An example of hypergraph approximation and some prospective differential privacy questions

Alexandre Louvet

2026-01-22

INRIA Lille

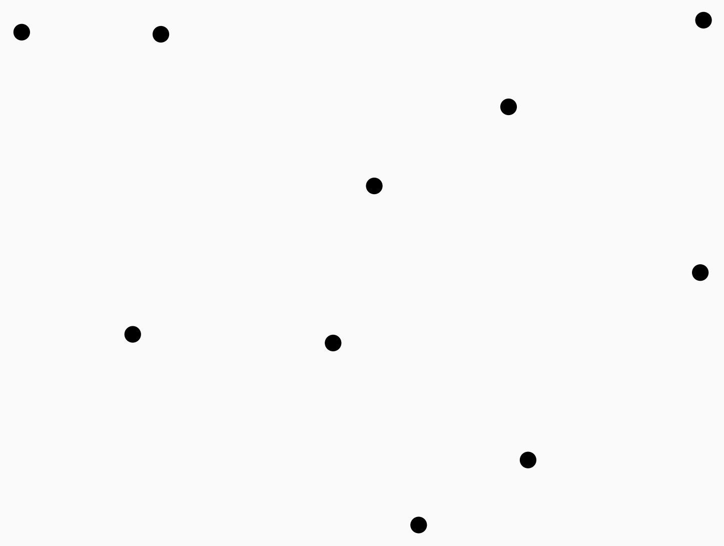
Set systems/Hypergraphs

Set systems/Hypergraphs

$X \longleftarrow$ ground set

Set systems/Hypergraphs

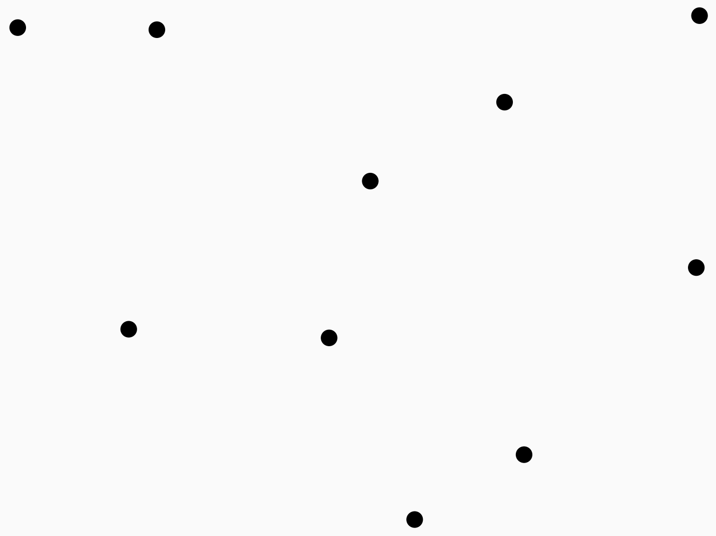
$X \longleftarrow$ ground set



Set systems/Hypergraphs

$X \longleftarrow$ ground set

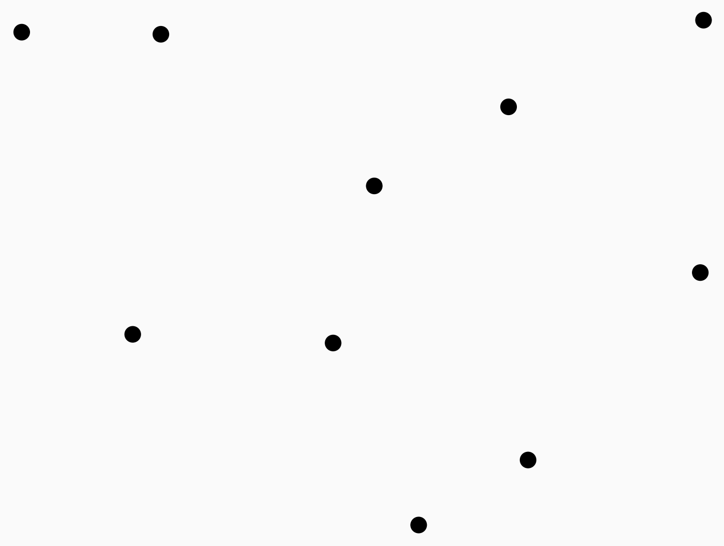
$\hookrightarrow n := |X|$



$X \longleftarrow$ ground set

$\hookrightarrow n := |X|$

$\mathcal{F} \longleftarrow$ collection of subsets of X

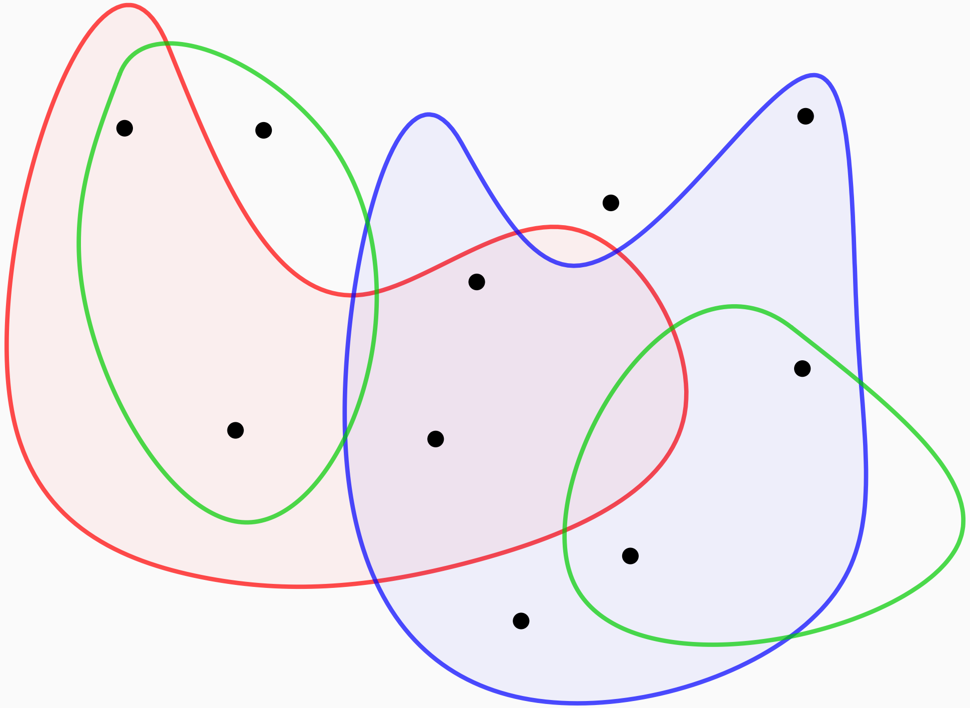


Set systems/Hypergraphs

$X \longleftarrow$ ground set

$\hookrightarrow n := |X|$

$\mathcal{F} \longleftarrow$ collection of subsets of X



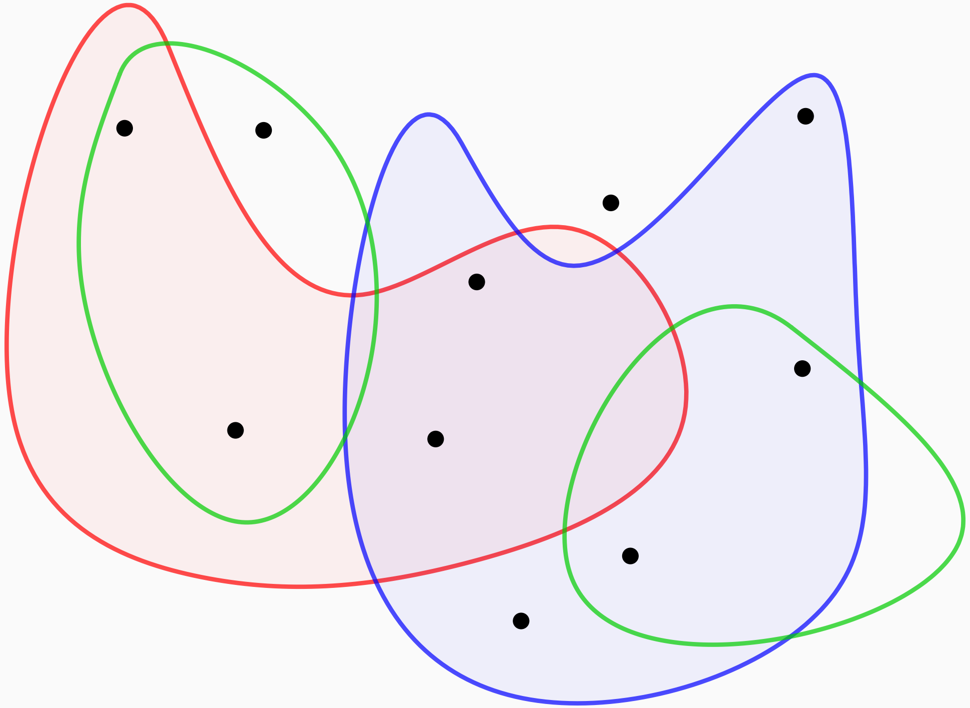
Set systems/Hypergraphs

$X \longleftarrow$ ground set

$\hookrightarrow n := |X|$

$\mathcal{F} \longleftarrow$ collection of subsets of X

$\hookrightarrow m := |\mathcal{F}|$



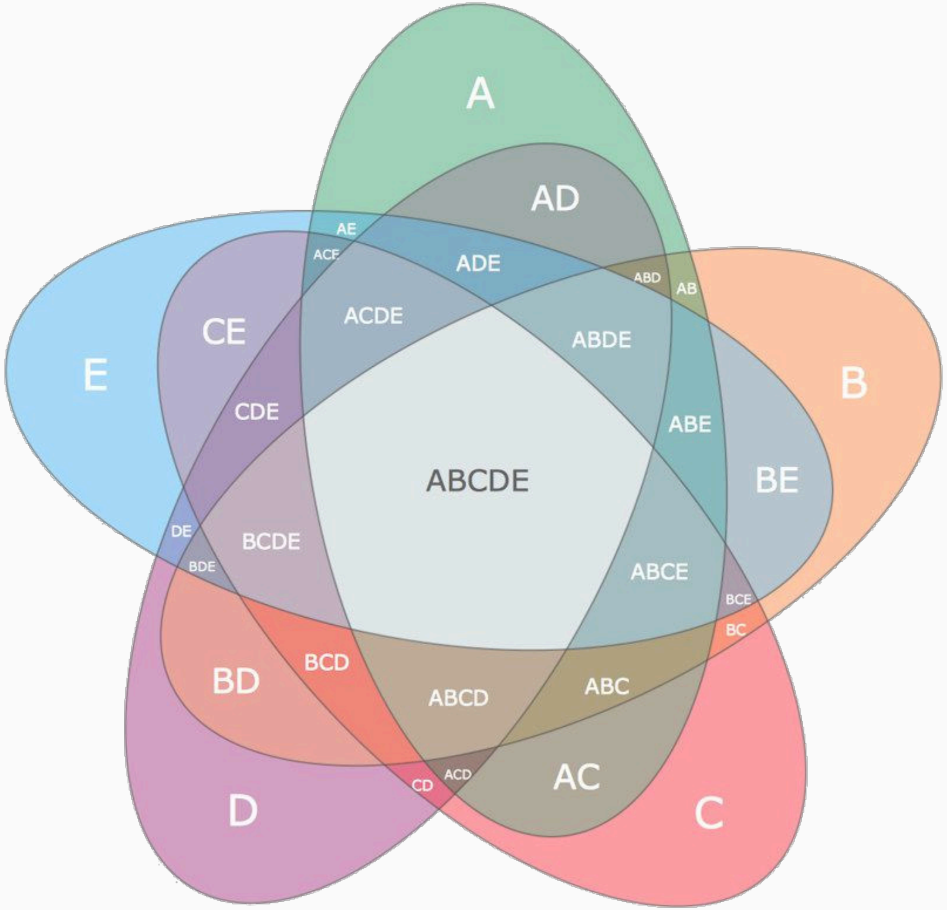
Different flavors of set systems

Different flavors of set systems

- Combinatorial set systems (Powerset, finite VC-dimension set systems,...)

Different flavors of set systems

- Combinatorial set systems (Powerset, finite VC-dimension set systems,...)

