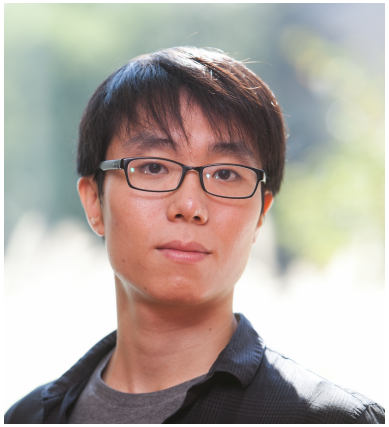
Clustering

How hard is it to Classify Data?

Karthik C. S.
(Tel Aviv University)

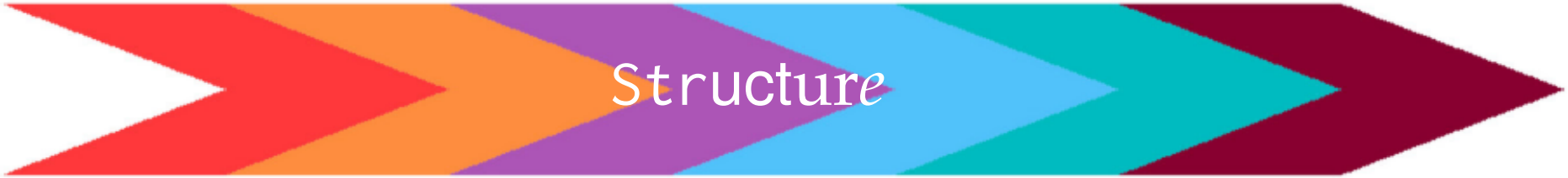
Joint work(s) with

Vincent Cohen-Addad
(Sorbonne University)



Euiwoong Lee
(New York University)

