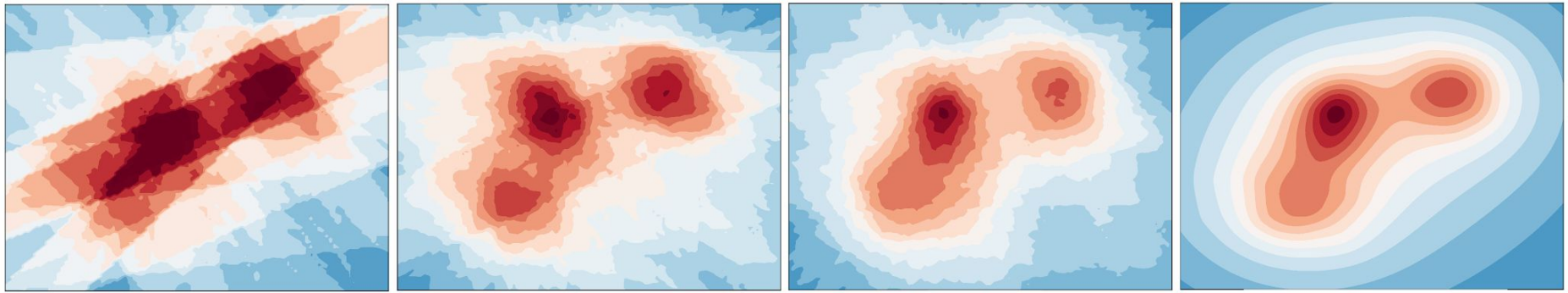


RACE Sketch for KDE

Genomic metric space: Jaccard string distance

Similarity kernel: MinHash Kernel

Cheap and small way to estimate the propensity



Team and Contact

Ben Coleman: benjamin.ray.coleman@gmail.com
PhD student (systems and theory for large-scale ML)

Benito Geordie: benito@thirdai.com
Software engineer (efficient ML implementations)

Todd Treangen: treangen@rice.edu
Assistant professor at Rice University (Bioinformatics)

Anshumali Shrivastava: anshumali@rice.edu
Associate professor at Rice University (Machine Learning)