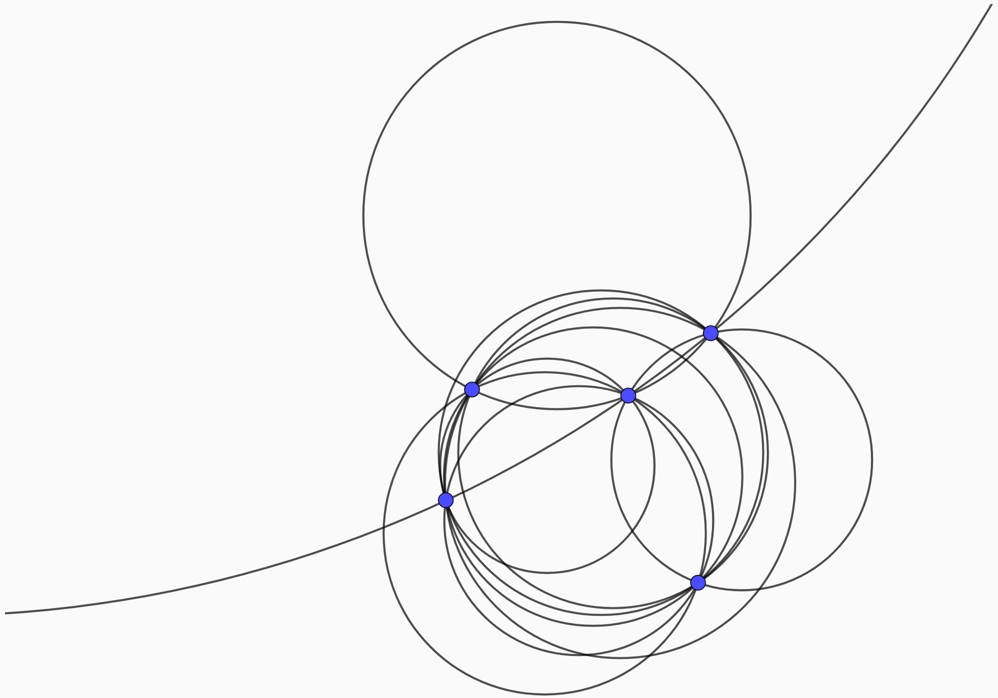


Different flavors of set systems

- Combinatorial set systems (Powerset, finite VC-dimension set systems,...)
- Geometric set systems (Halfspaces, Balls,...)

Different flavors of set systems

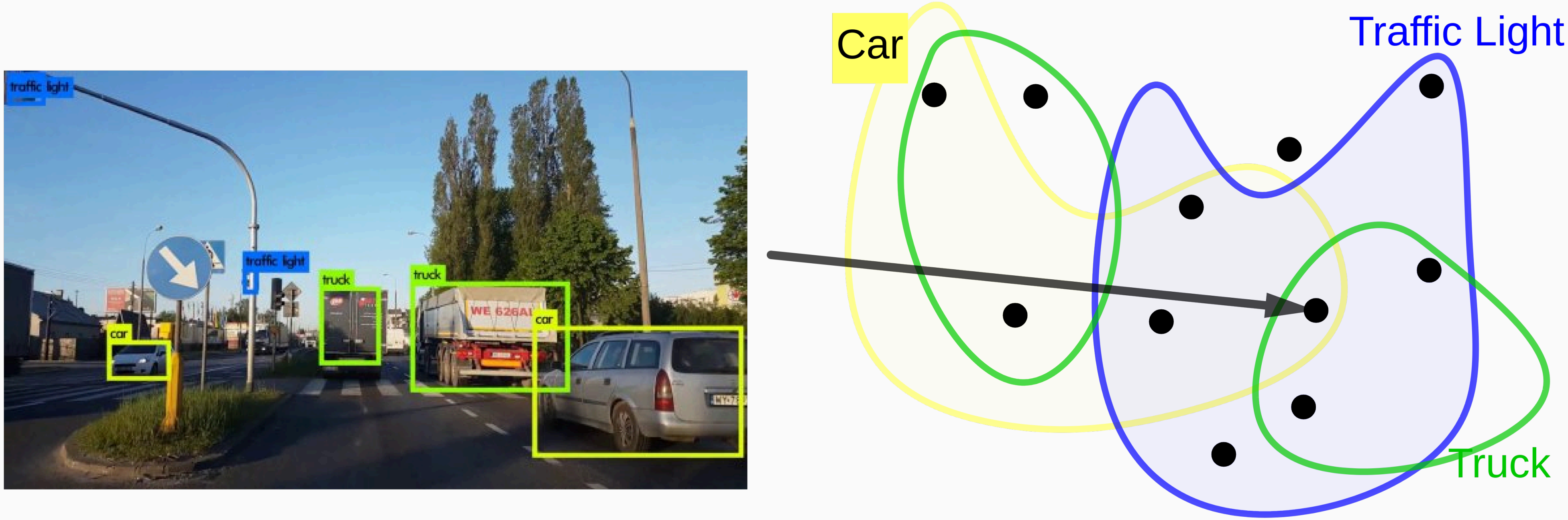
- Combinatorial set systems (Powerset, finite VC-dimension set systems,...)
- Geometric set systems (Halfspaces, Balls,...)



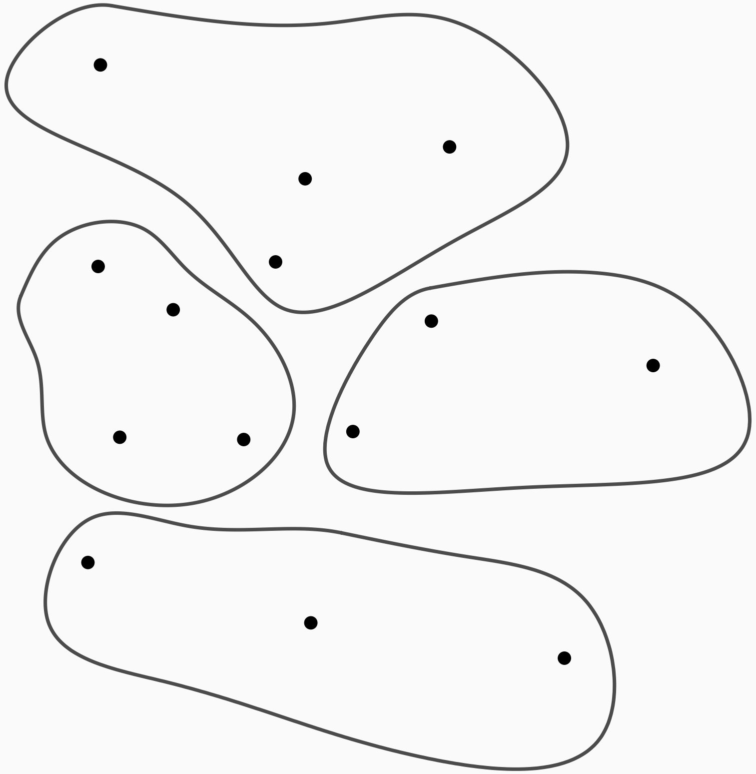
Different flavors of set systems

- Combinatorial set systems (Powerset, finite VC-dimension set systems,...)
- Geometric set systems (Halfspaces, Balls,...)
- None of the above (might be coming from your data)

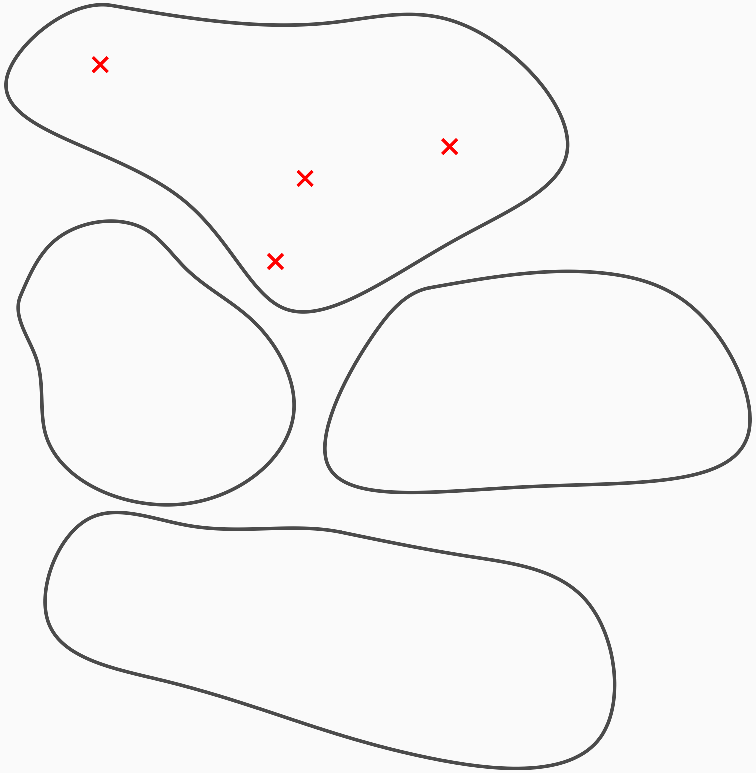
- Combinatorial set systems (Powerset, finite VC-dimension set systems,...)
- Geometric set systems (Halfspaces, Balls,...)
- None of the above (might be coming from your data)



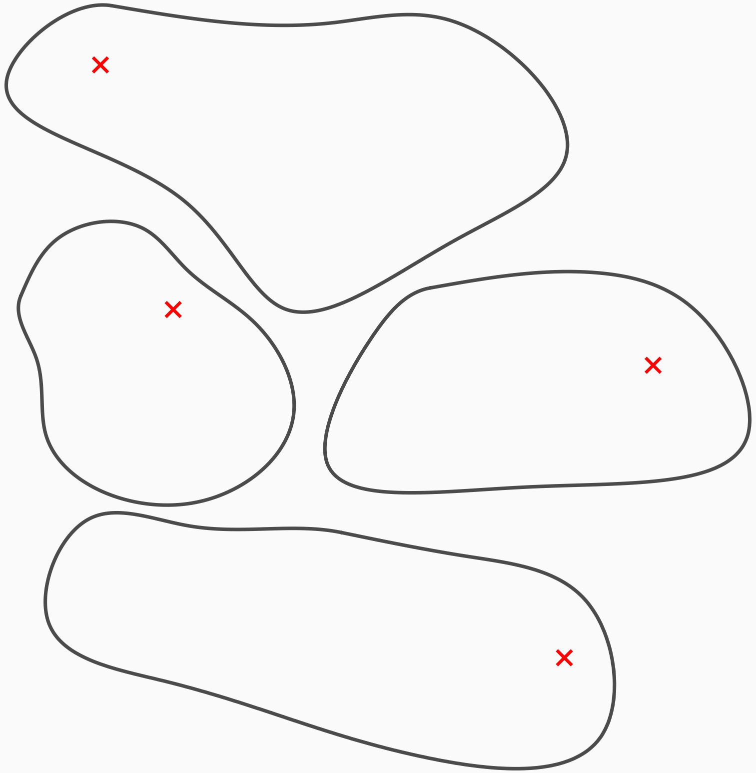
(ε-)approximations



(ε-)approximations



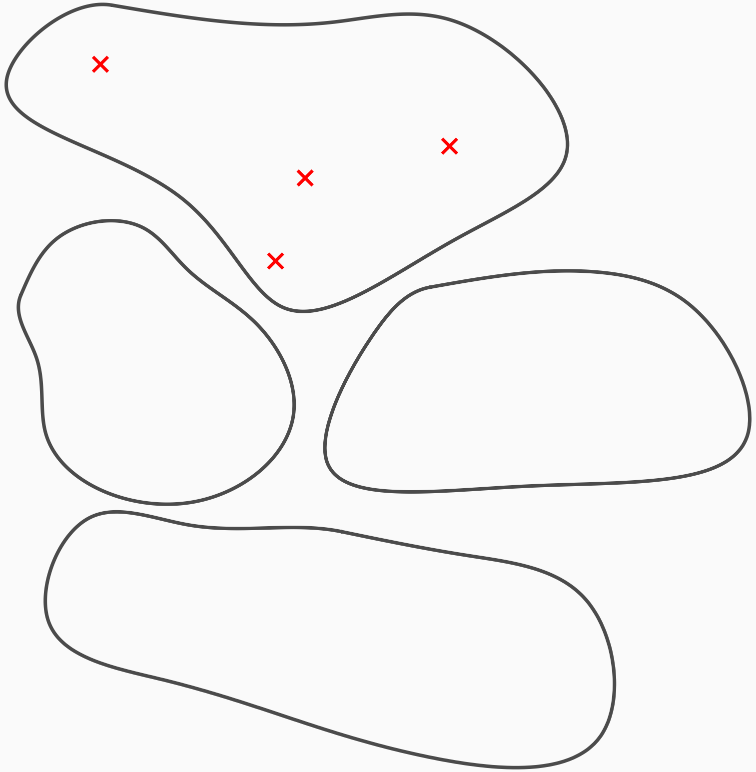
(ε-)approximations



(ε-)approximations

ε-approximations: $A \subseteq X$ s.t. $\forall F \in \mathcal{F}, \left| \frac{|F|}{n} - \frac{|F \cap A|}{|A|} \right| \leq \varepsilon$

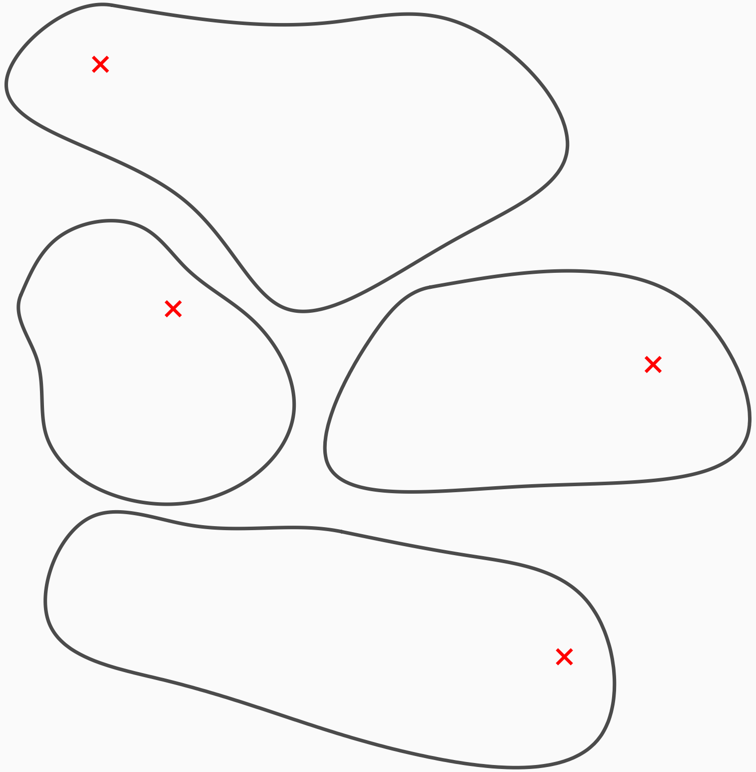
ε-approximations: $A \subseteq X$ s.t. $\forall F \in \mathcal{F}, \left| \frac{|F|}{n} - \frac{|F \cap A|}{|A|} \right| \leq \varepsilon$