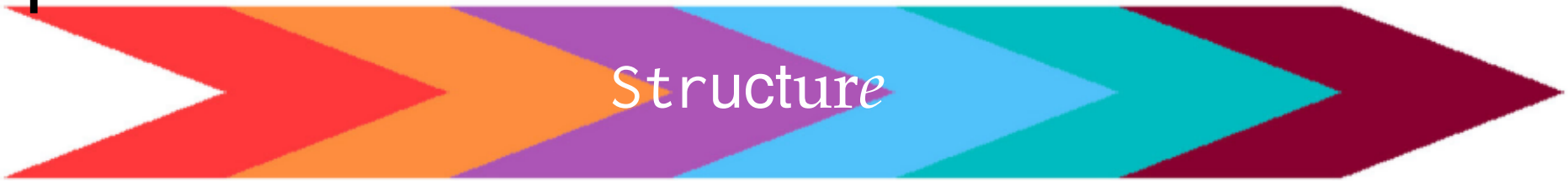
Spectrum of Computational Problems



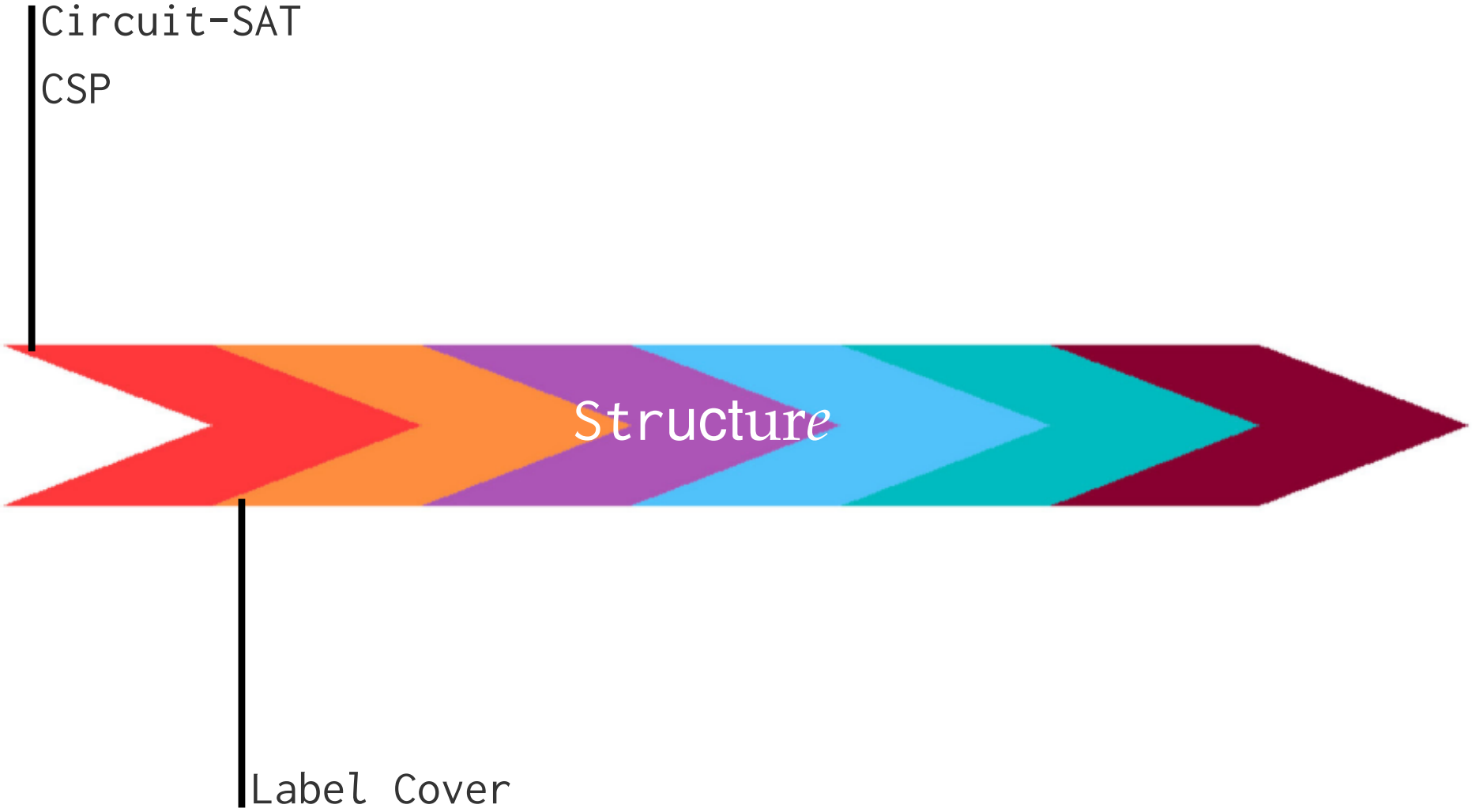
Spectrum of Computational Problems

Circuit-SAT

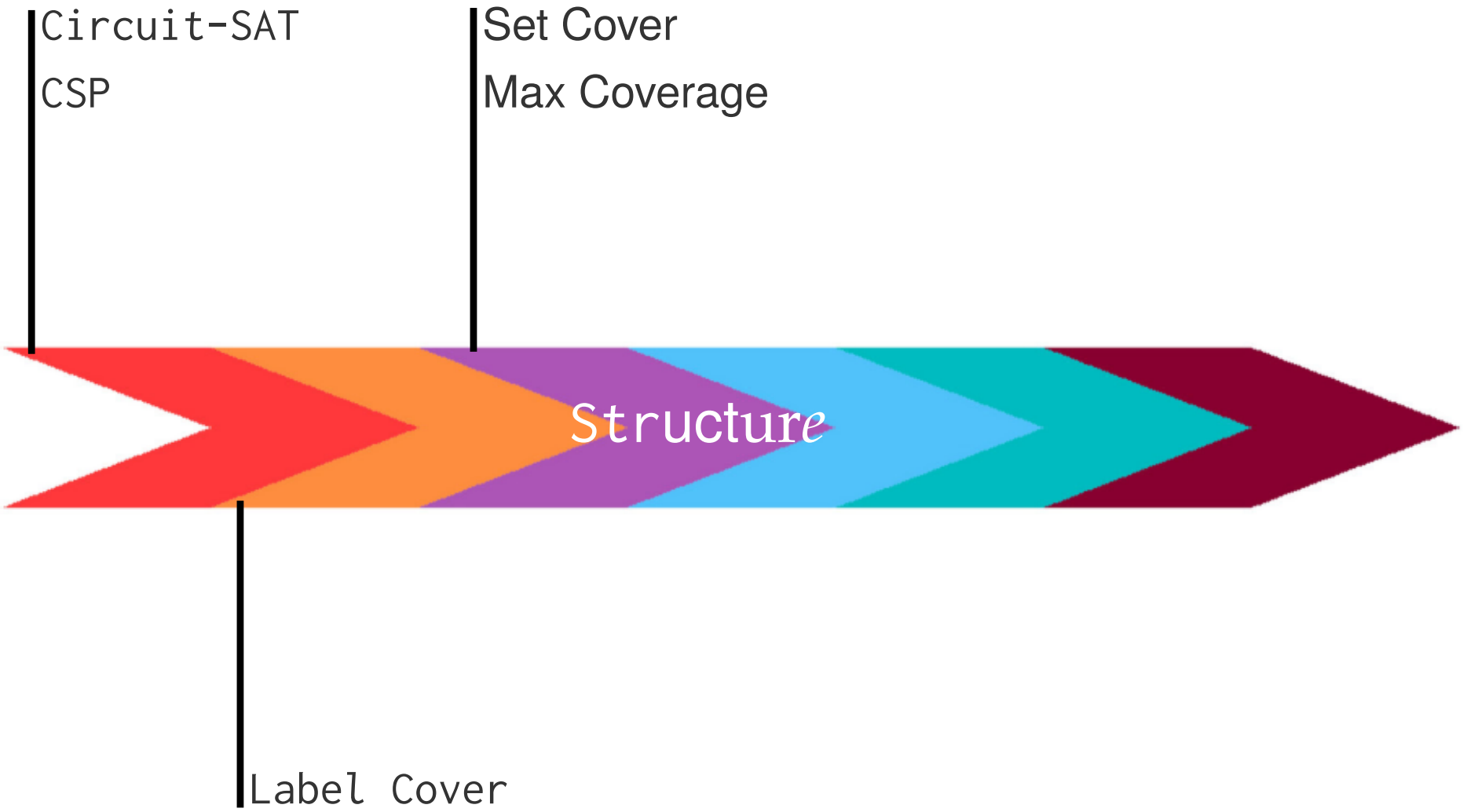
CSP



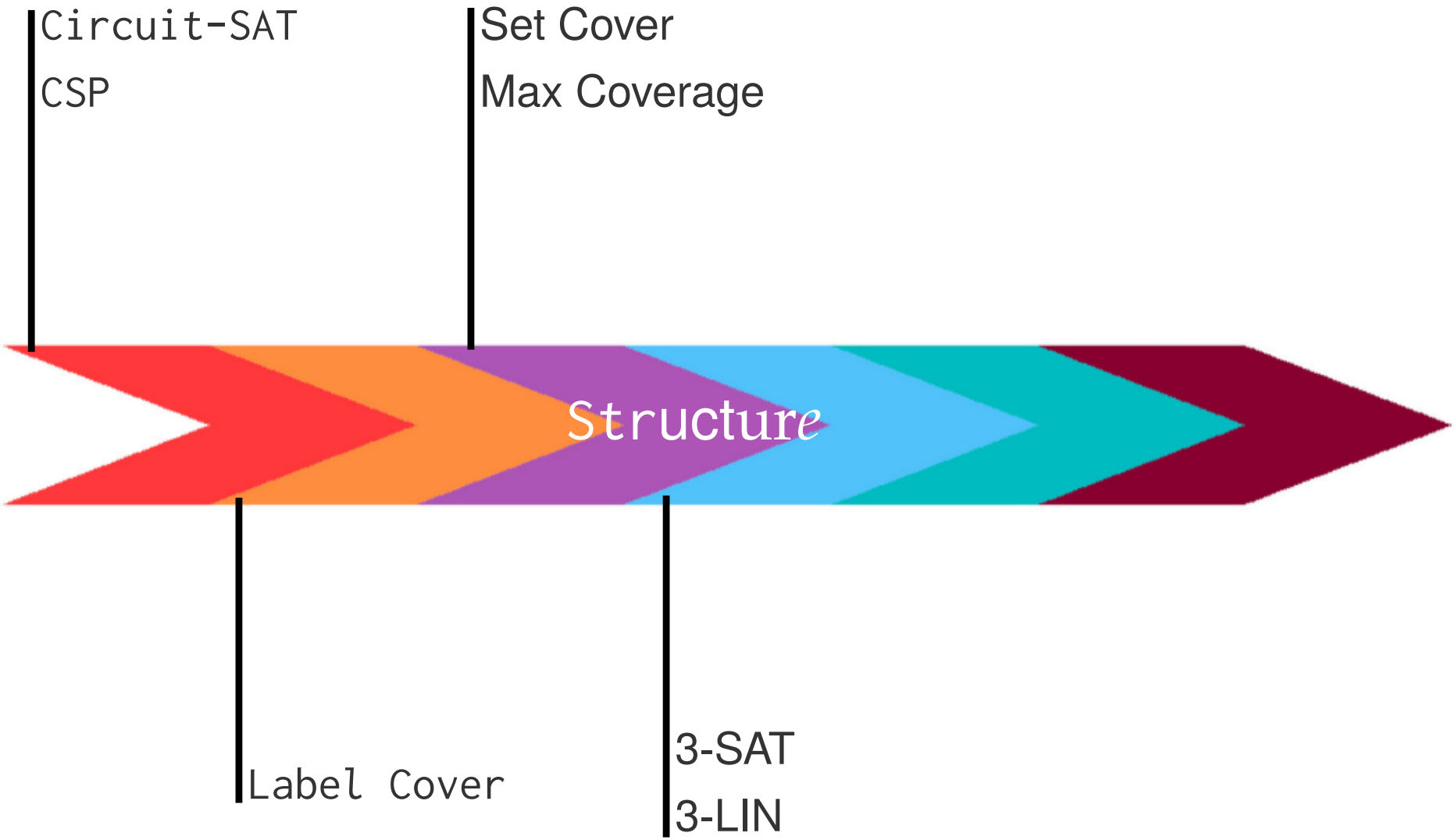
Spectrum of Computational Problems