$\varepsilon = \frac{10}{14}$

(ε-)approximations

ε-approximations: $A \subseteq X$ s.t. $\forall F \in \mathcal{F}, \left| \frac{|F|}{n} - \frac{|F \cap A|}{|A|} \right| \leq \varepsilon$



$$\varepsilon = \frac{1}{21}$$

Combinatorial discrepancy

Problem statement

Problem statement

2-coloring: $\chi : X \to \{-1, +1\}$

Problem statement

2-coloring: $\chi : X \to \{-1, +1\}$

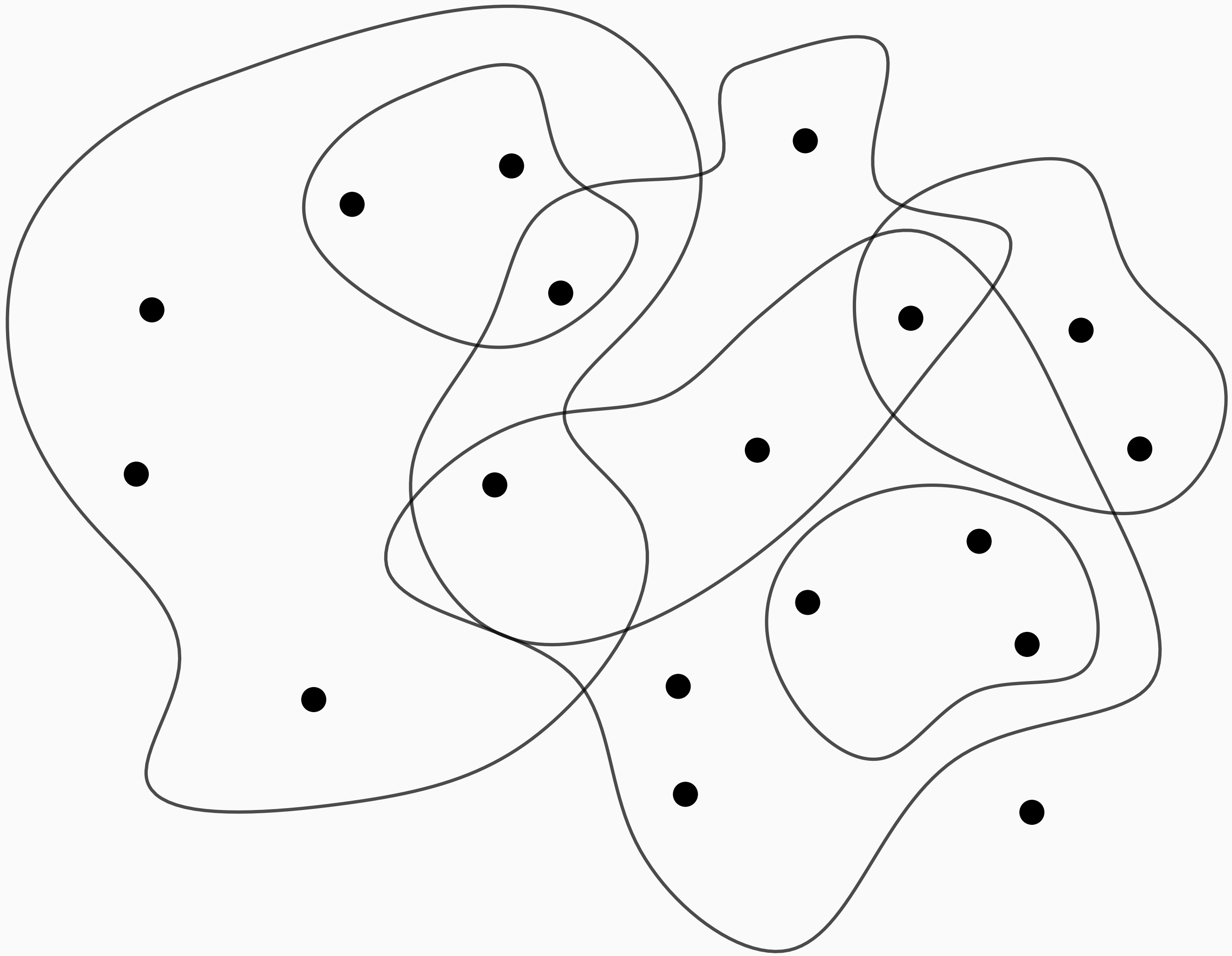
discrepancy w.r.t. χ: $\mathrm{disc}_\chi(X, \mathcal{F}) = \max_{F \in \mathcal{F}} \underbrace{\left| \sum_{x \in F} \chi(x) \right|}_{|\chi(F)|}$

Problem statement

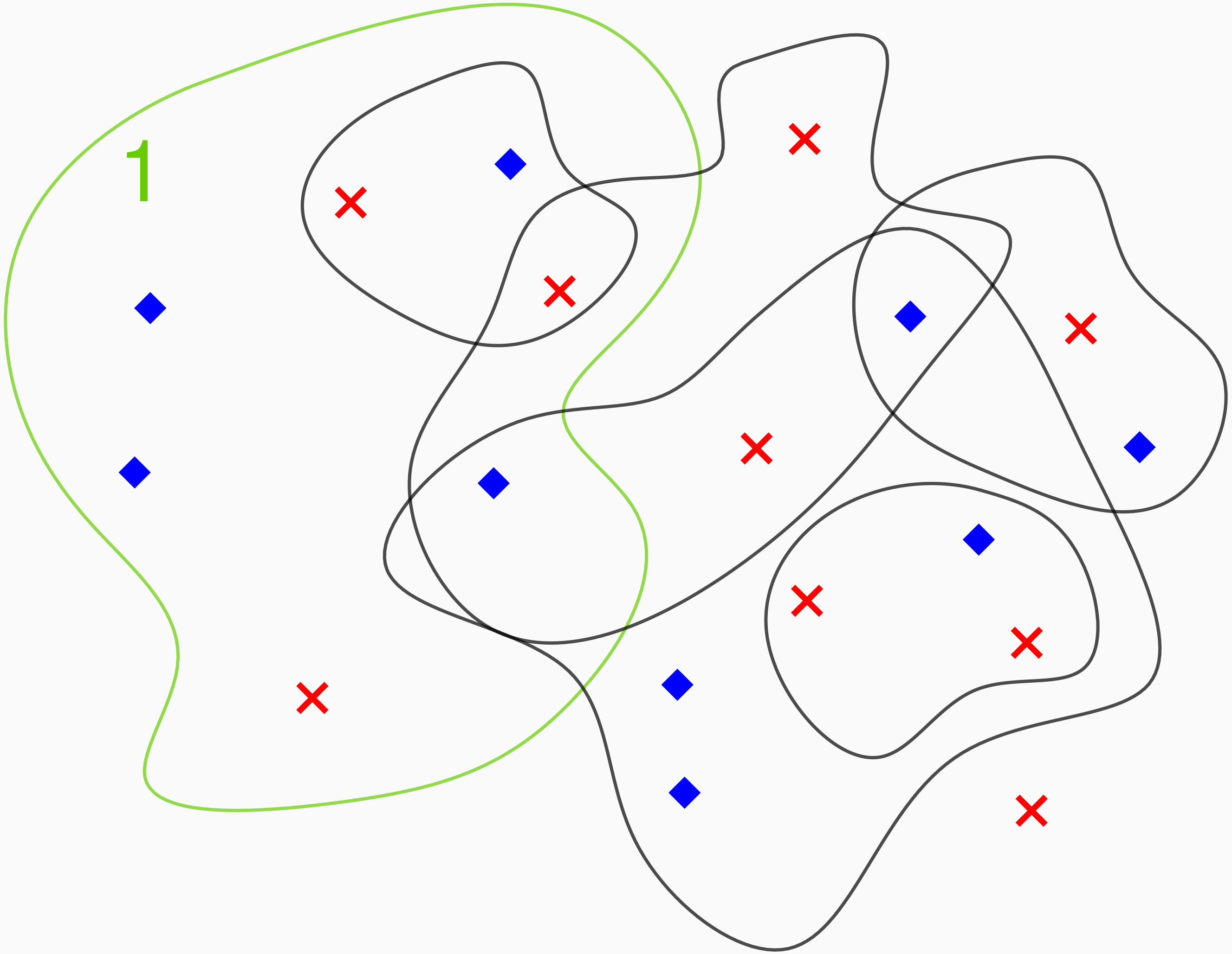
2-coloring: $\chi : X \to \{-1, +1\}$

discrepancy w.r.t. χ: $\mathrm{disc}_\chi(X, \mathcal{F}) = \max_{F \in \mathcal{F}} \underbrace{\left| \sum_{x \in F} \chi(x) \right|}_{|\chi(F)|}$ $\quad (- \text{ - } +)\ (\leq |F|)$

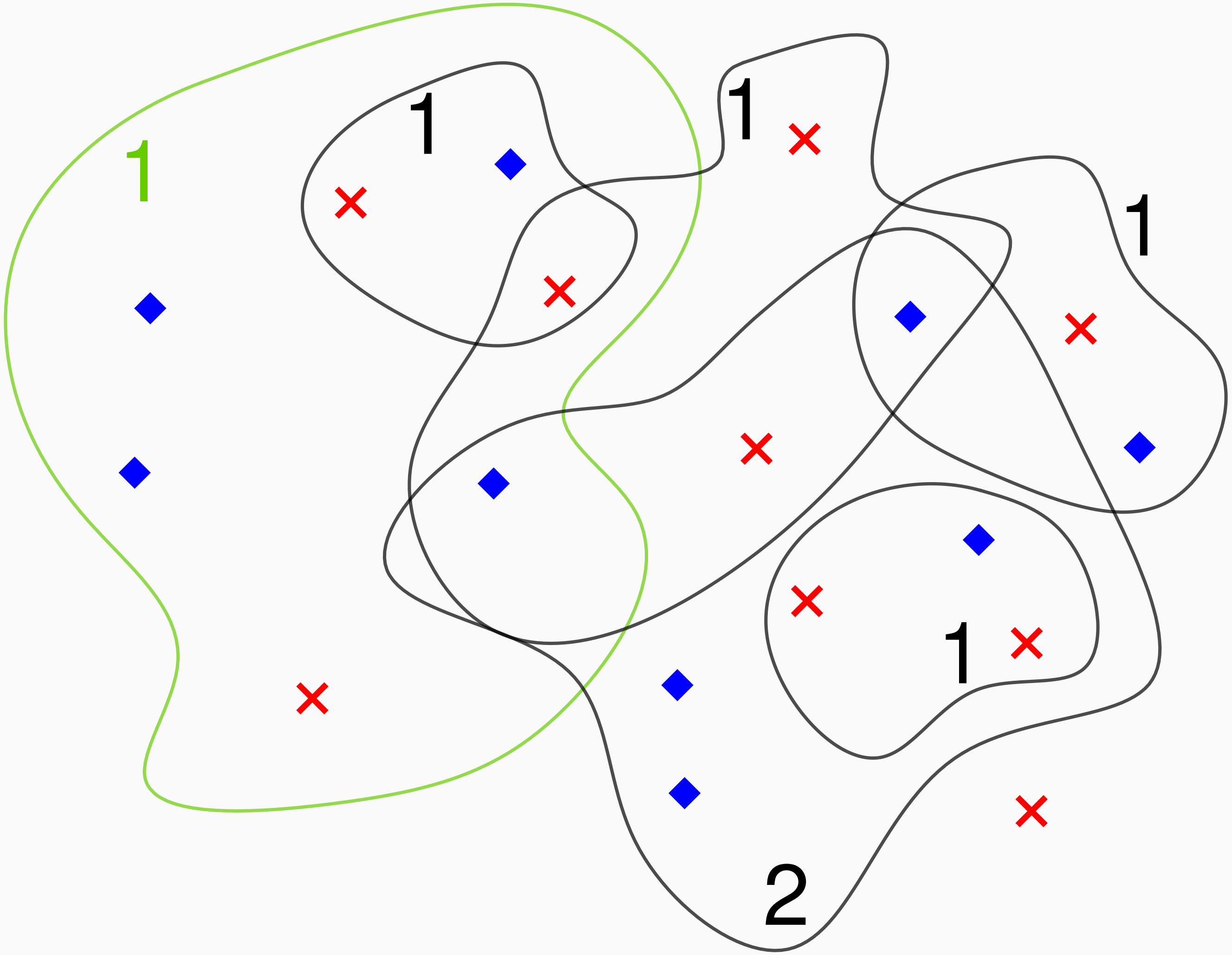
Problem statement

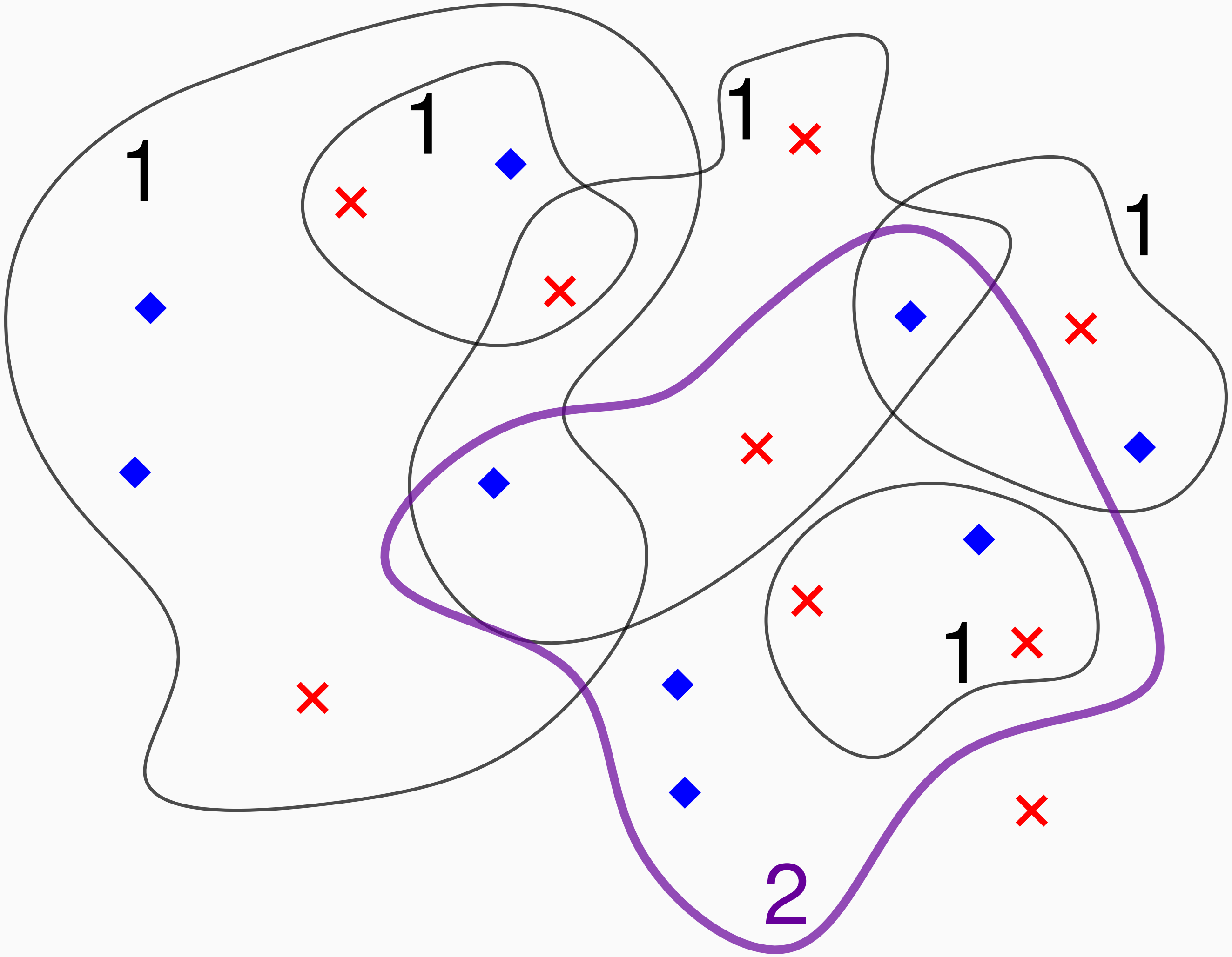


Problem statement



Problem statement




1
1
1
1
1
2

Problem statement

2-coloring: $\chi : X \to \{-1, 1\}$

discrepancy w.r.t. χ: $\operatorname{disc}_\chi(X, \mathcal{F}) = \max_{F \in \mathcal{F}} \left| \sum_{x \in F} \chi(x) \right|$

discrepancy of $(X, \mathcal{F})$: $\min_{\chi : X \to \{-1,1\}} \operatorname{disc}_\chi(X, \mathcal{F})$

2-coloring: $\chi : X \to \{-1, 1\}$

discrepancy w.r.t. χ: $\mathrm{disc}_\chi(X, \mathcal{F}) = \max_{F \in \mathcal{F}} \left| \sum_{x \in F} \chi(x) \right|$

discrepancy of $(X, \mathcal{F})$: $\min_{\chi : X \to \{-1,1\}} \mathrm{disc}_\chi(X, \mathcal{F})$

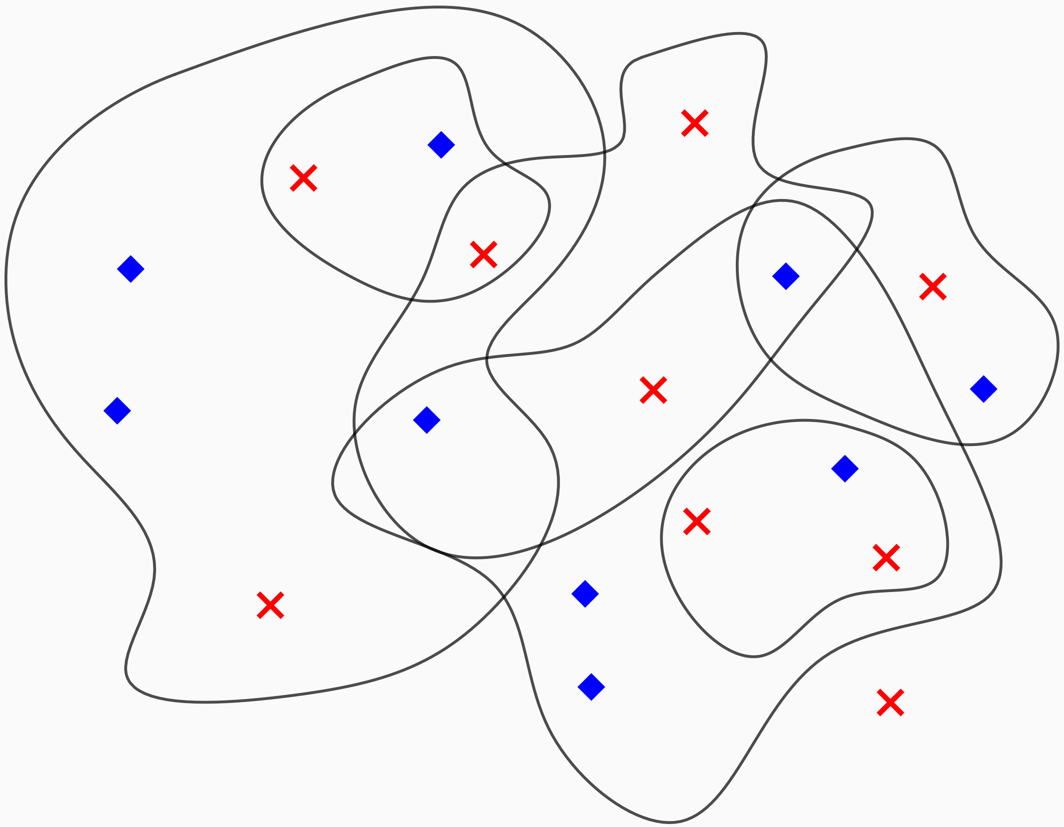
Goal: Compute a small discrepancy coloring

(ε-)approximations

(ε-)approximations

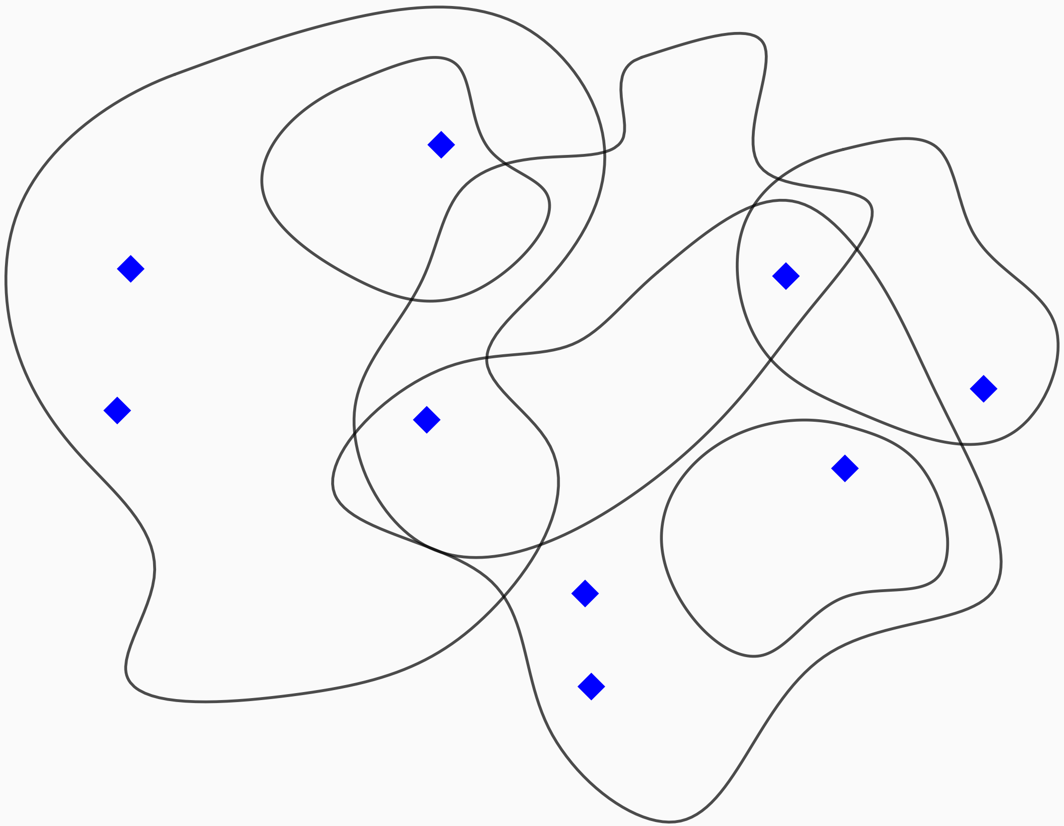
ε-approximations: $A \subseteq X$ s.t. $\forall F \in \mathcal{F}, \left| \frac{|F|}{n} - \frac{|F \cap A|}{|A|} \right| \leq \varepsilon$

ε-approximations: $A \subseteq X$ s.t. $\forall F \in \mathcal{F}, \left| \frac{|F|}{n} - \frac{|F \cap A|}{|A|} \right| \leq \varepsilon$



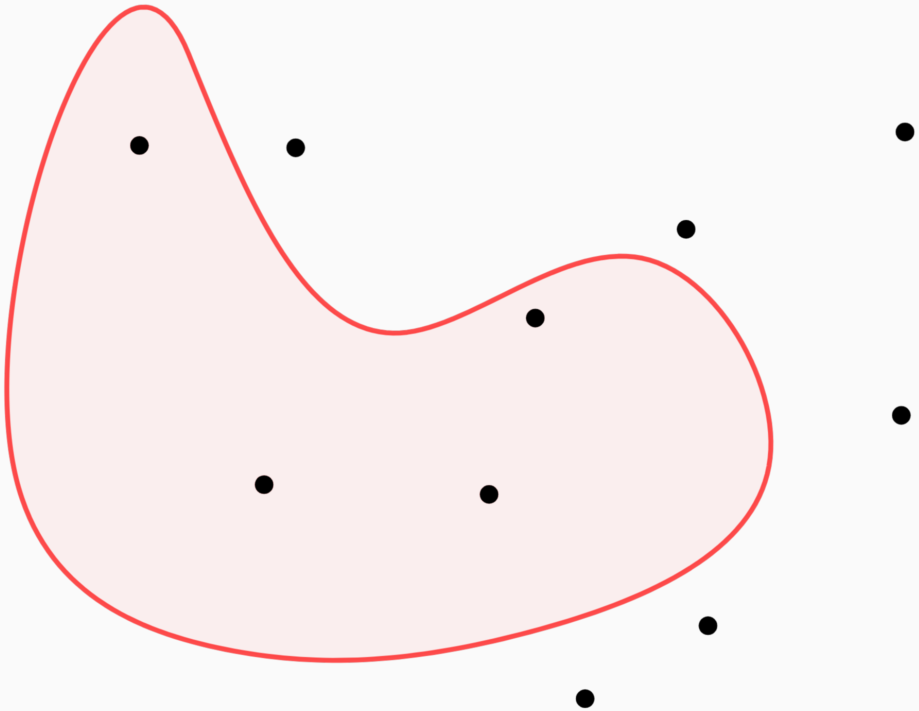
(ε-)approximations

ε-approximations: $A \subseteq X$ s.t. $\forall F \in \mathcal{F}, \left| \frac{|F|}{n} - \frac{|F \cap A|}{|A|} \right| \leq \varepsilon$

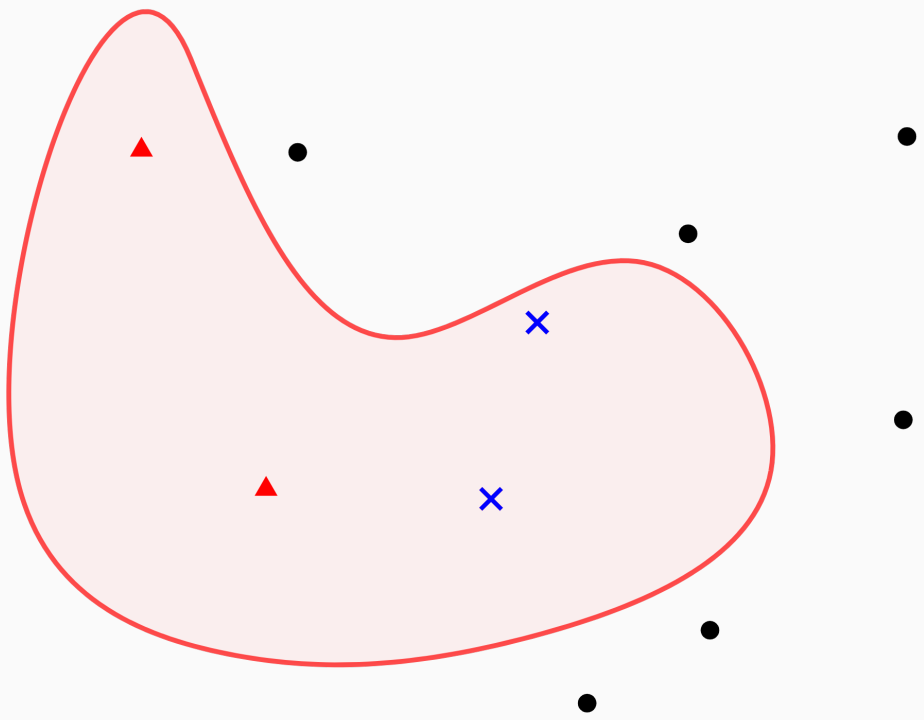


When does it get difficult?