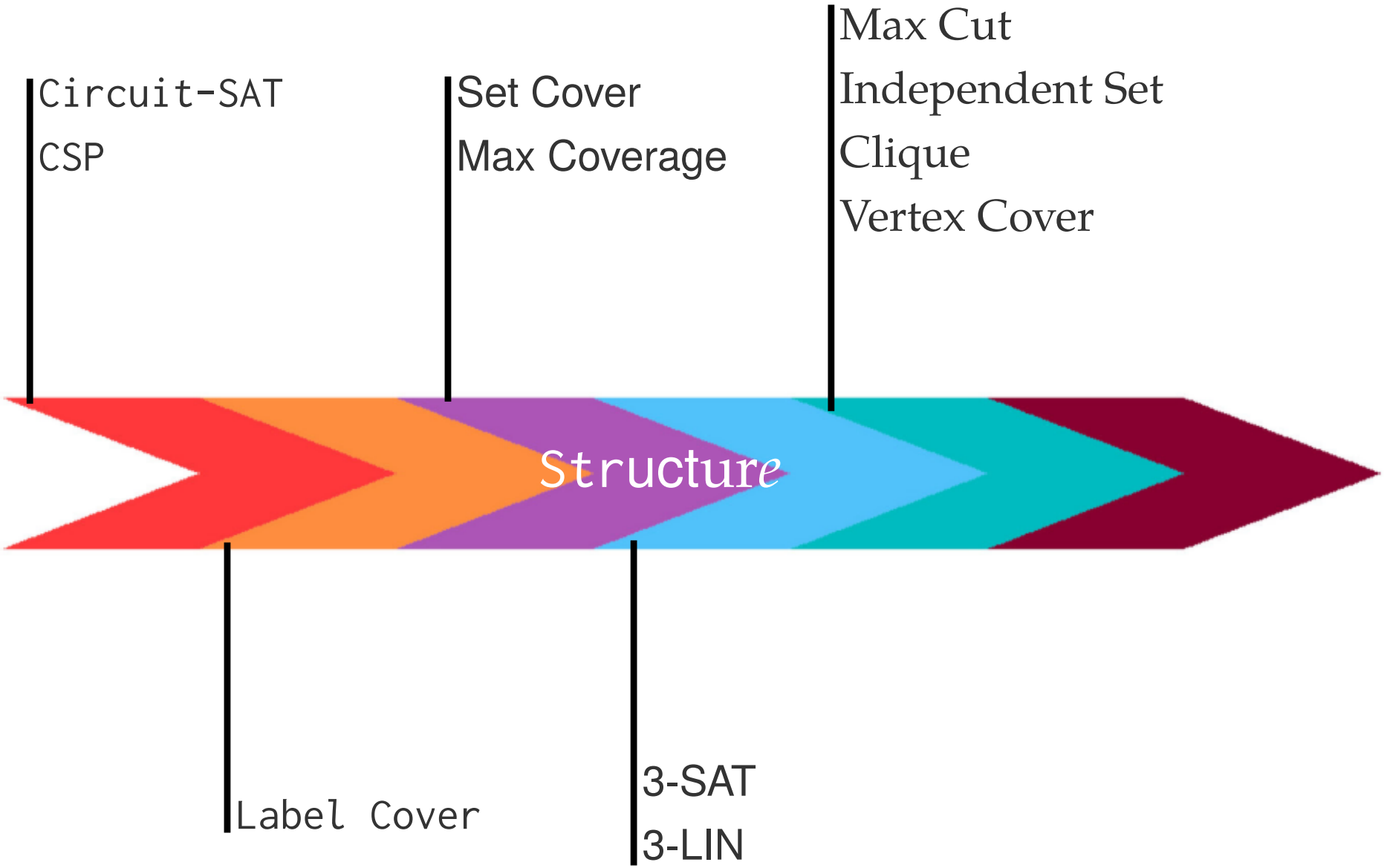
Spectrum of Computational Problems



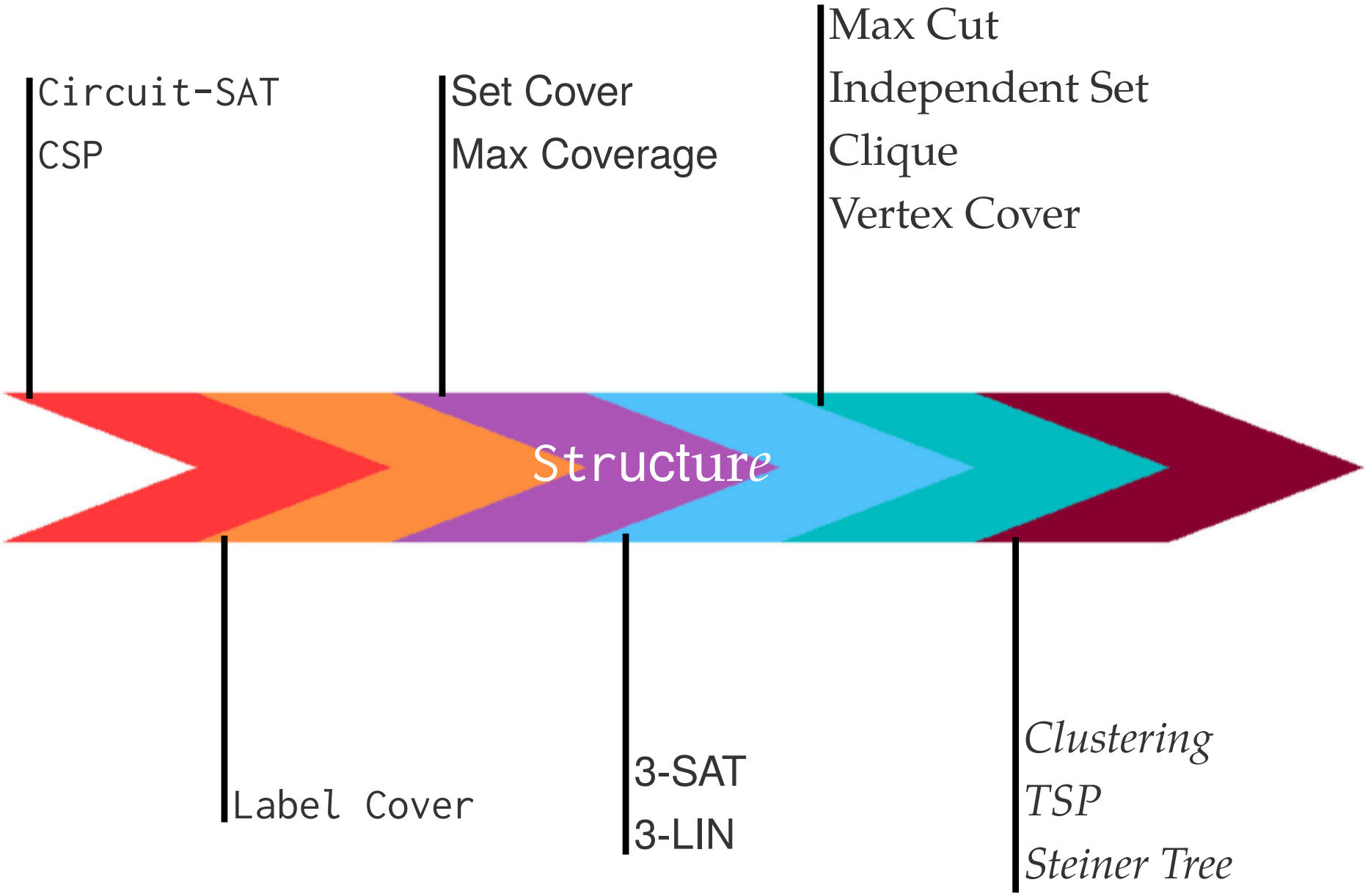
Spectrum of Computational Problems



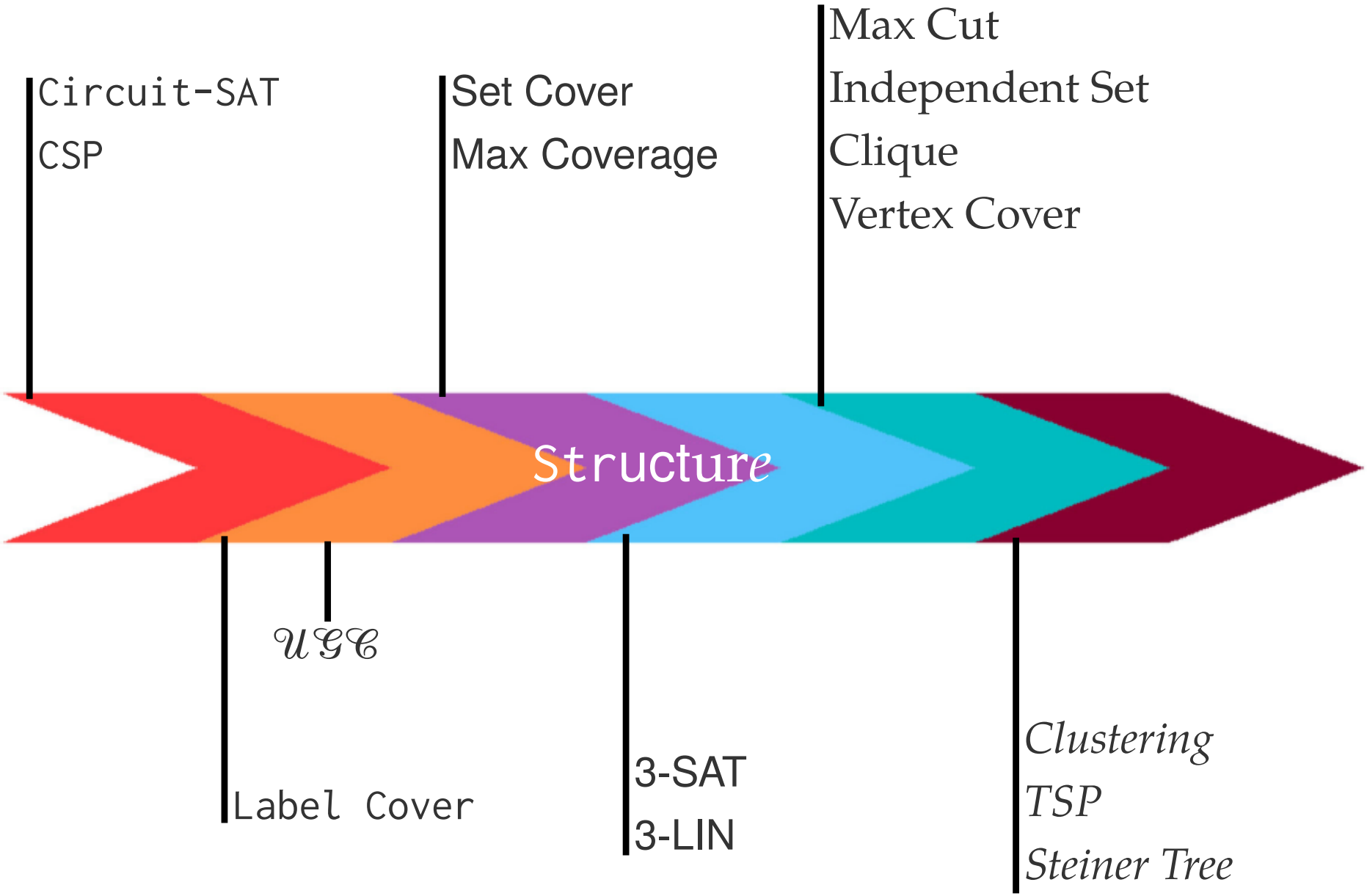
Spectrum of Computational Problems



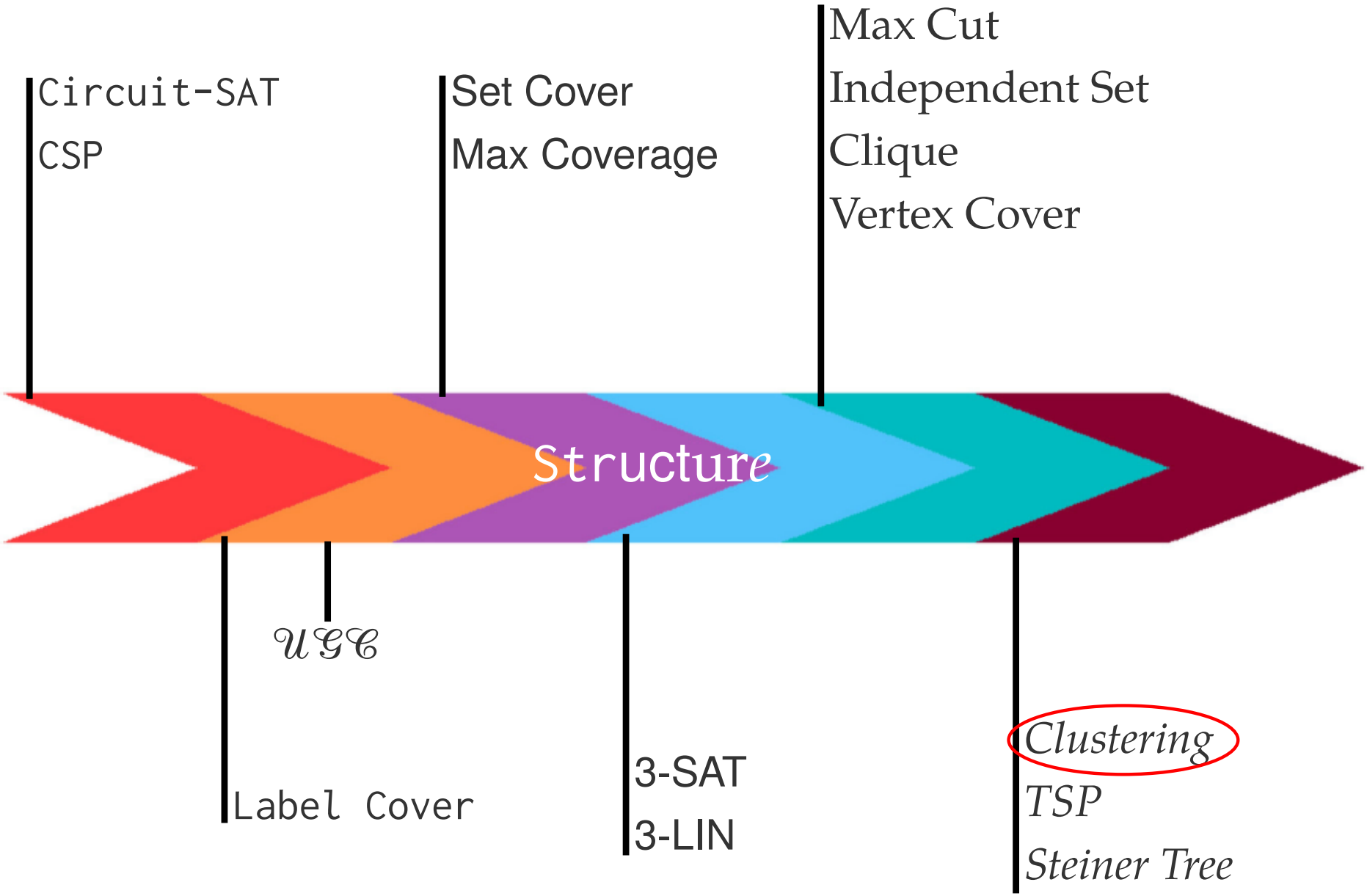
Spectrum of Computational Problems



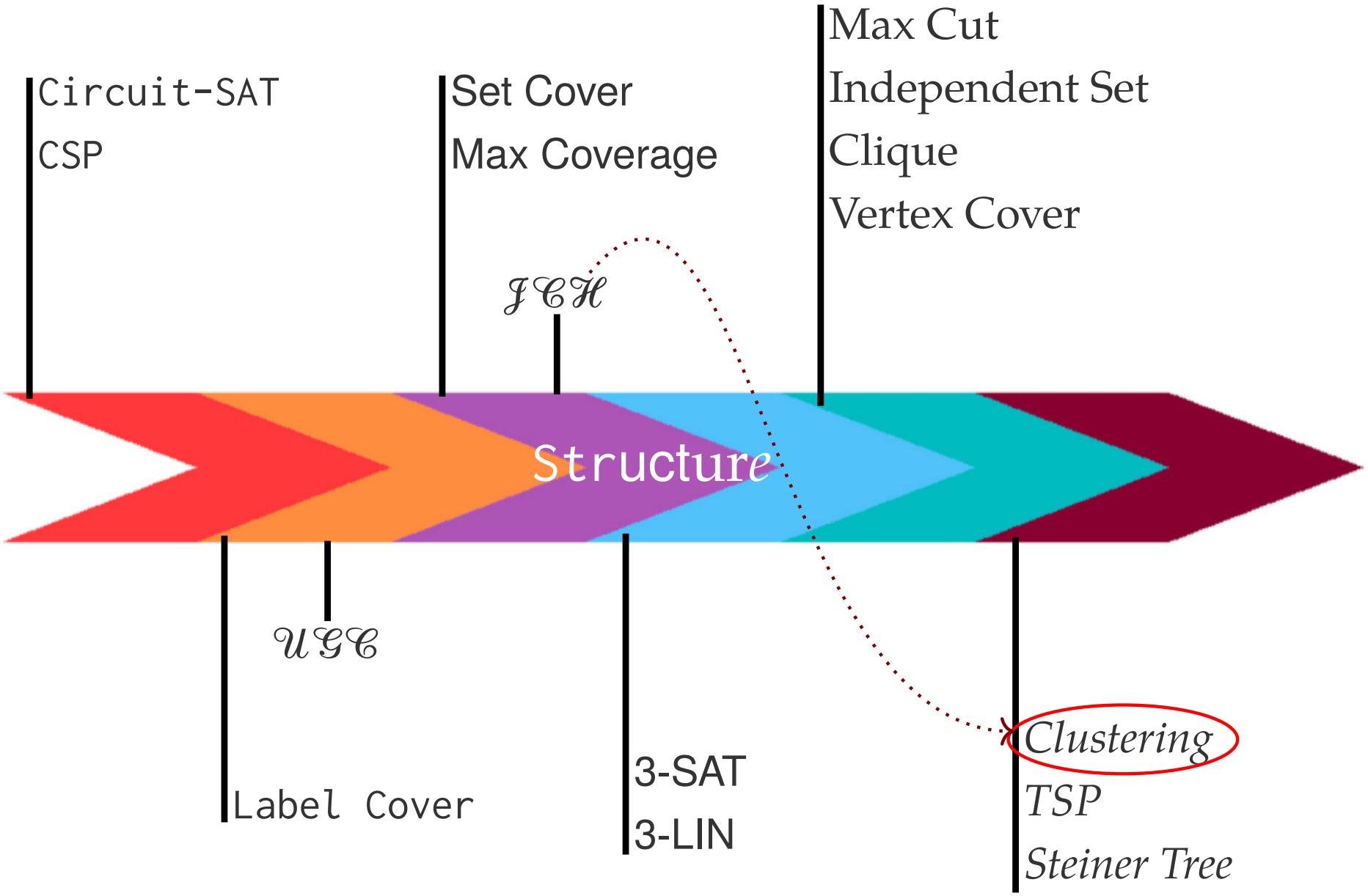