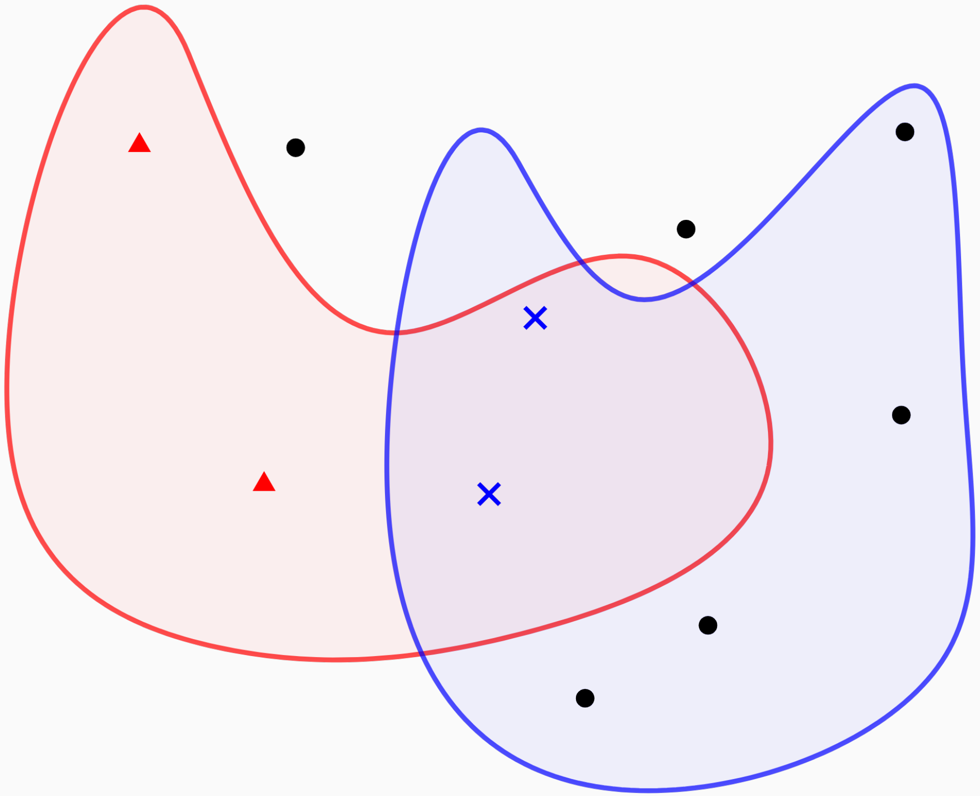
When does it get difficult?



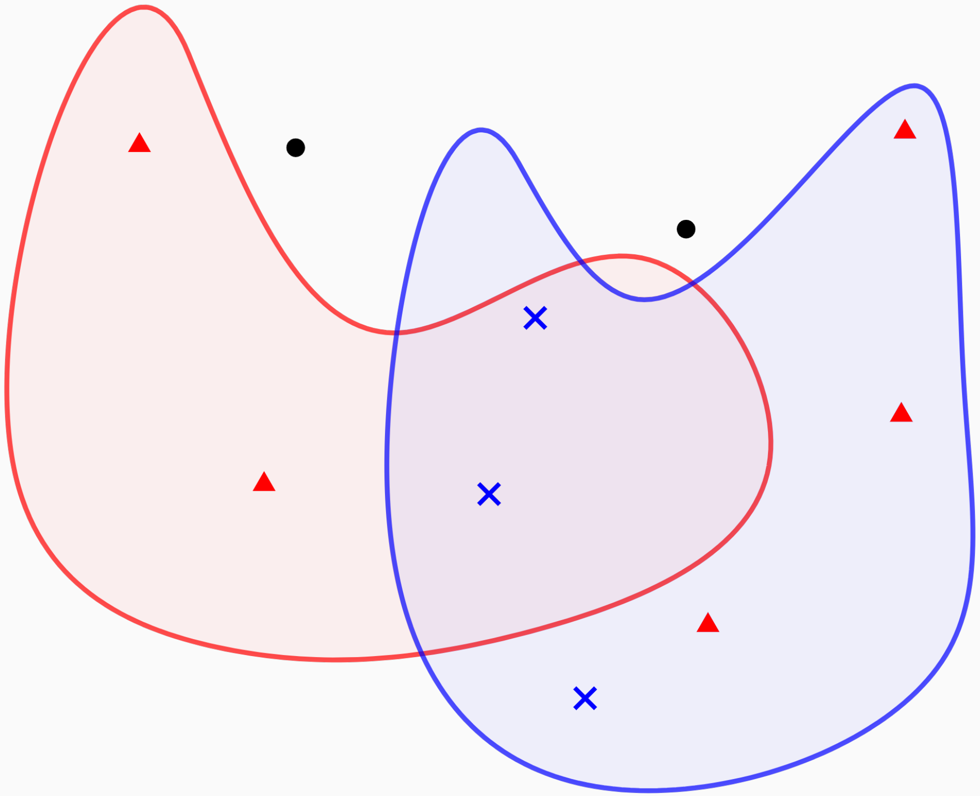
When does it get difficult?



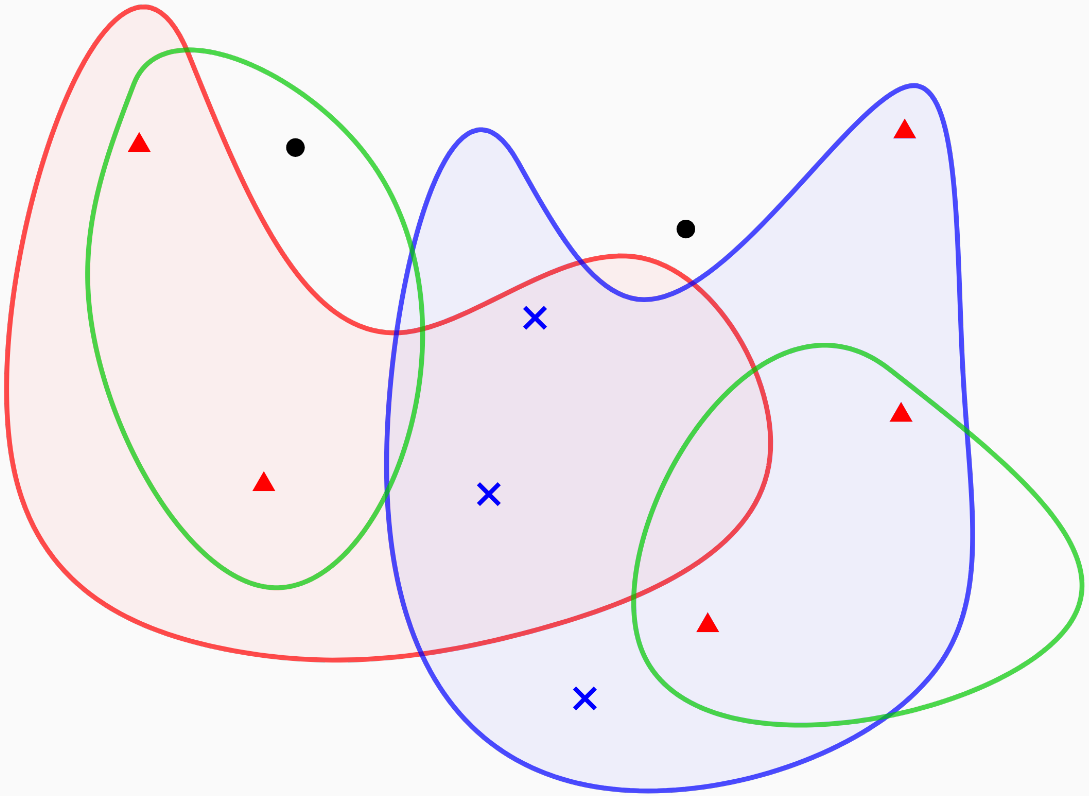
When does it get difficult?



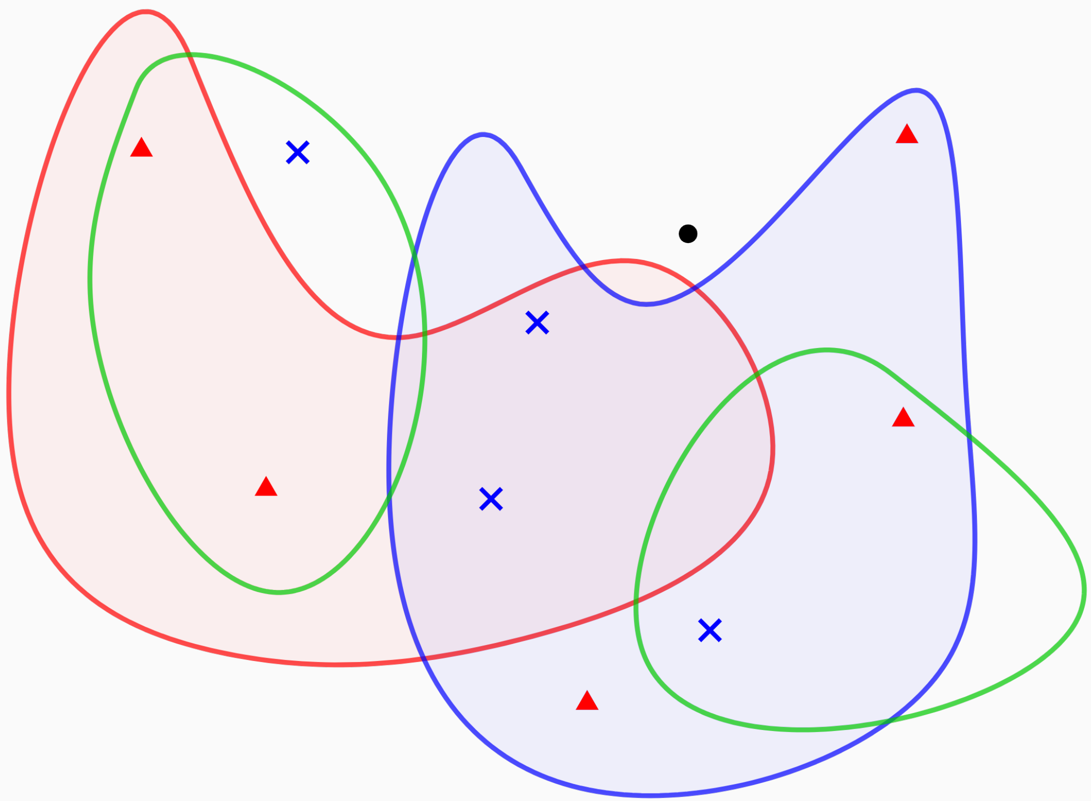
When does it get difficult?