Spectrum of Computational Problems



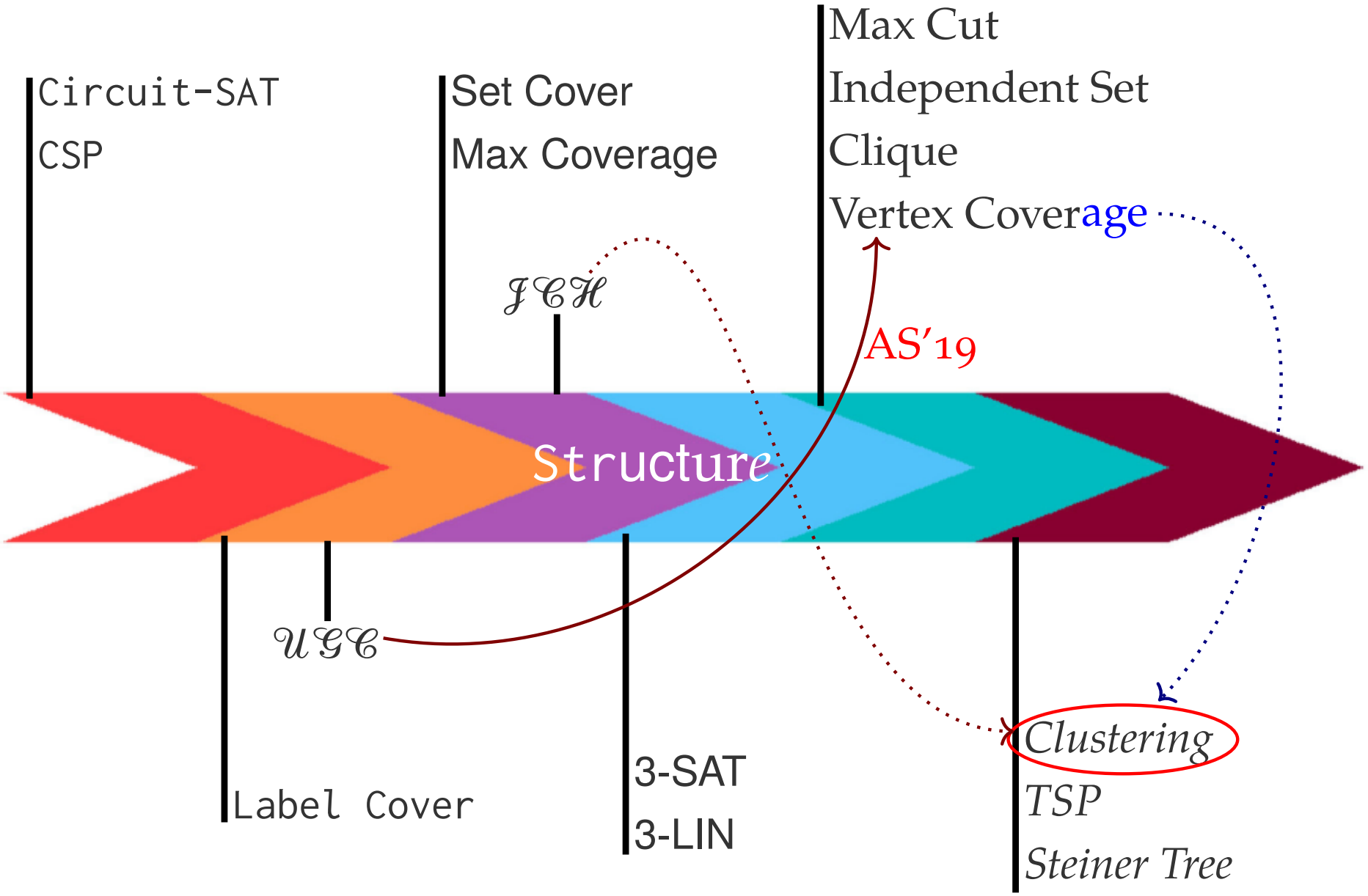
Spectrum of Computational Problems



Spectrum of Computational Problems

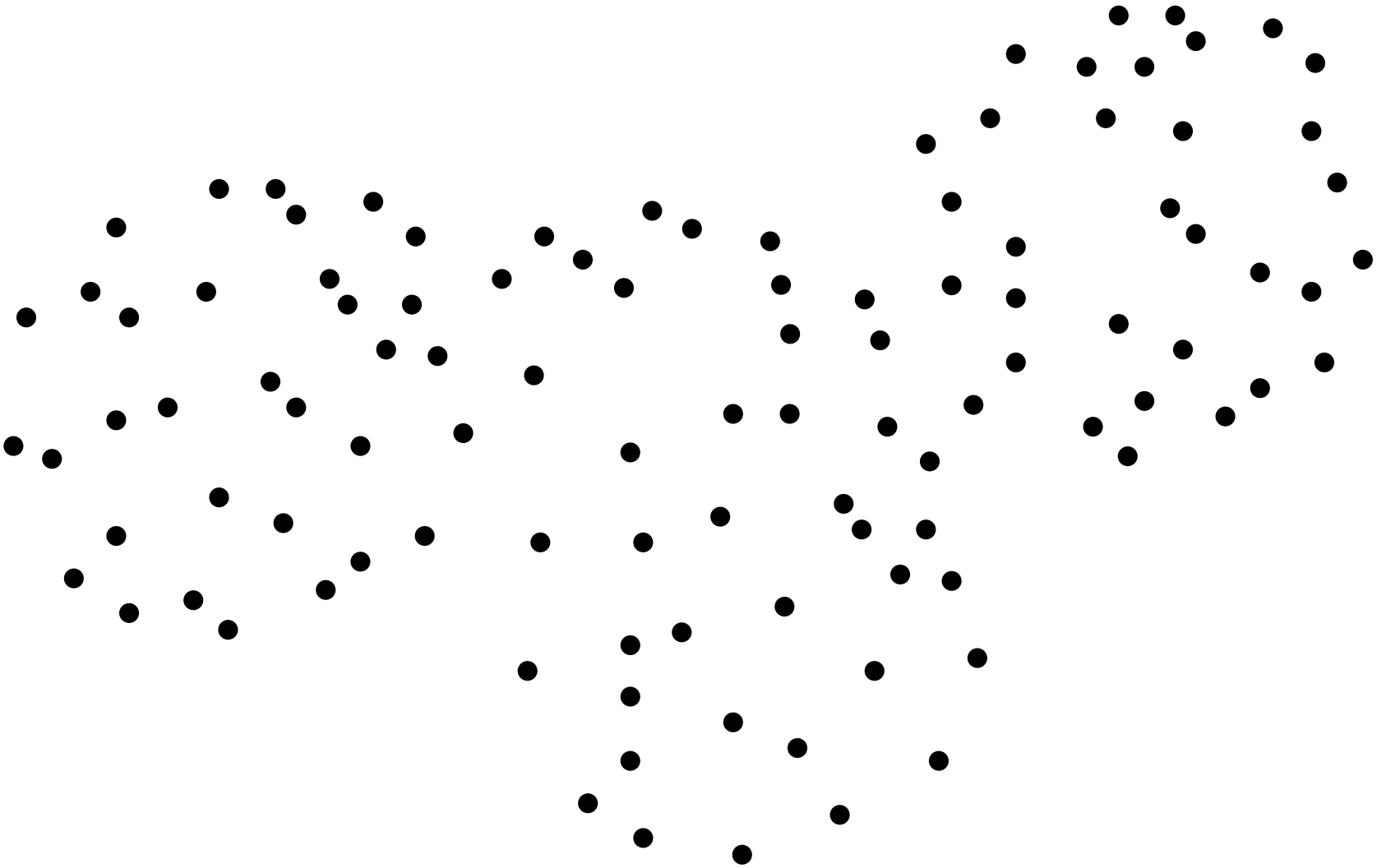


Spectrum of Computational Problems

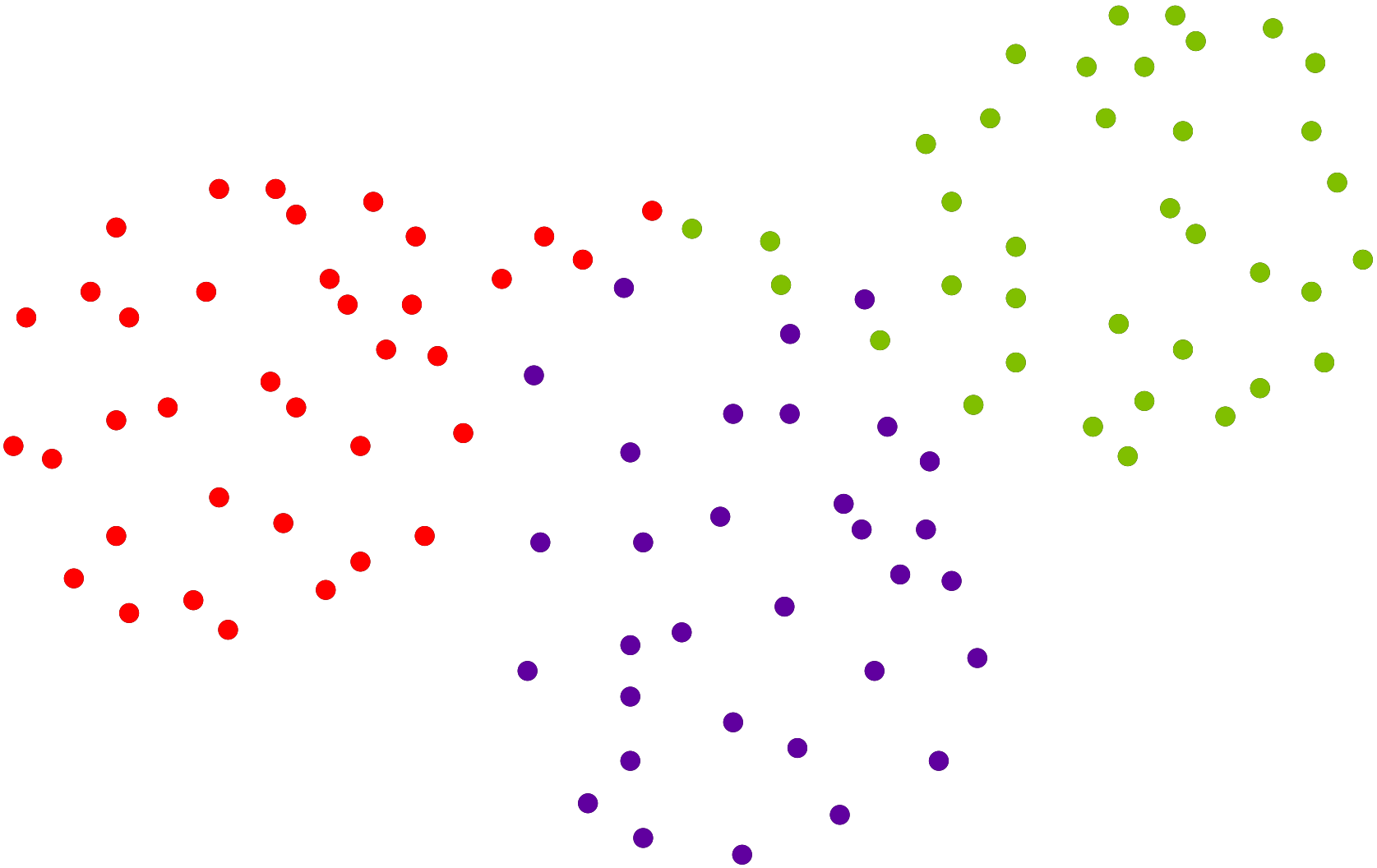


What is Clustering?

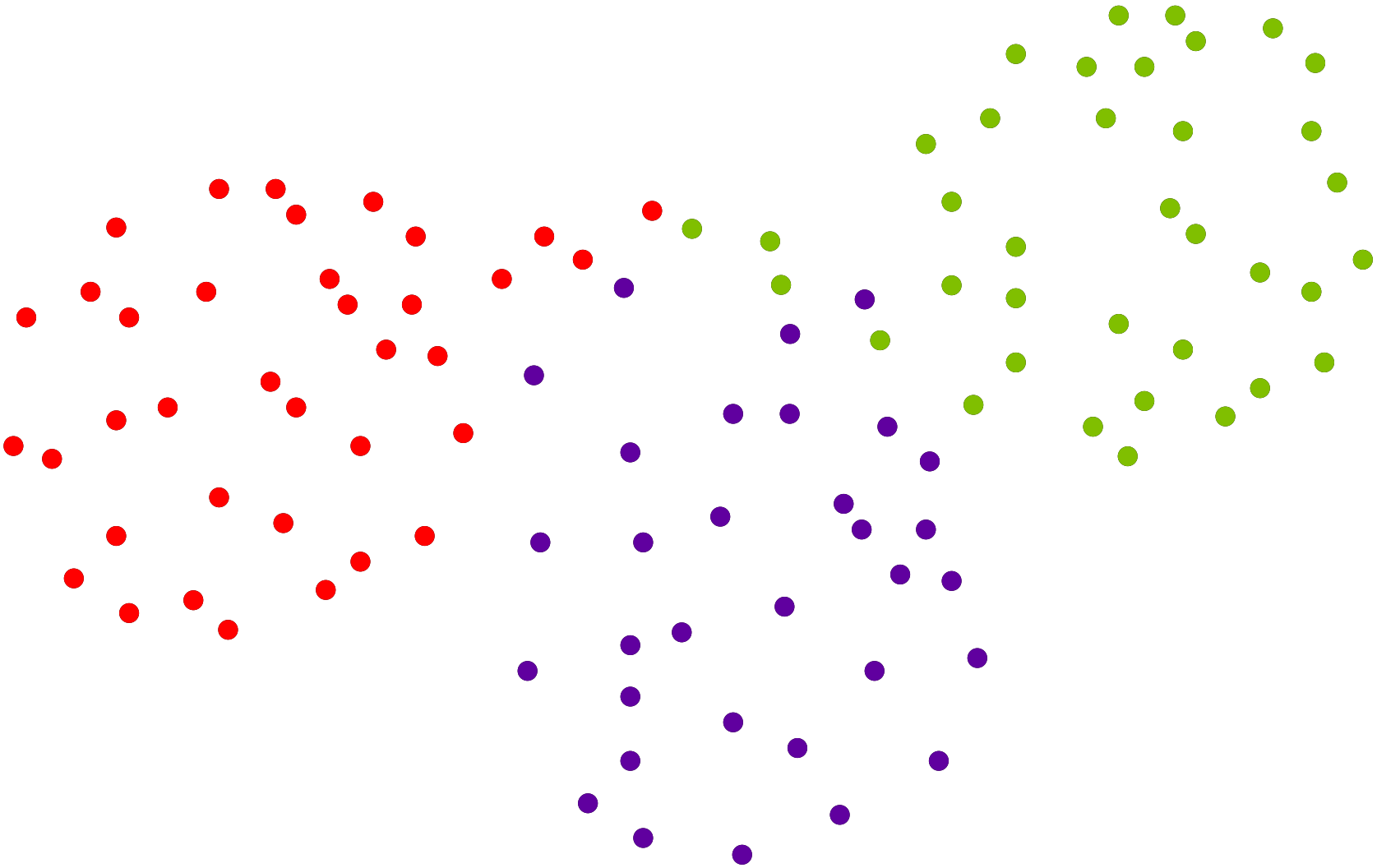
What is Clustering?



What is Clustering?



What is Clustering?



Task of Classifying Input Data

What is Clustering?

- (Γ, Δ) is a metric space

What is Clustering?

- (Γ, Δ) is a metric space
- <u>Input</u>: $X \subseteq \Gamma$, $k \in \mathbb{N}$

What is Clustering?

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):

What is Clustering?

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq \Gamma$ and $|C| = k$

What is Clustering?

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \to C$

What is Clustering?

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \to C$
  - σ is *good*

Continuous Version

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \to C$
  - σ is *good*

Discrete
~~Continuous~~ Version

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \to C$
  - σ is *good*

Discrete
~~Continuous~~ Version

- (Γ, Δ) is a metric space
- <u>Input</u>: $X \subseteq \Gamma$, $k \in \mathbb{N}$ and $\mathcal{S} \subseteq \Gamma$
- <u>Output</u>: A classification (C, σ):
  - $C \subseteq \cancel{\Gamma}\ \mathcal{S}$ and $|C| = k$
  - $\sigma : X \to C$
  - σ is *good*

What is Good Classification?

- k-means, k-median, k-center, min-sum, etc.

What is Good Classification?

- k-means, k-median, k-center, min-sum, etc.
- k-median value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))$$

What is Good Classification?

- k-means, k-median, k-center, min-sum, etc.
- k-median value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))$$

- k-means value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))^2$$

What is Good Classification?

- k-means, k-median, k-center, min-sum, etc.
- k-median value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))$$

- k-means value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))^2$$

Clustering Problem for objective Λ

What is Good Classification?

- k-means, k-median, k-center, min-sum, etc.
- k-median value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))$$

- k-means value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))^2$$

Clustering Problem for objective Λ

Yes: There is classification (C^*, σ^*), such that $\Lambda(X, \sigma^*) \leq \beta$

What is Good Classification?

- k-means, k-median, k-center, min-sum, etc.
- k-median value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))$$

- k-means value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))^2$$

Clustering Problem for objective Λ

Yes: There is classification (C^*, σ^*), such that $\Lambda(X, \sigma^*) \leq \beta$

No: For all classification (C, σ), we have $\Lambda(X, \sigma) > (1 + \delta) \cdot \beta$

Exact Computation

- NP-hard when $k = 2$ (Dasgupta'07)

Exact Computation

- NP-hard when $k = 2$ (Dasgupta'07)
- NP-hard in Euclidean plane (Megiddo–Supowit'84, Mahajan–Nimbhorkar–Varadarajan'12)

Exact Computation

- NP-hard when $k = 2$ (Dasgupta'07)
- NP-hard in Euclidean plane (Megiddo–Supowit'84, Mahajan–Nimbhorkar–Varadarajan'12)
- W[2]-hard in general metric (Guha-Khuller'99)

Approximation Algorithms

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)

Approximation Algorithms

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)

- General metric: k-median ≥ 2.67
  (Byrka–Pensyl–Rybicki–Srinivasan–Trinh'17)

Approximation Algorithms

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
- General metric: k-median ≥ 2.67
  (Byrka–Pensyl–Rybicki–Srinivasan–Trinh'17)
- Euclidean metric k-means:

Approximation Algorithms

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)

- General metric: k-median ≥ 2.67
  (Byrka–Pensyl–Rybicki–Srinivasan–Trinh'17)

- Euclidean metric k-means:

  - Poly time approximation ≈ 6.357
    (Ahmadian–Norouzi-Fard–Svensson–Ward'17)

Approximation Algorithms

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
- General metric: k-median ≥ 2.67
  (Byrka–Pensyl–Rybicki–Srinivasan–Trinh'17)
- Euclidean metric k-means:
  - Poly time approximation ≈ 6.357
    (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
  - Fixed Dimension: PTAS (Cohen-Addad'18)

Approximation Algorithms

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
- General metric: k-median ≥ 2.67
  (Byrka–Pensyl–Rybicki–Srinivasan–Trinh'17)
- Euclidean metric k-means:
  - Poly time approximation ≈ 6.357
    (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
  - Fixed Dimension: PTAS (Cohen-Addad'18)
  - Fixed k: PTAS (Kumar–Sabharwal–Sen'10)

Discrete Version:

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)
- ℓ_2-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)
- ℓ_1-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)
- ℓ_2-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)
- ℓ_1-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)
- ℓ_∞-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Guruswami-Indyk'03)

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)
- ℓ_2-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)
- ℓ_1-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)
- ℓ_∞-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Guruswami-Indyk'03)

Continuous Version:

k-means in Euclidean metric < 1.0013
(Lee-Schmidt-Wright'17)

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)
- ℓ_2-metric: k-means $\ll$ ~~1.01~~ 1.73, k-median $\ll$ ~~1.01~~ 1.27 (Trevisan'00)
- ℓ_1-metric: k-means $\ll$ ~~1.01~~ 3.94, k-median $\ll$ ~~1.01~~ 1.73 (Trevisan'00)
- ℓ_∞-metric: k-means $\ll$ ~~1.01~~ 3.94, k-median $\ll$ ~~1.01~~ 1.73 (Guruswami-Indyk'03)

Continuous Version:

k-means in Euclidean metric $<$ ~~1.0013~~ 1.36
(Lee-Schmidt-Wright'17)

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)
- ℓ_2-metric: k-means $\ll$ ~~1.01~~ 1.73, k-median $\ll$ ~~1.01~~ 1.27, 1.17, 1.06 (Trevisan'00)
- ℓ_1-metric: k-means $\ll$ ~~1.01~~ 3.94, k-median $\ll$ ~~1.01~~ 1.73, 1.56, 1.14 (Trevisan'00)
- ℓ_∞-metric: k-means $\ll$ ~~1.01~~ 3.94, k-median $\ll$ ~~1.01~~ 1.73, 3.94, 1.73 (Guruswami-Indyk'03)

Continuous Version:

k-means in Euclidean metric $<$ ~~1.0013~~ 1.36, 1.07
(Lee-Schmidt-Wright'17)

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94^*	1.73^*

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94^*	1.73^*

Continuous Version

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94^*	1.73^*

Continuous Version

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

A New Embedding Framework to potentially get Strong (tight?) Inapproximability results!

Warm up: General Metrics

Warm up: General Metrics

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$

Warm up: General Metrics

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Warm up: General Metrics

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

Warm up: General Metrics

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

Warm up: General Metrics

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is at most $1 - 1/e + \varepsilon$

Warm up: General Metrics

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is at most $1 - 1/e + \varepsilon$

$$\Downarrow$$

Warm up: General Metrics

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is at most $1 - 1/e + \varepsilon$

$$\Downarrow$$

Theorem (Guha-Khuller'99)

Fix $\varepsilon > 0$. Given input X. It is NP-hard to distinguish:

Warm up: General Metrics

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is at most $1 - 1/e + \varepsilon$

$$\Downarrow$$

Theorem (Guha-Khuller'99)

Fix $\varepsilon > 0$. Given input X. It is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \Delta(x, \sigma^*(x)) \leq |X|$

Warm up: General Metrics

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is at most $1 - 1/e + \varepsilon$

$$\Downarrow$$

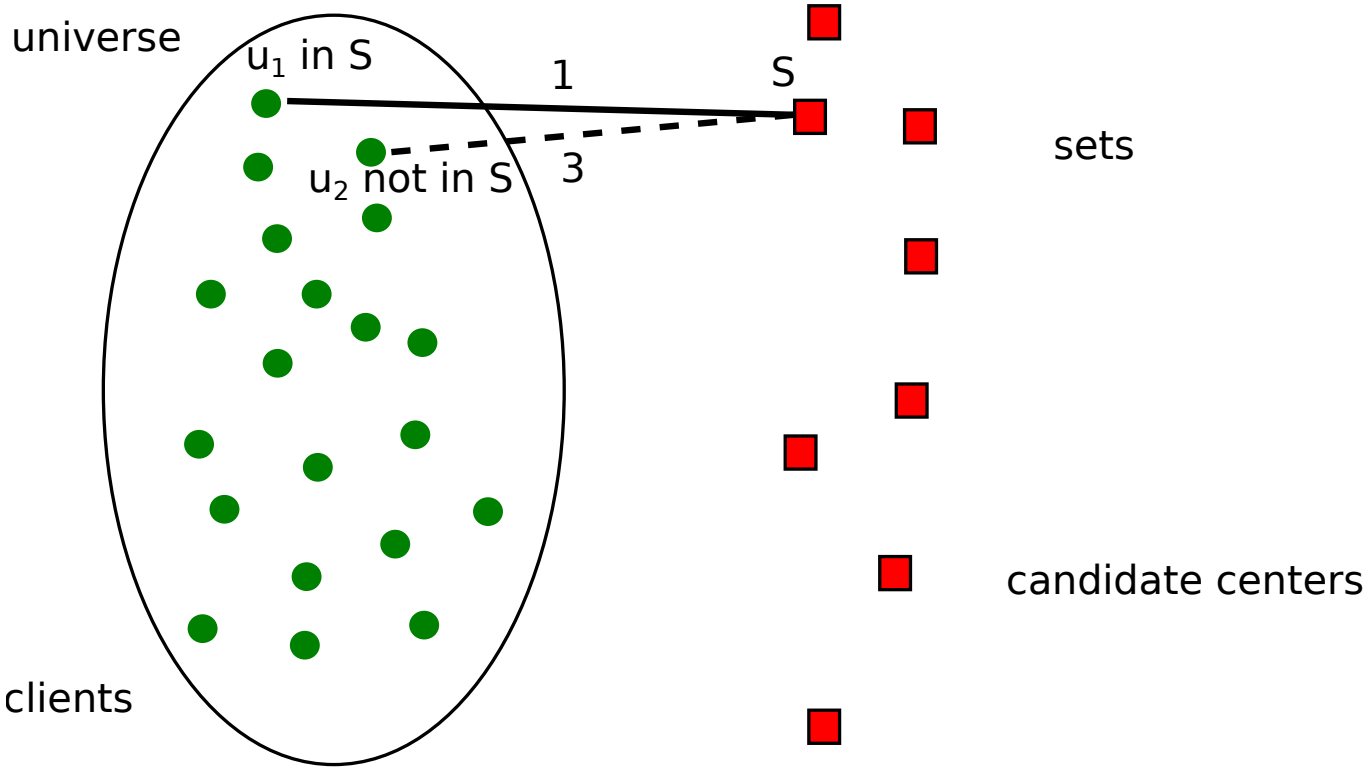
Theorem (Guha-Khuller'99)