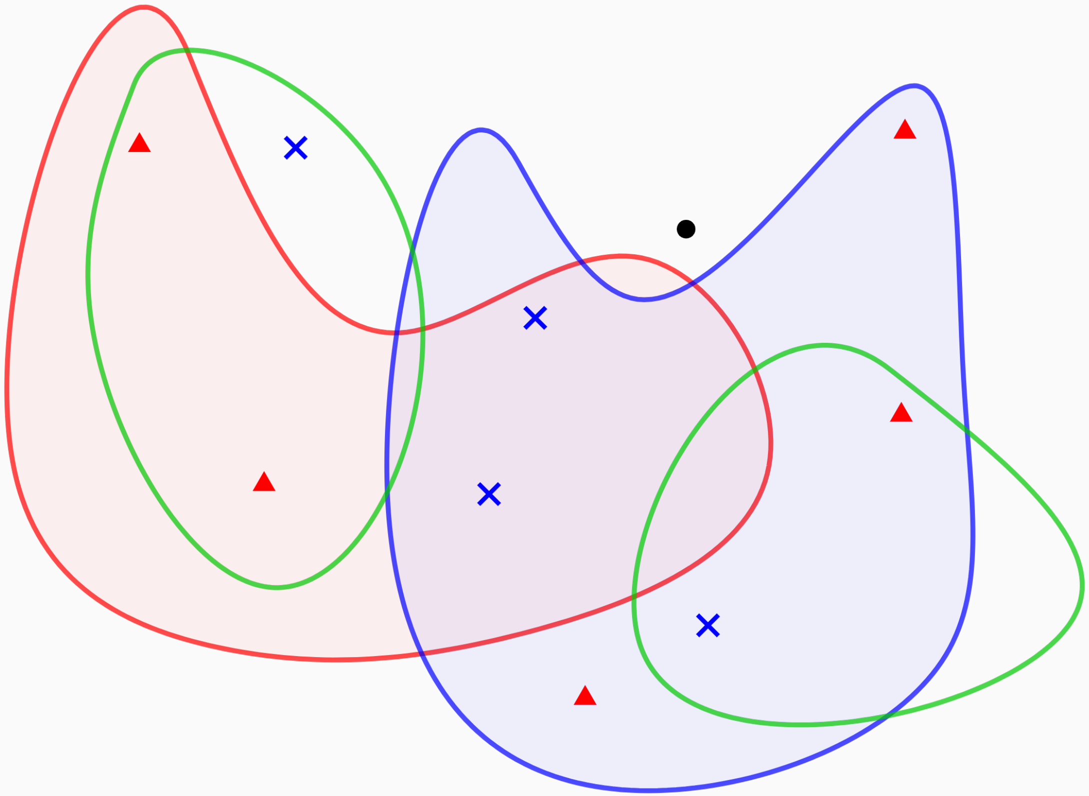
When does it get difficult?



When does it get difficult?

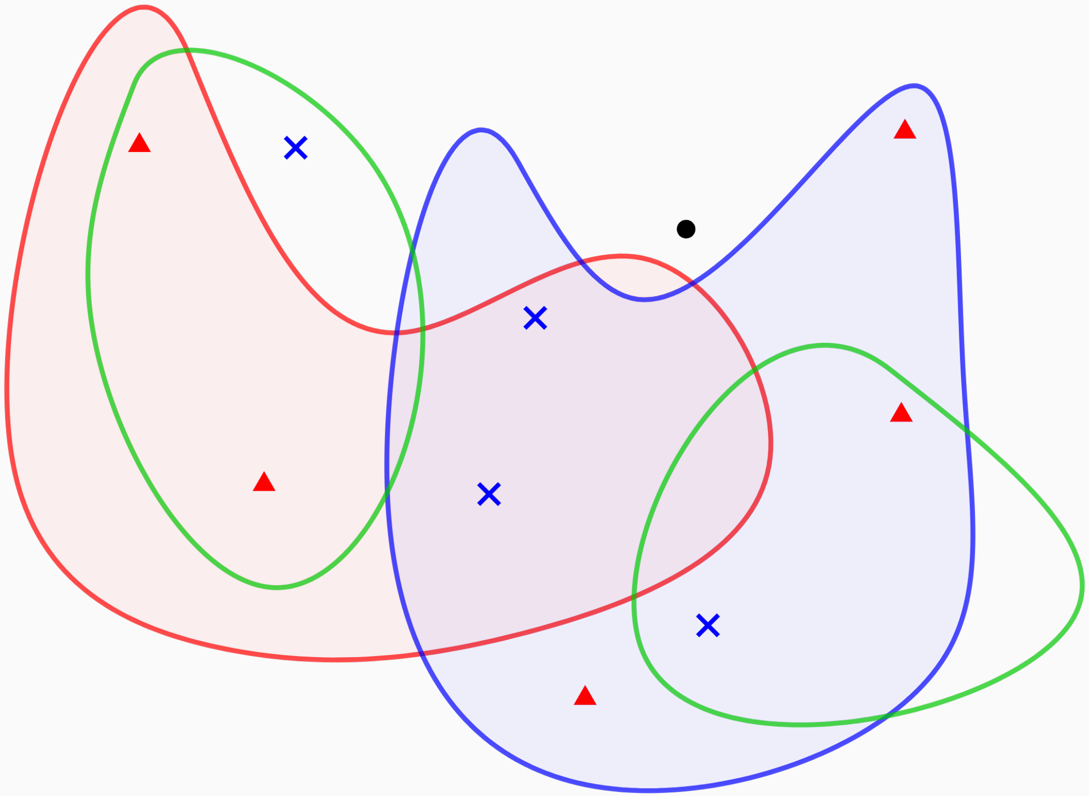


When does it get difficult?



Easy until $\Theta(n)$ sets: For $\Theta(n)$ sets, one can find colorings with discrepancy $\Theta(1)$!

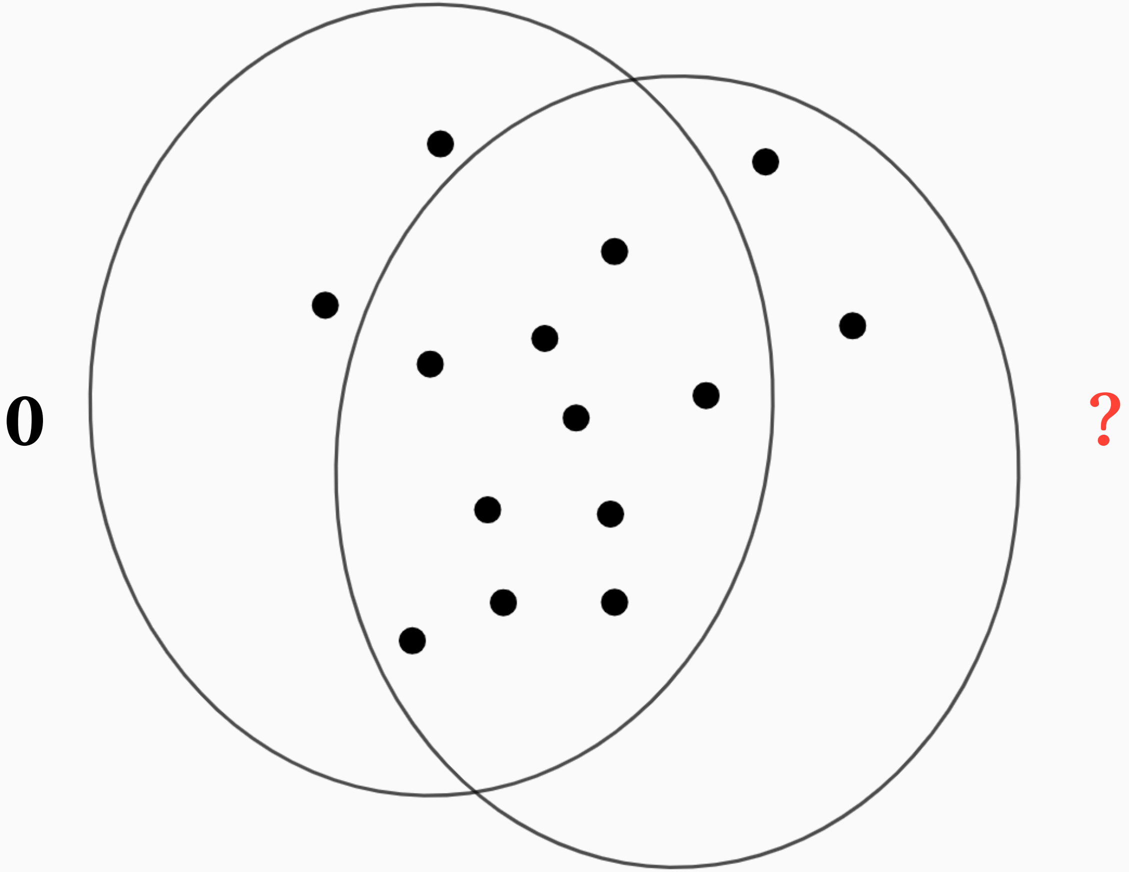
When does it get difficult?

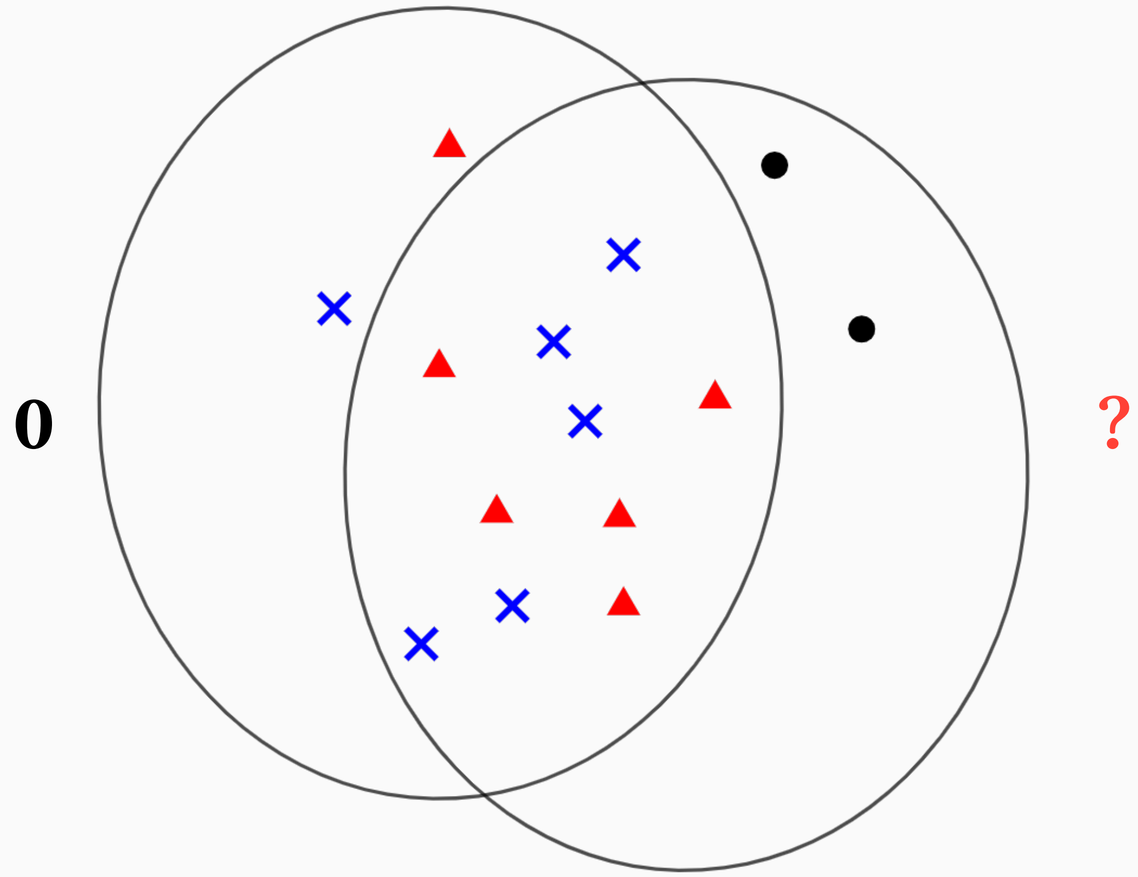


Easy until $\Theta(n)$ sets: For $\Theta(n)$ sets, one can find colorings with discrepancy $\Theta(1)$!
Hard for $m = \Omega(n)$: $\Omega\left(\sqrt{n \log\left(\frac{m}{n}\right)}\right)$ in general, $\Omega\left(n^{\frac{1}{2}-\frac{1}{2d}}\right)$ for VC-dim d