Fix $\varepsilon > 0$. Given input X. It is NP-hard to distinguish:

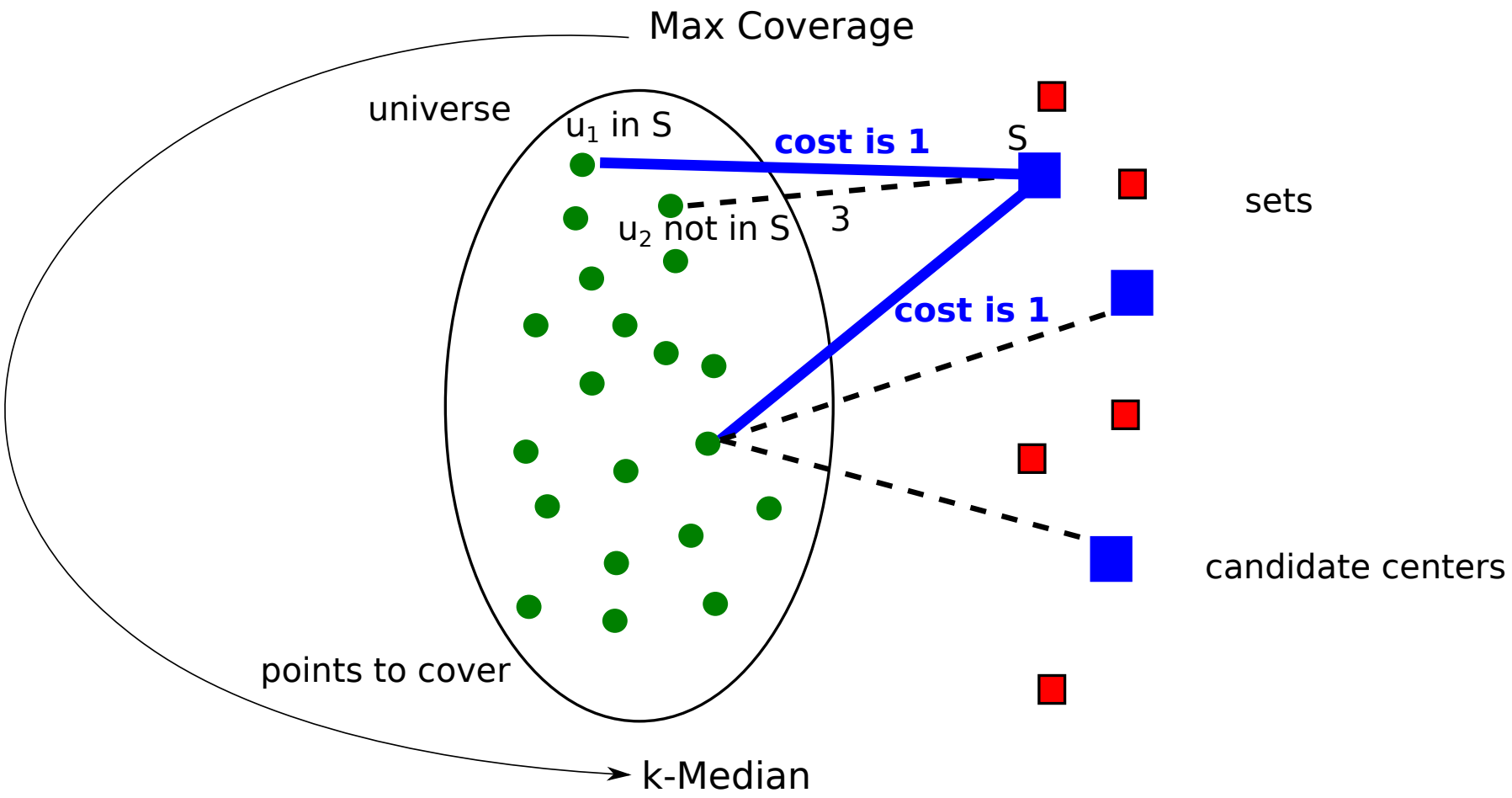
YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \Delta(x, \sigma^*(x)) \leq |X|$

NO: For all (C, σ) we have $\sum_{x \in X} \Delta(x, \sigma(x)) \geq (1 + 2/e - \varepsilon) \cdot |X|$

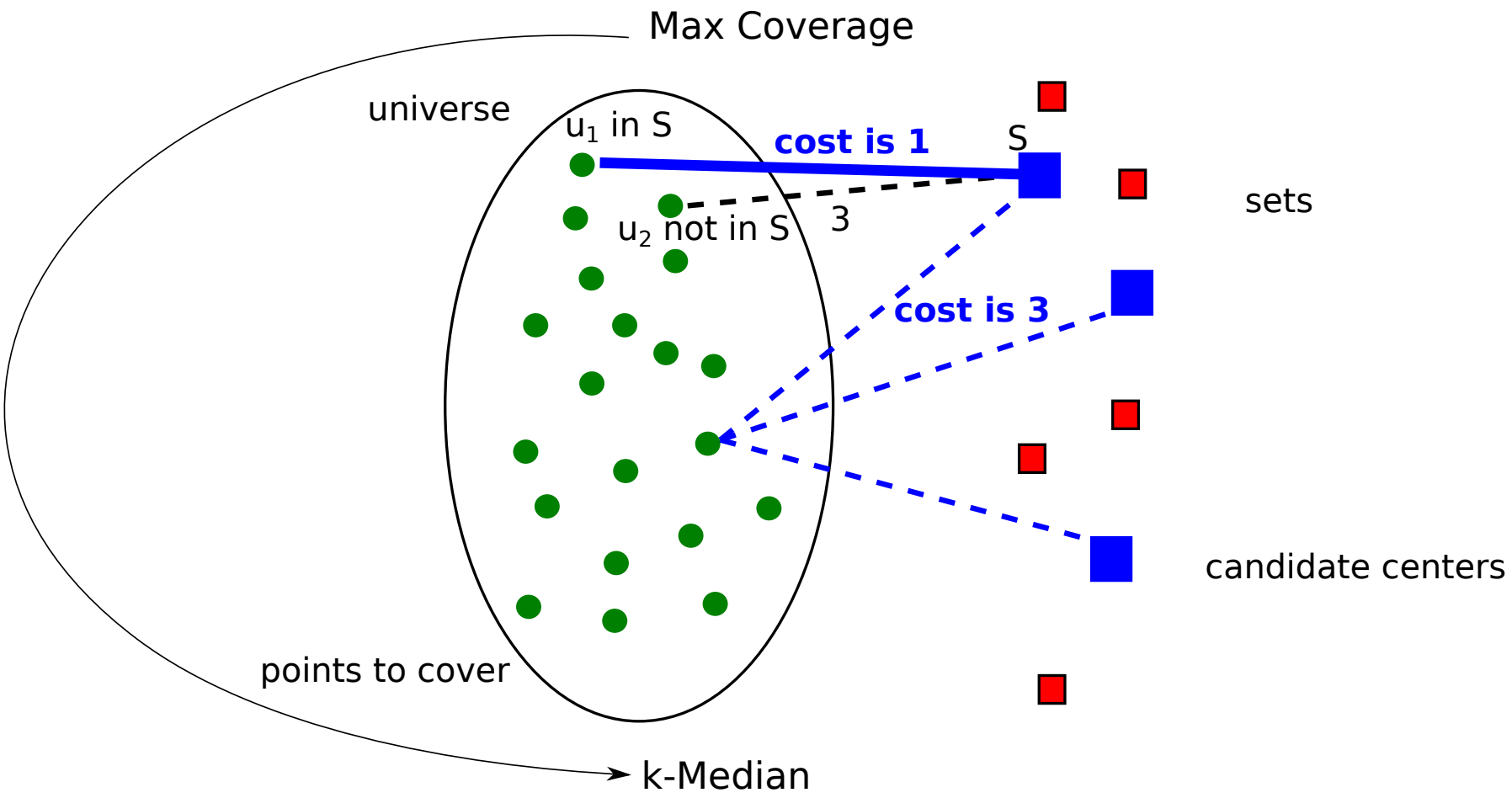
Warm up: General Metrics



Warm up: General Metrics



Warm up: General Metrics



Johnson Coverage Hypothesis

(α, t)-Johnson Coverage Problem

Given $E \subseteq \binom{[n]}{t}$, and k as input, distinguish between:

Completeness: There exists $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ such that

$$\forall T \in E,\ \exists S_i \in \mathscr{C},\ S_i \subset T.$$

Soundness: For every $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ we have

$$\Pr_{T \sim E}[\exists S_i,\ S_i \subset T] \leq \alpha.$$

Johnson Coverage Hypothesis

(α, t)-Johnson Coverage Problem

Given $E \subseteq \binom{[n]}{t}$, and k as input, distinguish between:

Completeness: There exists $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ such that

$$\forall T \in E,\ \exists S_i \in \mathscr{C},\ S_i \subset T.$$

Soundness: For every $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ we have

$$\Pr_{T \sim E}[\exists S_i,\ S_i \subset T] \leq \alpha.$$

Johnson Coverage Hypothesis (Cohen-Addad–K–Lee)

$\forall \varepsilon > 0, \exists t_\varepsilon \in \mathbb{N}$ such that $(1 - \frac{1}{e} + \varepsilon, t_\varepsilon)$-Johnson Coverage problem is NP-hard.

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$
- LP Integrality gap:

Determine smallest collection in $\binom{[n]}{t-1}$ that hits all of $\binom{[n]}{t}$

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$
- LP Integrality gap:

Determine smallest collection in $\binom{[n]}{t-1}$ that hits all of $\binom{[n]}{t}$

  - Hypergraph Turán number: Open since 1940s!

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$
- LP Integrality gap:

Determine smallest collection in $\binom{[n]}{t-1}$ that hits all of $\binom{[n]}{t}$

- Hypergraph Turán number: Open since 1940s!
- Recently resolved for $t = 3$
- Improved SDP gaps for Clustering

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming $(1\text{-}\frac{1}{e}, t)$-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming $(1\text{-}\frac{1}{e}, t)$-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq \left(1 + \frac{8}{e}\right) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming $(0.93,2)$-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming $(0.93,2)$-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

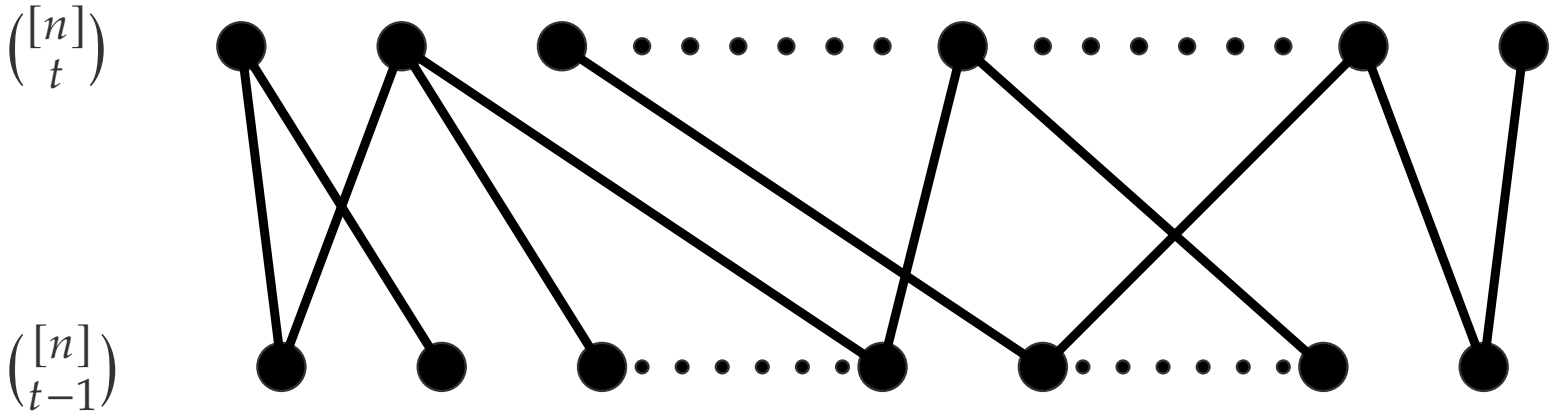
YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

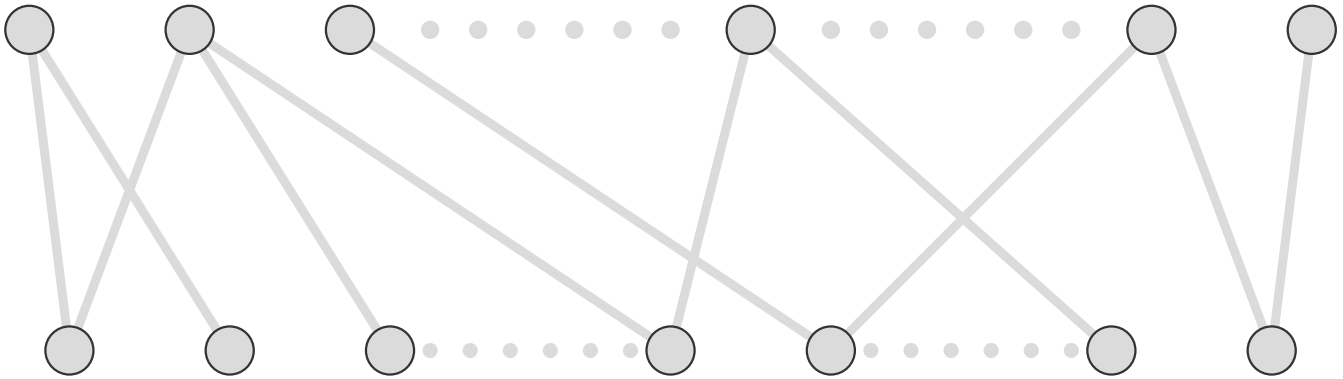
$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq 1.56 \cdot n'.$$

Johnson Graph Embedding



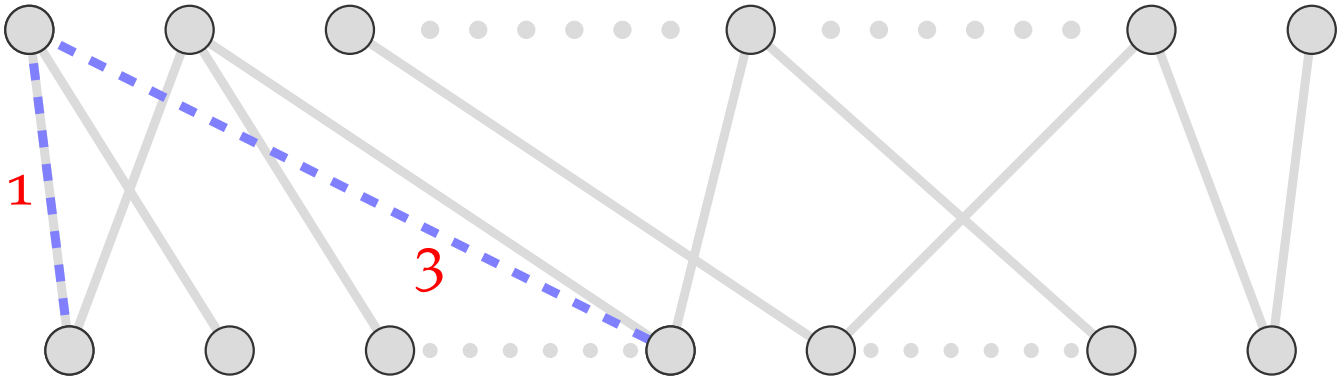
Johnson Graph Embedding

Points in $\{0,1\}^d$



Johnson Graph Embedding

Points in $\{0,1\}^d$

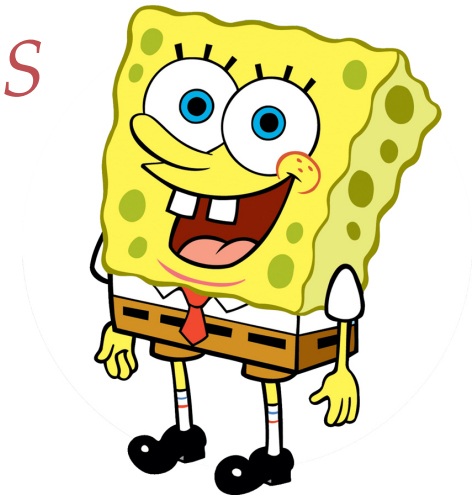


Containment Game

Containment Game

$T \in \binom{[n]}{t}$

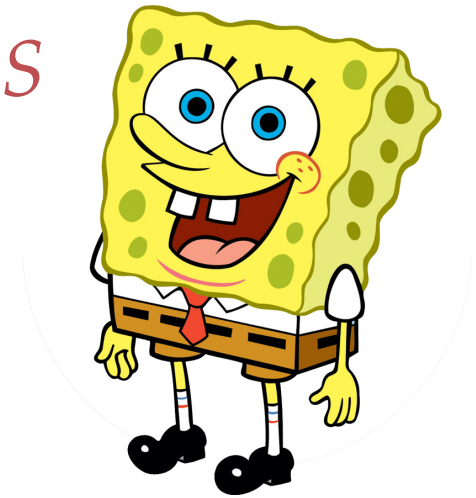
$\binom{[n]}{t-1} \ni S$



Containment Game

$T \in \binom{[n]}{t}$

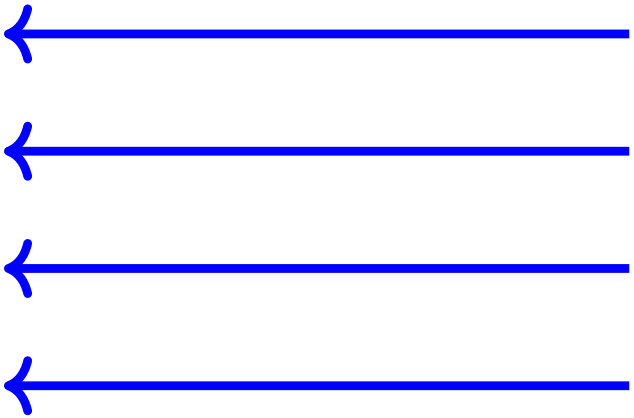
$\binom{[n]}{t-1} \ni S$



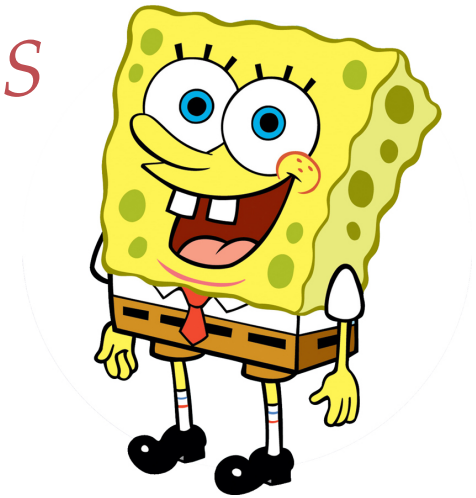
Containment Game

$T \in \binom{[n]}{t}$

$\binom{[n]}{t-1} \ni S$



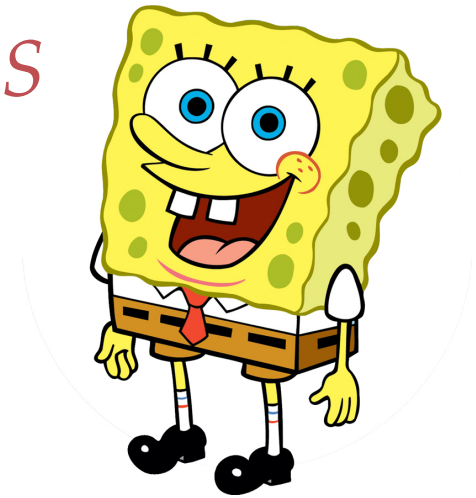
Public Randomness



Containment Game

$T \in \binom{[n]}{t}$

$\binom{[n]}{t-1} \ni S$



Public Randomness

GOAL

Determine if $S \subset T$

Containment Game: Protocols

- Deterministic Protocol:
  - Message length: $O(t \log n)$ bits
  - Completeness: 1, Soundness: 0

Containment Game: Protocols

- Deterministic Protocol:
  - Message length: $O(t \log n)$ bits
  - Completeness: 1, Soundness: 0
- Randomized Protocol:
  - Message length: $O_{\varepsilon,t}(1)$ bits

Containment Game: Protocols

- Deterministic Protocol:
  - Message length: $O(t \log n)$ bits
  - Completeness: 1, Soundness: 0
- Randomized Protocol:
  - Message length: $O_{\varepsilon,t}(1)$ bits
  - Completeness: 1, Soundness: ε

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $S_i := \{\mathscr{C}(u)_i \mid u \in S\}$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $S_i := \{\mathscr{C}(u)_i \mid u \in S\}$
- Alice checks if $S_i \subseteq T_i := \{\mathscr{C}(u)_i \mid u \in T\}$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $S_i := \{\mathscr{C}(u)_i \mid u \in S\}$
- Alice checks if $S_i \subseteq T_i := \{\mathscr{C}(u)_i \mid u \in T\}$
- Message length: $(t-1) \cdot \log_2 q$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $S_i := \{\mathscr{C}(u)_i \mid u \in S\}$
- Alice checks if $S_i \subseteq T_i := \{\mathscr{C}(u)_i \mid u \in T\}$
- Message length: $(t-1) \cdot \log_2 q$
- Soundness: $t \cdot (1 - \Delta(\mathscr{C}))$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $S_i := \{\mathscr{C}(u)_i \mid u \in S\}$
- Alice checks if $S_i \subseteq T_i := \{\mathscr{C}(u)_i \mid u \in T\}$
- Message length: $(t-1) \cdot \log_2 q$
- Soundness: $t \cdot (1 - \Delta(\mathscr{C})) \approx O_t(1/\sqrt{q})$ (for AG codes)

Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$

$$\vdots$$

Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

Embedding Transcript into Hamming metric

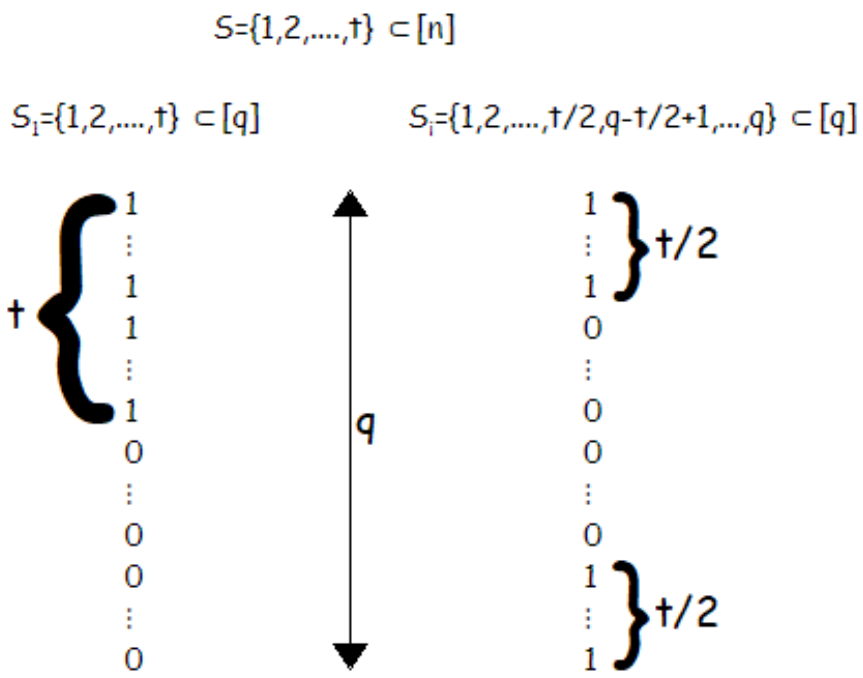
- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

$$\tau(S)_i = e_{S_i}, \quad \text{where } S_i = \{\mathscr{C}(u)_i \mid u \in S\} \subseteq [q]$$

Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

$$\tau(S)_i = e_{S_i}, \quad \text{where } S_i = \{\mathscr{C}(u)_i \mid u \in S\} \subseteq [q]$$



Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

$$\tau(S)_i = e_{S_i}, \quad \text{where } S_i = \{\mathscr{C}(u)_i \mid u \in S\} \subseteq [q]$$

- $X = \{\tau(T) \mid T \in E\}$

Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

$$\tau(S)_i = e_{S_i}, \quad \text{where } S_i = \{\mathscr{C}(u)_i \mid u \in S\} \subseteq [q]$$

- $X = \{\tau(T) \mid T \in E\}$
- $\mathcal{S} = \left\{\tau(S) \,\middle|\, S \in \binom{[n]}{t-1}\right\}$

Structural Observations

Suppose $S \subset T$

For every block i, we have $S_i \subset T_i$

$S_i = \{1,2,\ldots,t/2,q-t/2+1,\ldots,q\} \subset [q]$ $\qquad$ $T_i = S_i \cup \{t+1\} \subset [q]$

1	1
⋮	⋮
1	1
0	0
⋮	⋮
0	0
0	1
0	0
⋮	⋮
0	0
1	1
⋮	⋮
1	1

$$\boxed{|\tau(T\_i)-\tau(S\_i)|=1}$$

Structural Observations

Suppose $S \not\subset T$

For most blocks i, we have $S_i \not\subset T_i$

$S_i \setminus T_i = \{q\}$	$T_i \setminus S_i = \{t+1, t+2\}$
1	1
⋮	⋮
1	1
0	0
⋮	⋮
0	0
0	1
0	1
0	0
⋮	⋮
0	0
1	1
⋮	⋮
1	1
1	0

$$\boxed{|\tau(T\_i)-\tau(S\_i)| \geq 3}$$

Completeness of Reduction

- $\mathcal{S}' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ be a cover of $E \subseteq \binom{[n]}{t}$

Completeness of Reduction

- $\mathcal{S}' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ be a cover of $E \subseteq \binom{[n]}{t}$
- Build $\sigma : X \to C \subseteq \mathcal{S}$:

Completeness of Reduction

- $\mathcal{S}' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ be a cover of $E \subseteq \binom{[n]}{t}$
- Build $\sigma : X \to C \subseteq \mathcal{S}$:

$$\sigma(\tau(T)) = \tau(S_i), \text{ where } S_i \subset T$$

Completeness of Reduction

- $\mathcal{S}' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ be a cover of $E \subseteq \binom{[n]}{t}$
- Build $\sigma : X \to C \subseteq \mathcal{S}$:

$$\sigma(\tau(T)) = \tau(S_i), \quad \text{where } S_i \subset T$$

- Fix $T \in E$ and $i \in [c \cdot \log n]$

Distance between $\tau(T)$ and $\sigma(\tau(T))$ on block i is 1

Completeness of Reduction

- $\mathcal{S}' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ be a cover of $E \subseteq \binom{[n]}{t}$
- Build $\sigma : X \to C \subseteq \mathcal{S}$:

$$\sigma(\tau(T)) = \tau(S_i), \quad \text{where } S_i \subset T$$

- Fix $T \in E$ and $i \in [c \cdot \log n]$

Distance between $\tau(T)$ and $\sigma(\tau(T))$ on block i is 1

- k-means objective is:

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 = (c \cdot \log n)^2 \cdot |X|$$

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $\mathcal{S}' \subseteq \binom{[n]}{t-1}$ of size k:

$$S \in \mathcal{S}' \iff \tau(S) \in C$$

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $\mathcal{S}' \subseteq \binom{[n]}{t-1}$ of size k:

$$S \in \mathcal{S}' \iff \tau(S) \in C$$

- $\exists E' \subseteq E$, s.t. $\forall T \in E'$, T does not contain any subset in $\mathcal{S}'$

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $\mathcal{S}' \subseteq \binom{[n]}{t-1}$ of size k:

$$S \in \mathcal{S}' \iff \tau(S) \in C$$

- $\exists E' \subseteq E$, s.t. $\forall T \in E'$, T does not contain any subset in $\mathcal{S}'$
- Fix $\tau(T) \in X_{E'}$ and $i \in [c \cdot \log n]$

Distance between $\tau(T)$ and $\sigma(\tau(T))$ on block i is mostly 3

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $\mathcal{S}' \subseteq \binom{[n]}{t-1}$ of size k:

$$S \in \mathcal{S}' \iff \tau(S) \in C$$

- $\exists E' \subseteq E$, s.t. $\forall T \in E'$, T does not contain any subset in $\mathcal{S}'$
- Fix $\tau(T) \in X_{E'}$ and $i \in [c \cdot \log n]$

Distance between $\tau(T)$ and $\sigma(\tau(T))$ on block i is mostly 3

- k-means objective is:

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 = (c \cdot \log n)^2 \cdot |X \setminus X_{E'}| + 9 \cdot (c \cdot \log n)^2 \cdot |X_{E'}|$$

Our Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94^*	1.73^*

Continuous Version

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Johnson graph Embedding

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94^*	1.73^*

Continuous Version

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

Johnson graph Embedding

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94^*	1.73^*

Continuous Version

Use Feige's Instance

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

Johnson graph Embedding

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94^*	1.73^*

Continuous Version

Use Feige's Instance

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

Decoding Vertex Cover

Key Takeaways

- Improved Inapproximability of

Key Takeaways

- Improved Inapproximability of
- k-means and k-median

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics
- Using Transcript of Containment Protocol

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics
- Using Transcript of Containment Protocol
- And Geometric Realization of Johnson Graphs

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics
- Using Transcript of Containment Protocol
- And Geometric Realization of Johnson Graphs

Open: Is JCH true?

Bonus: Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

Bonus: Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \geq 4 \cdot n'$.

Bonus: Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \geq 4 \cdot n'$.

- k-median: 2 inapproximability

Bonus: Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \geq 4 \cdot n'$.

- k-median: 2 inapproximability

Continuous is harder than Discrete!

Bonus: Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \geq 4 \cdot n'$.

- k-median: 2 inapproximability

Continuous is harder than Discrete!

- Constant Bicriteria inapproximability

Bonus: Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \geq 4 \cdot n'$.

- k-median: 2 inapproximability

Continuous is harder than Discrete!

- Constant Bicriteria inapproximability
- Assuming UGC, hardness for $k = 2$!

Bonus: Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \geq 4 \cdot n'$.

- k-median: 2 inapproximability

Continuous is harder than Discrete!

- Constant Bicriteria inapproximability
- Assuming UGC, hardness for $k = 2$!
- Dependency on d, k, and ℓ_∞ tight

Bonus: Minsum (Definition)

- (Γ, Δ) is a metric space
- <u>Input</u>: $X \subseteq \Gamma$, $k \in \mathbb{N}$

Bonus: Minsum (Definition)

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ that minimizes:

Bonus: Minsum (Definition)

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ that minimizes:

$$\boxed{\sum_{i \in [k]} \sum_{x, y \in X_i} \Delta(x, y)}$$

Bonus: Minsum (Definition)

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ that minimizes:

$$\boxed{\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y)}$$

- Approximation: $O(\log n)$ [Behsaz et al.'15]

Bonus: Minsum (Definition)

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ that minimizes:

$$\boxed{\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y)}$$

- Approximation: $O(\log n)$ [Behsaz et al.'15]
- Hardness: $1 + \varepsilon$ [Guruswami-Indyk'03]

Bonus: Minsum (Result)

Theorem (Cohen-Addad–K–Lee)

Given input (X, k), it is NP-hard to distinguish:

Bonus: Minsum (Result)

Theorem (Cohen-Addad–K–Lee)

Given input (X, k), it is NP-hard to distinguish:

YES: There exists a partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ such that

$$\sum_{i \in [k]} \sum_{x, y \in X_i} \Delta(x, y) \leq n',$$

Bonus: Minsum (Result)

Theorem (Cohen-Addad–K–Lee)

Given input (X, k), it is NP-hard to distinguish:

YES: There exists a partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ such that

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y) \leq n',$$

NO: For every partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ we have

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y) \geq 1.41 \cdot n'.$$

Bonus: Minsum (Result)

Theorem (Cohen-Addad–K–Lee)

Given input (X, k), it is NP-hard to distinguish:

YES: There exists a partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ such that

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y) \leq n',$$

NO: For every partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ we have

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y) \geq 1.41 \cdot n'.$$

Key Ingredient: Hard Instances of Max-Coverage with large girth

Bonus: State-of-the-art Table

Metric	Discrete k-means	Discrete k-median	Continuous k-means	Continuous k-median	k-Minsum
General	3.94 [GK99]	1.73 [GK99]	4	2	1.415
ℓ_0	3.94*	1.73*	2.10*	1.36*	APX-hard [GI03]
ℓ_1	3.94*	1.73*	1.46*	1.36*	APX-hard [GI03]
ℓ_2	1.73*	1.27*	1.36*	1.04*	APX-hard [GI03]
ℓ_∞	3.94	1.73	4	2	1.415

*Under JCH

THANK YOU!