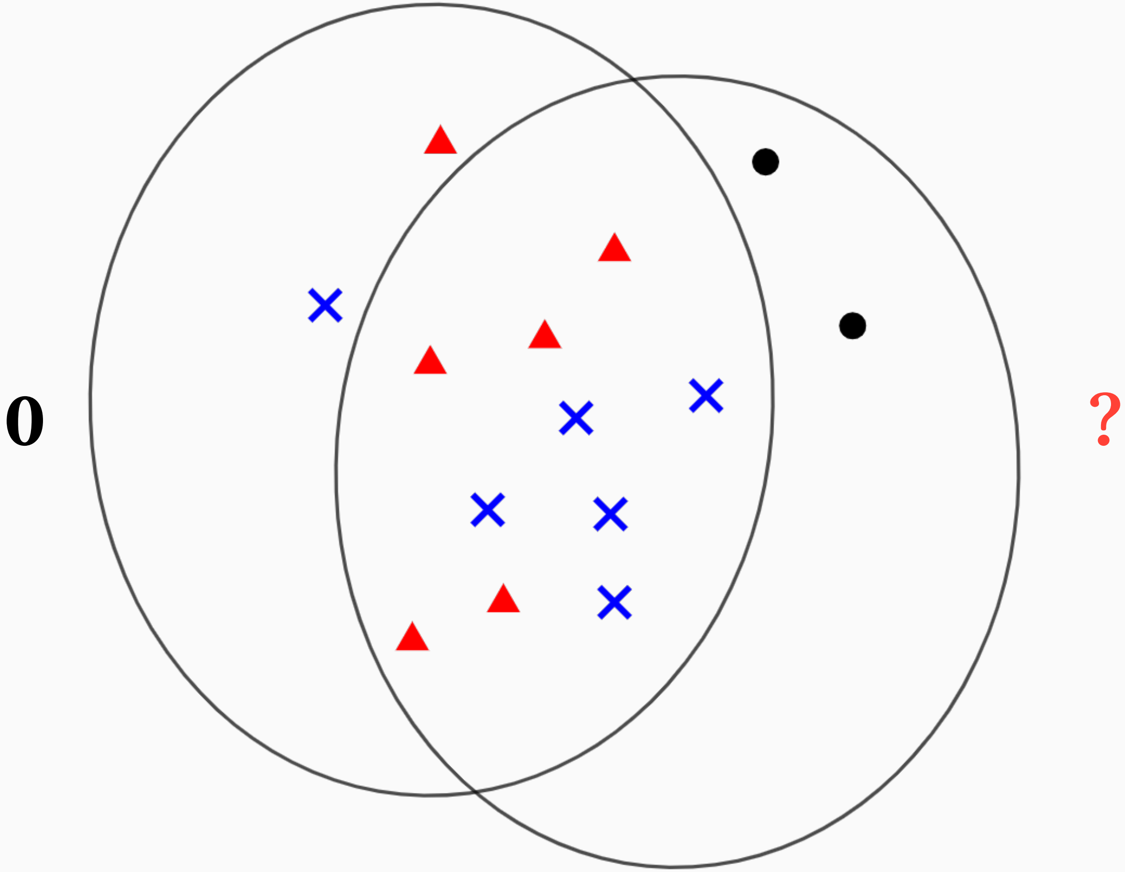
Discrepancy and symmetric difference

Discrepancy and symmetric difference

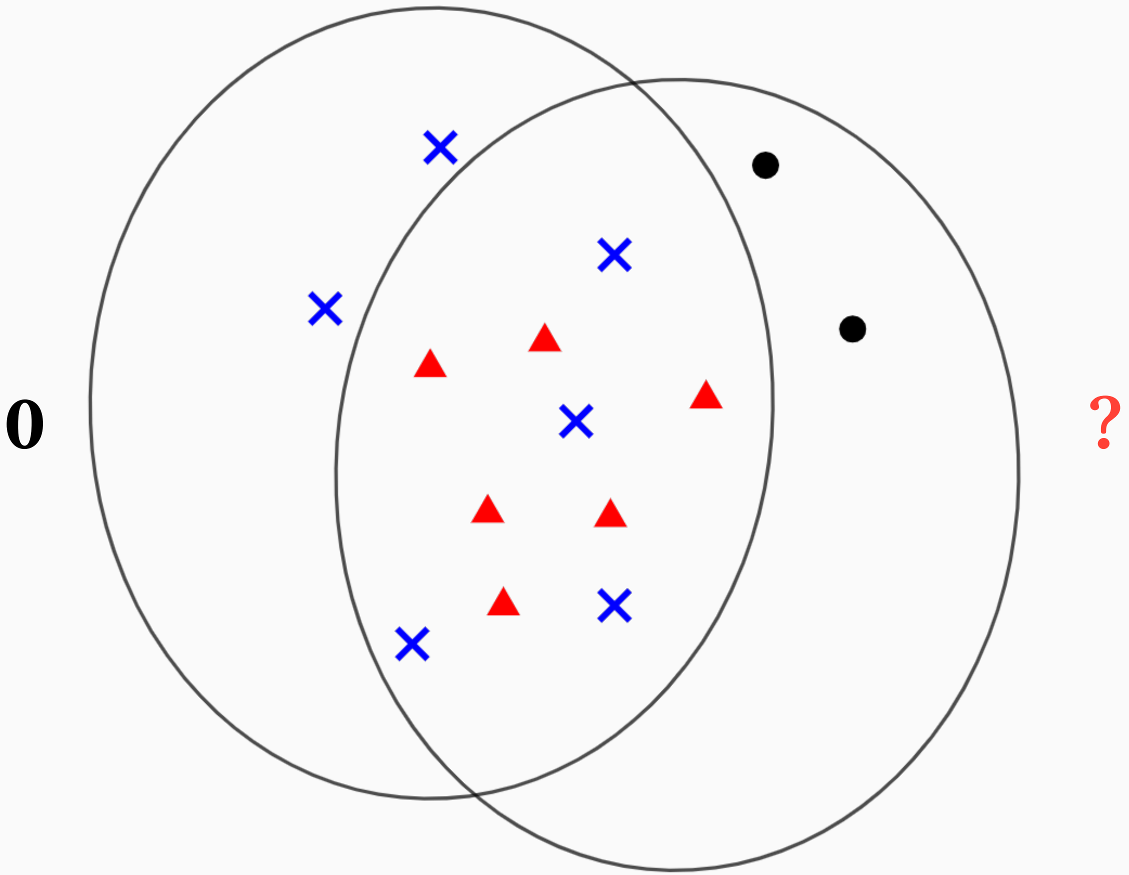



0
?

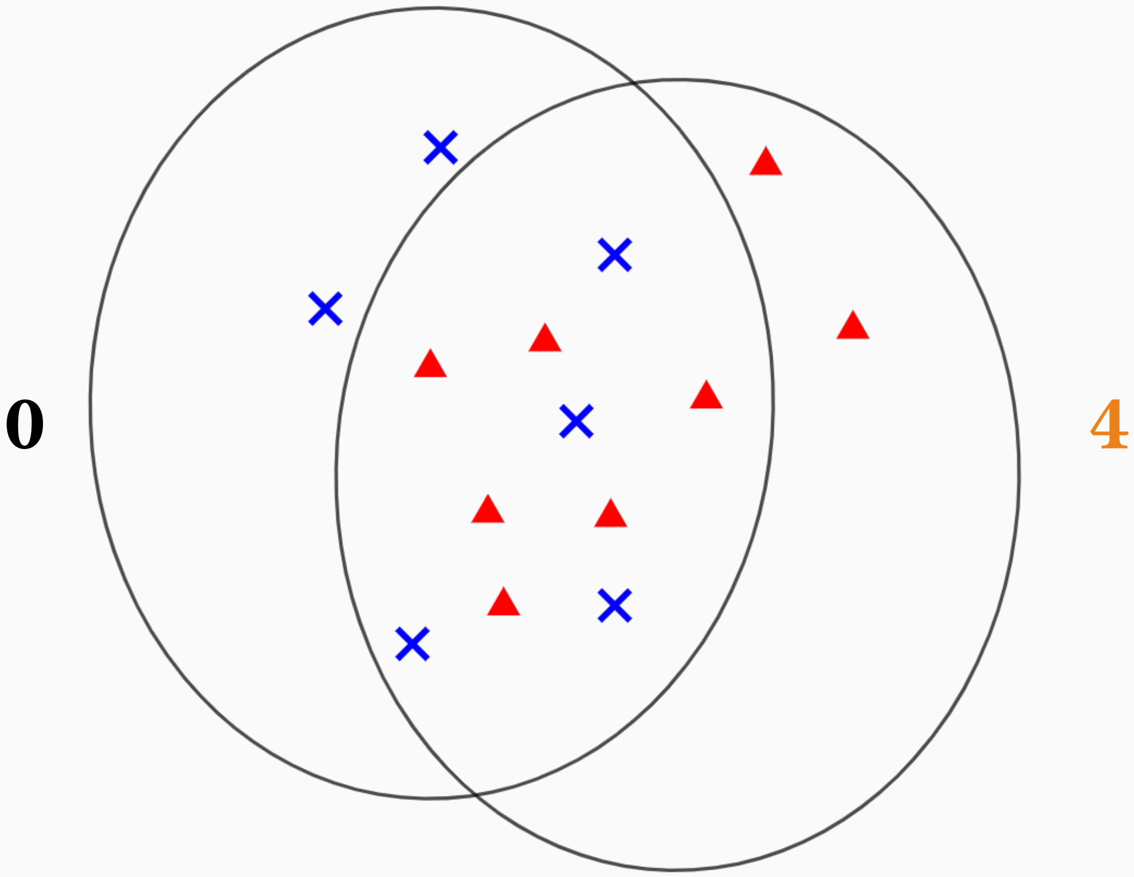
Discrepancy and symmetric difference



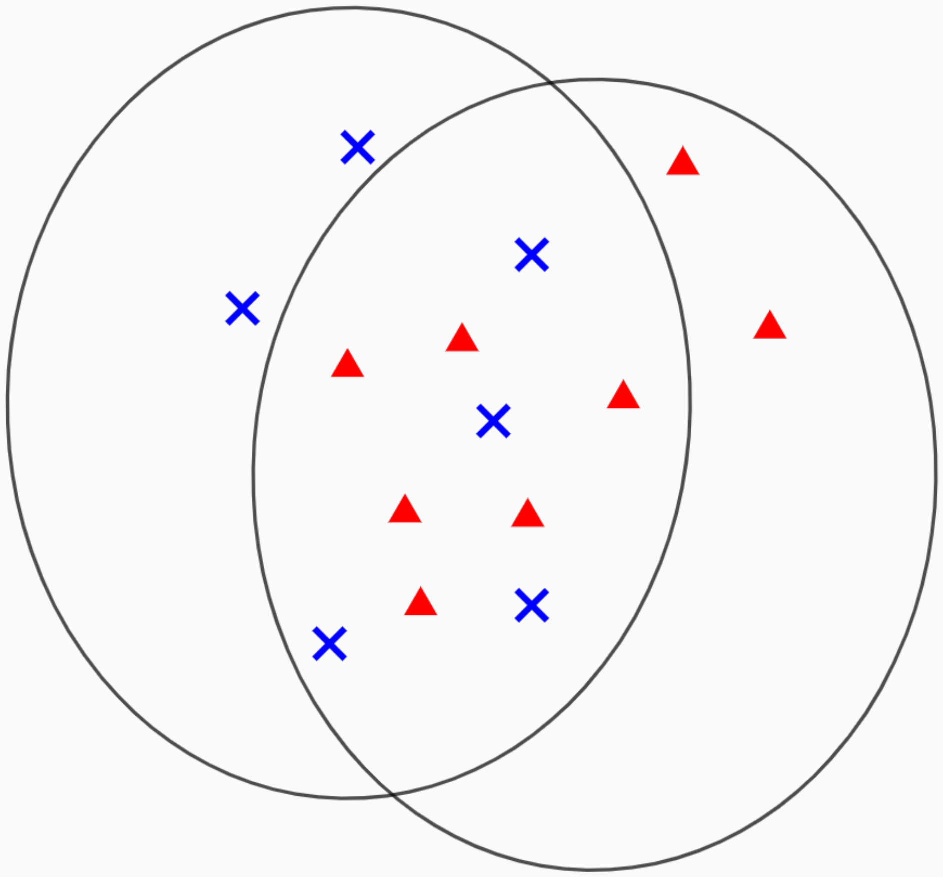
Discrepancy and symmetric difference



Discrepancy and symmetric difference



Discrepancy and symmetric difference



$$\forall F, F' \in \mathcal{F}, |\chi(F)| \leq |\chi(F')| + |\chi(F \setminus F')| + |\chi(F' \setminus F)|$$

Recap

Recap

- To compute ε-approximations, we solve the combinatorial discrepancy problem

Recap

- To compute ε-approximations, we solve the combinatorial discrepancy problem
- Discrepancy is easy for a few ranges

Recap

- To compute ε-approximations, we solve the combinatorial discrepancy problem
- Discrepancy is easy for a few ranges
- Discrepancy is easy for ranges close to ranges with low discrepancy

Recap

- To compute ε-approximations, we solve the combinatorial discrepancy problem
- Discrepancy is easy for a few ranges
- Discrepancy is easy for ranges close to ranges with low discrepancy

↳ **Let's find a (small) set of ranges such that all ranges are close to one of them!**

δ-covering

δ-covering

$$\mathcal{C} \subseteq \mathcal{F} \text{ s.t. } \forall F \in \mathcal{F}, \exists C \in \mathcal{C} \text{ with } \underbrace{|\Delta(F, C)|}_{=|(F \setminus C) \cup (C \setminus F)|} \leq \delta$$

δ-covering

$$\mathcal{C} \subseteq \mathcal{F} \text{ s.t. } \forall F \in \mathcal{F}, \exists C \in \mathcal{C} \text{ with } \underbrace{|\Delta(F, C)|}_{= |(F \setminus C) \cup (C \setminus F)|} \leq \delta$$

In general: $O\left(e^{\frac{n}{\delta}}\right)$ **large!**

δ-covering

$$\mathcal{C} \subseteq \mathcal{F} \text{ s.t. } \forall F \in \mathcal{F}, \exists C \in \mathcal{C} \text{ with } \underbrace{|\Delta(F, C)|}_{= |(F \setminus C) \cup (C \setminus F)|} \leq \delta$$

In general: $O\left(e^{\frac{n}{\delta}}\right)$ **large!**

BUT

δ-covering

$$\mathcal{C} \subseteq \mathcal{F} \text{ s.t. } \forall F \in \mathcal{F}, \exists C \in \mathcal{C} \text{ with } \underbrace{|\Delta(F,C)|}_{=|(F \setminus C) \cup (C \setminus F)|} \leq \delta$$

In general: $O\left(e^{\frac{n}{\delta}}\right)$ **large!**

BUT

VC-dimension d set systems: $O\left(\left(\frac{n}{\delta}\right)^d\right)$ (Haussler 1995)

δ-covering

$$\mathcal{C} \subseteq \mathcal{F} \text{ s.t. } \forall F \in \mathcal{F}, \exists C \in \mathcal{C} \text{ with } \underbrace{|\Delta(F, C)|}_{= |(F \setminus C) \cup (C \setminus F)|} \leq \delta$$

In general: $O\left(e^{\frac{n}{\delta}}\right)$ **large!**

BUT

VC-dimension d set systems: $O\left(\left(\frac{n}{\delta}\right)^d\right)$ (Haussler 1995)

Can be polynomially computed starting from a random sample of X (ε-net) (Matousek et al. 1993, Louvet 2025)

Putting it all together

Putting it all together

$n^{1-\frac{1}{d}}$-covering of $(X, \mathcal{F})$ of size $O\left(\left(\frac{n}{n^{1-\frac{1}{d}}}\right)^d\right) = O(n)$

Putting it all together

$n^{1-\frac{1}{d}}$-covering of $(X, \mathcal{F})$ of size $O\left(\left(\frac{n}{n^{1-\frac{1}{d}}}\right)^d\right) = O(n)$

↳ Coloring χ with $O(1)$ discrepancy on these ranges

Putting it all together

$n^{1-\frac{1}{d}}$-covering of $(X, \mathcal{F})$ of size $O\left(\left(\frac{n}{n^{1-\frac{1}{d}}}\right)^d\right) = O(n)$

↳ Coloring χ with $O(1)$ discrepancy on these ranges

Other ranges are at distance at most $n^{1-\frac{1}{d}}$

Putting it all together

$n^{1-\frac{1}{d}}$-covering of $(X, \mathcal{F})$ of size $O\left(\left(\frac{n}{n^{1-\frac{1}{d}}}\right)^d\right) = O(n)$

↳ Coloring χ with $O(1)$ discrepancy on these ranges

Other ranges are at distance at most $n^{1-\frac{1}{d}}$

↳ $|\chi(F)| \leq |\chi(C)| + |\chi(F \setminus C)| + |\chi(C \setminus F)|$

Putting it all together

$n^{1-\frac{1}{d}}$-covering of $(X, \mathcal{F})$ of size $O\left(\left(\frac{n}{n^{1-\frac{1}{d}}}\right)^d\right) = O(n)$

↳ Coloring χ with $O(1)$ discrepancy on these ranges

Other ranges are at distance at most $n^{1-\frac{1}{d}}$

↳ $|\chi(F)| \leq |\chi(C)| + |\chi(F \setminus C)| + |\chi(C \setminus F)| = O\left(\sqrt{n^{1-\frac{1}{d}} \log(m)}\right)$

Putting it all together

$n^{1-\frac{1}{d}}$-covering of $(X, \mathcal{F})$ of size $O\left(\left(\frac{n}{n^{1-\frac{1}{d}}}\right)^d\right) = O(n)$

↳ Coloring χ with $O(1)$ discrepancy on these ranges

Other ranges are at distance at most $n^{1-\frac{1}{d}}$

↳ $|\chi(F)| \leq |\chi(C)| + |\chi(F \setminus C)| + |\chi(C \setminus F)| = O\left(\sqrt{n^{1-\frac{1}{d}} \log(m)}\right)$

With a little more effort (chaining), one can obtain $O\left(n^{\frac{1}{2}-\frac{1}{2d}}\right)$ (optimal)

Summary

Summary

We have a difficult problem to solve

Summary

We have a difficult problem to solve **(combinatorial discrepancy)**

Summary

We have a difficult problem to solve **(combinatorial discrepancy)**

We compute an approximation of the data

Summary

We have a difficult problem to solve **(combinatorial discrepancy)**

We compute an approximation of the data **(δ-covering)**

Summary

We have a difficult problem to solve **(combinatorial discrepancy)**

We compute an approximation of the data **(δ-covering)**

We solve the problem really well on the approximation

Summary

We have a difficult problem to solve **(combinatorial discrepancy)**

We compute an approximation of the data **(δ-covering)**

We solve the problem really well on the approximation **($\Theta(1)$ discrepancy)**

Summary

We have a difficult problem to solve **(combinatorial discrepancy)**

We compute an approximation of the data **(δ-covering)**

We solve the problem really well on the approximation **($\Theta(1)$ discrepancy)**

We obtain a good guarantee on the original set system

Summary

We have a difficult problem to solve **(combinatorial discrepancy)**

We compute an approximation of the data **(δ-covering)**

We solve the problem really well on the approximation **($\Theta(1)$ discrepancy)**

We obtain a good guarantee on the original set system **(Optimal order worst-case discrepancy)**

Differential privacy?

Membership inference

Consider the problem of membership inference (Does an element x belong to a set $\mathcal{Q}$?)

Membership inference

Consider the problem of membership inference (Does an element x belong to a set $\mathcal{Q}$?)

To make it DP: Compute true answer and then randomize it (coin toss example)

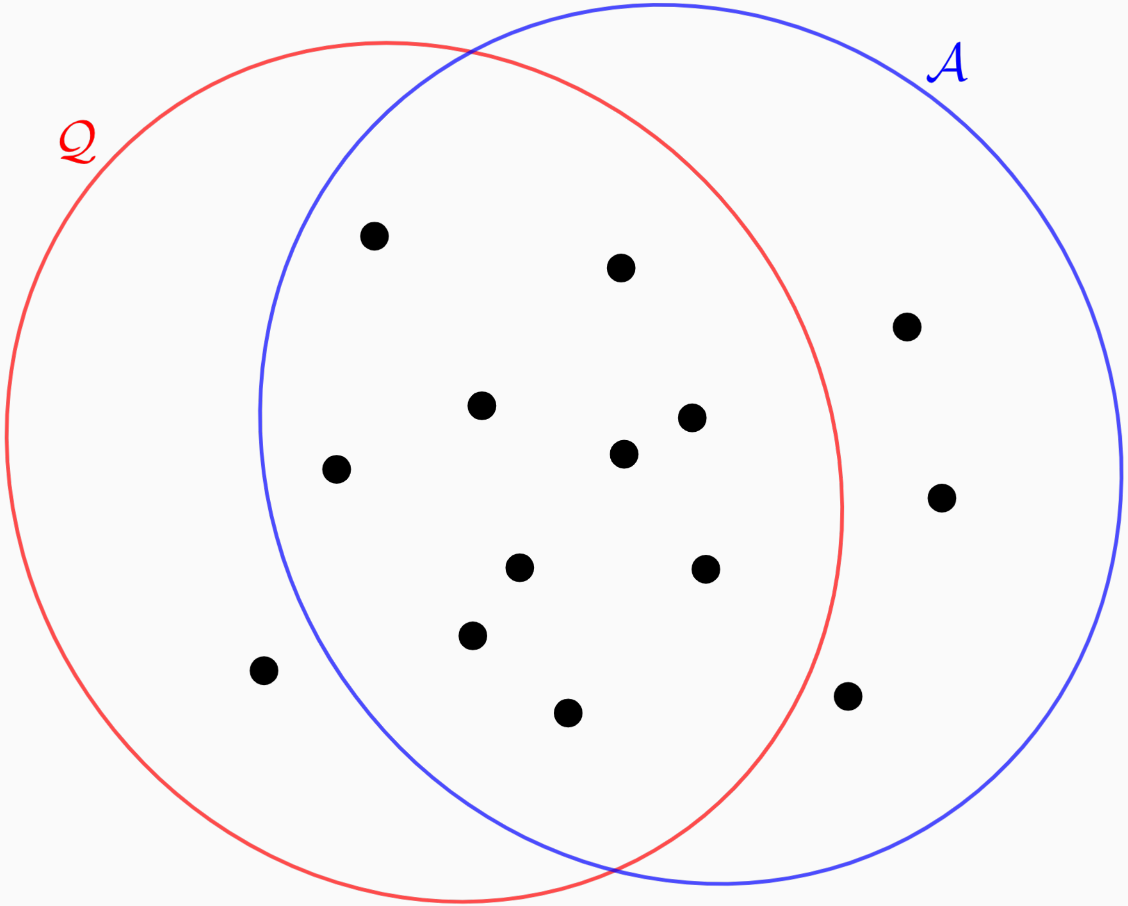
Membership inference

Consider the problem of membership inference (Does an element x belong to a set $\mathcal{Q}$?)

To make it DP: Compute true answer and then randomize it (coin toss example)

What if instead, one answers this membership query by computing inference with the range $\mathcal{A}$ covering $\mathcal{Q}$ in a δ-covering.

Membership inference



Membership inference

Consider the problem of membership inference (Does an element x belong to a set $\mathcal{Q}$?)

To make it DP: Compute true answer and then randomize it (coin toss example)

What if instead, one answers this membership query by computing inference with the range $\mathcal{A}$ covering $\mathcal{Q}$ in a δ-covering.

Membership inference

Consider the problem of membership inference (Does an element x belong to a set $\mathcal{Q}$?)

To make it DP: Compute true answer and then randomize it (coin toss example)

What if instead, one answers this membership query by computing inference with the range $\mathcal{A}$ covering $\mathcal{Q}$ in a δ-covering.

Aim: Improving runtime (less ranges to query from) and maybe utility

Counting query

Consider the counting query (How many x have property y?)

Counting query

Consider the counting query (How many x have property y?)

To make it DP: compute true answer and then randomize it (Laplacian mechanism)

Counting query

Consider the counting query (How many x have property y?)

To make it DP: compute true answer and then randomize it (Laplacian mechanism)

What if instead, one answers by making a counting query on an ε-approximation of the set system and computing proportionality.

Counting query

Consider the counting query (How many x have property y?)

To make it DP: compute true answer and then randomize it (Laplacian mechanism)

What if instead, one answers by making a counting query on an ε-approximation of the set system and computing proportionality.

Aim: Improving runtime (less elements in the counting query) and maybe utility

A prospective paradigm for DP

A prospective paradigm for DP

Current DP paradigm:

Data

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}}$$

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer}$$

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}}$$

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}} \text{Noised answer (DP)}$$

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}} \text{Noised answer (DP)}$$

Prospective DP paradigm:

Data

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}} \text{Noised answer (DP)}$$

Prospective DP paradigm:

$$\text{Data} \xrightarrow[\text{intermediate complexity}]{\text{Approximation}}$$

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}} \text{Noised answer (DP)}$$

Prospective DP paradigm:

$$\text{Data} \xrightarrow[\text{intermediate complexity}]{\text{Approximation}} \text{Approximated/Noised data}$$

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}} \text{Noised answer (DP)}$$

Prospective DP paradigm:

$$\text{Data} \xrightarrow[\text{intermediate complexity}]{\text{Approximation}} \text{Approximated/Noised data} \xrightarrow[\text{intermediate complexity}]{\text{Algorithmic process}}$$

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}} \text{Noised answer (DP)}$$

Prospective DP paradigm:

$$\text{Data} \xrightarrow[\text{intermediate complexity}]{\text{Approximation}} \text{Approximated/Noised data} \xrightarrow[\text{intermediate complexity}]{\text{Algorithmic process}} \text{Noised answer (DP?)}$$

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}} \text{Noised answer (DP)}$$

Prospective DP paradigm:

$$\text{Data} \xrightarrow[\text{intermediate complexity}]{\text{Approximation}} \text{Approximated/Noised data} \xrightarrow[\text{intermediate complexity}]{\text{Algorithmic process}} \text{Noised answer (DP?)}$$

Some technical challenges:

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}} \text{Noised answer (DP)}$$

Prospective DP paradigm:

$$\text{Data} \xrightarrow[\text{intermediate complexity}]{\text{Approximation}} \text{Approximated/Noised data} \xrightarrow[\text{intermediate complexity}]{\text{Algorithmic process}} \text{Noised answer (DP?)}$$

Some technical challenges:

- Characterizing the distribution of errors where most results focus only on worst case error.

A prospective paradigm for DP

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}} \text{Noised answer (DP)}$$

Prospective DP paradigm:

$$\text{Data} \xrightarrow[\text{intermediate complexity}]{\text{Approximation}} \text{Approximated/Noised data} \xrightarrow[\text{intermediate complexity}]{\text{Algorithmic process}} \text{Noised answer (DP?)}$$

Some technical challenges:

- Characterizing the distribution of errors where most results focus only on worst case error.
- Extending some algorithms to return “out-data” results.

Current DP paradigm:

$$\text{Data} \xrightarrow[\text{high complexity}]{\text{Algorithmic process}} \text{Exact answer} \xrightarrow[\text{low complexity}]{\text{Noise addition}} \text{Noised answer (DP)}$$

Prospective DP paradigm:

$$\text{Data} \xrightarrow[\text{intermediate complexity}]{\text{Approximation}} \text{Approximated/Noised data} \xrightarrow[\text{intermediate complexity}]{\text{Algorithmic process}} \text{Noised answer (DP?)}$$

Some technical challenges:

- Characterizing the distribution of errors where most results focus only on worst case error.
- Extending some algorithms to return “out-data” results.

Thank you.

Bibliography

Haussler D (1995) Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. Journal of Combinatorial Theory, Series A 69:

Louvet A (2025) Faster Algorithms for Approximating Combinatorial and Geometric Data. Doctoral dissertation

Matousek J, Welzl E, Wernisch L (1993) Discrepancy and ε-approximations for bounded VC-dimension. Combinatorica 13: