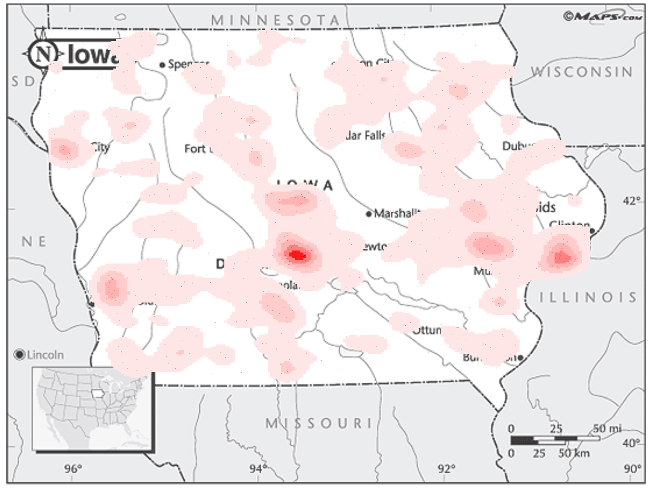
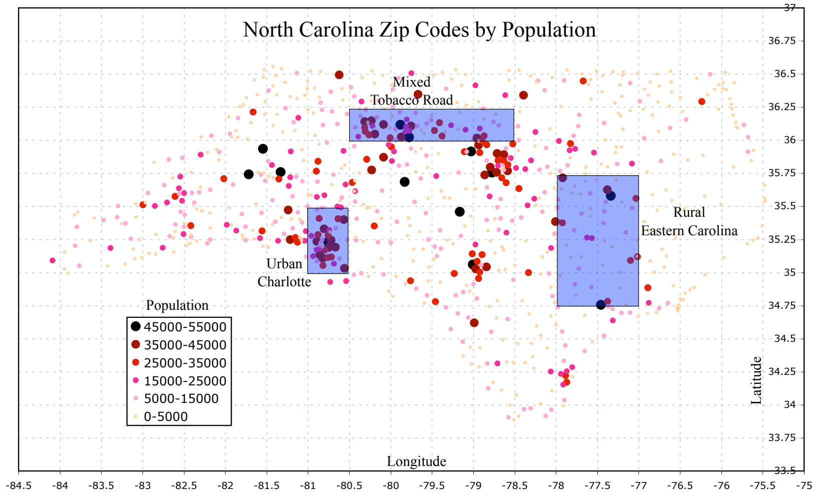
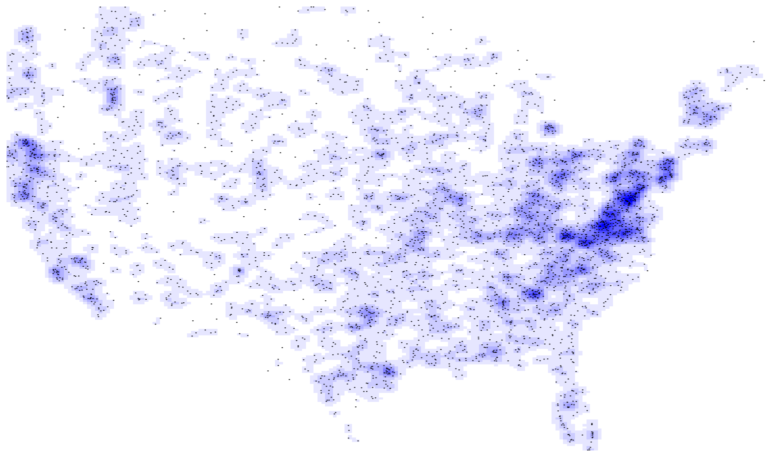
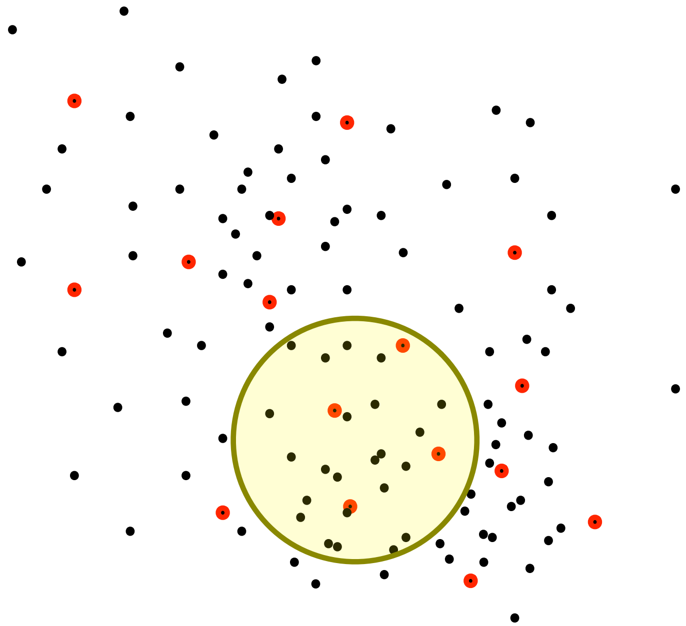


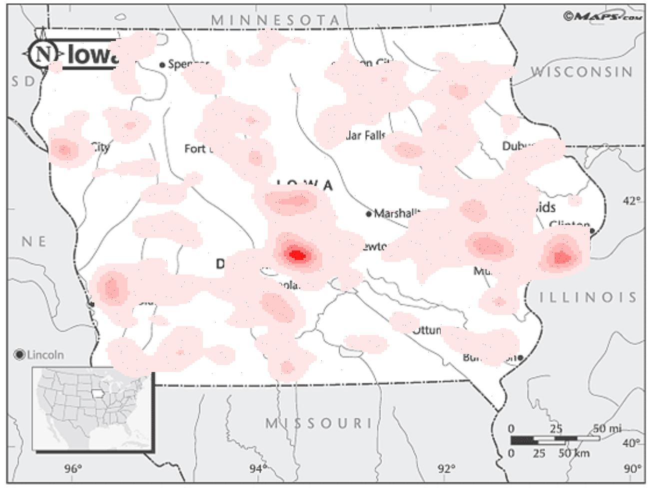


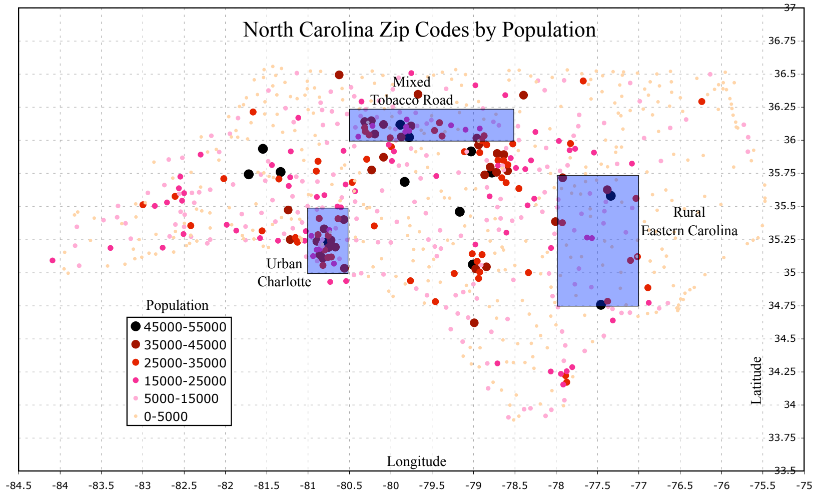
Scalable Spatial Scan Statistics with Coresets



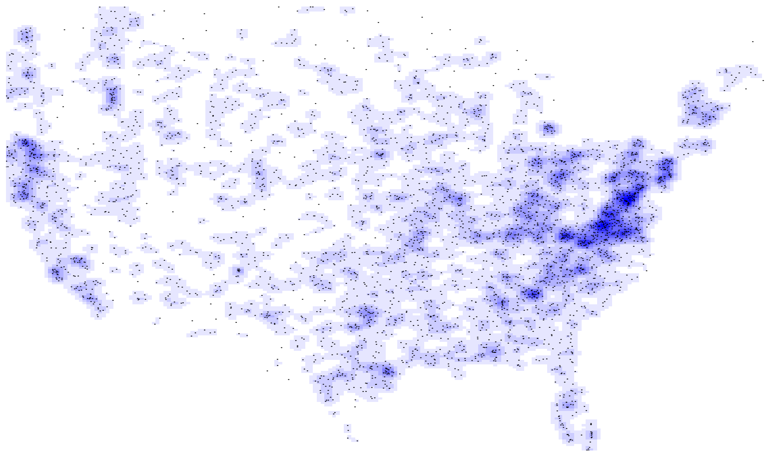
Jeff M. Phillips
School of Computing
University of Utah





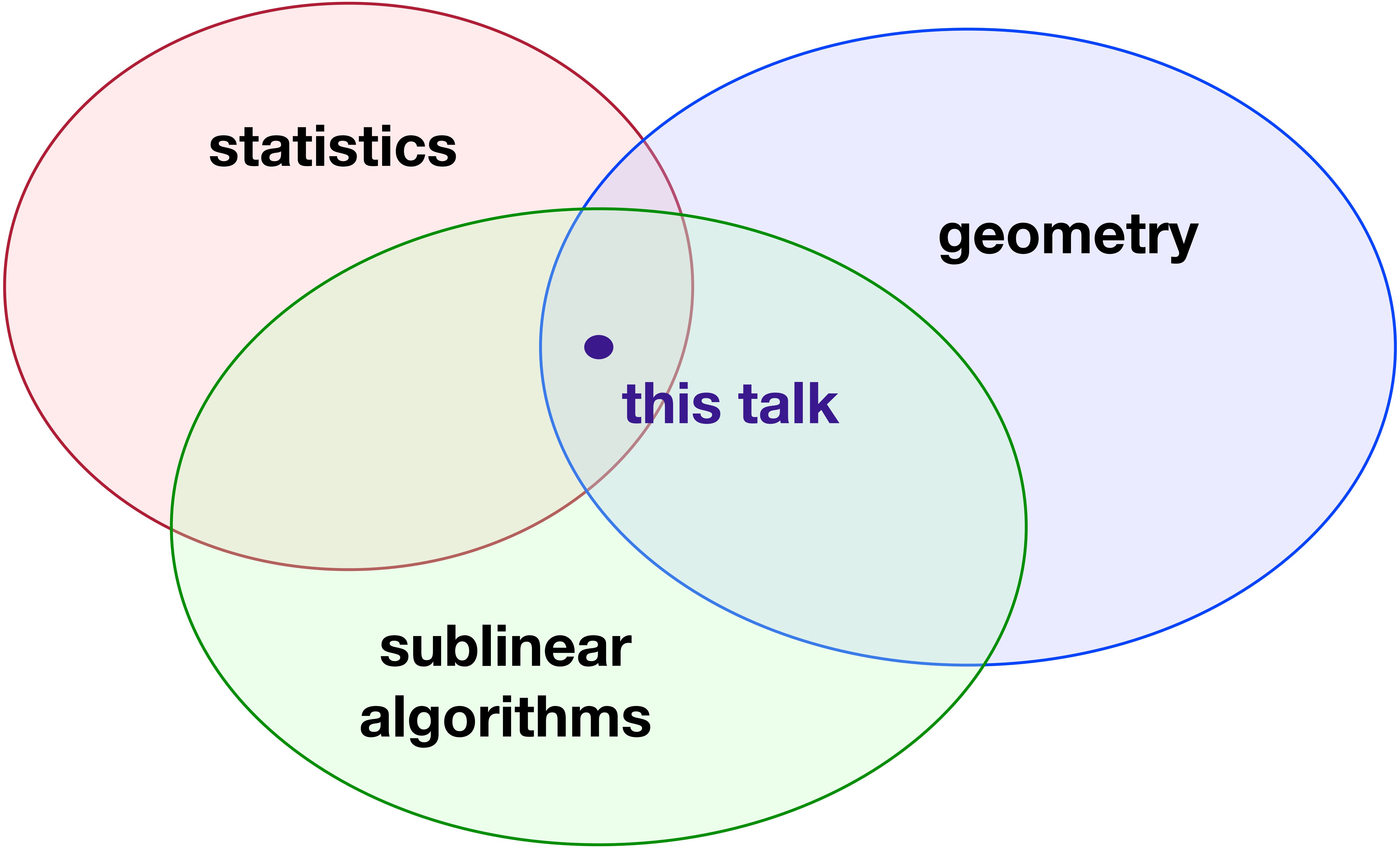


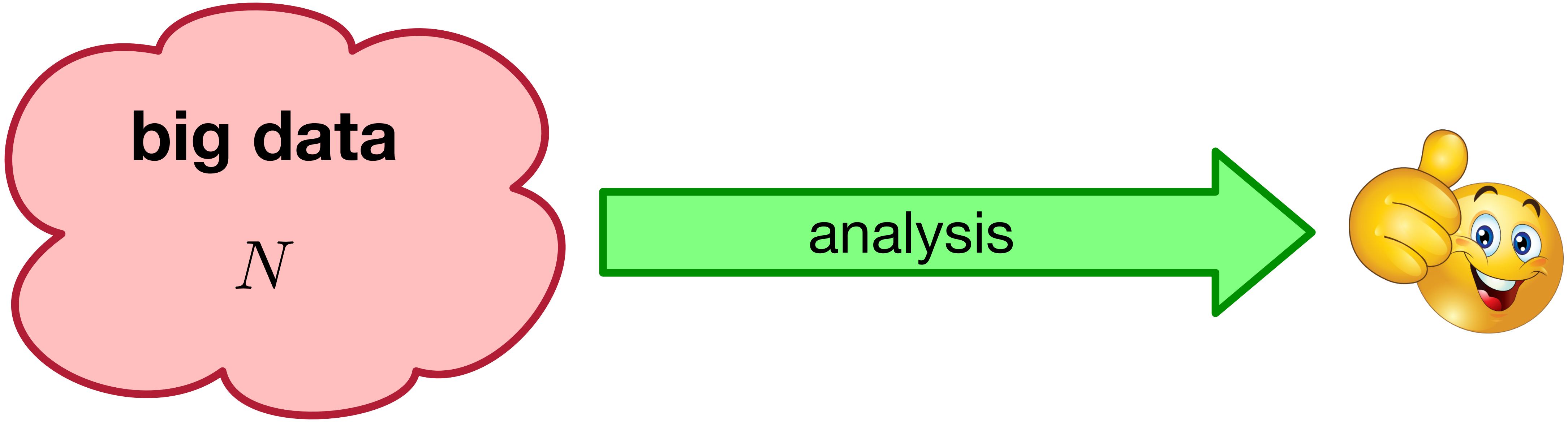
Scalable Spatial Scan Statistics with Coresets

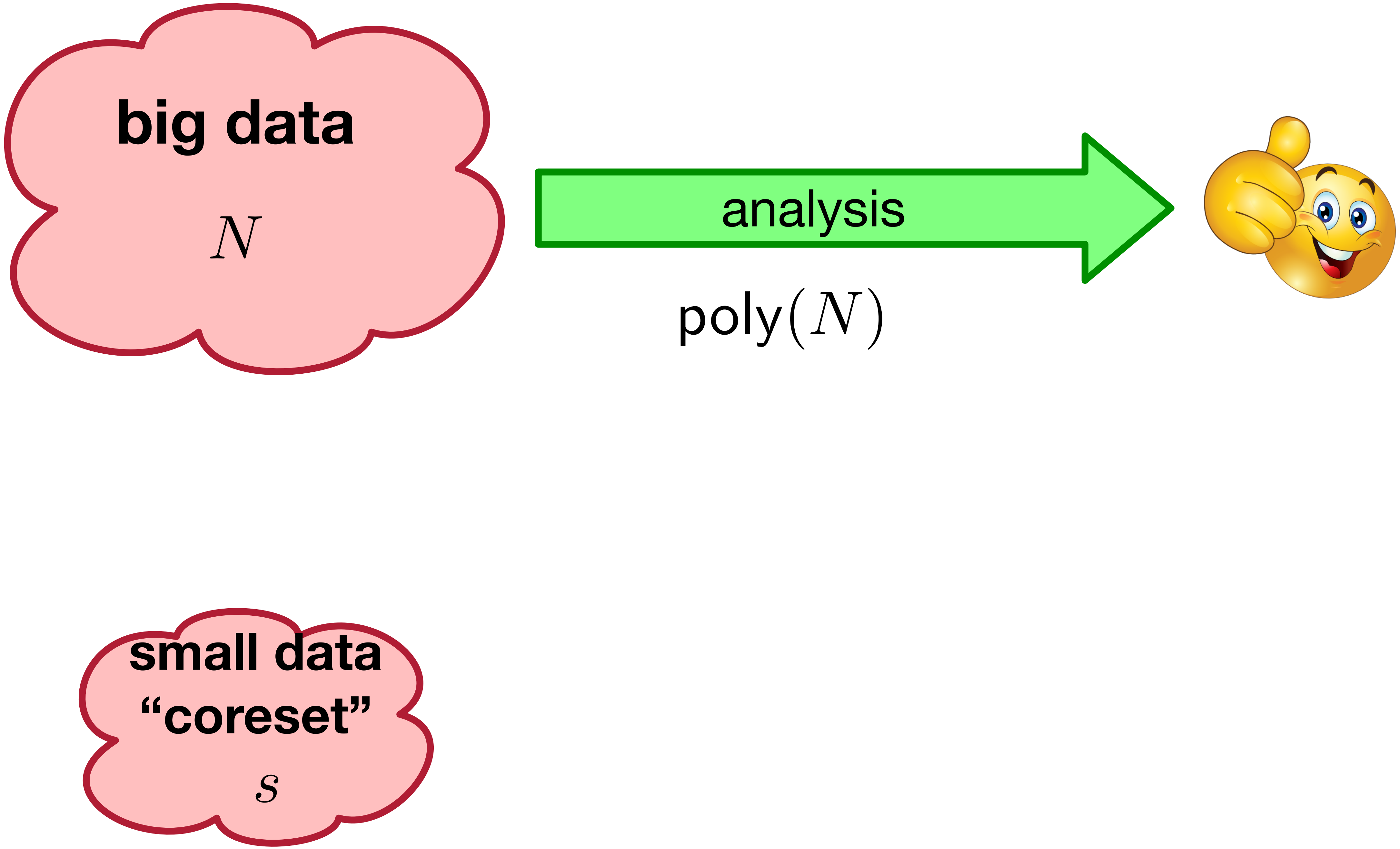


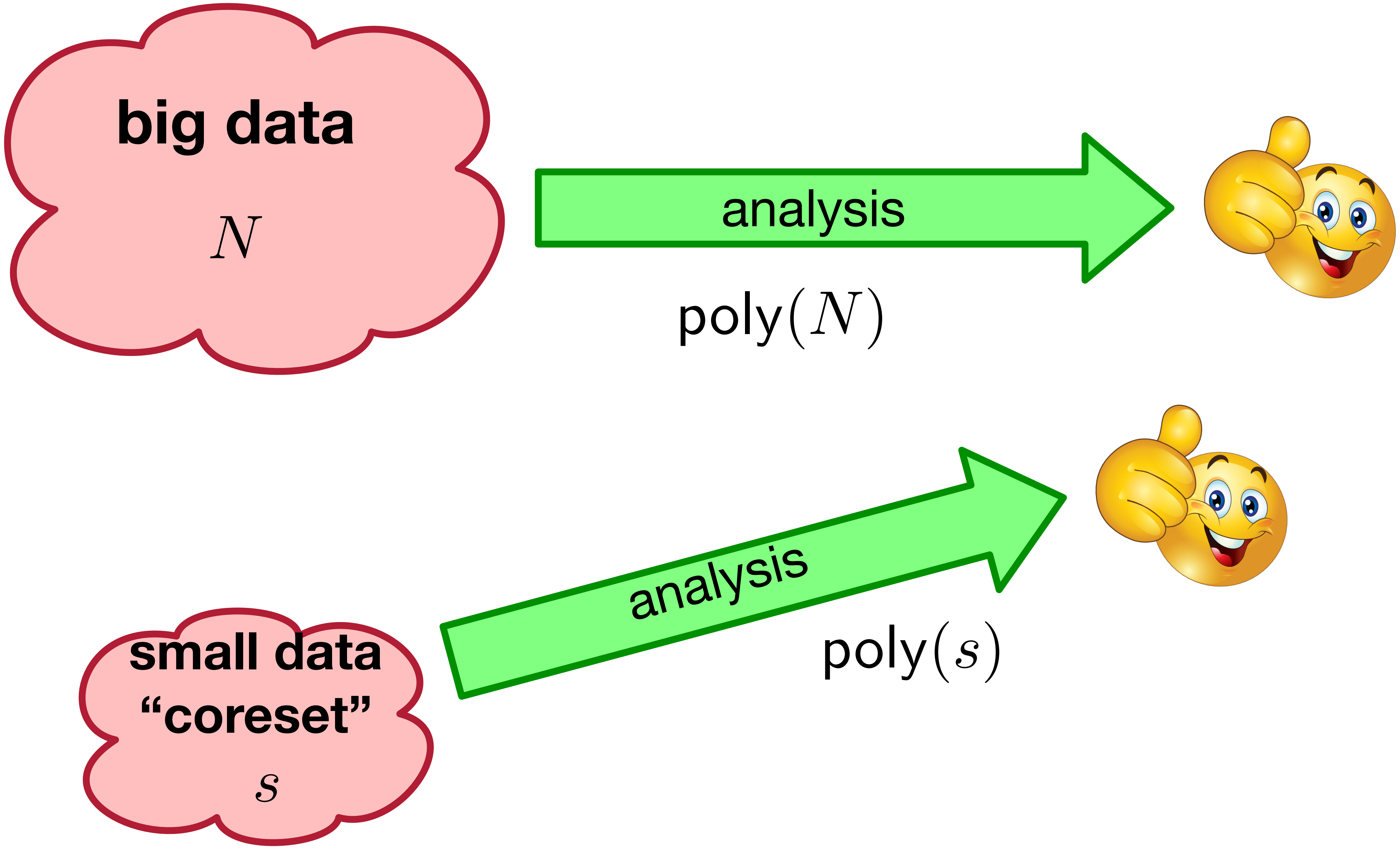
Jeff M. Phillips
School of Computing
University of Utah

Michael Matheny

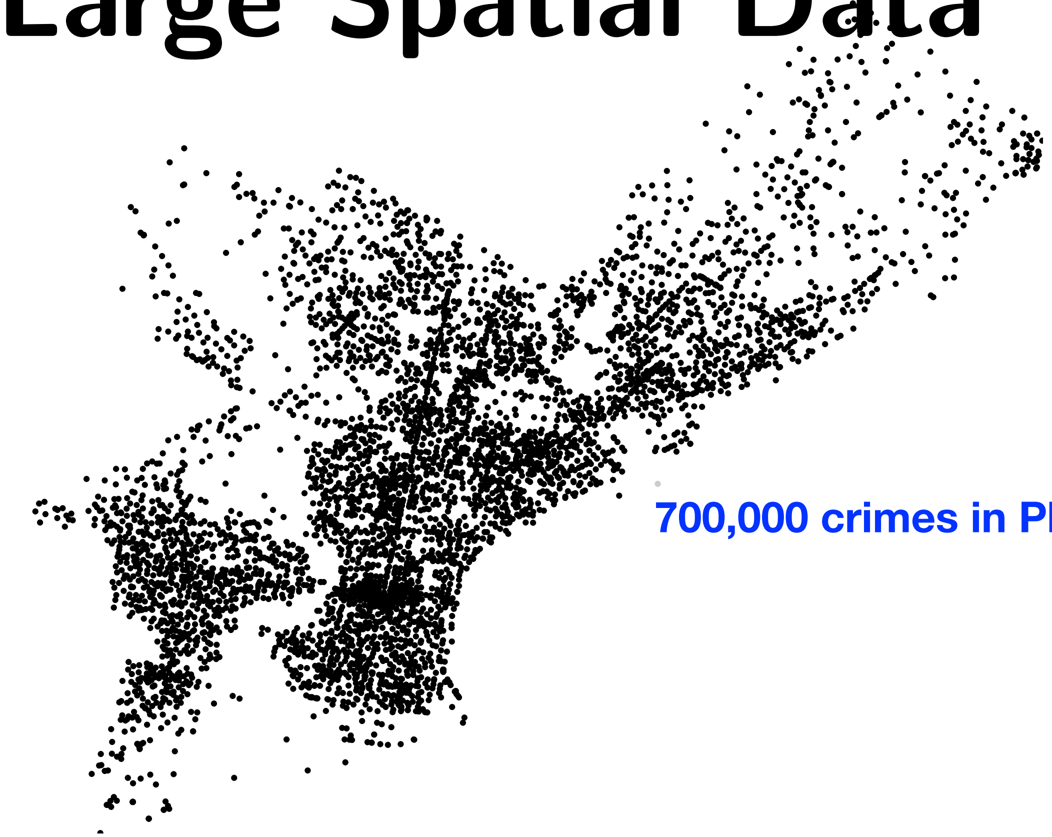

statistics
geometry
this talk
sublinear
algorithms


big data
N
analysis


big data
N
analysis
poly(N)
small data
“coreset”
s

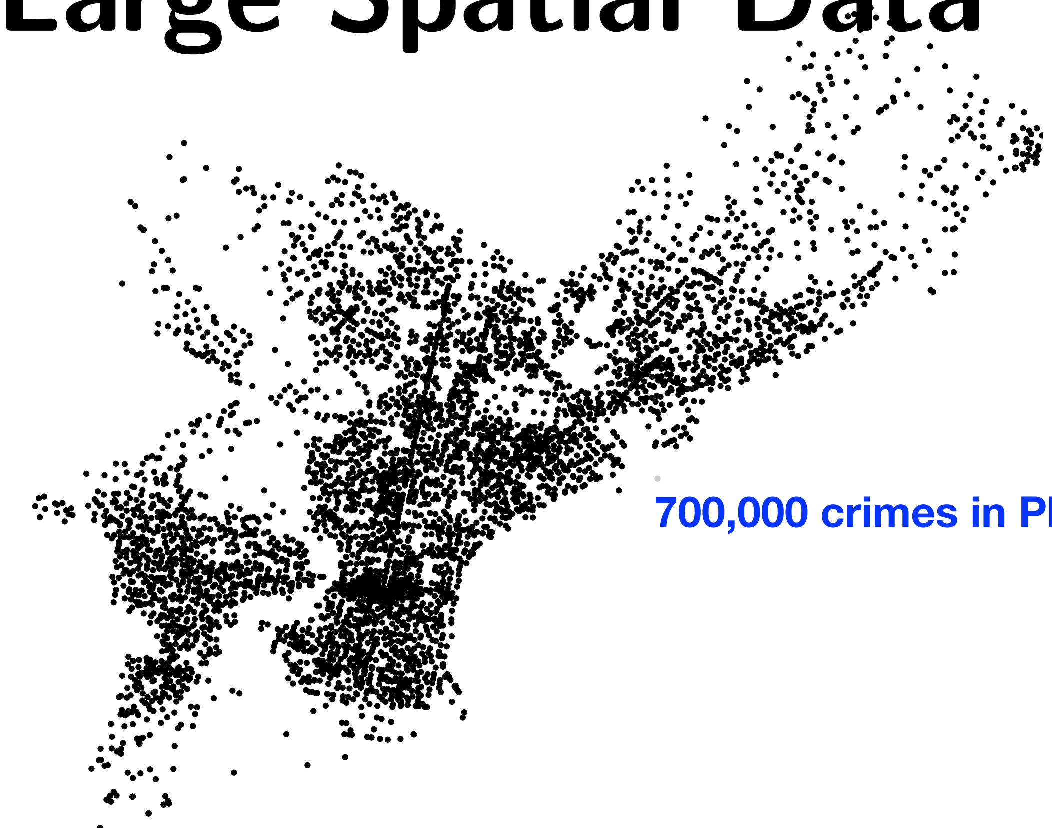

big data
N
analysis
poly(N)
small data
“coreset”
s
analysis
poly(s)

Large Spatial Data



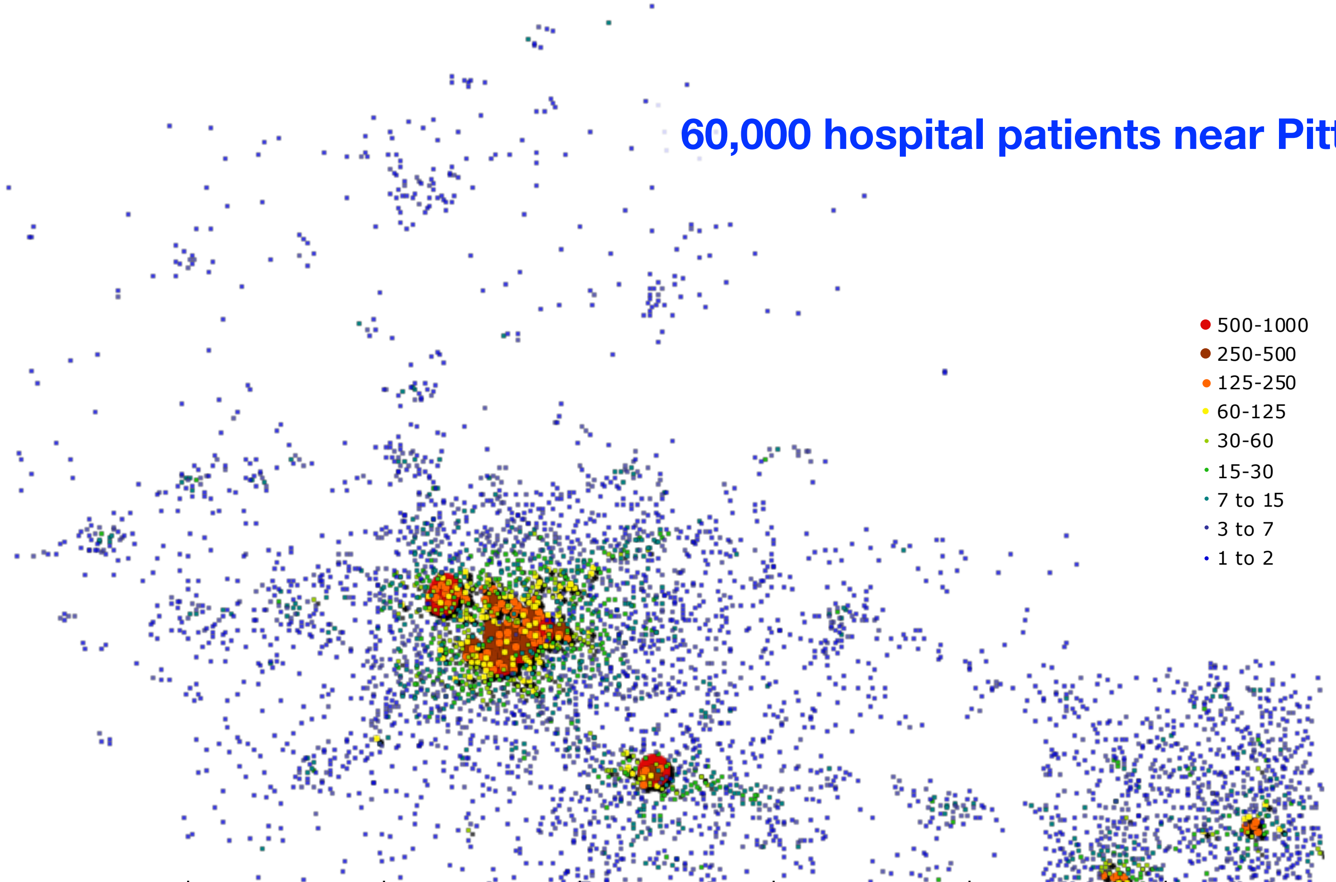
700,000 crimes in Philadelphia

Large Spatial Data

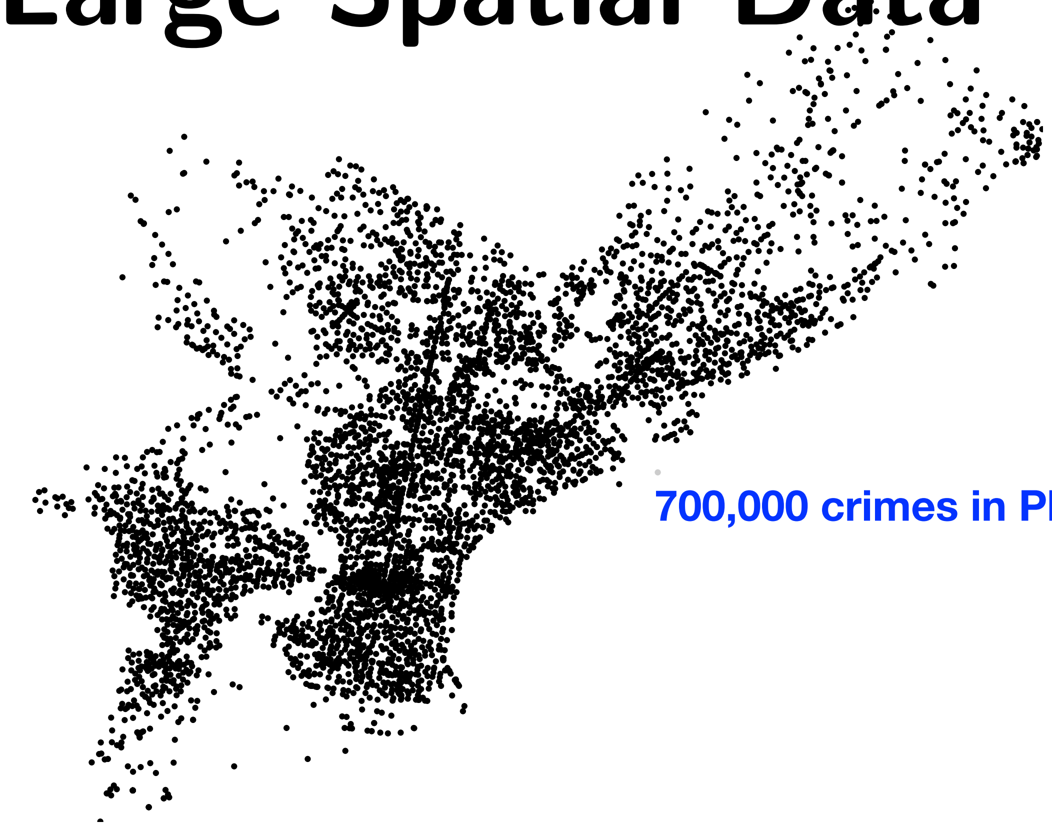


700,000 crimes in Philadelphia

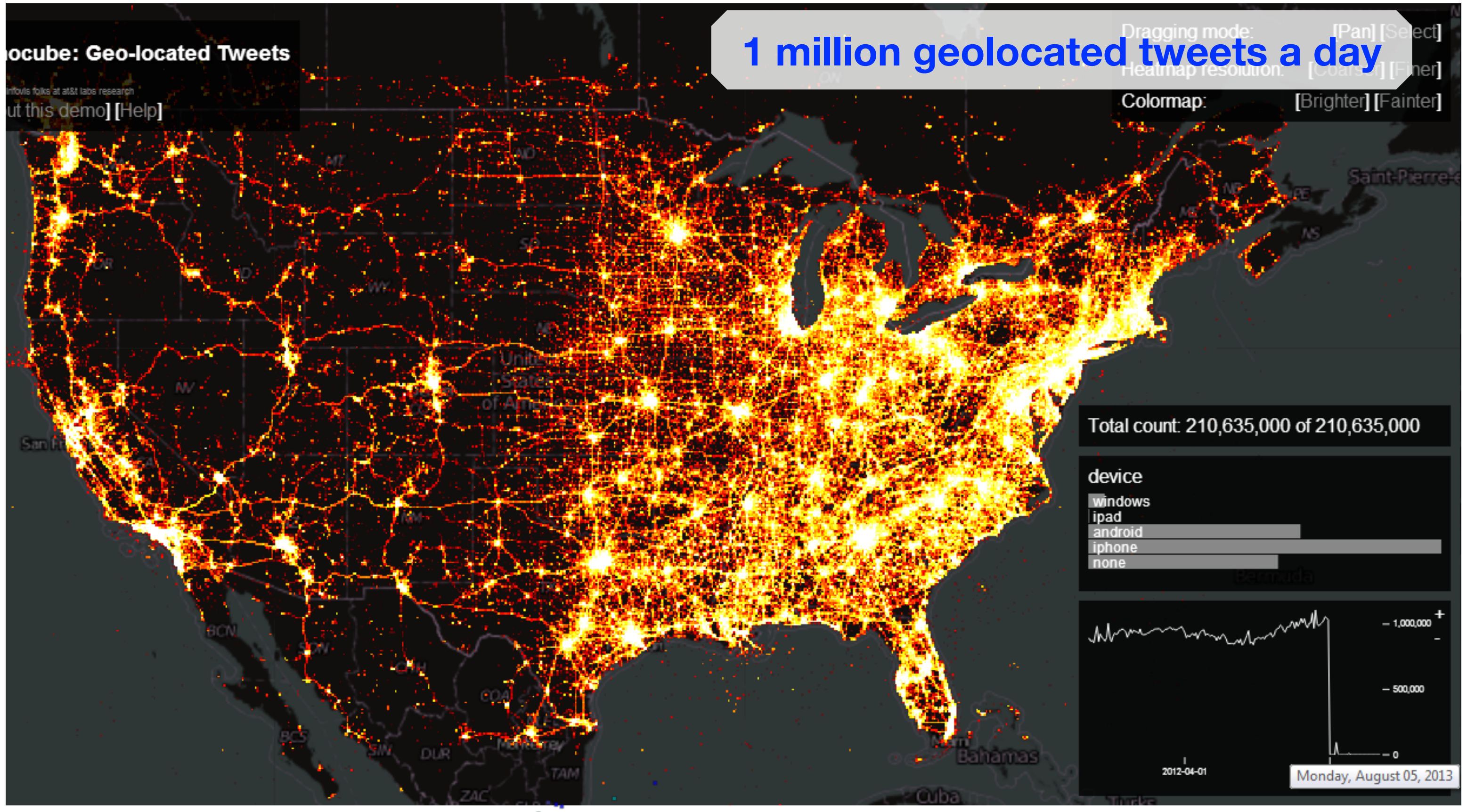
60,000 hospital patients near Pittsburgh



Large Spatial Data

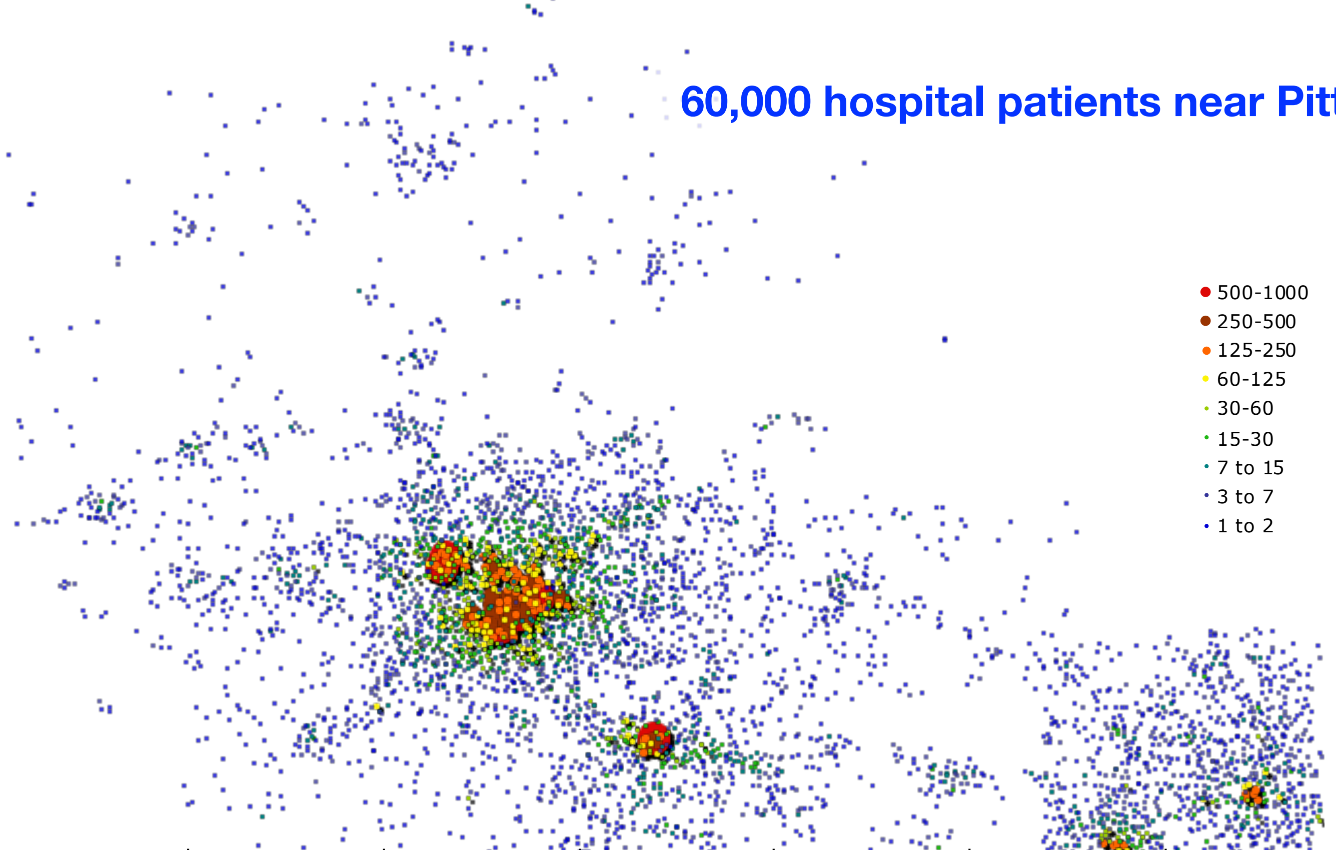


700,000 crimes in Philadelphia

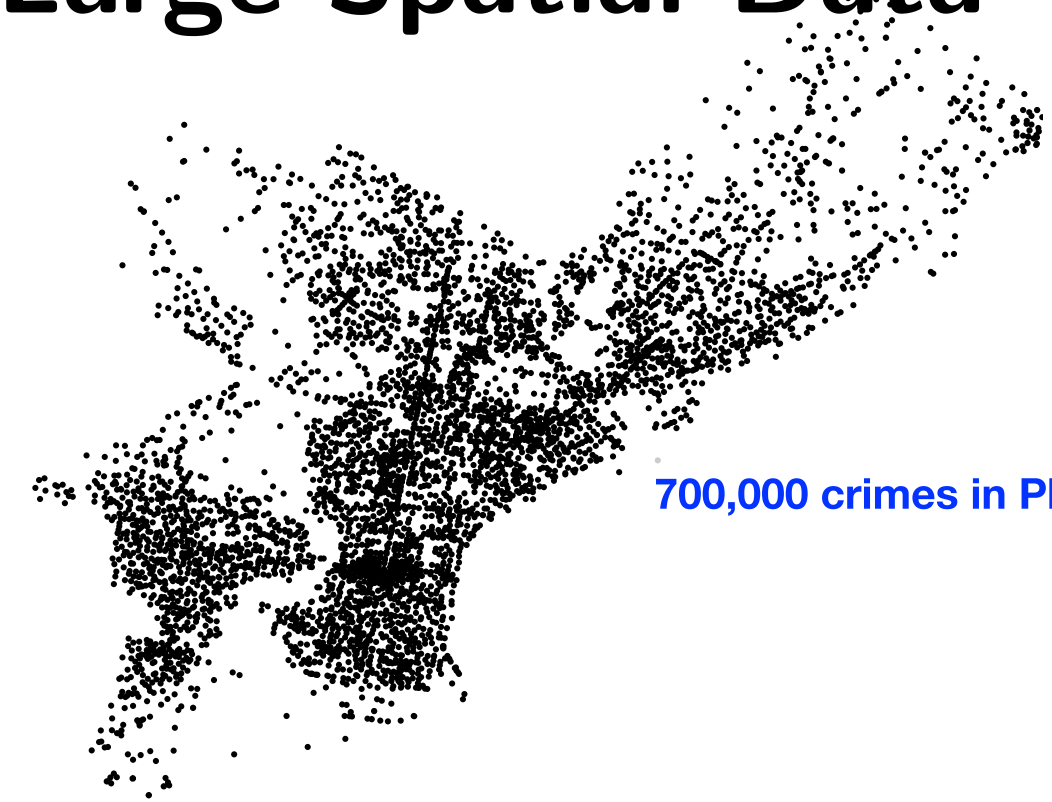


1 million geolocated tweets a day

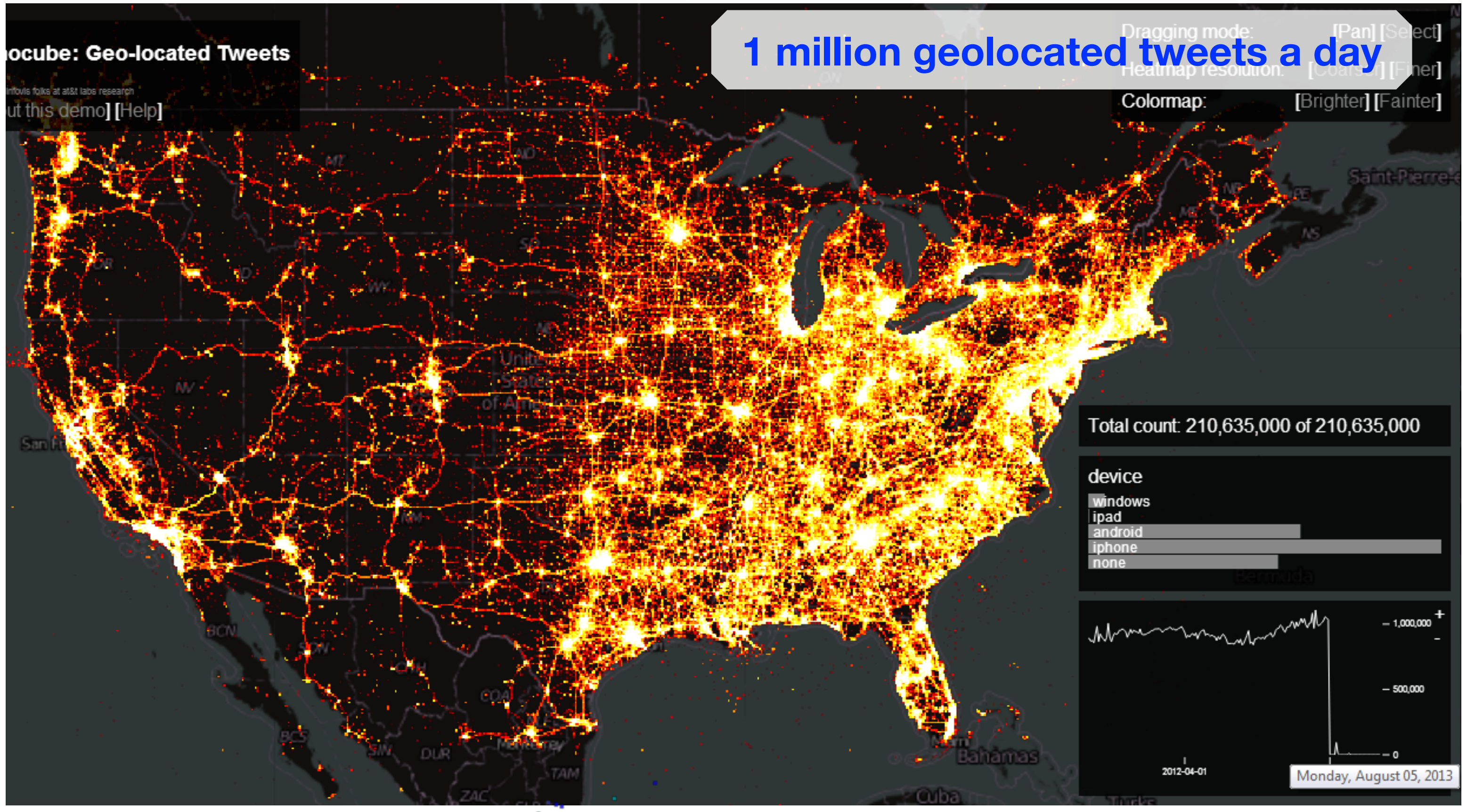
60,000 hospital patients near Pittsburgh



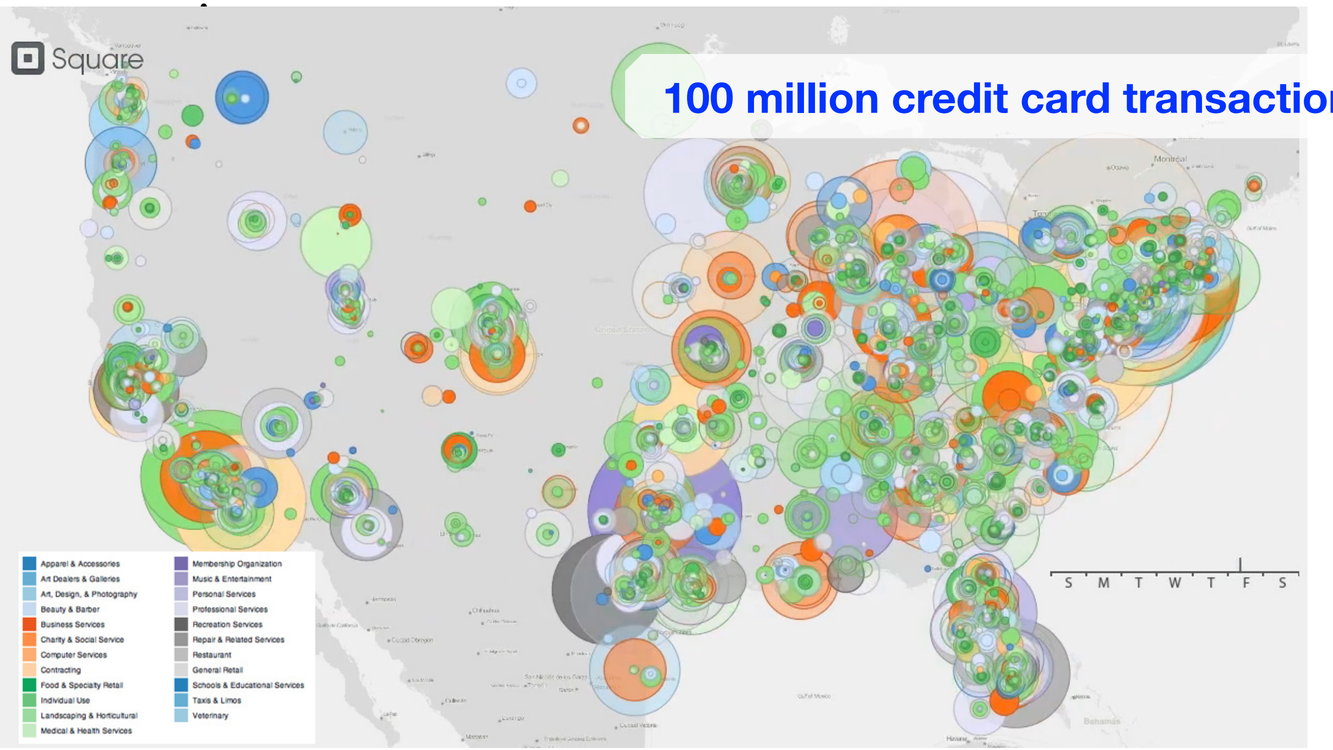
Large Spatial Data



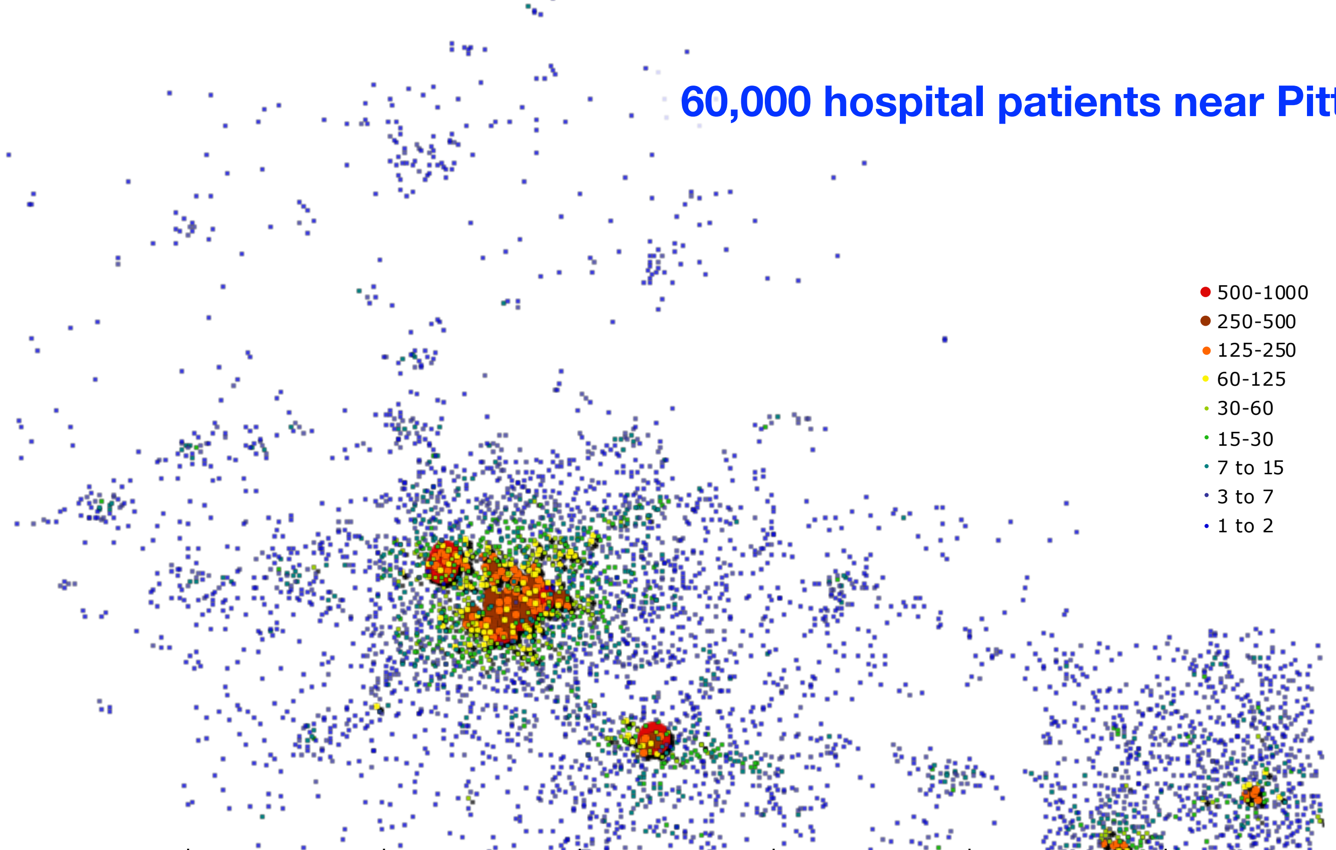
700,000 crimes in Philadelphia



1 million geolocated tweets a day

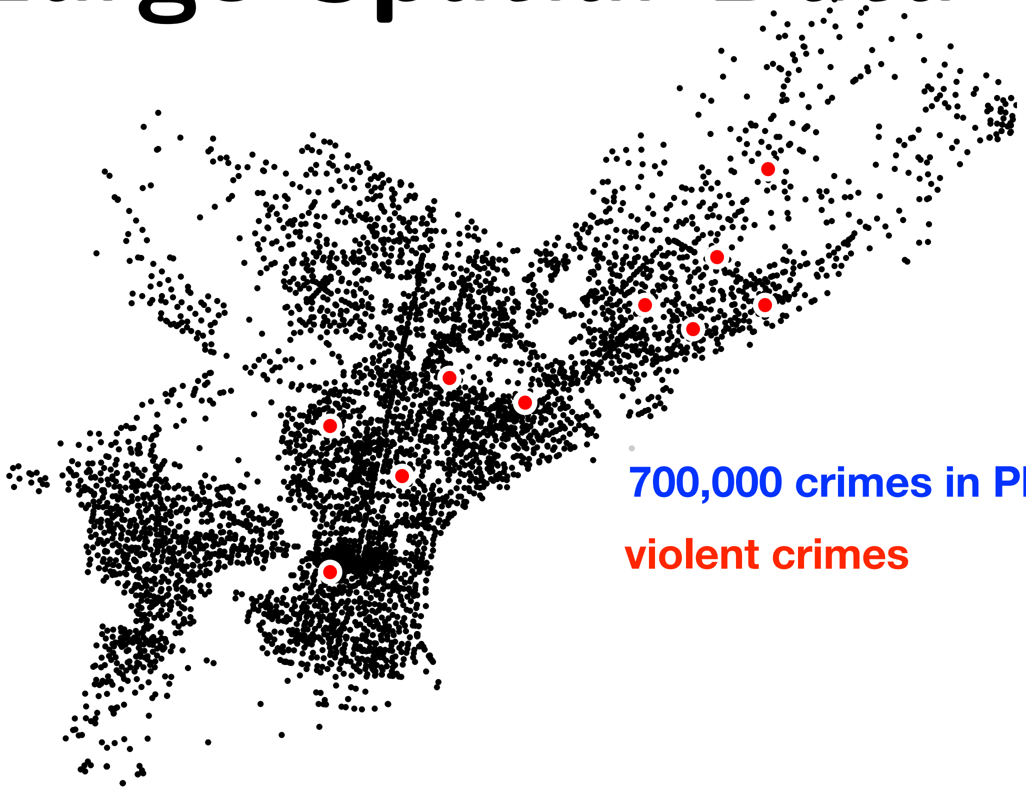


100 million credit card transactions a day



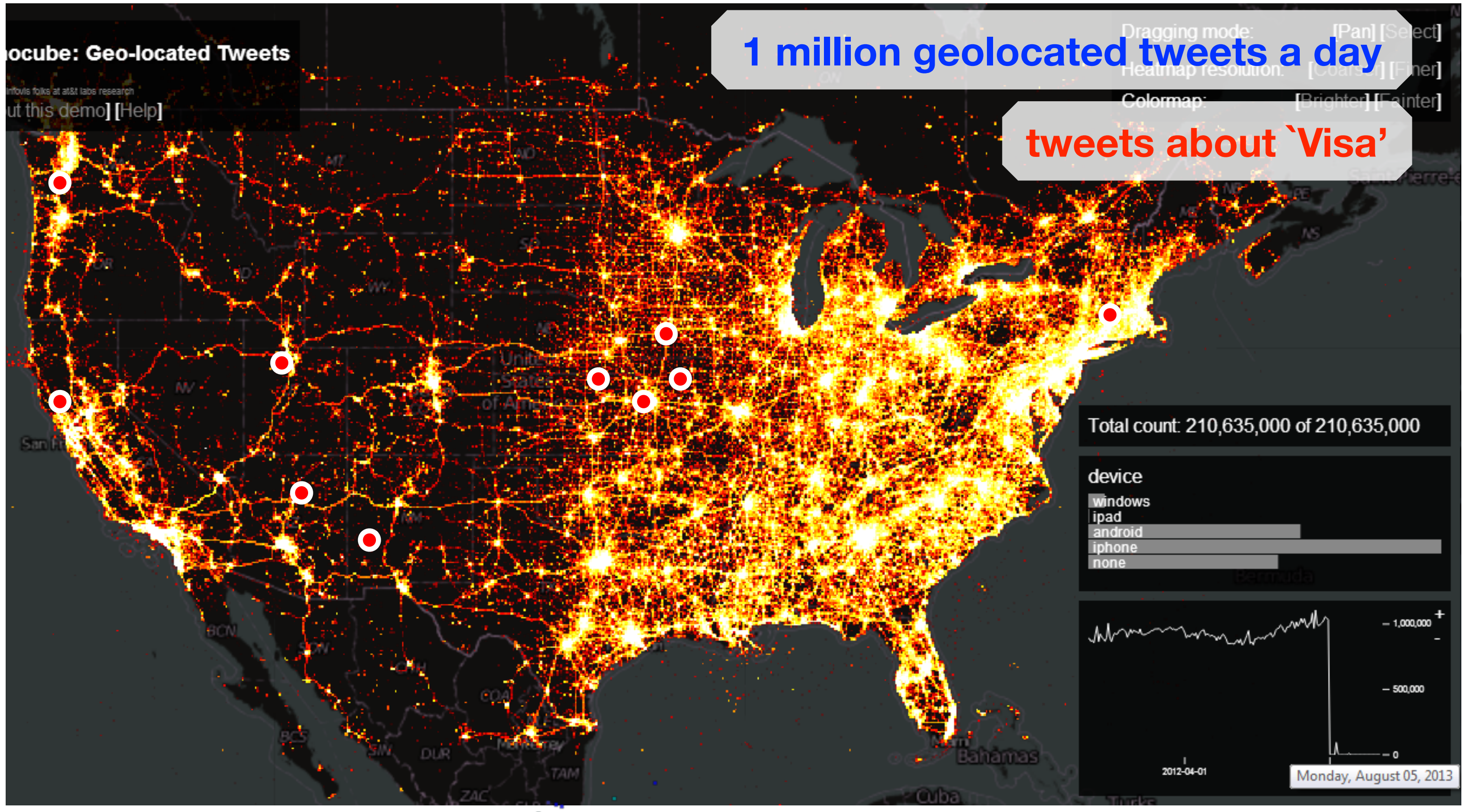
60,000 hospital patients near Pittsburgh

Large Spatial Data



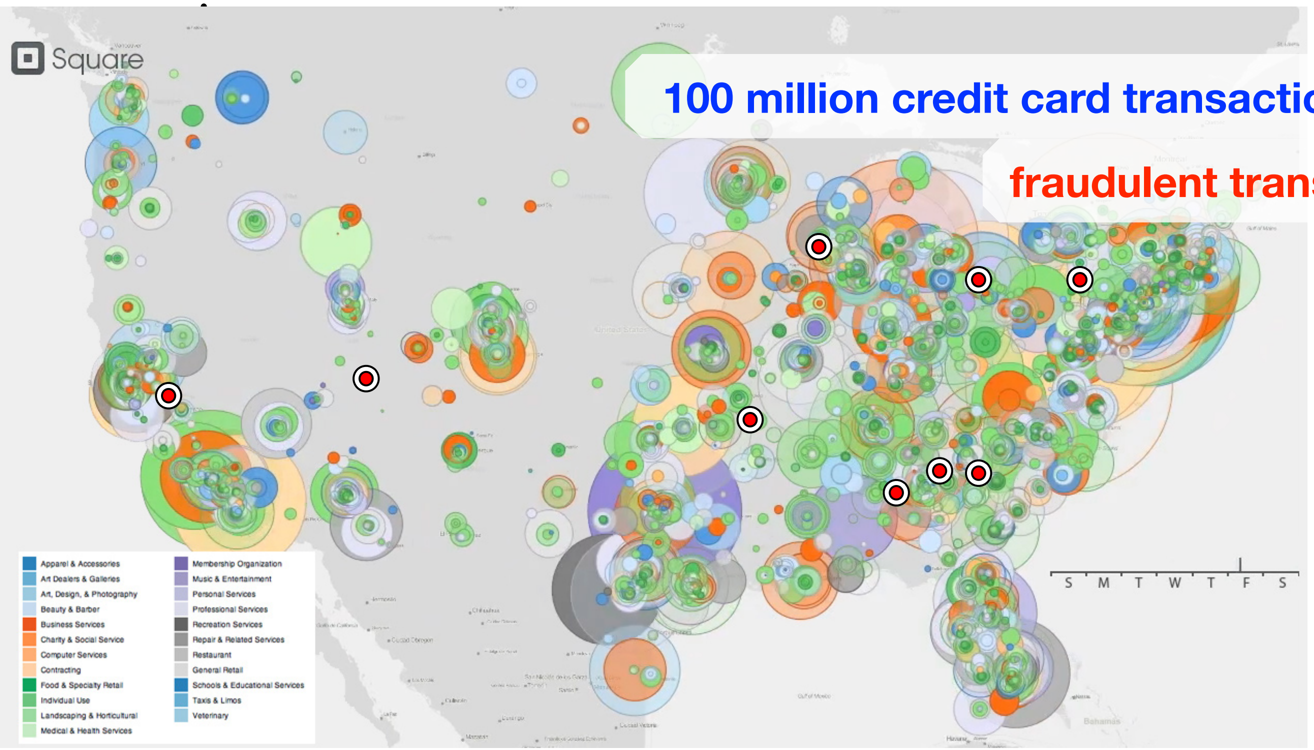
700,000 crimes in Philadelphia

violent crimes



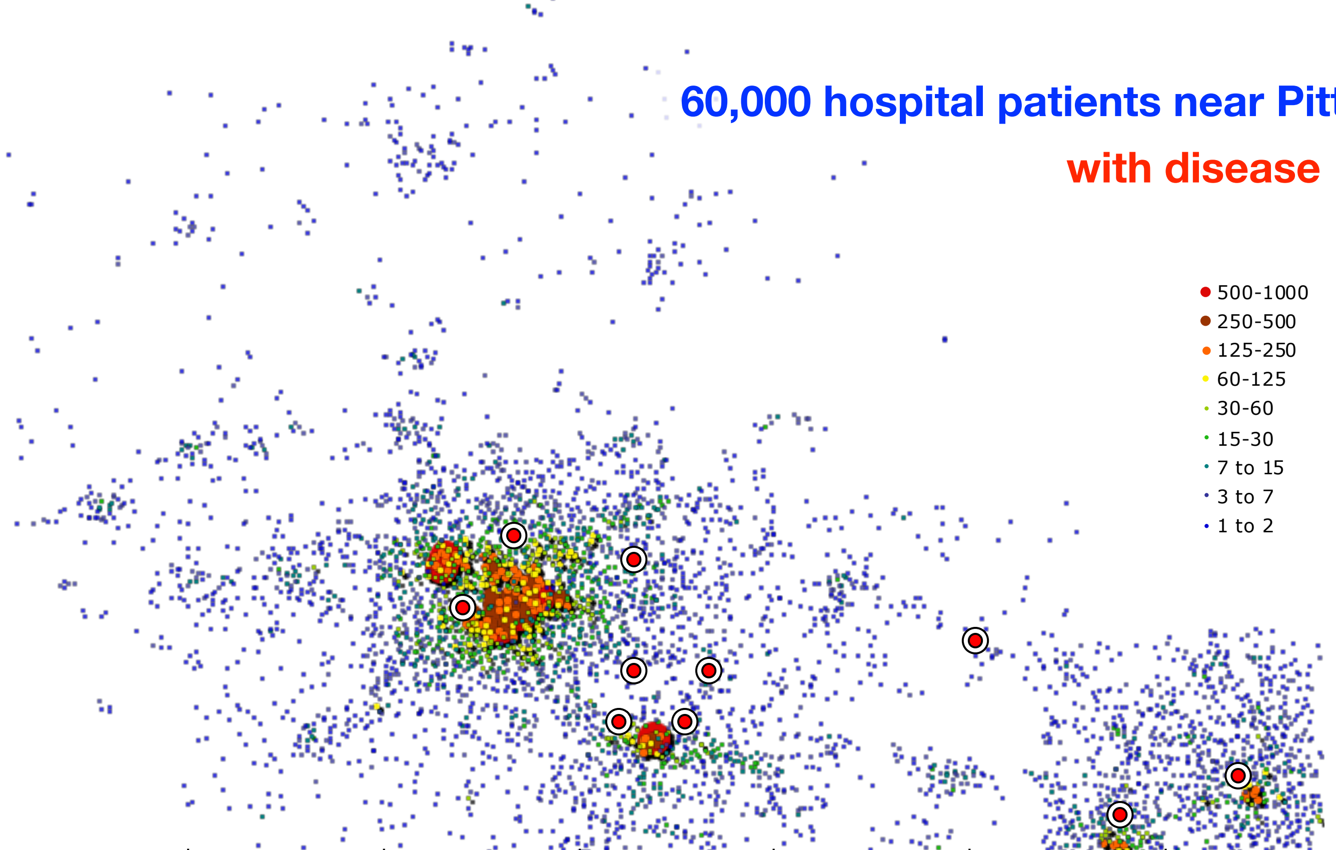
1 million geolocated tweets a day

tweets about `Visa'



100 million credit card transactions a day

fraudulent transactions



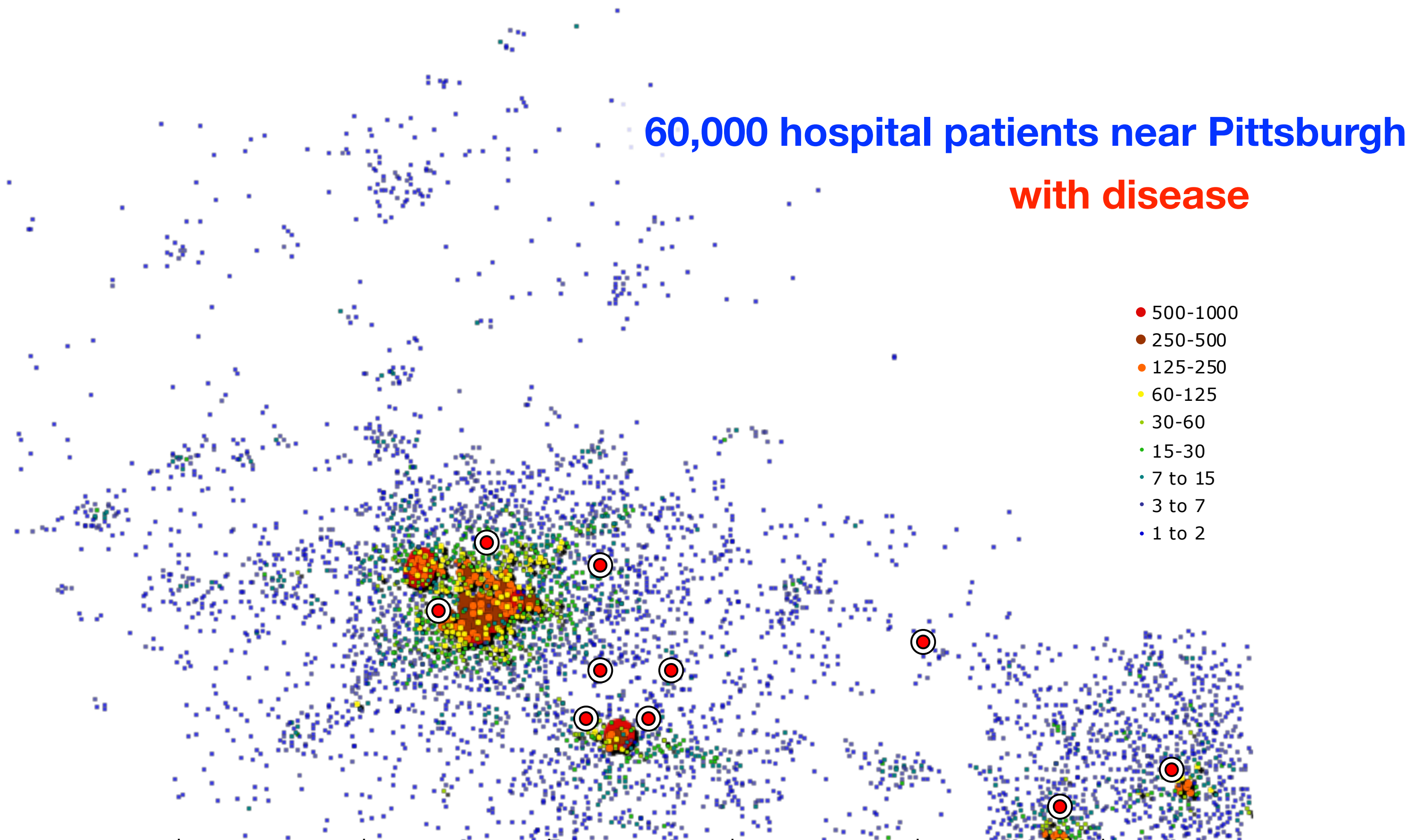
60,000 hospital patients near Pittsburgh

with disease

Large Spatial Data

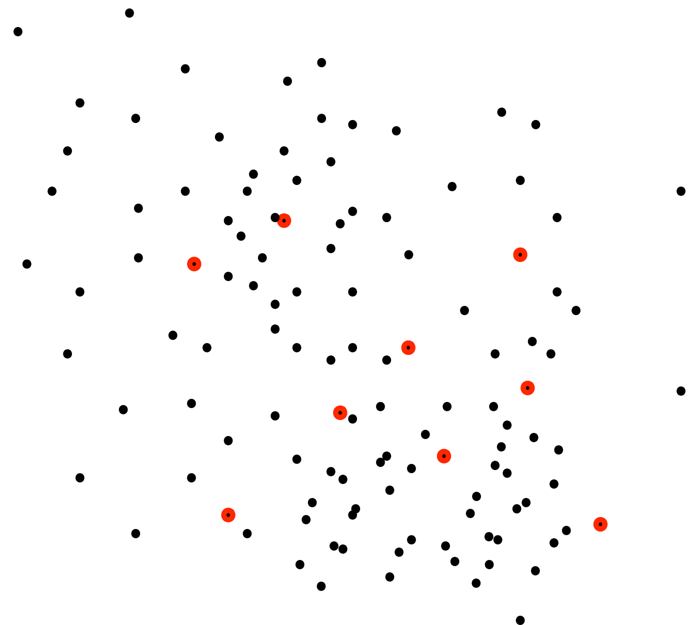
What is common?

- Data Sets have gotten **really** large!
- Spatial position is not uniform, its clustered
- Grouping of measured data describes anomalies



Spatial Scan Statistics

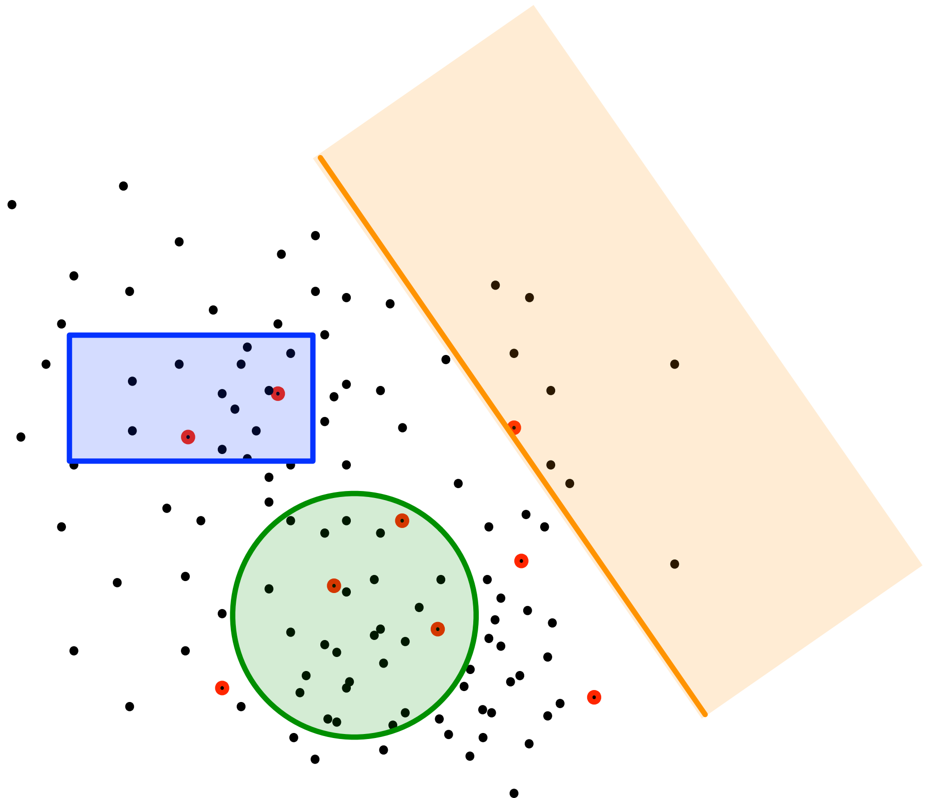
Find a region where **measured data** is significantly denser than **background data**.
(Martin Kulldorff 1997)



Spatial Scan Statistics

Find a region where **measured data** is significantly denser than **background data**. (Martin Kulldorff 1997)

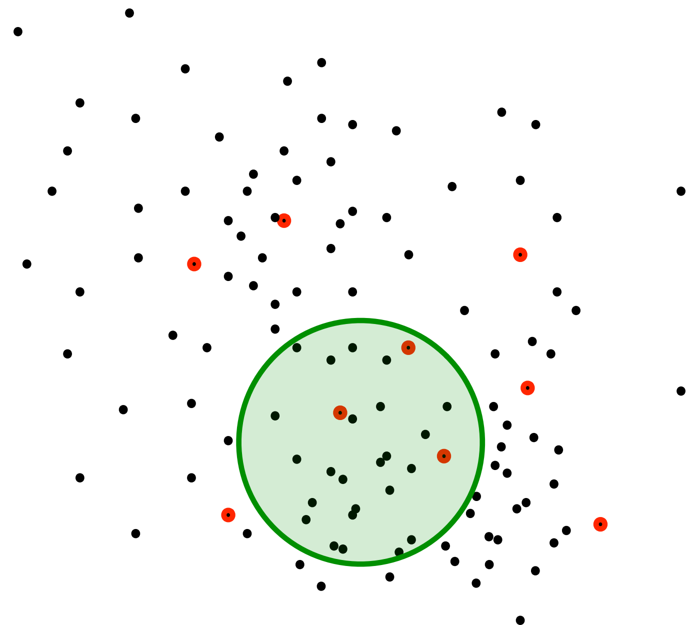
- Family of regions $\mathcal{C}$ e.g., disks, axis-aligned rectangles, halfspaces



Spatial Scan Statistics

Find a region where **measured data** is significantly denser than **background data**. (Martin Kulldorff 1997)

- Family of regions $\mathcal{C}$ e.g., disks, axis-aligned rectangles, halfspaces
- Data set $X \subset \mathbb{R}^2$, measured set $R \subset X$.



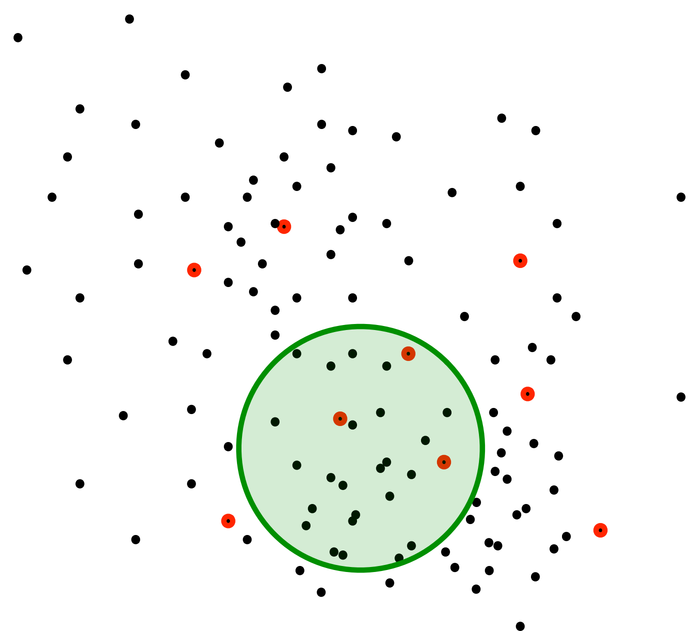
Spatial Scan Statistics

Find a region where **measured data** is significantly denser than **background data**. (Martin Kulldorff 1997)

- Family of regions $\mathcal{C}$ e.g., disks, axis-aligned rectangles, halfspaces
- Data set $X \subset \mathbb{R}^2$, measured set $R \subset X$.
- Define statistic $\Phi(C, X, R) = \Phi(C)$: log-likelihood ratio $\Phi(C) = \log(\frac{\Pr(\mathcal{H}_0 | C, X, R)}{\Pr(\mathcal{H}_1 | C, X, R)})$

  $\mathcal{H}_0$: no anomaly, rate of measured points same inside than outside

  $\mathcal{H}_1$: cluster C has *higher* rate of measured points than outside C



Spatial Scan Statistics

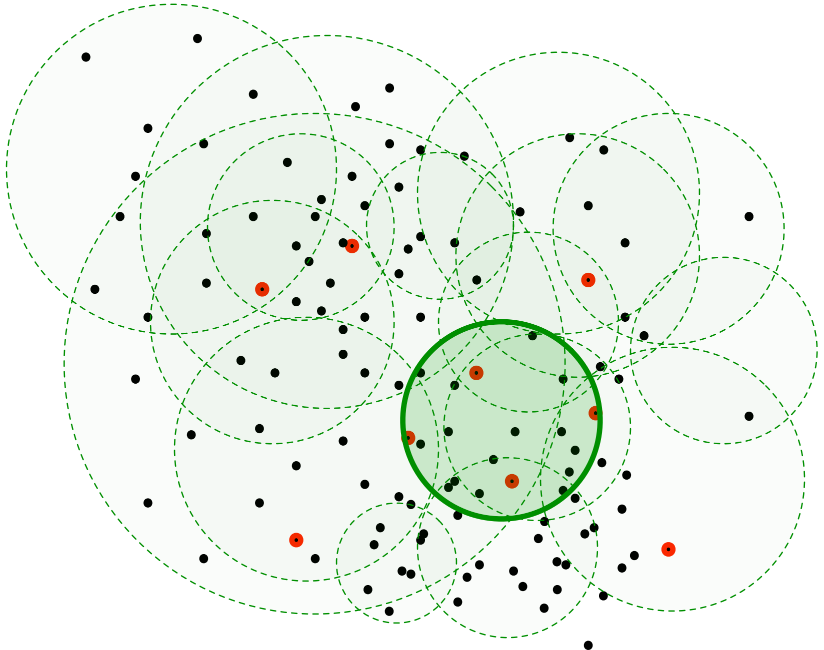
Find a region where **measured data** is significantly denser than **background data**. (Martin Kulldorff 1997)

- Family of regions $\mathcal{C}$ e.g., disks, axis-aligned rectangles, halfspaces
- Data set $X \subset \mathbb{R}^2$, measured set $R \subset X$.
- Define statistic $\Phi(C, X, R) = \Phi(C)$: log-likelihood ratio $\Phi(C) = \log(\frac{\Pr(\mathcal{H}_0|C,X,R)}{\Pr(\mathcal{H}_1|C,X,R)})$

  $\mathcal{H}_0$: no anomaly, rate of measured points same inside than outside

  $\mathcal{H}_1$: cluster C has *higher* rate of measured points than outside C

- Scan **all** $C \in \mathcal{C}$ to find $C^* = \arg\max_{C \in \mathcal{C}} \Phi(C)$



Spatial Scan Statistics

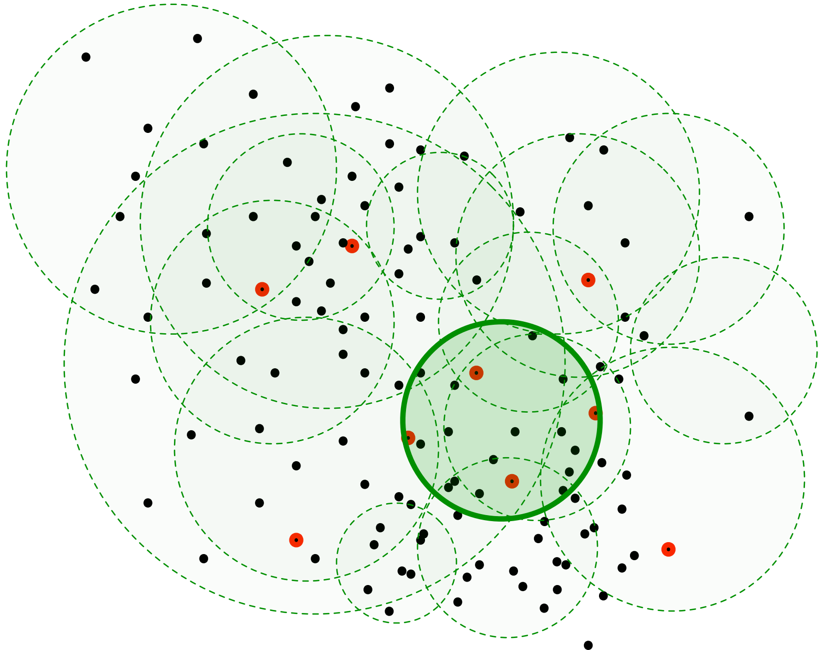
Find a region where **measured data** is significantly denser than **background data**. (Martin Kulldorff 1997)

- Family of regions $\mathcal{C}$ e.g., disks, axis-aligned rectangles, halfspaces
- Data set $X \subset \mathbb{R}^2$, measured set $R \subset X$.
- Define statistic $\Phi(C, X, R) = \Phi(C)$: log-likelihood ratio $\Phi(C) = \log(\frac{\Pr(\mathcal{H}_0 | C, X, R)}{\Pr(\mathcal{H}_1 | C, X, R)})$

  $\mathcal{H}_0$: no anomaly, rate of measured points same inside than outside

  $\mathcal{H}_1$: cluster C has *higher* rate of measured points than outside C
- Scan **all** $C \in \mathcal{C}$ to find $C^* = \arg\max_{C \in \mathcal{C}} \Phi(C)$
- Run 1000 permutation tests to measure significance

  Repeat **scan** on random data, 1000x



FiveThirtyEight

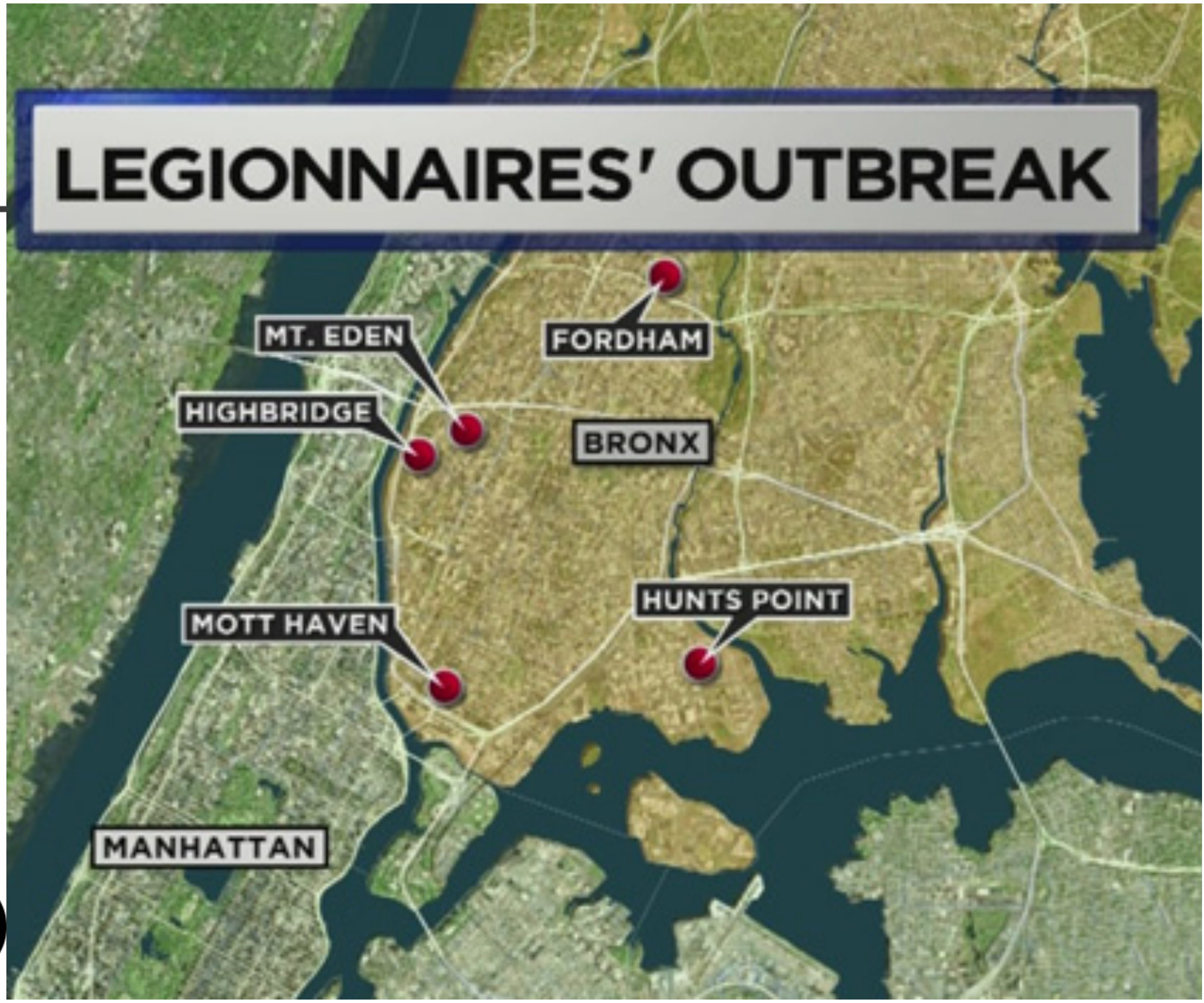
Politics Sports **Science & Health** Economics Culture

NOV. 16, 2016 AT 8:00 AM

How New York Hunts For Early Signs Of Disease Outbreaks

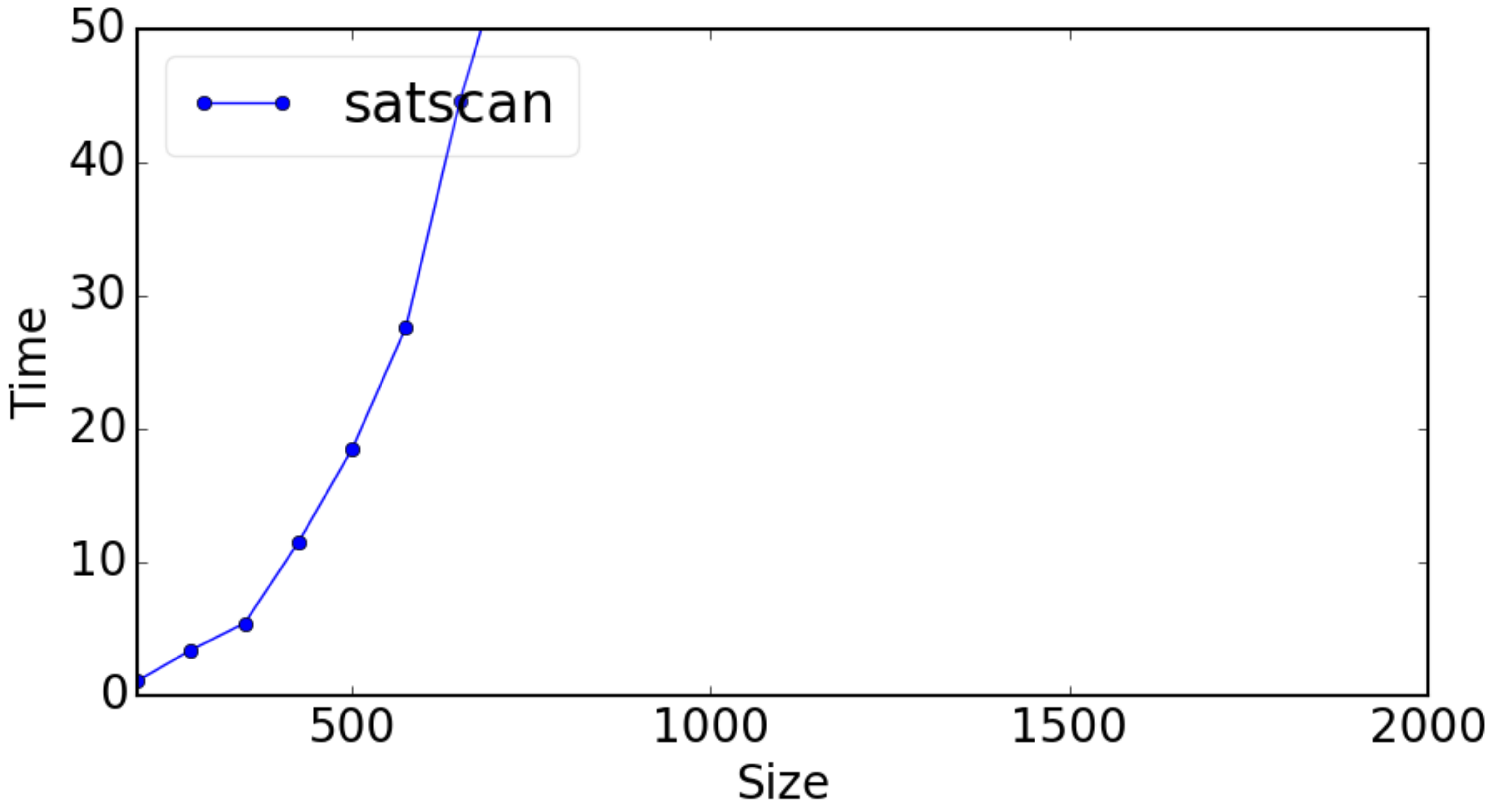
By Ian Evans

Filed under Public Health

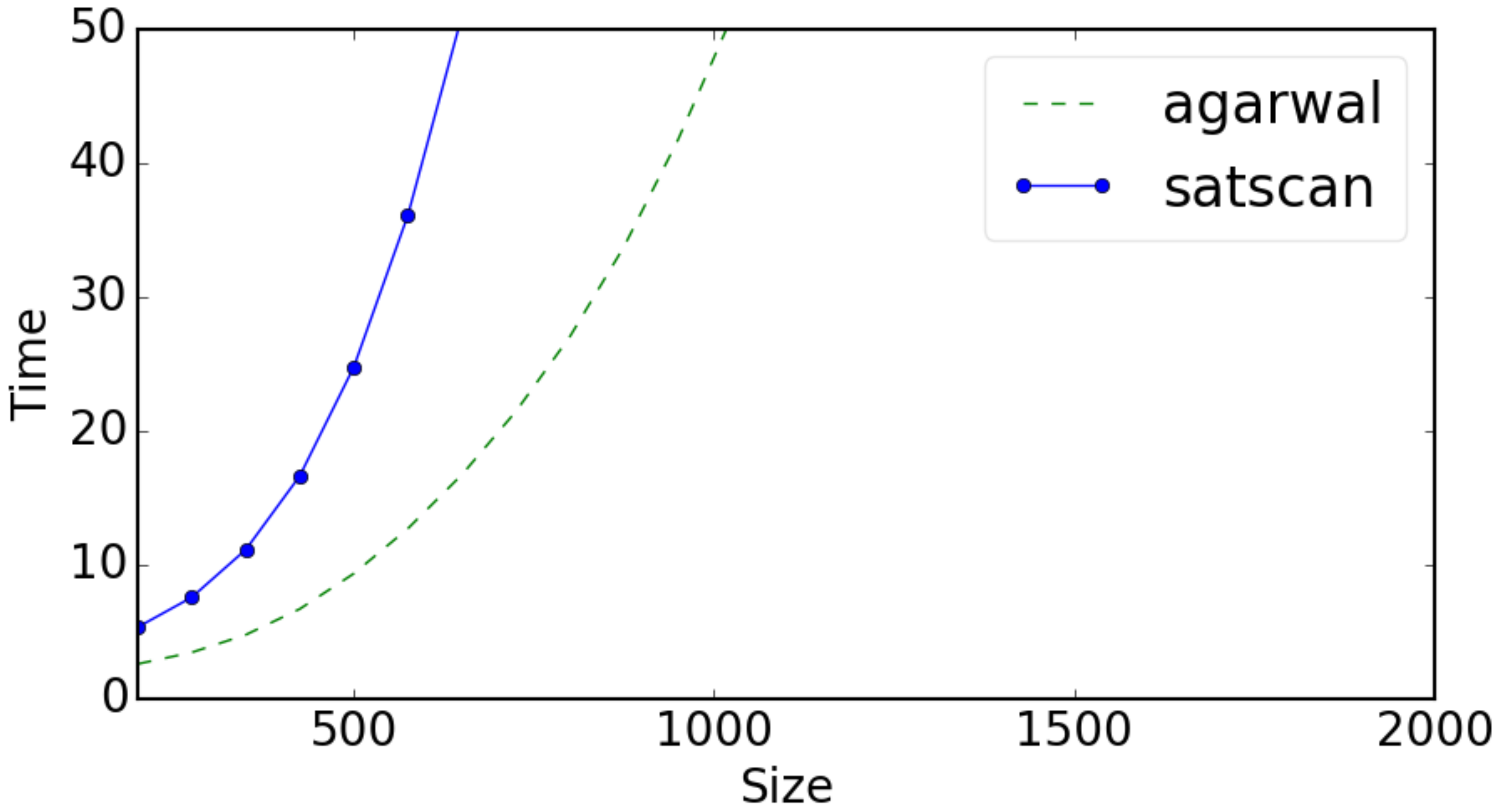


On July 29, 2015, the New York City Department of Health and Mental Hygiene sent out an alert — 31 people in the South Bronx had contracted Legionnaires' disease, a lung infection from waterborne bacteria that kills about 1 out of every 10 people who get it. By the time officials found the source (a cooling tower) and contained the spread, 128 people had contracted Legionnaires' and 12 people had died. It was the largest outbreak of Legionnaires' disease in the city's history — an outbreak that was first detected by a computer program.

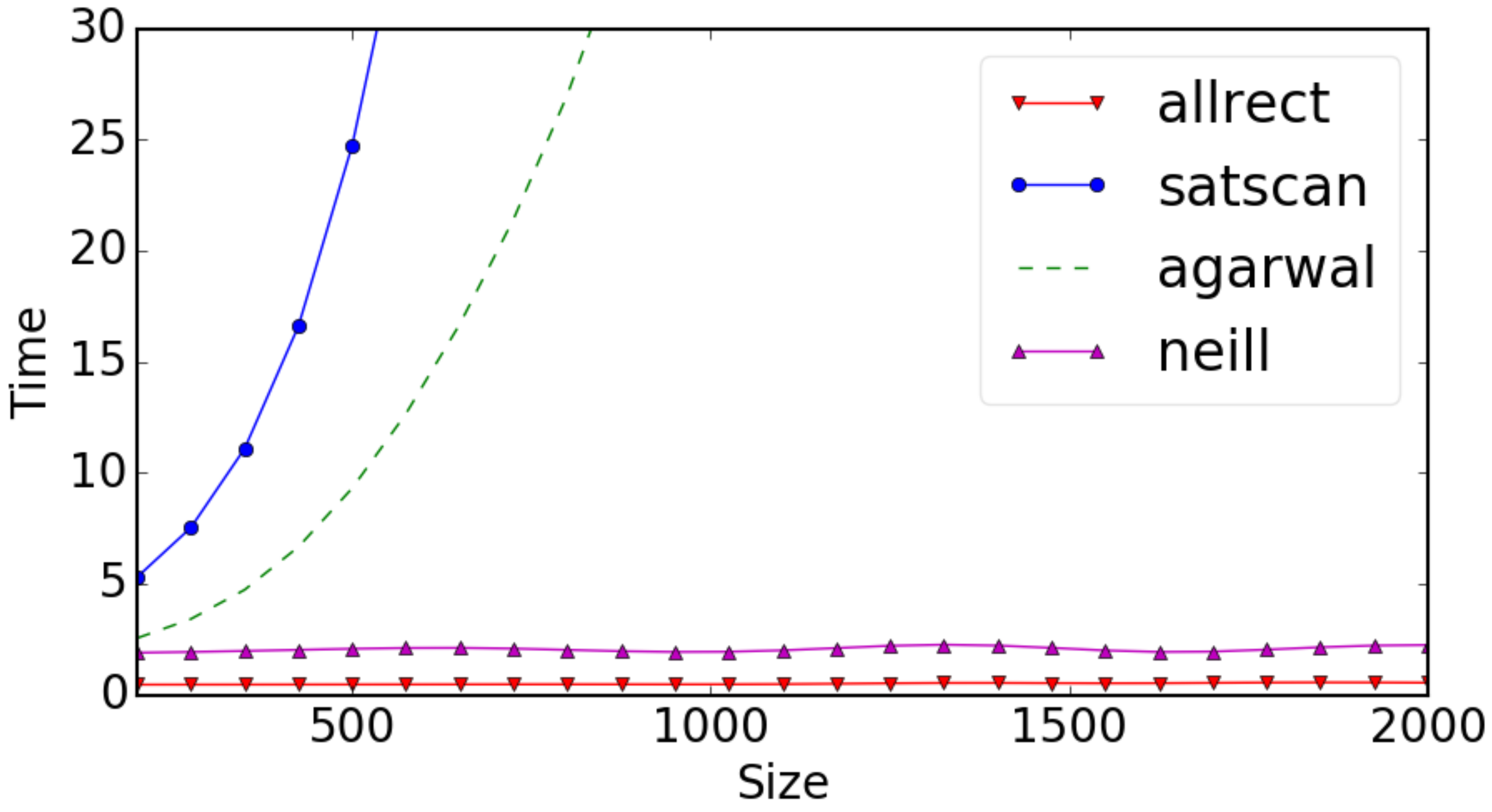
SatScan is not Scalable



SatScan is not Scalable

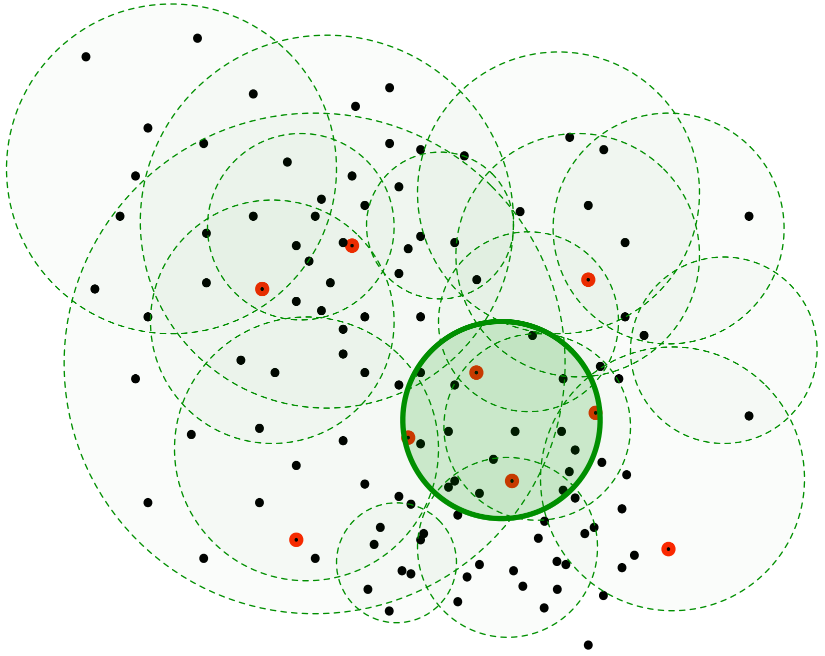


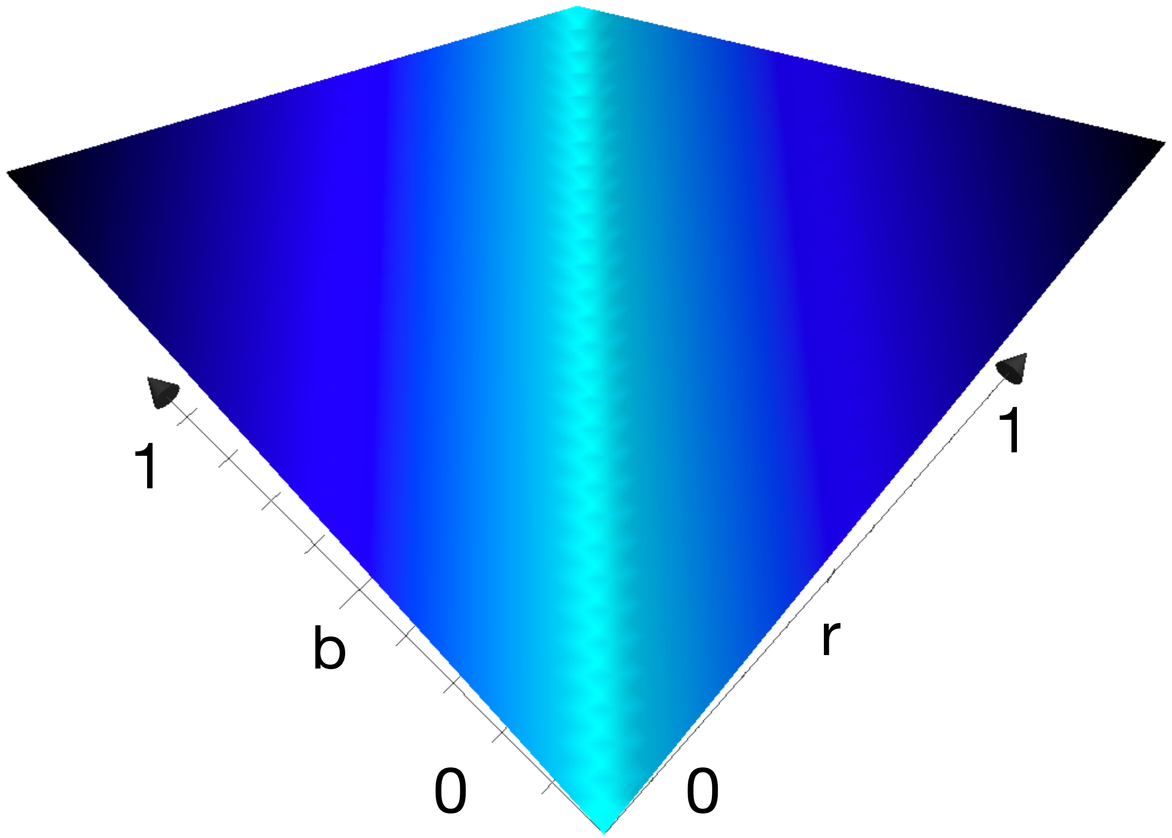
SatScan is not Scalable



Statistic Function Φ

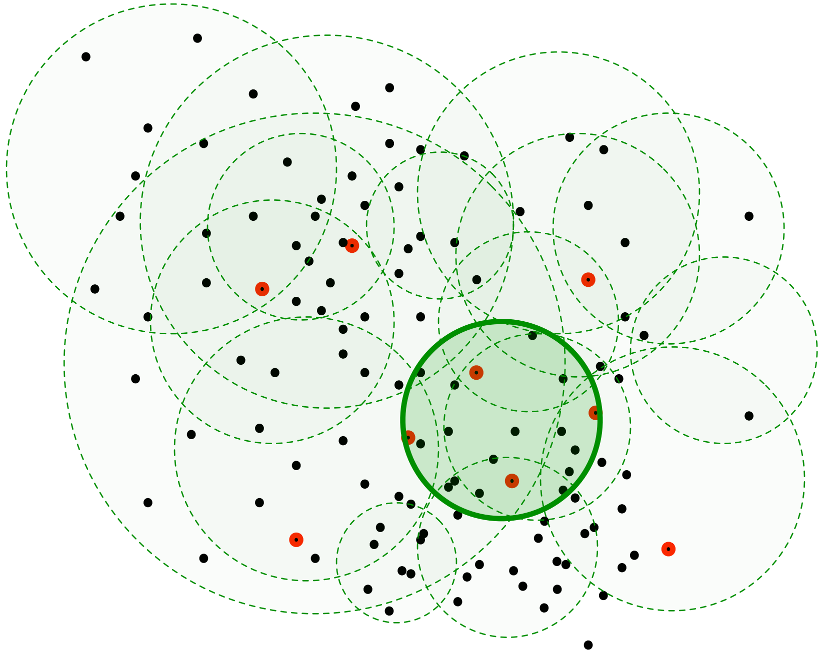
- baseline $b(C) = \frac{|X \cap C|}{|X|}$ measured $r(C) = \frac{|R \cap C|}{|R|}$
- $\Phi(C) = \phi(b(C), r(C)) = |b(C) - r(C)|$

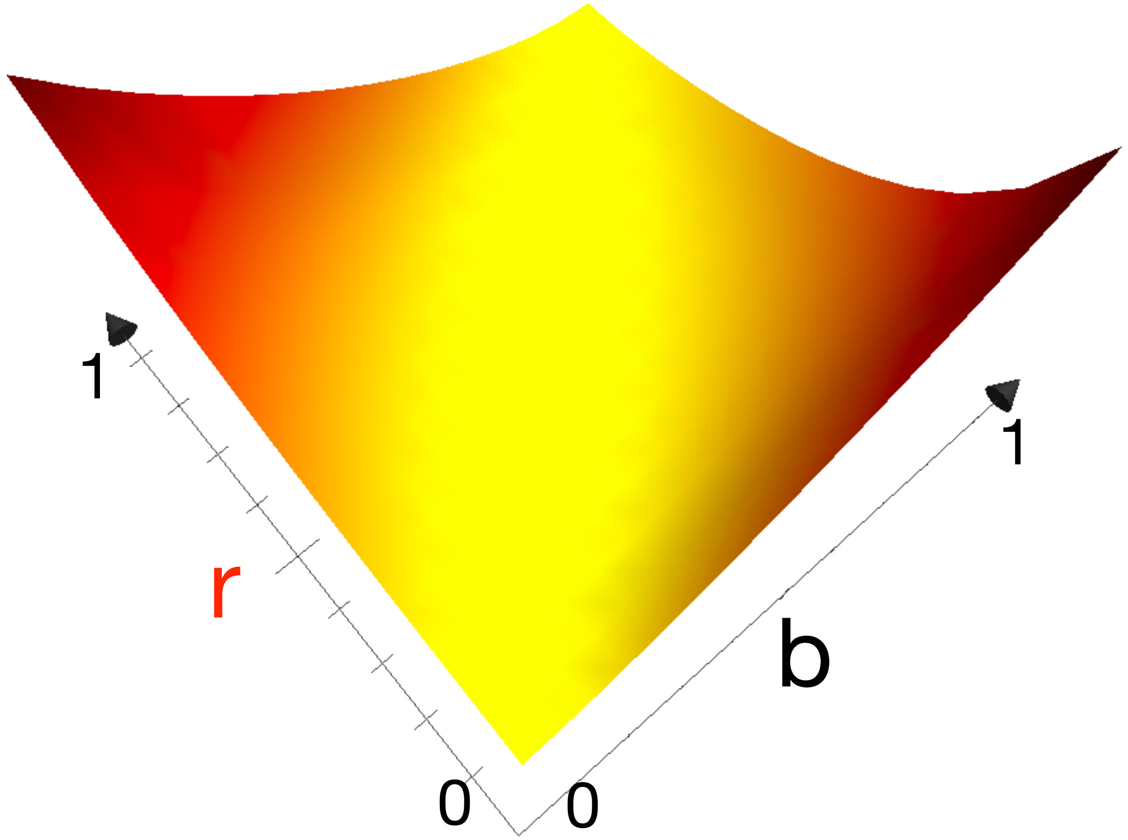




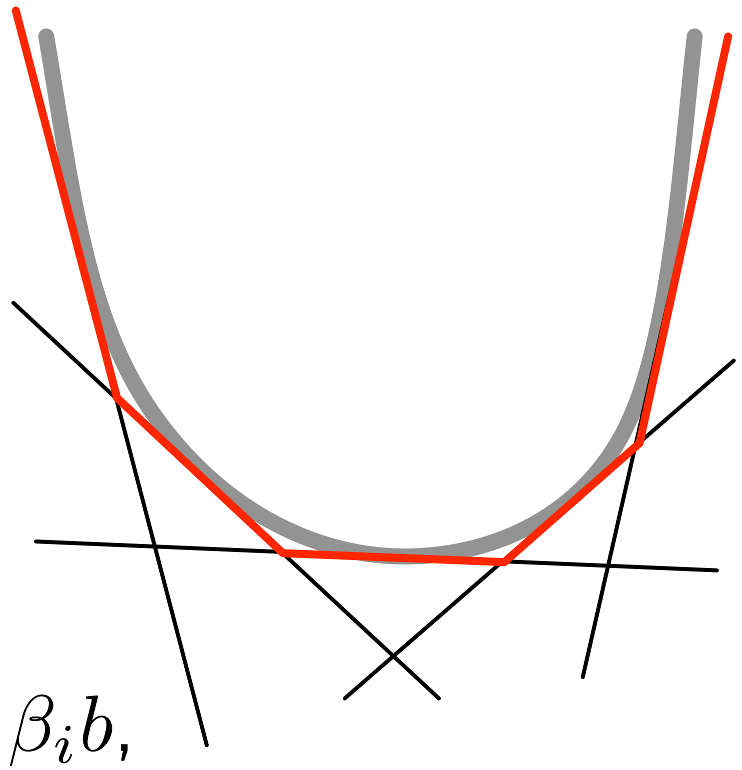
Statistic Function Φ

- baseline $b(C) = \frac{|X \cap C|}{|X|}$ measured $r(C) = \frac{|R \cap C|}{|R|}$
- $\Phi(C) = \phi(b(C), r(C)) = |b(C) - r(C)|$
- Kulldorff: $\phi_K(b, r) = r \ln \frac{r}{b} + (1 - r) \ln \frac{1-r}{1-b} = \mathsf{KL}(r, b)$





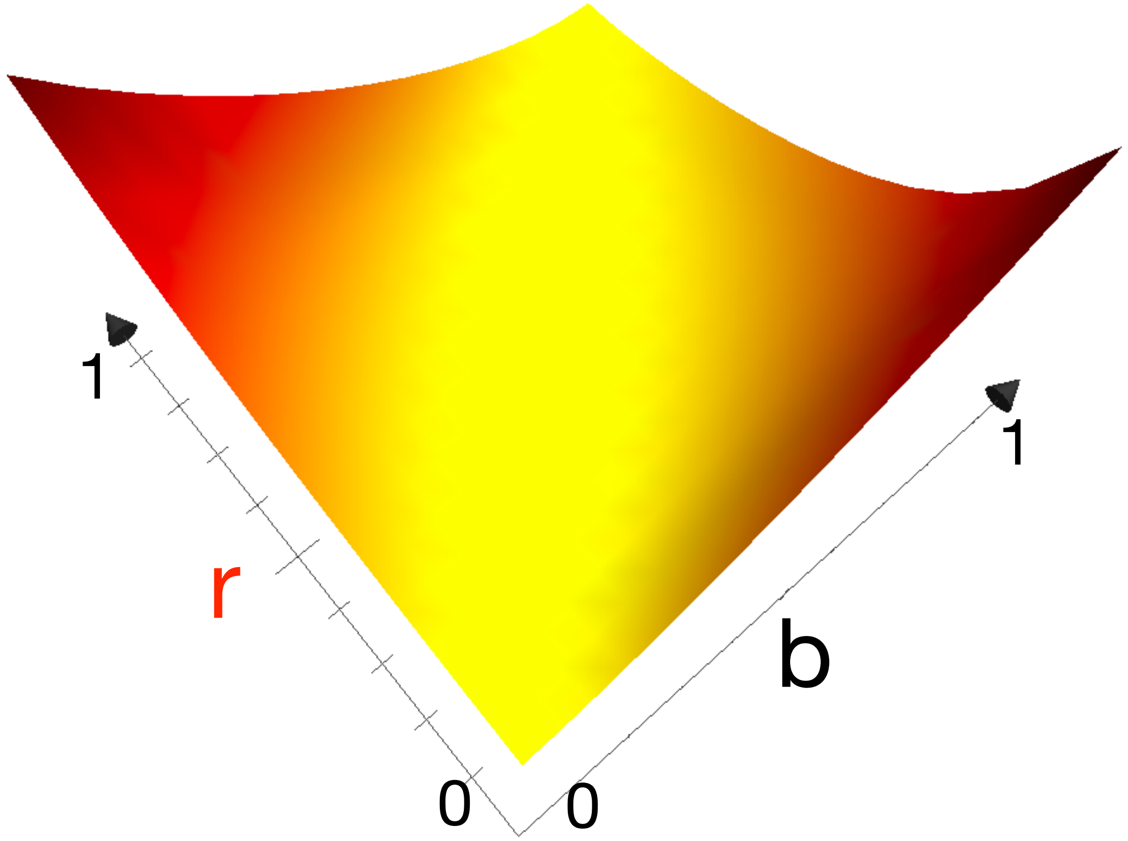
Statistic Function Φ



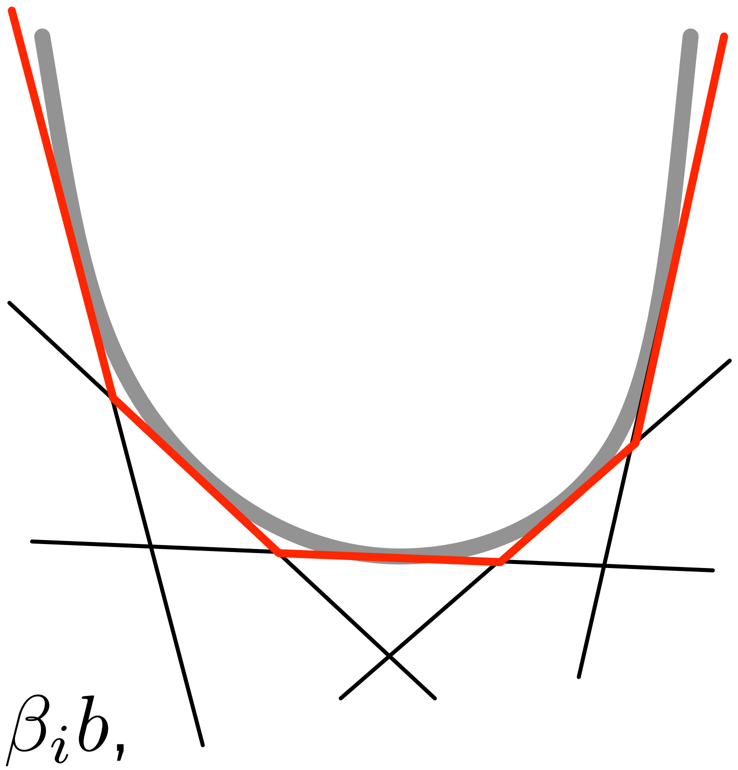
- baseline $b(C) = \frac{|X \cap C|}{|X|}$ measured $r(C) = \frac{|R \cap C|}{|R|}$
- $\Phi(C) = \phi(b(C), r(C)) = |b(C) - r(C)|$
- Kulldorff: $\phi_K(b, r) = r \ln \frac{r}{b} + (1-r) \ln \frac{1-r}{1-b} = \mathsf{KL}(r, b)$

Exists set of k linear functions $\{\phi_1, \ldots, \phi_k\}$ s.t. $\phi(r, b) = \alpha_i r + \beta_i b$, for $\varepsilon \in (0,1)$, so for all $(r, b) \in [\varepsilon, 1-\varepsilon]^2$ then

$$\phi_K(r, b) \geq \max_i \phi_i(r, b) \geq \phi_K(r, b) - \varepsilon.$$



Statistic Function Φ

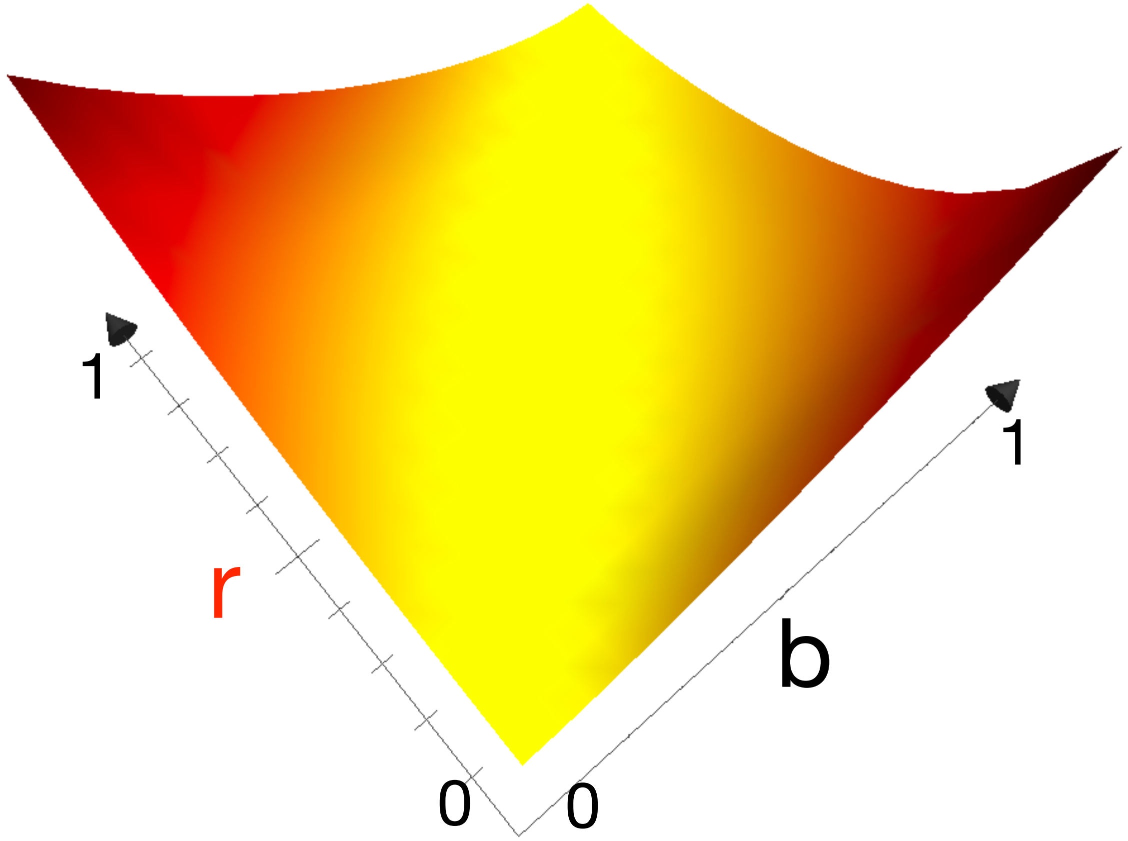


- baseline $b(C) = \frac{|X \cap C|}{|X|}$ measured $r(C) = \frac{|R \cap C|}{|R|}$
- $\Phi(C) = \phi(b(C), r(C)) = |b(C) - r(C)|$
- Kulldorff: $\phi_K(b,r) = r \ln \frac{r}{b} + (1-r) \ln \frac{1-r}{1-b} = \mathsf{KL}(r,b)$

Exists set of k linear functions $\{\phi_1, \ldots, \phi_k\}$ s.t. $\phi(r,b) = \alpha_i r + \beta_i b$, for $\varepsilon \in (0,1)$, so for all $(r,b) \in [\varepsilon, 1-\varepsilon]^2$ then

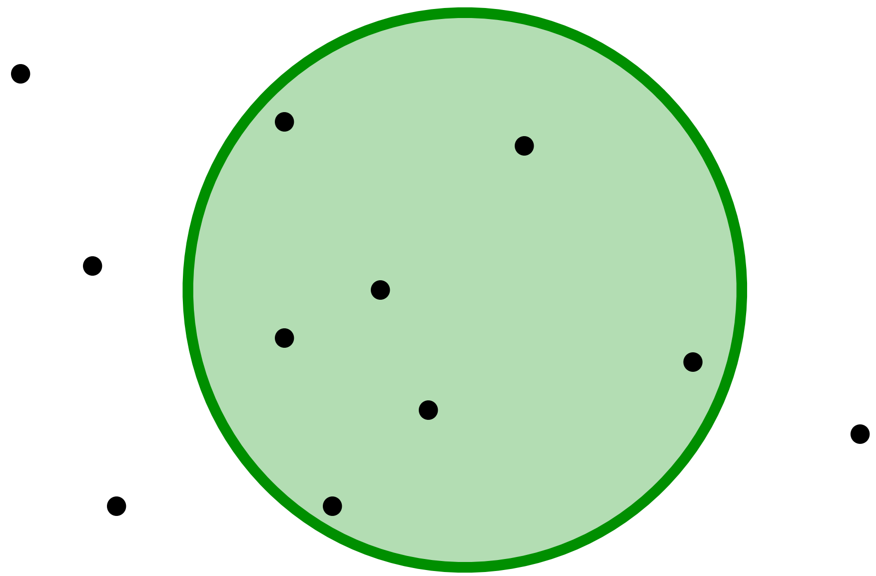
$$\phi_K(r,b) \geq \max_i \phi_i(r,b) \geq \phi_K(r,b) - \varepsilon.$$

- Agarwal etal 2006 : $k = O((1/\varepsilon)\log(1/\varepsilon))$
- Phillips and Matheny 2018 : $k = 1/\sqrt{\varepsilon}$



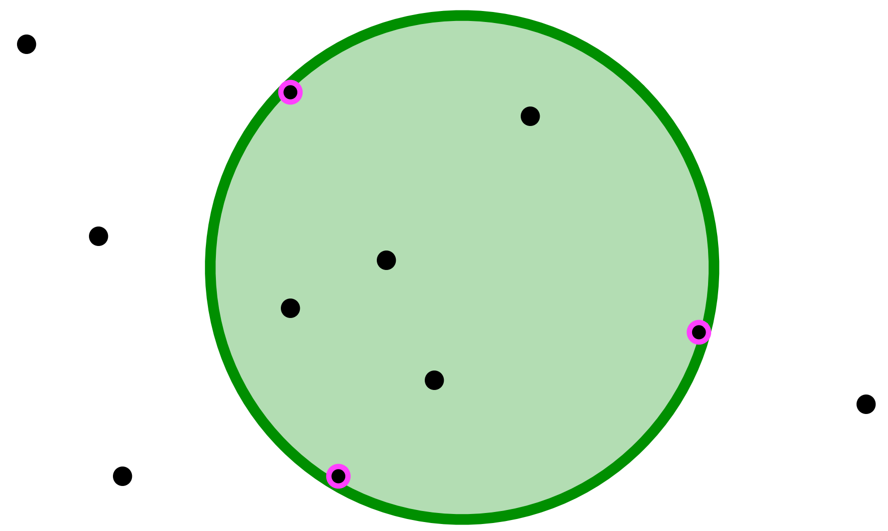
Scanning all Ranges

- Every disc is combinatorially defined by at most 3 points.



Scanning all Ranges

- Every disc is combinatorially defined by at most 3 points.

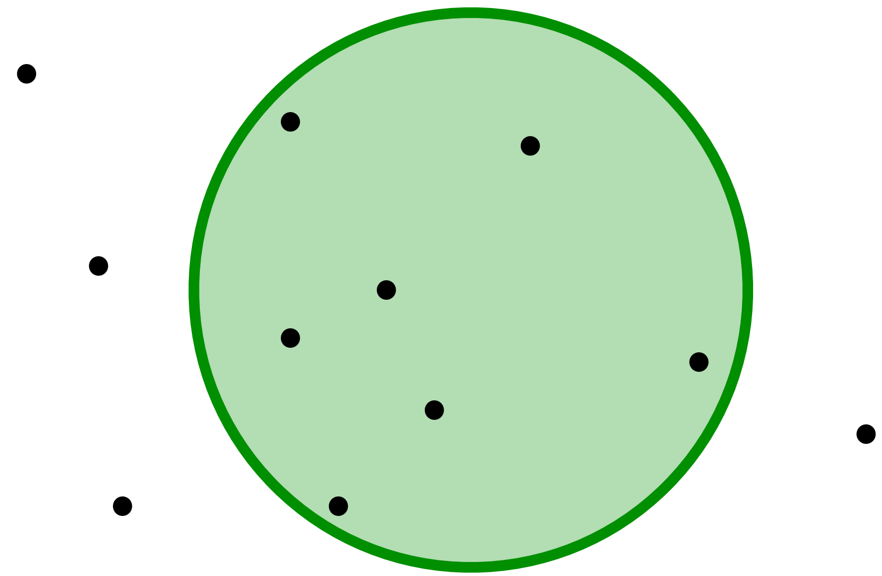


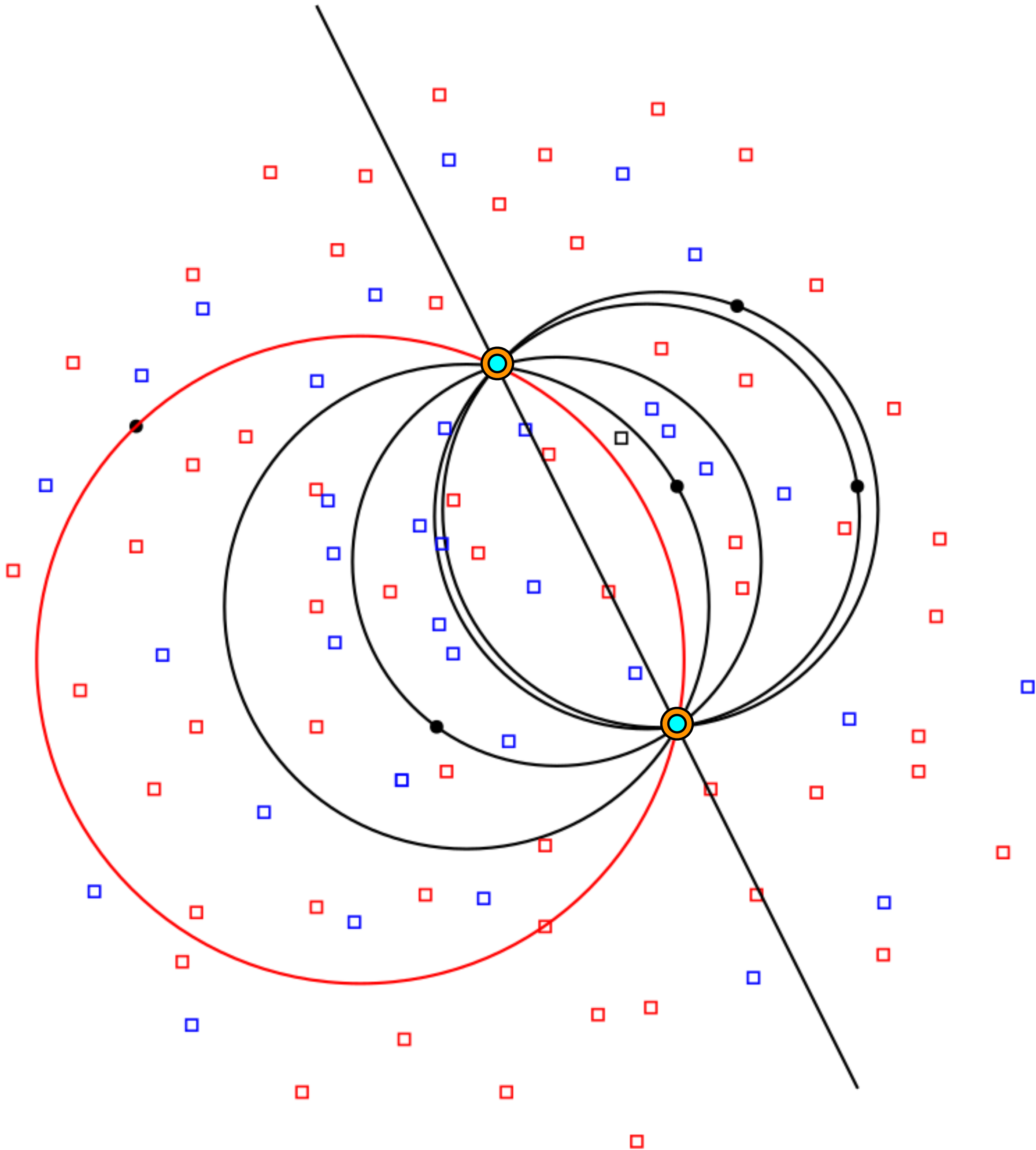
Scanning all Ranges

- Every disc is combinatorially defined by at most 3 points.

  $\Rightarrow$ Choose all $\binom{n}{2}$ pairs, scan through n points.

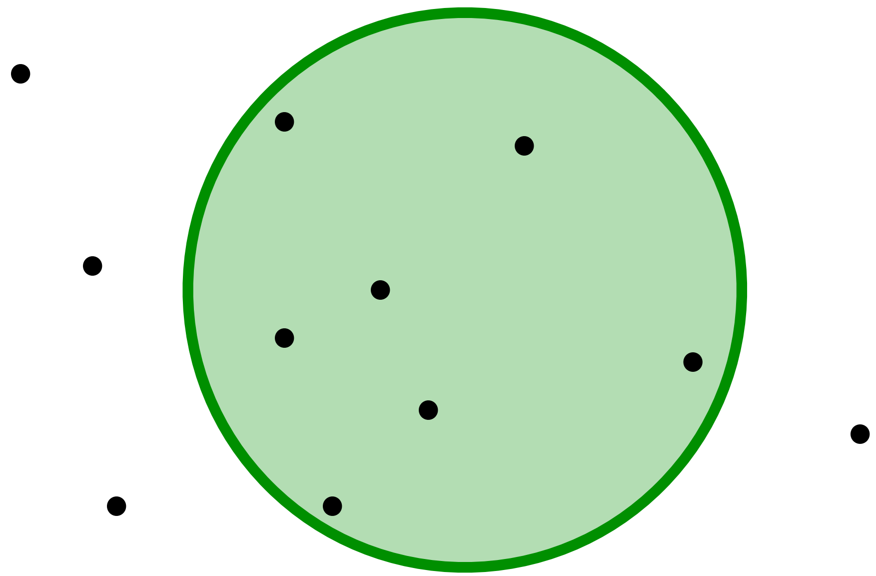
  $O(n^3)$ time to scan **all** disks!

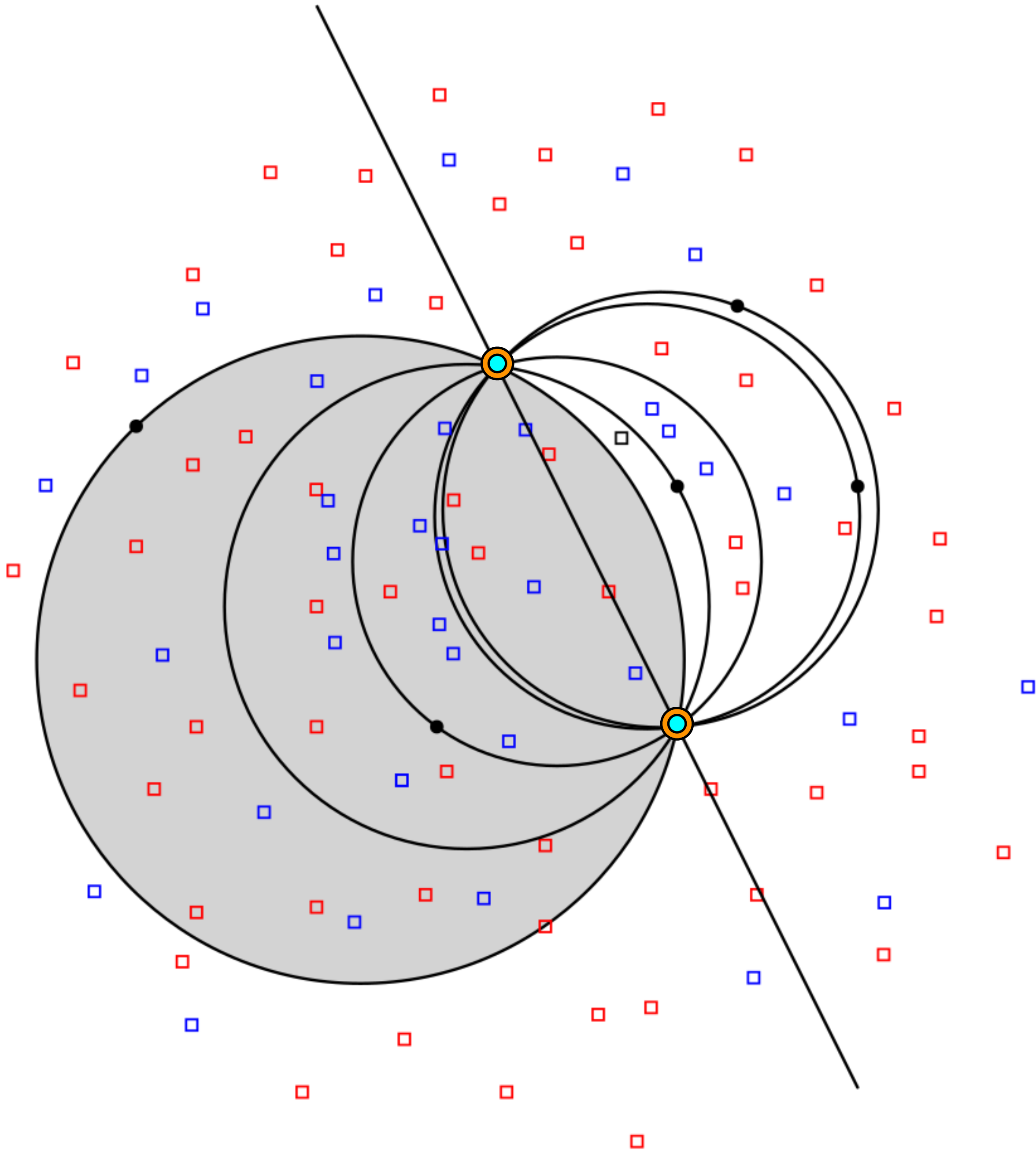




Scanning all Ranges

- Every disc is combinatorially defined by at most 3 points.
  - $\Rightarrow$ Choose all $\binom{n}{2}$ pairs, scan through n points.
    $O(n^3)$ time to scan **all** disks!



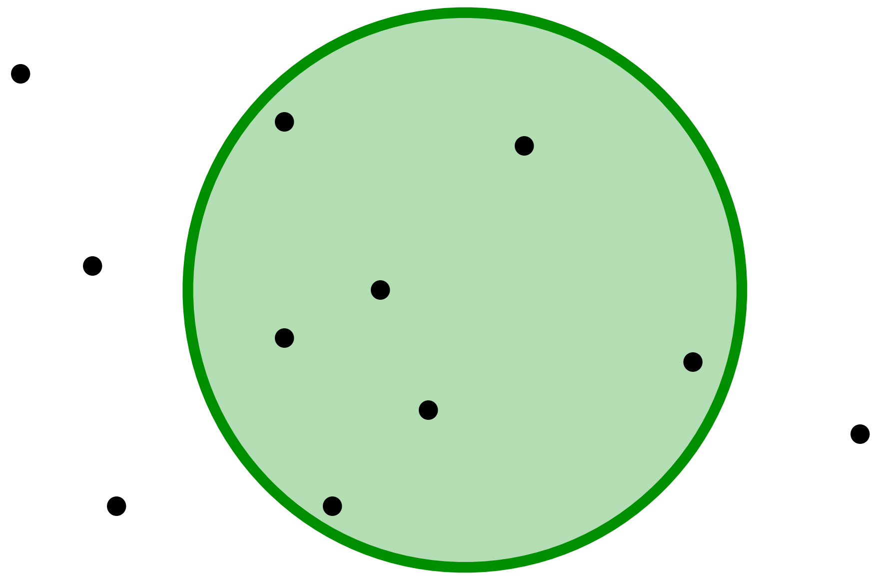


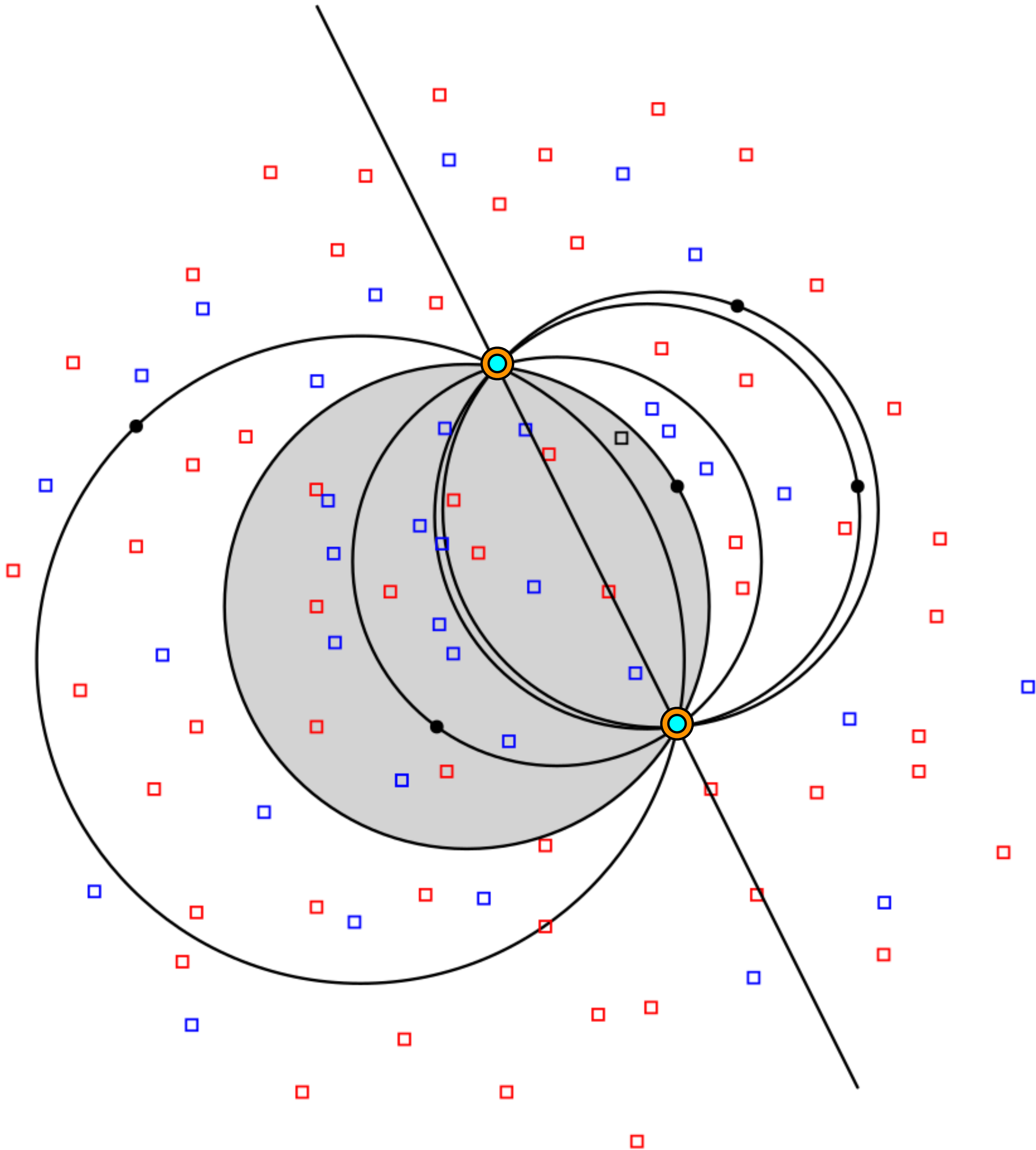
Scanning all Ranges

- Every disc is combinatorially defined by at most 3 points.

  $\Rightarrow$ Choose all $\binom{n}{2}$ pairs, scan through n points.

  $O(n^3)$ time to scan **all** disks!



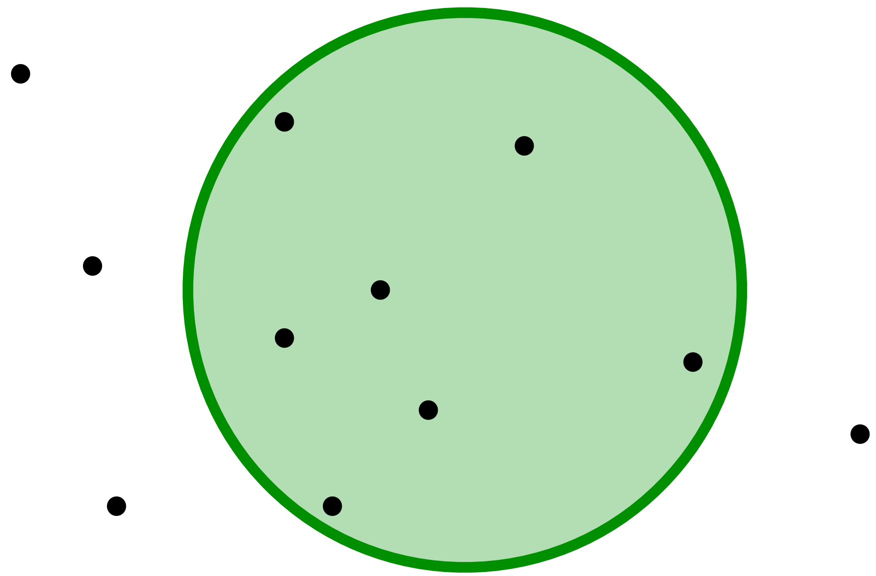


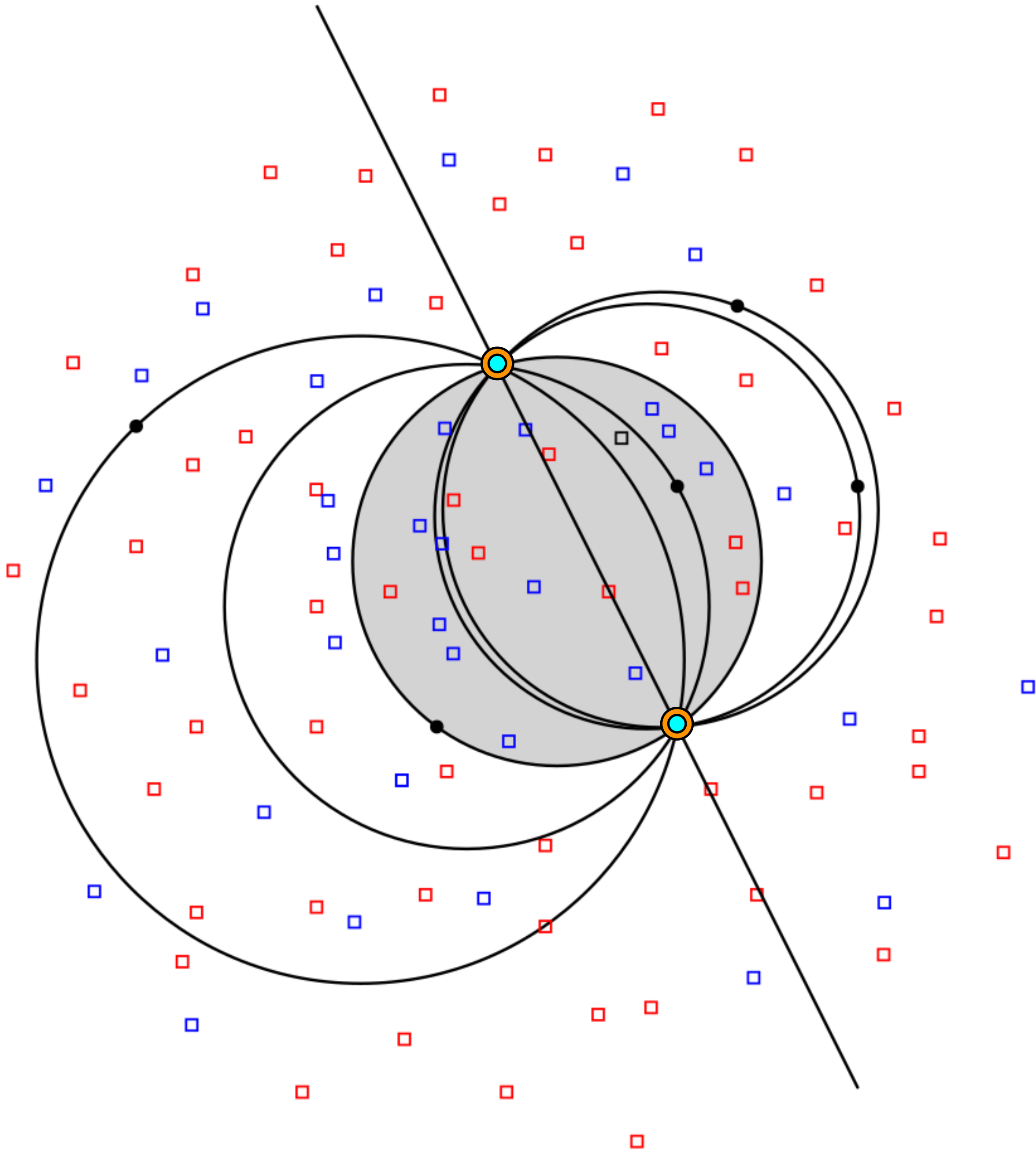
Scanning all Ranges

- Every disc is combinatorially defined by at most 3 points.

  $\Rightarrow$ Choose all $\binom{n}{2}$ pairs, scan through n points.

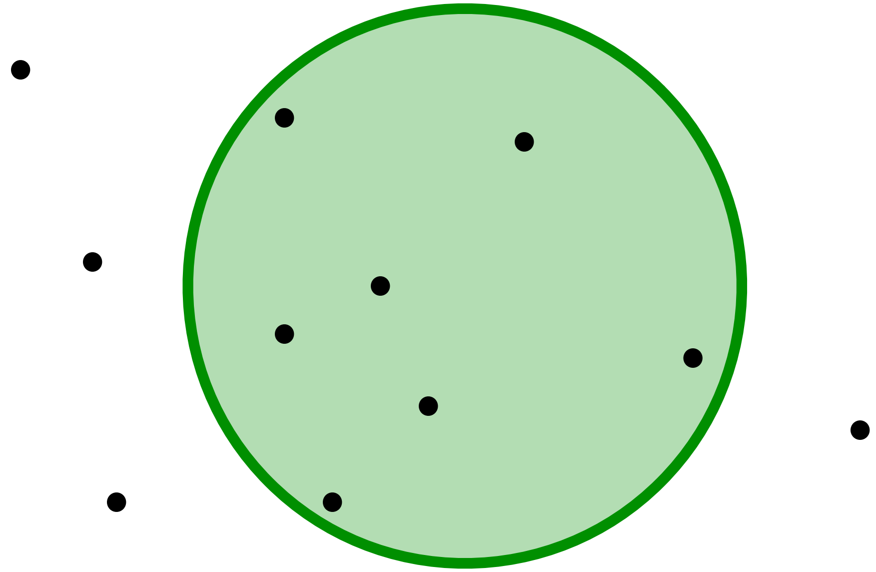
  $O(n^3)$ time to scan **all** disks!

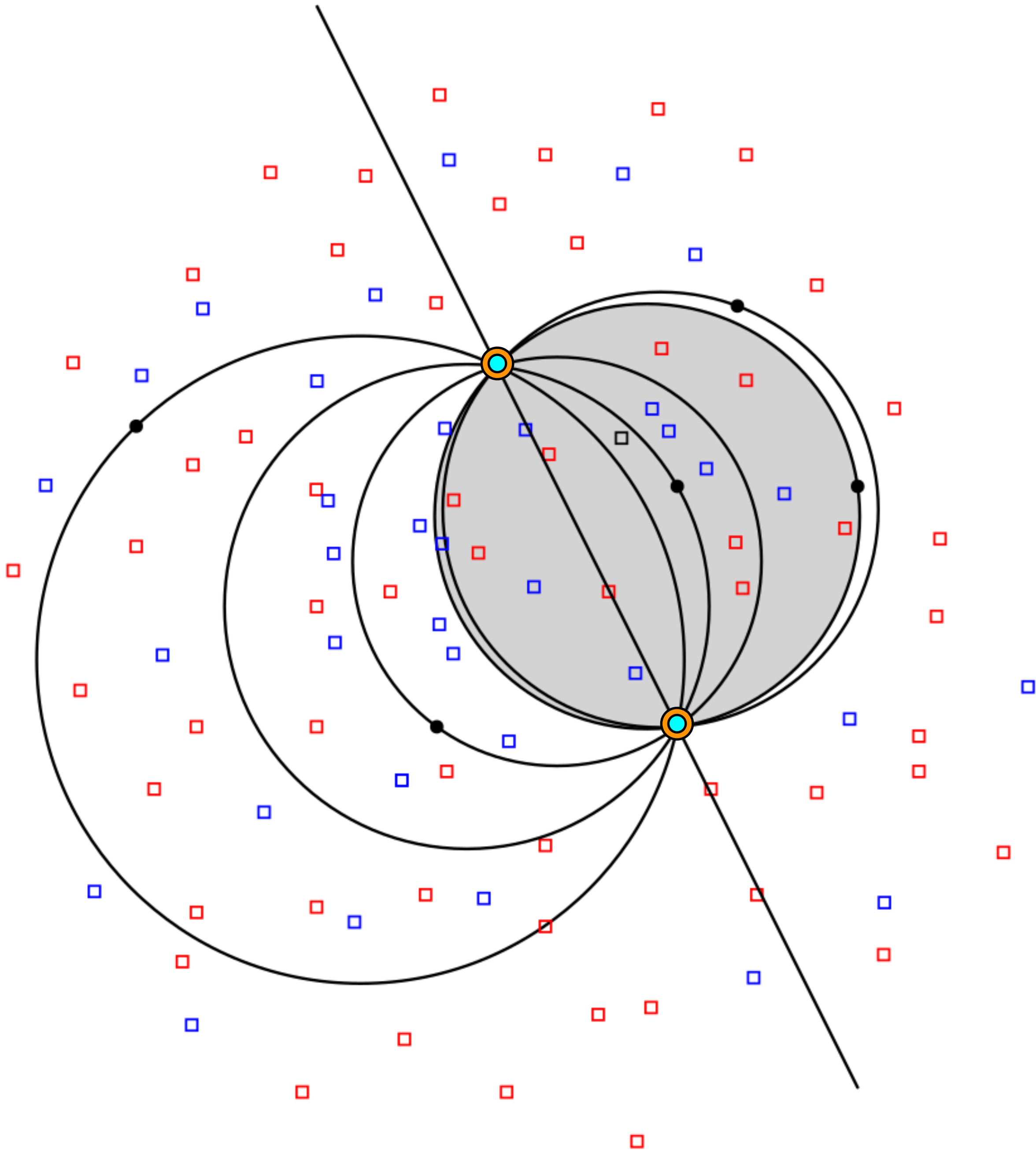




Scanning all Ranges

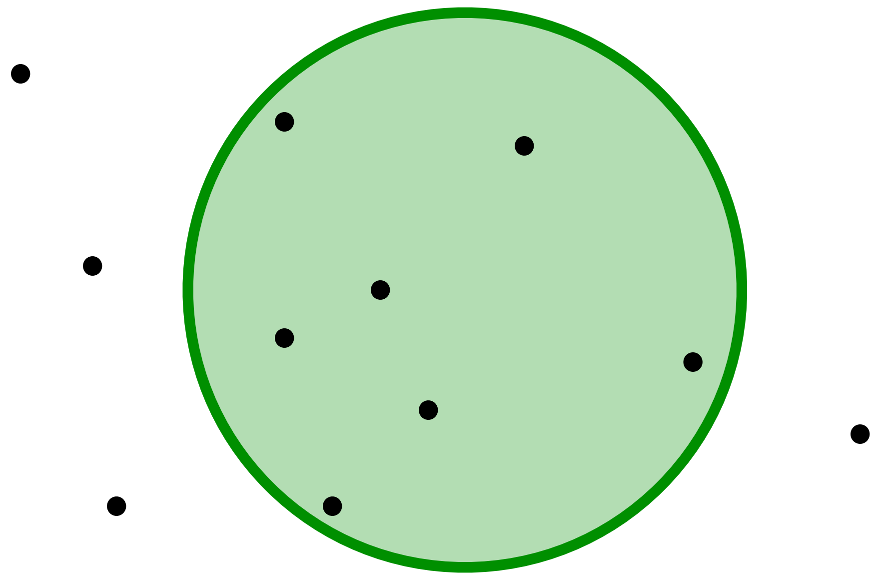
- Every disc is combinatorially defined by at most 3 points.
  $\Rightarrow$ Choose all $\binom{n}{2}$ pairs, scan through n points.
  $O(n^3)$ time to scan **all** disks!

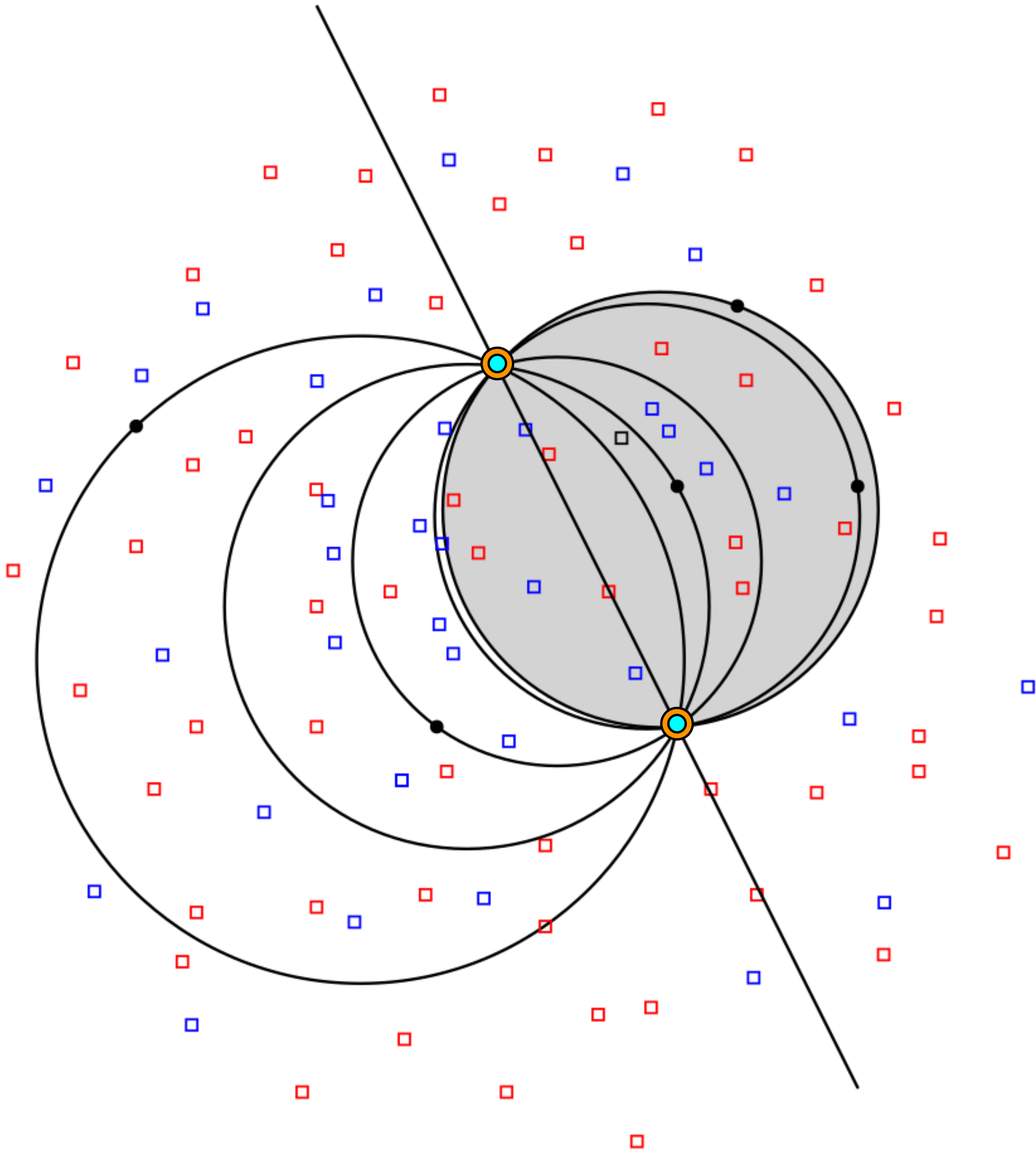




Scanning all Ranges

- Every disc is combinatorially defined by at most 3 points.
  $\Rightarrow$ Choose all $\binom{n}{2}$ pairs, scan through n points.
  $O(n^3)$ time to scan **all** disks!



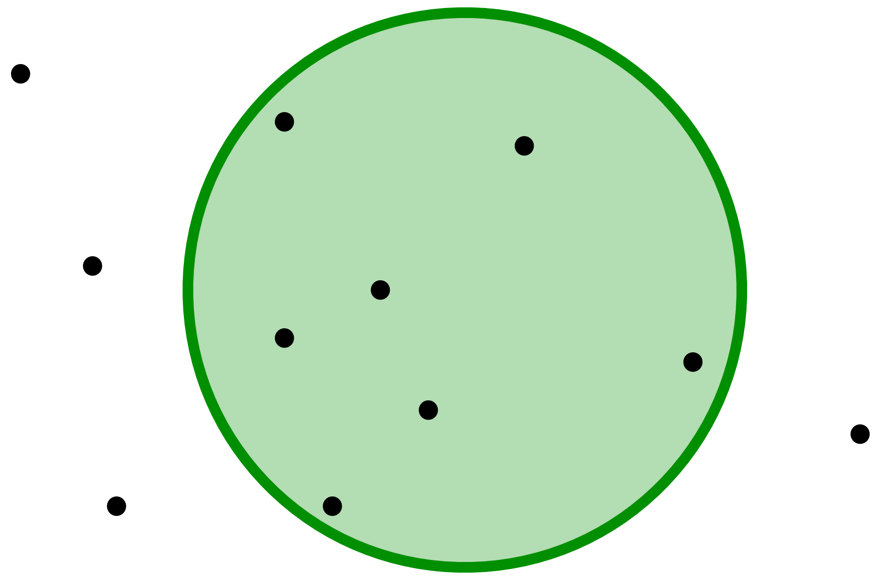


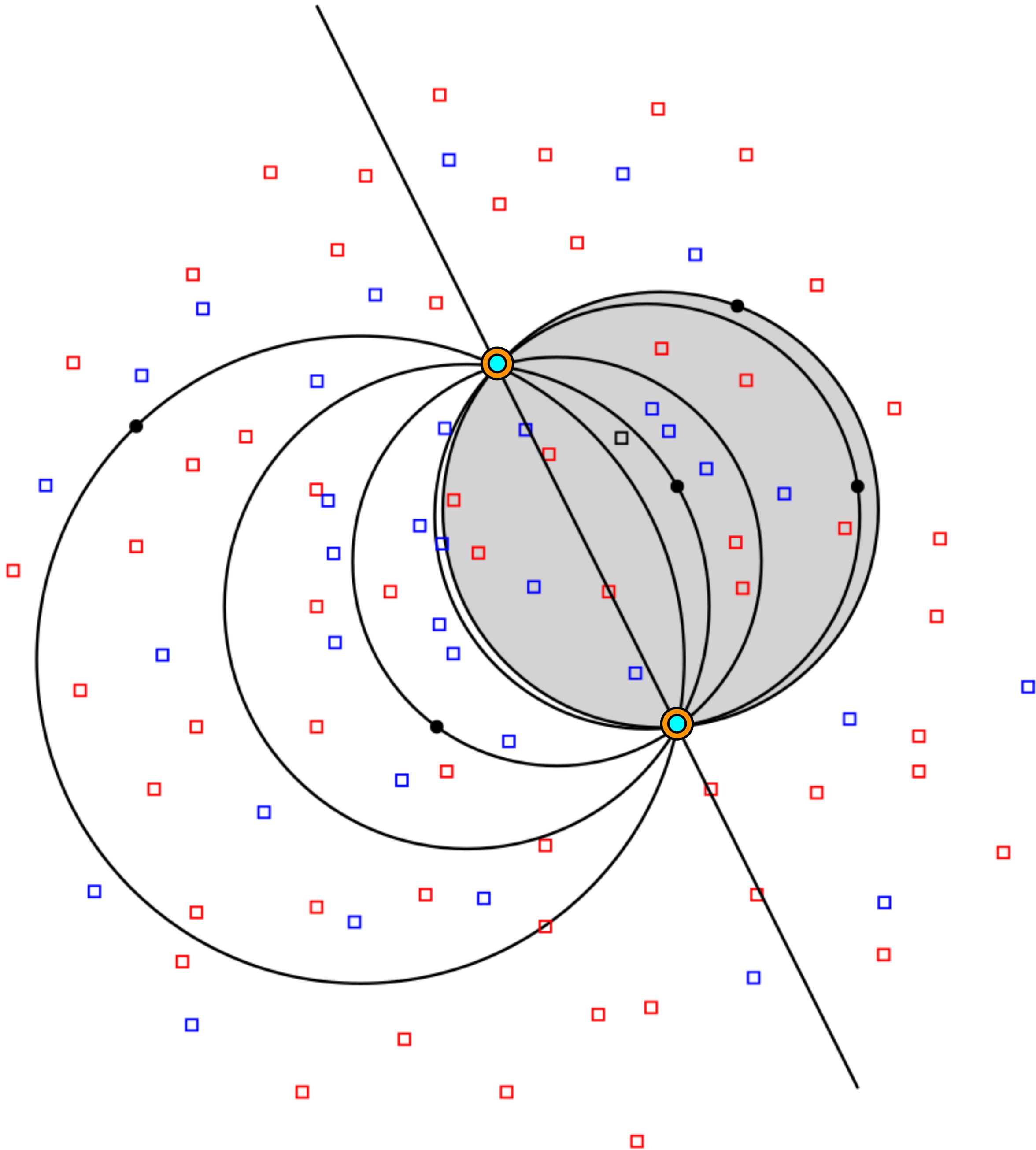
Scanning all Ranges

- Every disc is combinatorially defined by at most 3 points.
  $\Rightarrow$ Choose all $\binom{n}{2}$ pairs, scan through n points.
  $O(n^3)$ time to scan **all** disks!

- $O(n^d)$ halfspaces in $\mathbb{R}^d$

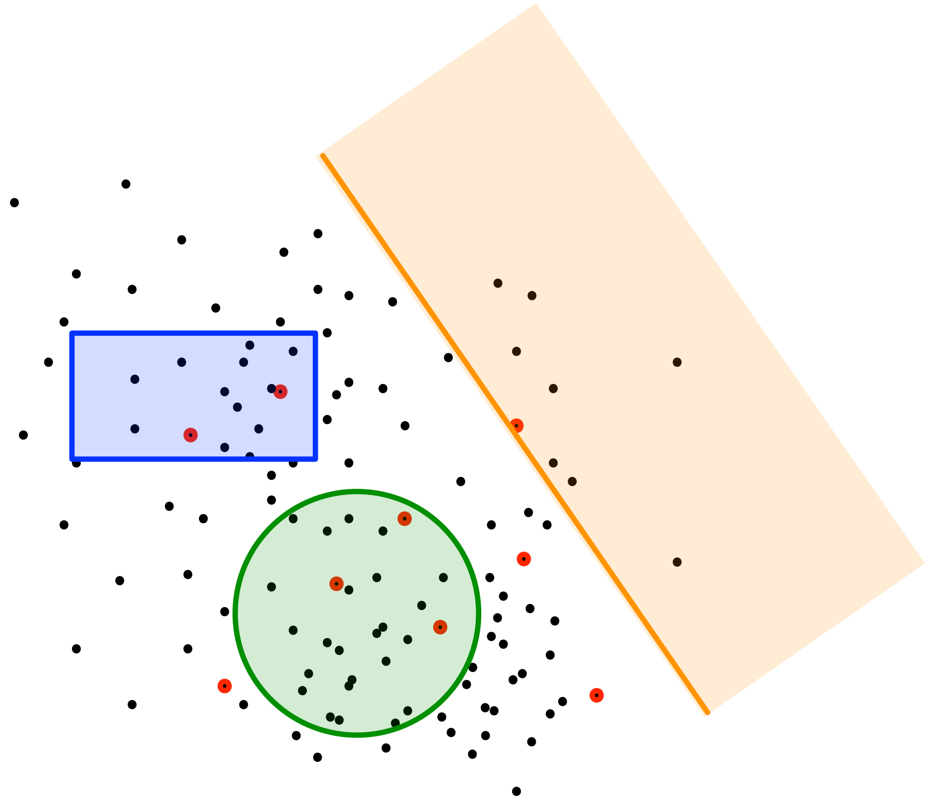
- $O(n^{2d})$ axis-aligned rectangles in $\mathbb{R}^d$





Conforming Range Space

- A *range space* is a pair $(X, \mathcal{A})$, where $\mathcal{A}$ is a set of subsets of set X.



Conforming Range Space

- A *range space* is a pair $(X, \mathcal{A})$, where $\mathcal{A}$ is a set of subsets of set X.

- A mapping $\psi_{\mathcal{A}}$ is *conforming* to $(X, \mathcal{A})$ if for any $N \subset X$

(recovery) :
(inclusion) :

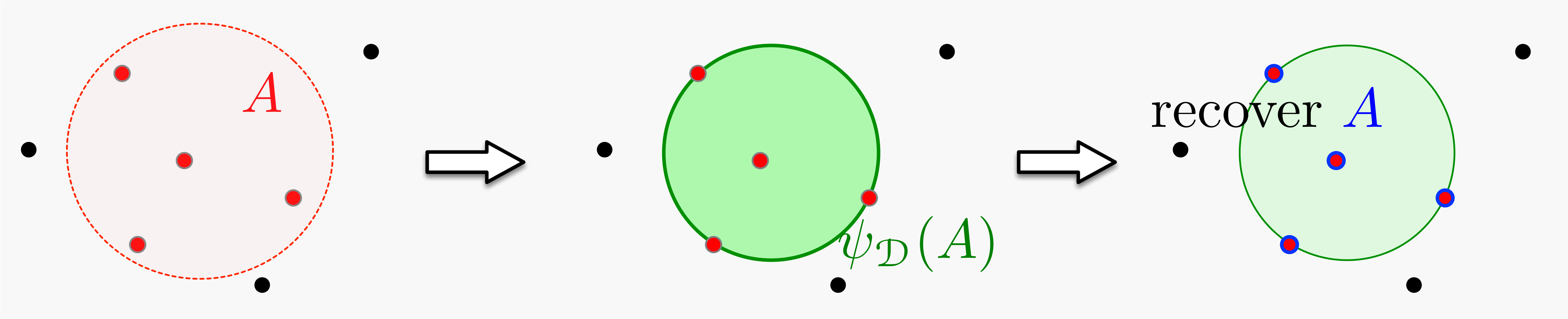
Conforming Range Space

- A *range space* is a pair $(X, \mathcal{A})$, where $\mathcal{A}$ is a set of subsets of set X.

- A mapping $\psi_{\mathcal{A}}$ is *conforming* to $(X, \mathcal{A})$ if for any $N \subset X$

(recovery) : any $A \in (N, \mathcal{A})$ then $\psi_{\mathcal{A}}(A) \cap N = A$

(inclusion) :



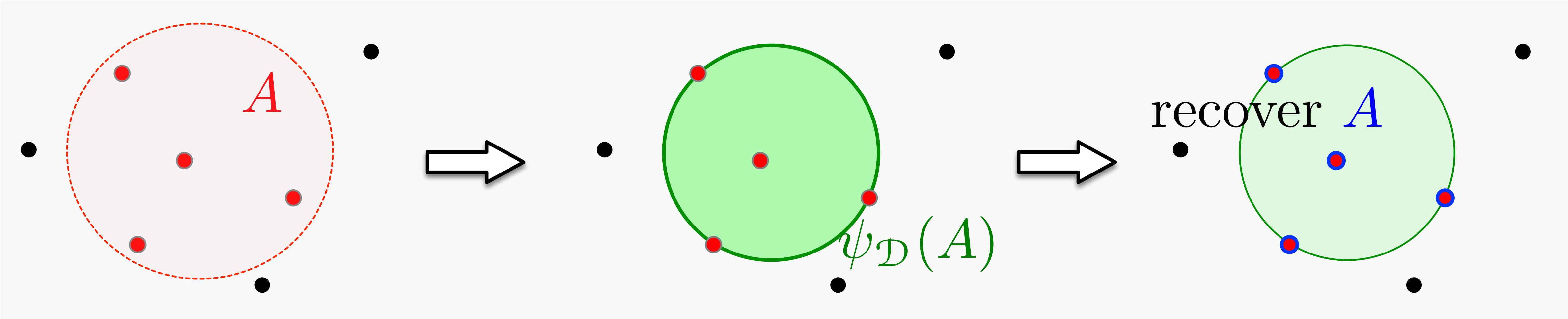
Conforming Range Space

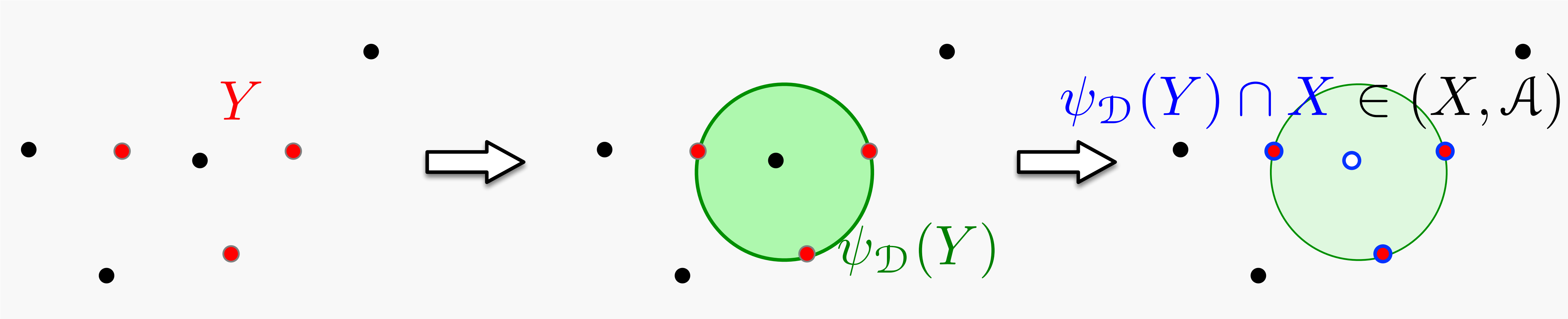
- A *range space* is a pair $(X, \mathcal{A})$, where $\mathcal{A}$ is a set of subsets of set X.

- A mapping $\psi_{\mathcal{A}}$ is *conforming* to $(X, \mathcal{A})$ if for any $N \subset X$

(recovery) : any $A \in (N, \mathcal{A})$ then $\psi_{\mathcal{A}}(A) \cap N = A$

(inclusion) : any $Y \subset X$ then $\psi_{\mathcal{A}}(Y) \cap X \in (X, \mathcal{A})$





Conforming Range Space

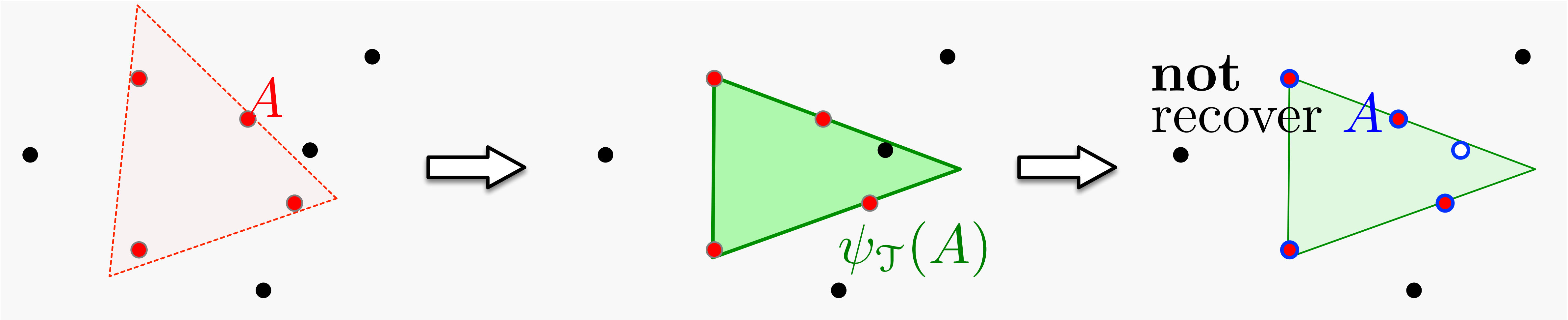
- A *range space* is a pair $(X, \mathcal{A})$, where $\mathcal{A}$ is a set of subsets of set X.

- A mapping $\psi_{\mathcal{A}}$ is *conforming* to $(X, \mathcal{A})$ if for any $N \subset X$

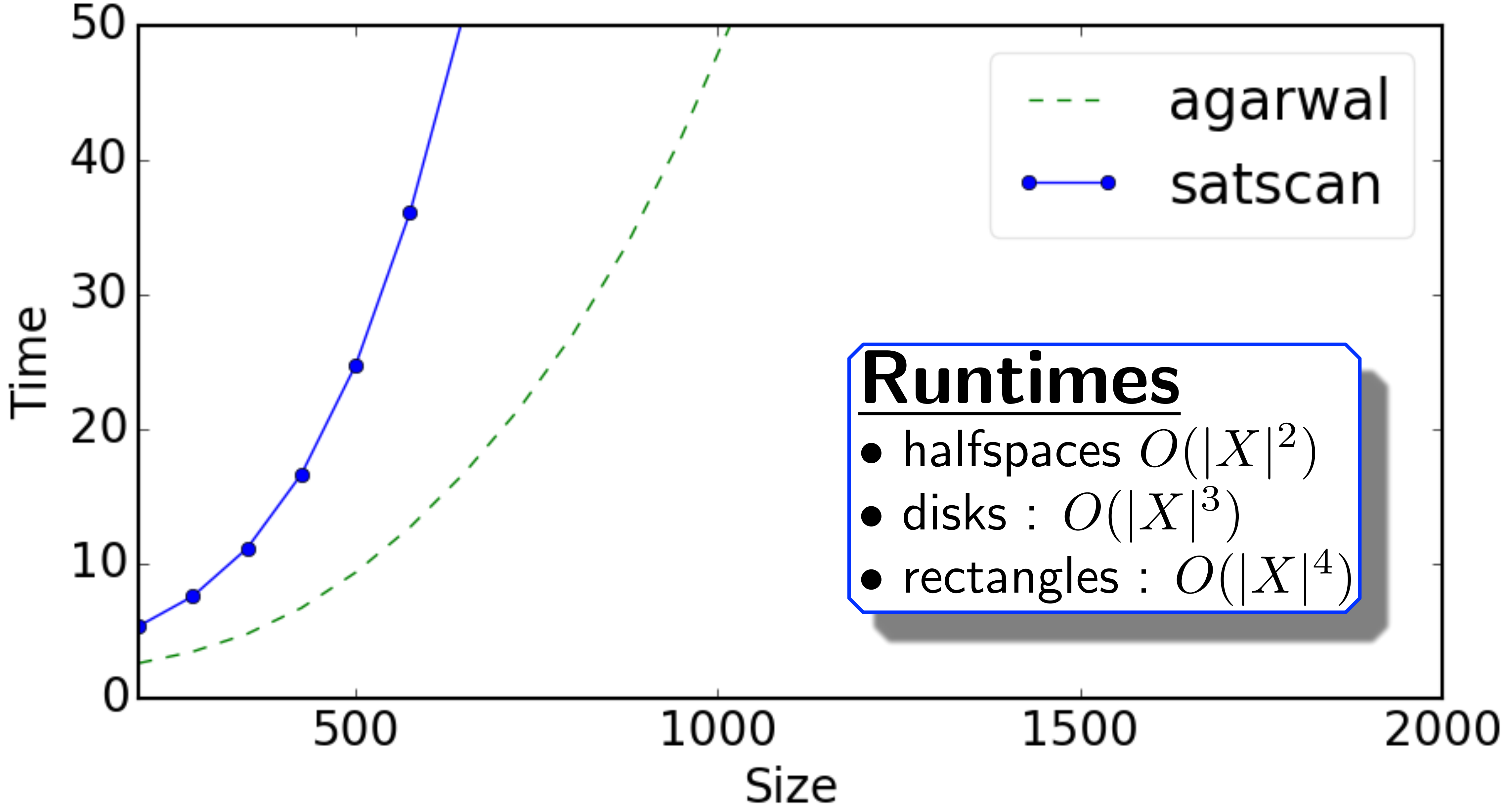
(recovery) : any $A \in (N, \mathcal{A})$ then $\psi_{\mathcal{A}}(A) \cap N = A$
(inclusion) : any $Y \subset X$ then $\psi_{\mathcal{A}}(Y) \cap X \in (X, \mathcal{A})$

- Not all $\phi_{\mathcal{A}}$ induced by small enclosing shapes are conforming.

  $\phi_{\mathcal{T}}$: the smallest enclosing triangle

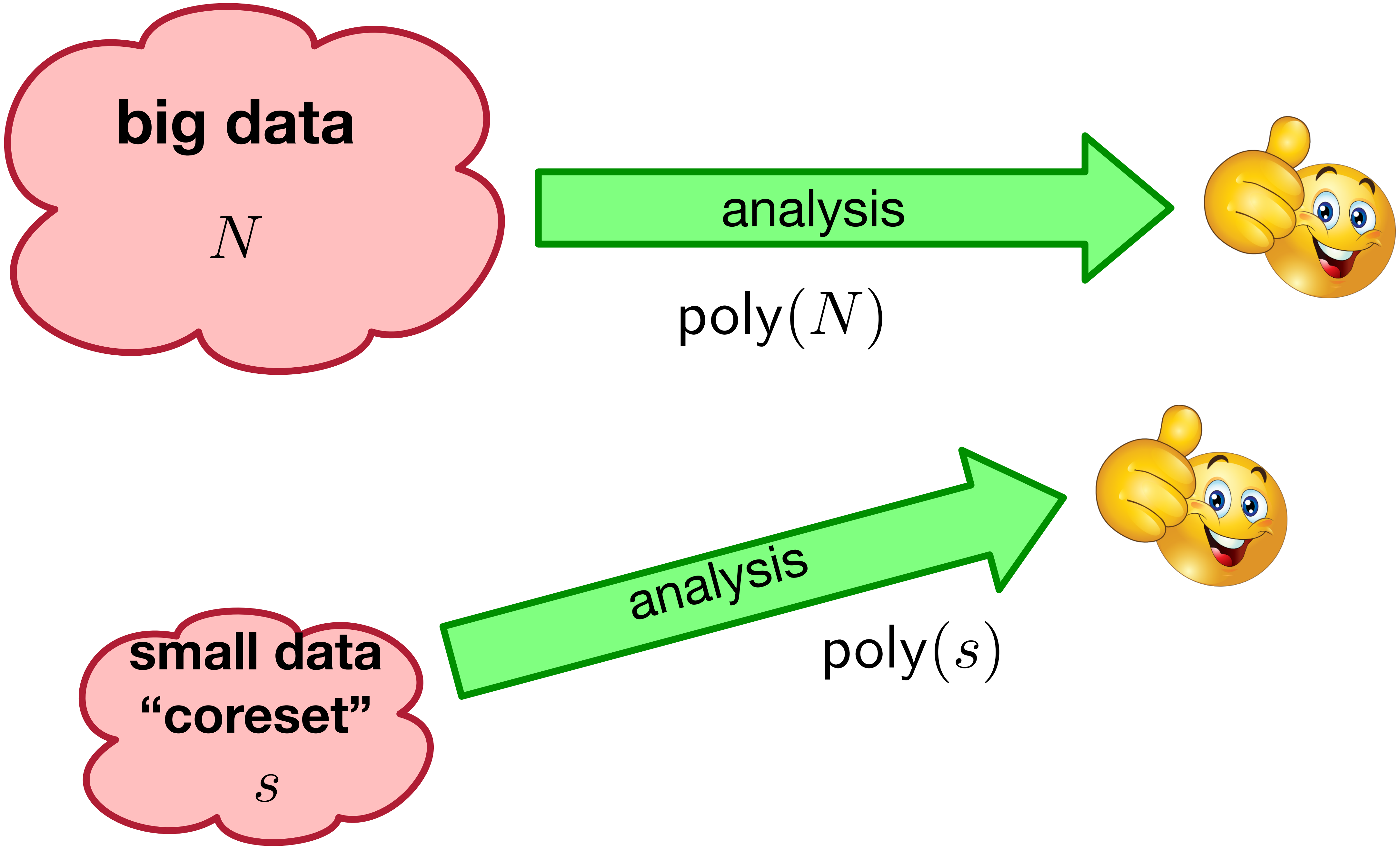


SatScan is not Scalable



Runtimes

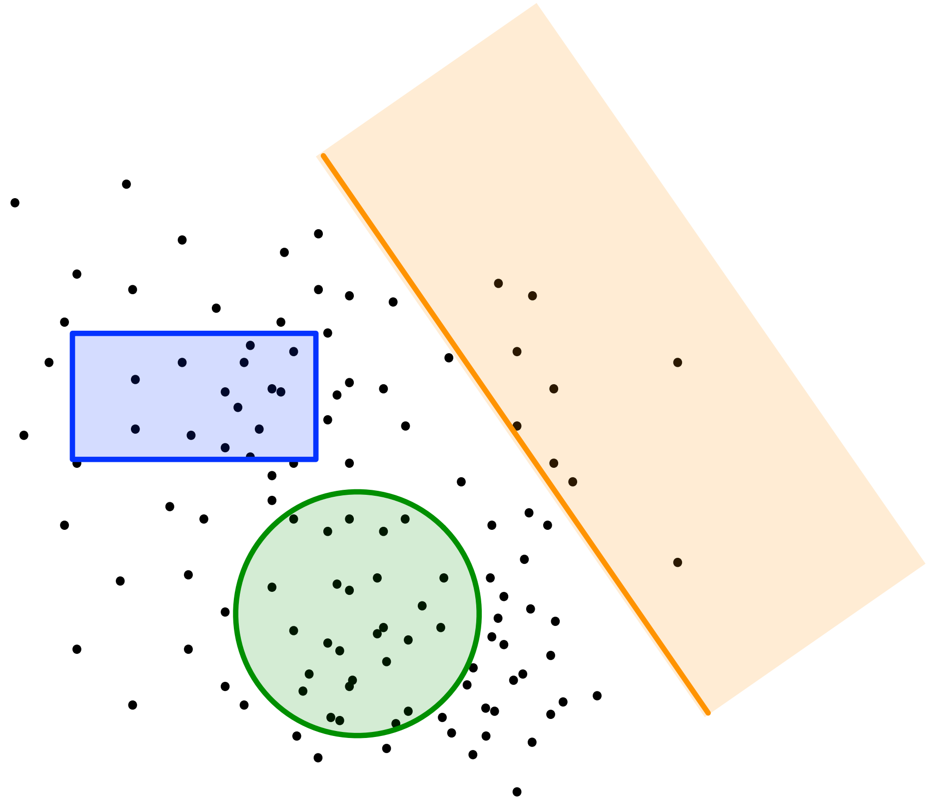
- halfspaces $O(|X|^2)$
- disks : $O(|X|^3)$
- rectangles : $O(|X|^4)$


big data
N
analysis
poly(N)
small data
“coreset”
s
analysis
poly(s)

ε-Samples and ε-Nets

- A *range space* is a pair $(X, \mathcal{C})$, where $\mathcal{C}$ is a set of subsets of set X.

- The *VC-dimension* of $(X, \mathcal{C})$ is the maximum size $Y \subset X$ so
  all subset $Z \subset Y$ are elements of $(Y, \mathcal{C})$.
  $\Rightarrow$ in $\mathbb{R}^2$: disks $\nu = 3$; halfspaces $\nu = 3$; rectangles $\nu = 4$.



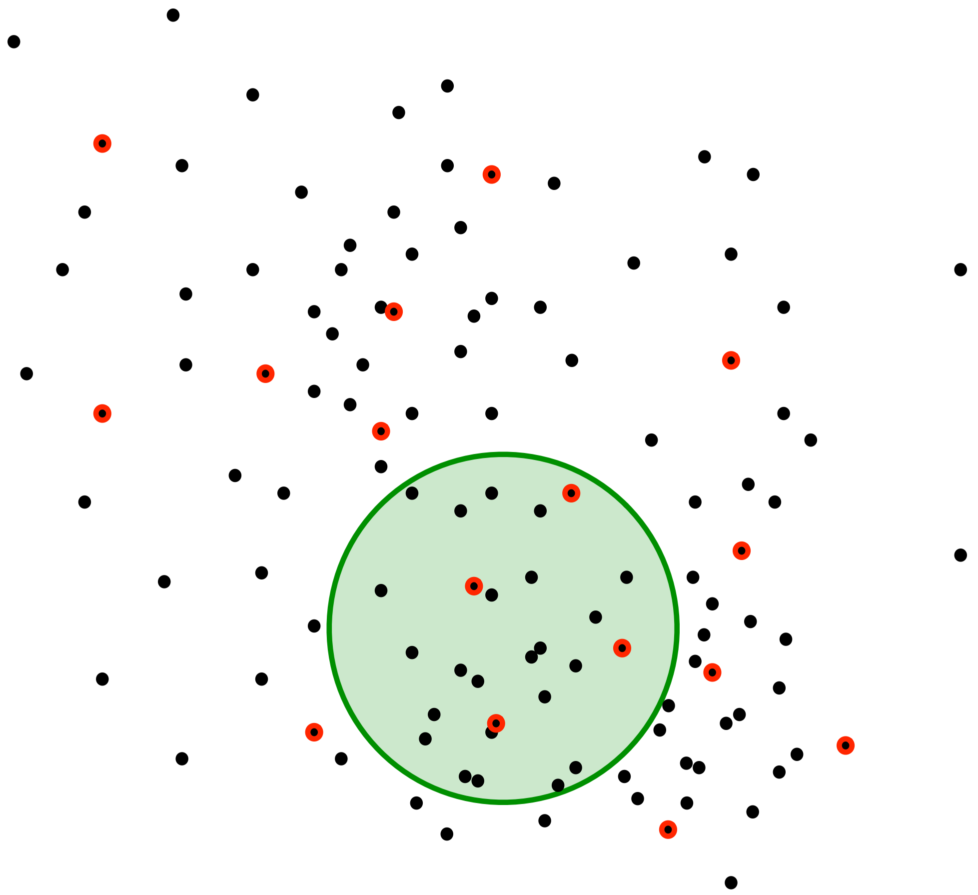
ε-Samples and ε-Nets

- A *range space* is a pair $(X, \mathcal{C})$, where $\mathcal{C}$ is a set of subsets of set X.

- The *VC-dimension* of $(X, \mathcal{C})$ is the maximum size $Y \subset X$ so
  all subset $Z \subset Y$ are elements of $(Y, \mathcal{C})$.
  $\Rightarrow$ in $\mathbb{R}^2$: disks $\nu = 3$; halfspaces $\nu = 3$; rectangles $\nu = 4$.

An ε-**sample** $S \subset X$ maintains density of $(X, \mathcal{C})$ so

for all $C \in \mathcal{C}$ $\quad \left| \frac{|C \cap X|}{|X|} - \frac{|C \cap S|}{|S|} \right| \leq \varepsilon.$



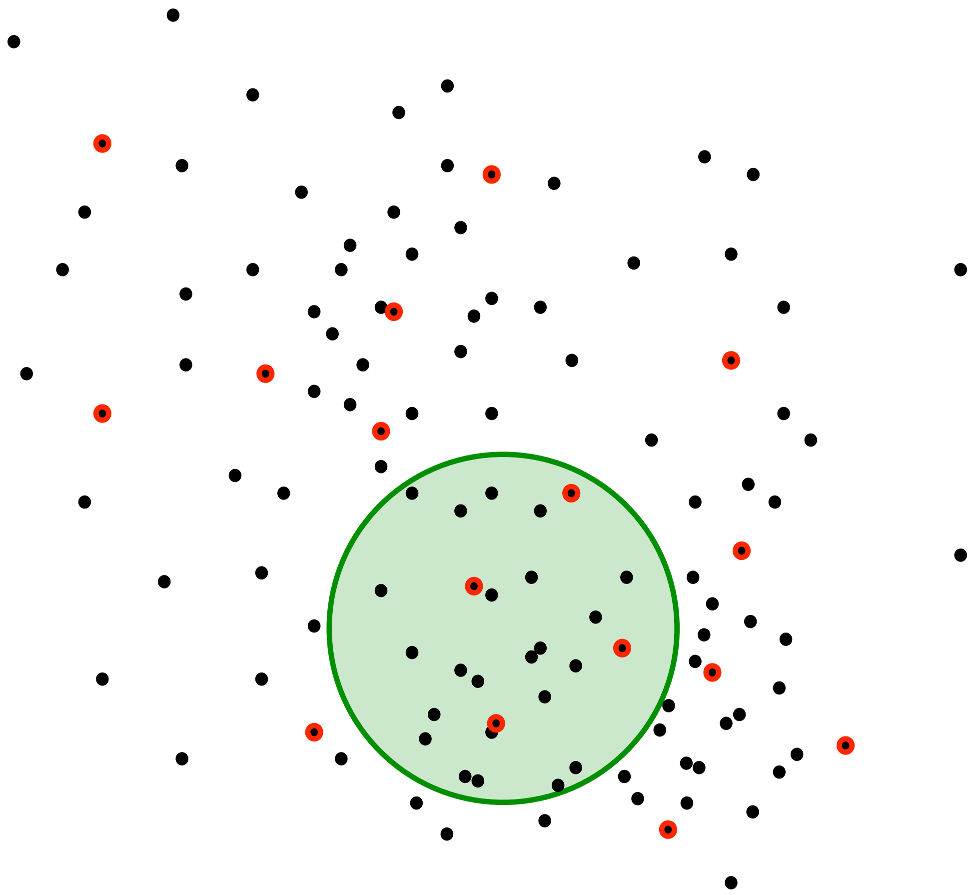
ε-Samples and ε-Nets

- A *range space* is a pair $(X, \mathcal{C})$, where $\mathcal{C}$ is a set of subsets of set X.
- The *VC-dimension* of $(X, \mathcal{C})$ is the maximum size $Y \subset X$ so
  all subset $Z \subset Y$ are elements of $(Y, \mathcal{C})$.
  $\Rightarrow$ in $\mathbb{R}^2$: disks $\nu = 3$; halfspaces $\nu = 3$; rectangles $\nu = 4$.

An ε-**sample** $S \subset X$ maintains density of $(X, \mathcal{C})$ so
for all $C \in \mathcal{C}$ $\quad \left| \frac{|C \cap X|}{|X|} - \frac{|C \cap S|}{|S|} \right| \leq \varepsilon.$

$$\frac{|C \cap X|}{|X|} = \frac{22}{119} = 0.227$$

$$\frac{|C \cap S|}{|S|} = \frac{4}{16} = 0.25$$

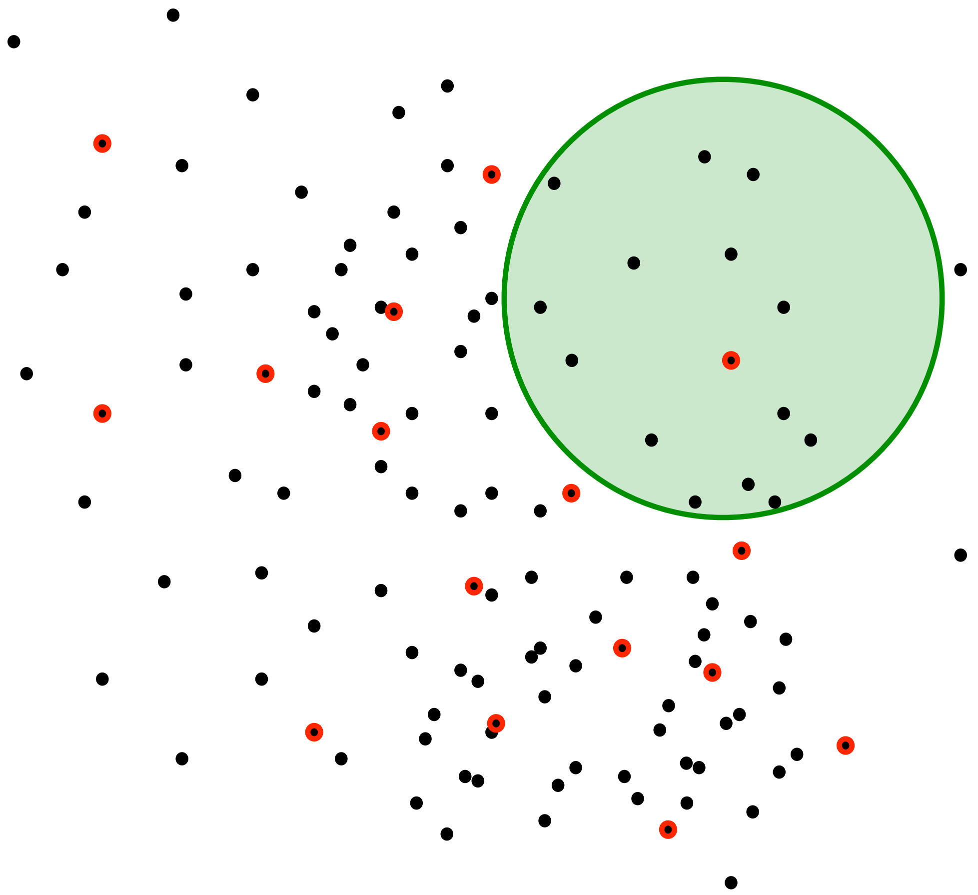


ε-Samples and ε-Nets

- A *range space* is a pair $(X, \mathcal{C})$, where $\mathcal{C}$ is a set of subsets of set X.
- The *VC-dimension* of $(X, \mathcal{C})$ is the maximum size $Y \subset X$ so
  all subset $Z \subset Y$ are elements of $(Y, \mathcal{C})$.
  $\Rightarrow$ in $\mathbb{R}^2$: disks $\nu = 3$; halfspaces $\nu = 3$; rectangles $\nu = 4$.

An ε-**net** $S \subset X$ hits every large enough subset $(X, \mathcal{C})$ so

for all $C \in \mathcal{C}$ with $\dfrac{|C \cap X|}{|X|} \geq \varepsilon$, then $C \cap S \neq \emptyset$.

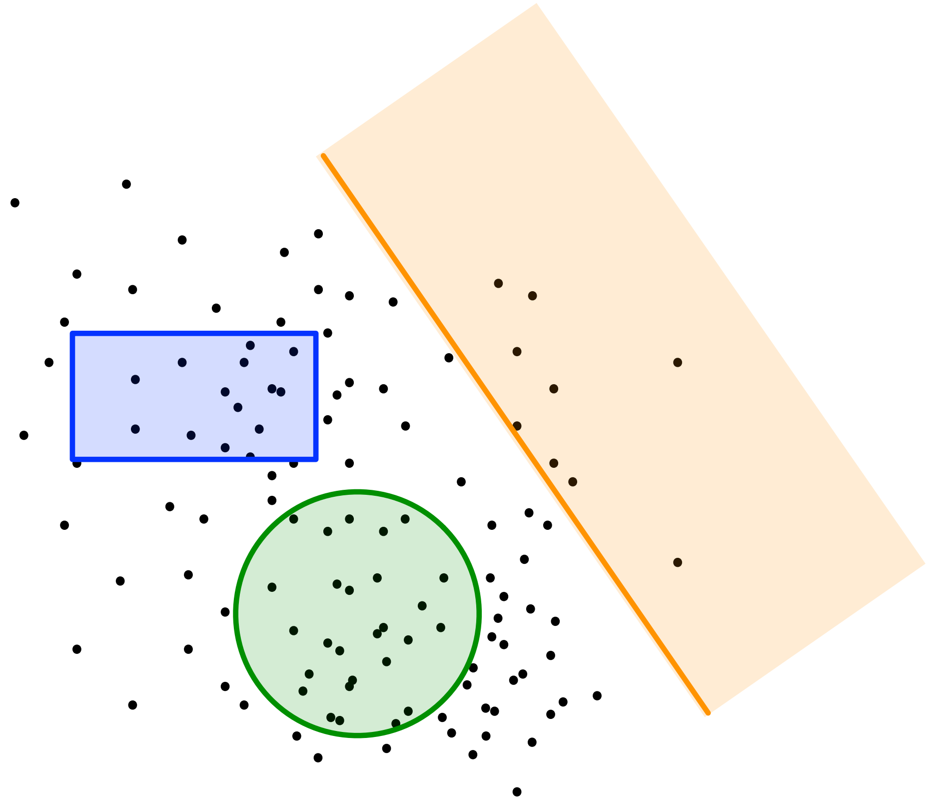


ε-Samples and ε-Nets

- A *range space* is a pair $(X, \mathcal{C})$, where $\mathcal{C}$ is a set of subsets of set X.

- The *VC-dimension* of $(X, \mathcal{C})$ is the maximum size $Y \subset X$ so
  all subset $Z \subset Y$ are elements of $(Y, \mathcal{C})$.
  $\Rightarrow$ in $\mathbb{R}^2$: disks $\nu = 3$; halfspaces $\nu = 3$; rectangles $\nu = 4$.

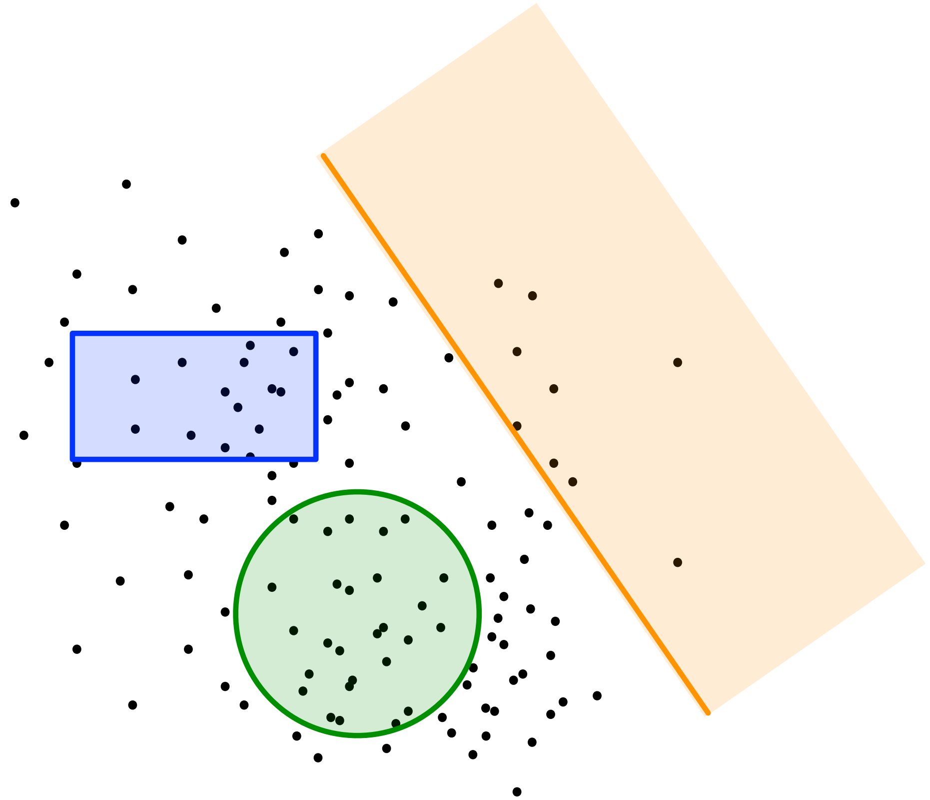
- A random sample $S \subset X$ of size k, with probability at least $1 - \delta$, is a
  $\Rightarrow$ ε-sample for $k = \Omega(\frac{1}{\varepsilon^2}(\nu + \log \frac{1}{\delta})) \approx \frac{1}{\varepsilon^2}$
  $\Rightarrow$ ε-net for $k = \Omega(\frac{\nu}{\varepsilon} \log \frac{1}{\varepsilon\delta}) \approx \frac{1}{\varepsilon} \log \frac{1}{\varepsilon}$



ε-Samples and ε-Nets

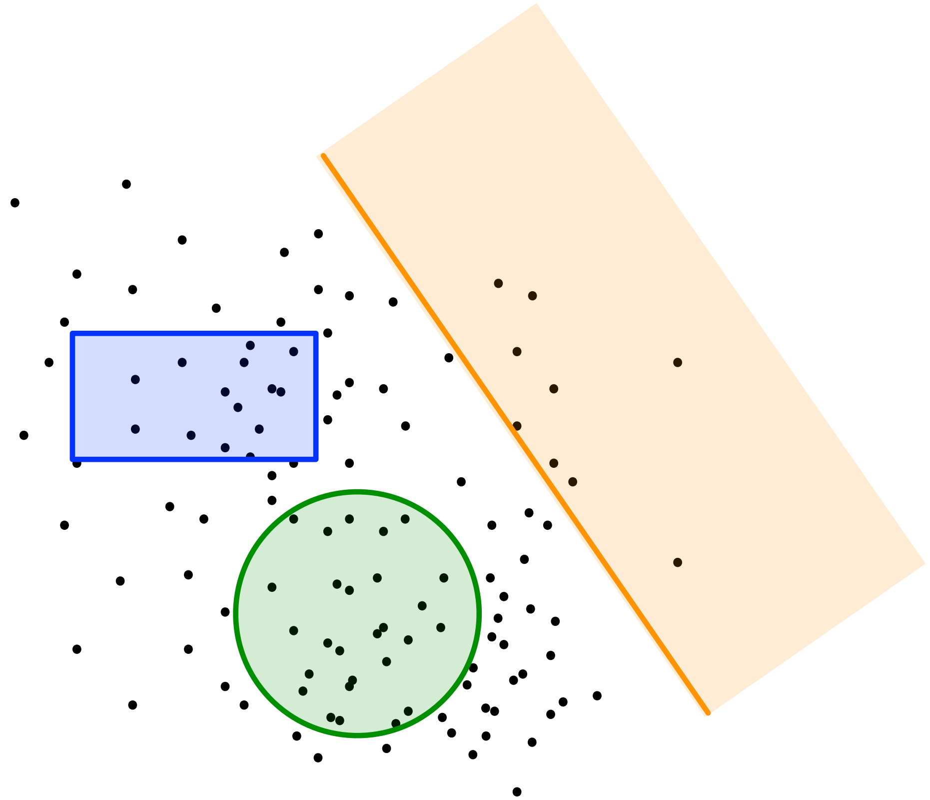
- A *range space* is a pair $(X, \mathcal{C})$, where $\mathcal{C}$ is a set of subsets of set X.
- The *VC-dimension* of $(X, \mathcal{C})$ is the maximum size $Y \subset X$ so all subset $Z \subset Y$ are elements of $(Y, \mathcal{C})$.
  $\Rightarrow$ in $\mathbb{R}^2$: disks $\nu = 3$; halfspaces $\nu = 3$; rectangles $\nu = 4$.
- A random sample $S \subset X$ of size k, with probability at least $1 - \delta$, is a
  $\Rightarrow$ ε-sample for $k = \Omega(\frac{1}{\varepsilon^2}(\nu + \log \frac{1}{\delta})) \approx \frac{1}{\varepsilon^2}$
  $\Rightarrow$ ε-net for $k = \Omega(\frac{\nu}{\varepsilon} \log \frac{1}{\varepsilon\delta}) \approx \frac{1}{\varepsilon} \log \frac{1}{\varepsilon}$

Idea: Sample-then-scan!



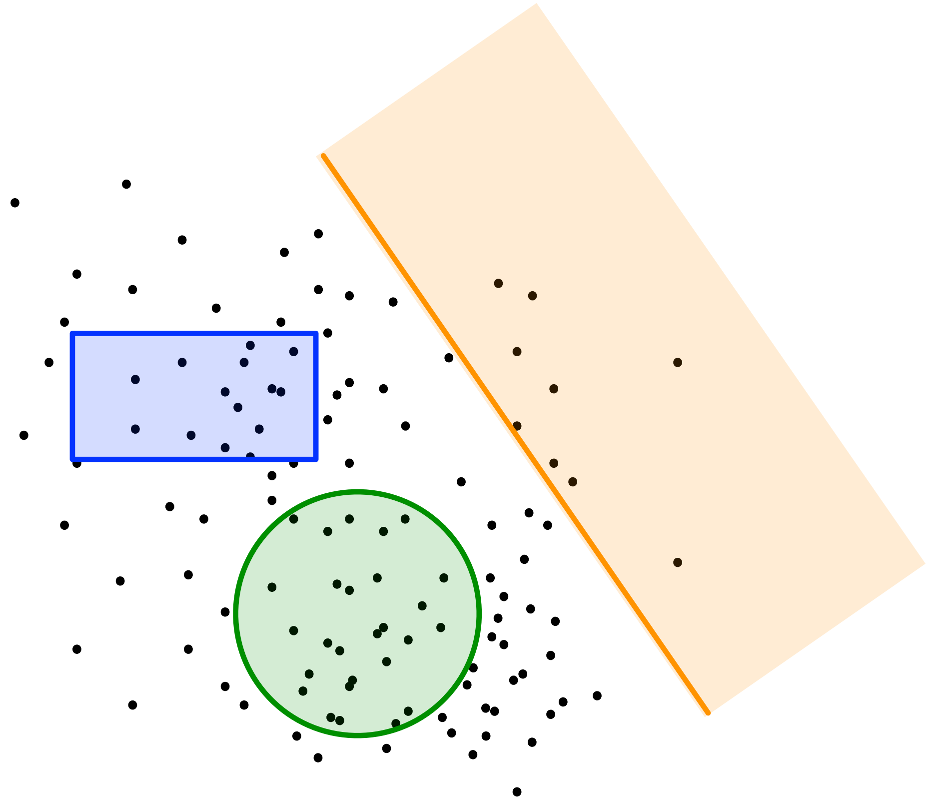
Sample then Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$.
  (do same for $R \to S_R$ independently)



Sample then Scan

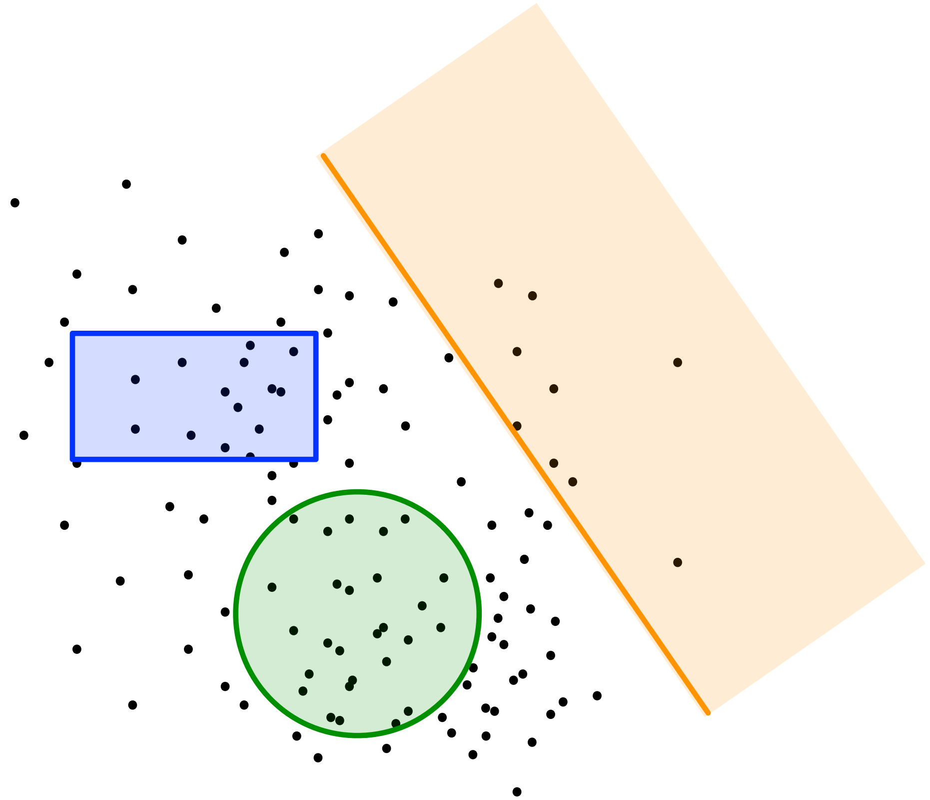
- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$.
  (do same for $R \to S_R$ independently)
  Now for all $C \in \mathcal{C}$ we have $|b(C) - b_S(C)| \leq \varepsilon \Rightarrow |\Phi(C) - \Phi_S(C)| \leq O(\varepsilon)$.



Sample then Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$.
  (do same for $R \to S_R$ independently)
  Now for all $C \in \mathcal{C}$ we have $|b(C) - b_S(C)| \leq \varepsilon \Rightarrow |\Phi(C) - \Phi_S(C)| \leq O(\varepsilon)$.

- Enumerate all s^ν ranges $C \in (S, \mathcal{C})$
  for each evaluate $\Phi(C)$ in $O(s)$ time.
  Total runtime $O(s^{\nu+1}) = 1/\varepsilon^{2\nu+2}$ time.

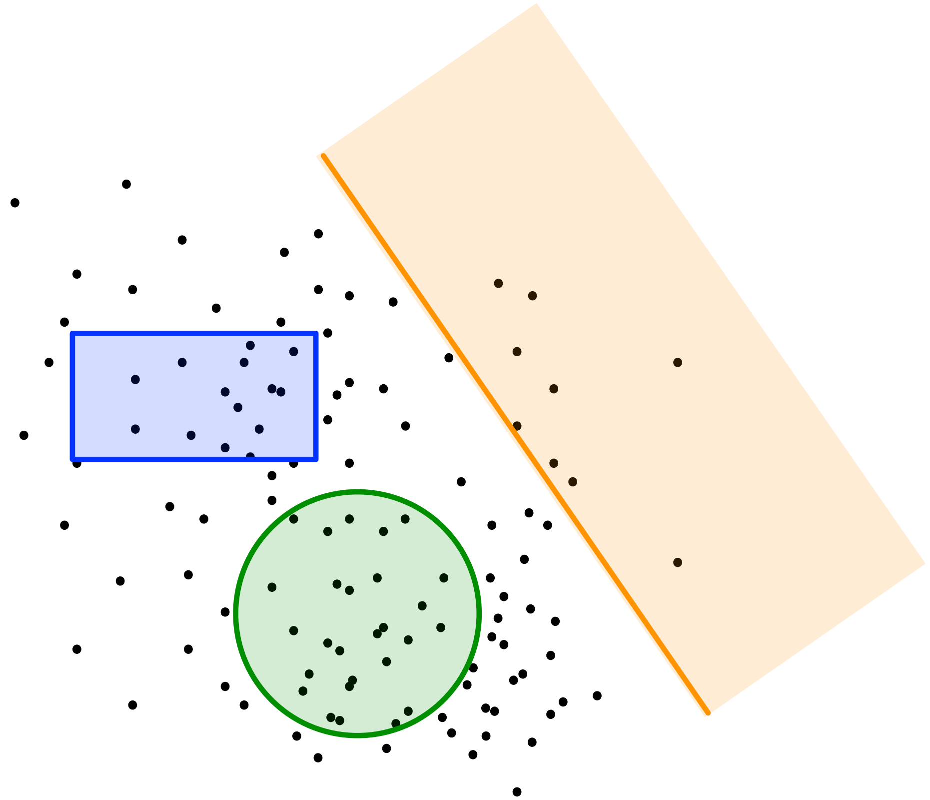


Sample then Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$.
  (do same for $R \to S_R$ independently)
  Now for all $C \in \mathcal{C}$ we have $|b(C) - b_S(C)| \leq \varepsilon \Rightarrow |\Phi(C) - \Phi_S(C)| \leq O(\varepsilon)$.

- Enumerate all s^ν ranges $C \in (S, \mathcal{C})$
  for each evaluate $\Phi(C)$ in $O(s)$ time.
  Total runtime $O(s^{\nu+1}) = 1/\varepsilon^{2\nu+2}$ time.

- Special cases have faster runtime:
  disks $O(s^3) = 1/\varepsilon^6$
  halfspaces $O(s^2) = 1/\varepsilon^4$
  rectangles $O(\frac{1}{\sqrt{\varepsilon}} s^2 \log s) = (1/\varepsilon^{4.5}) \log \frac{1}{\varepsilon}$



Sample then Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$.
  (do same for $R \to S_R$ independently)
  Now for all $C \in \mathcal{C}$ we have $|b(C) - b_S(C)| \leq \varepsilon \Rightarrow |\Phi(C) - \Phi_S(C)| \leq O(\varepsilon)$.

- Enumerate all s^ν ranges $C \in (S, \mathcal{C})$
  for each evaluate $\Phi(C)$ in $O(s)$ time.
  Total runtime $O(s^{\nu+1}) = 1/\varepsilon^{2\nu+2}$ time.

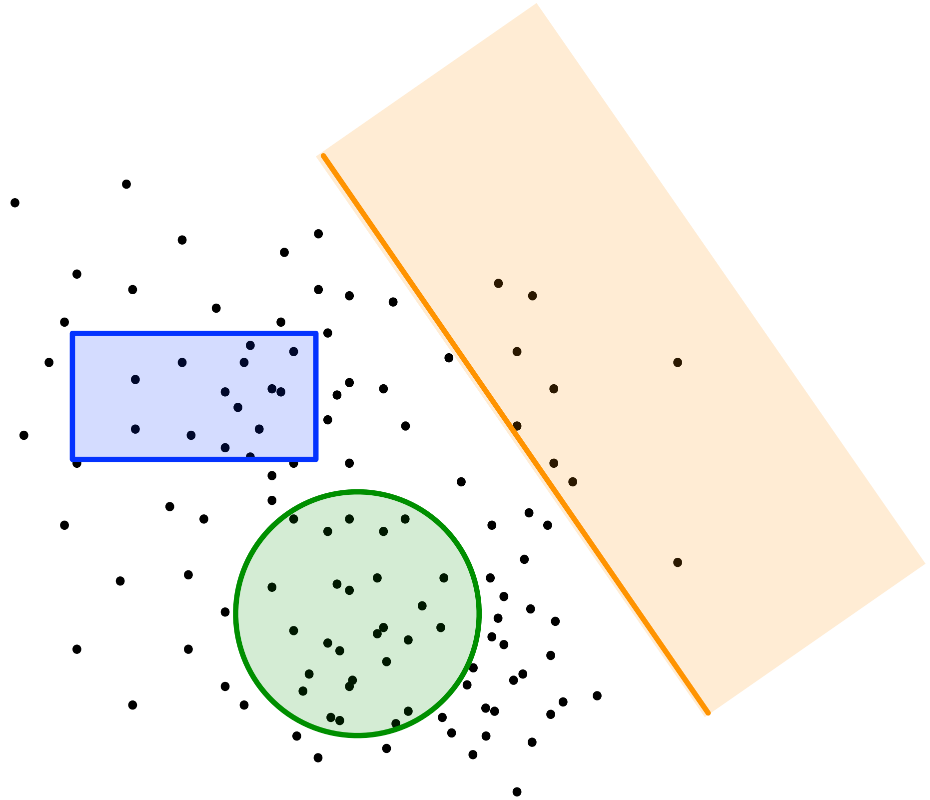
- Special cases have faster runtime:
  disks $O(s^3) = 1/\varepsilon^6$
  halfspaces $O(s^2) = 1/\varepsilon^4$
  rectangles $O(\frac{1}{\sqrt{\varepsilon}} s^2 \log s) = (1/\varepsilon^{4.5}) \log \frac{1}{\varepsilon}$

Setting $\varepsilon = \frac{1}{100} = 0.01$
$\to$ requires $s \approx 10{,}000$
$\to$ so $s^2 \approx 100$million
Still too slow!

Two-Level Sample-then-Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  1. create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$.
  2. create an ε-net $N \subset X$; $|N| = n = (1/\varepsilon)\log(1/\varepsilon)$.

  (do same for $R \to S_R$, $R \to N_R$ independently)

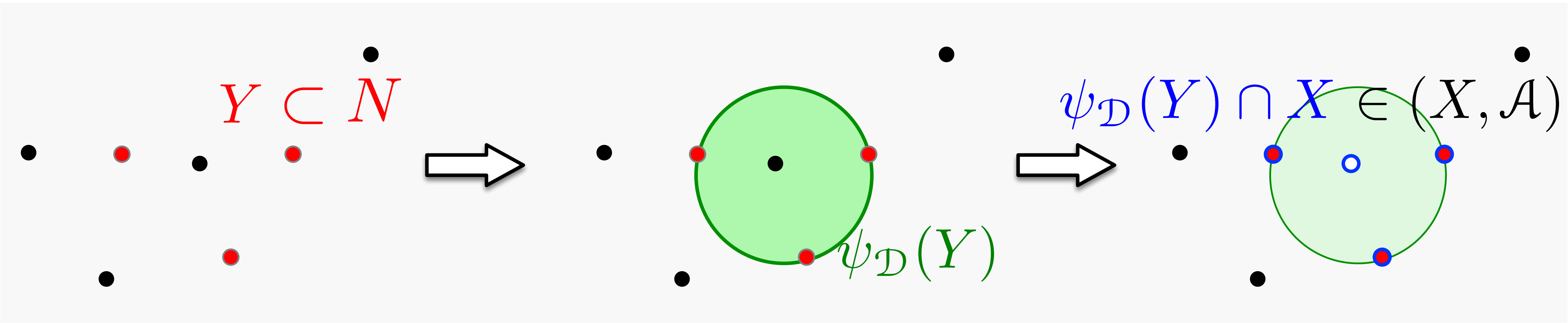


Two-Level Sample-then-Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  1. create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$.
  2. create an ε-net $N \subset X$; $|N| = n = (1/\varepsilon)\log(1/\varepsilon)$.

  (do same for $R \to S_R$, $R \to N_R$ independently)

- For induced range subsets $\mathcal{C}_{|N} = \{C \cap N \mid C \in \mathcal{C}\}$

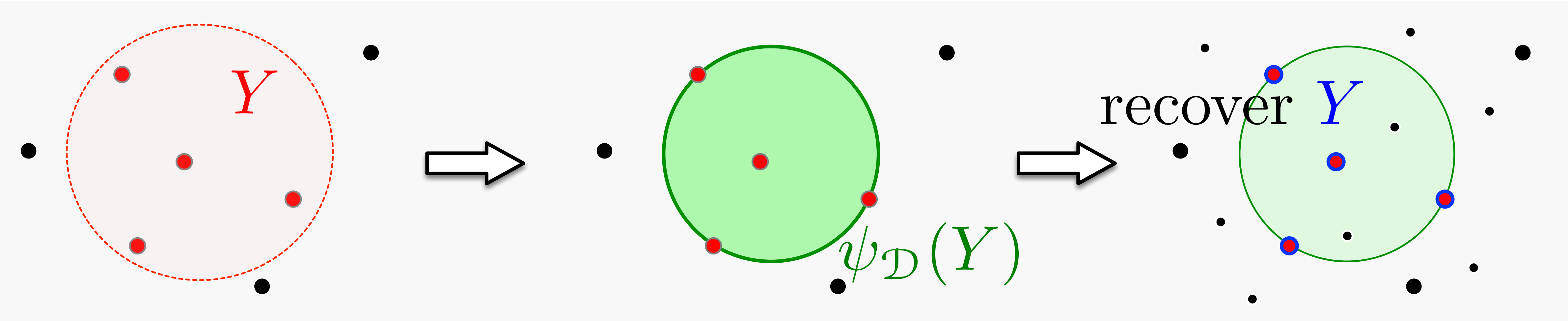


Two-Level Sample-then-Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  1. create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$.
  2. create an ε-net $N \subset X$; $|N| = n = (1/\varepsilon)\log(1/\varepsilon)$.

  (do same for $R \to S_R$, $R \to N_R$ independently)

- For induced range subsets $\mathcal{C}_{|N} = \{C \cap N \mid C \in \mathcal{C}\}$
  define reduced range space $(X, \mathcal{C}_{\Delta N}) = \{X \cap \psi_{\mathcal{C}}(Y) \mid Y \in (N, \mathcal{C}_{|N})\}$.

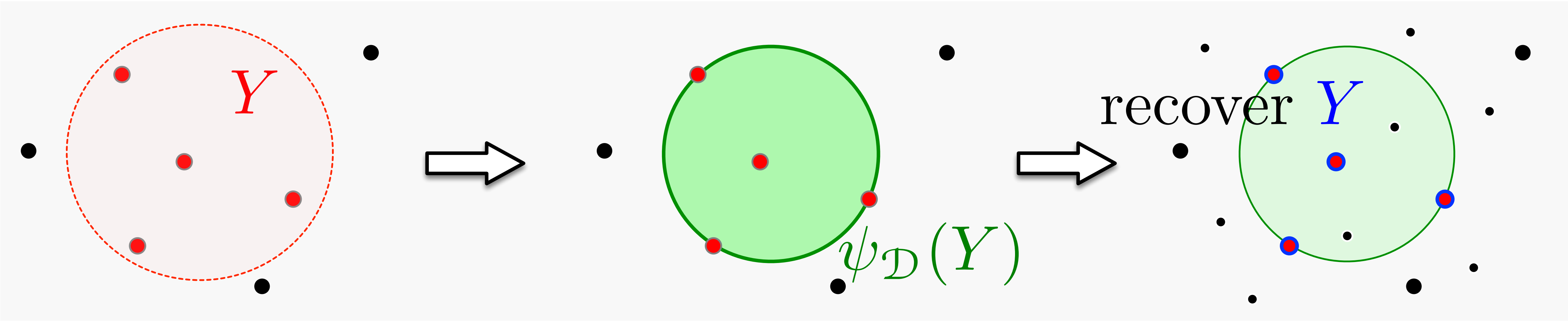


Two-Level Sample-then-Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  1. create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$.
  2. create an ε-net $N \subset X$; $|N| = n = (1/\varepsilon)\log(1/\varepsilon)$.

  (do same for $R \to S_R$, $R \to N_R$ independently)

- For induced range subsets $\mathcal{C}_{|N} = \{C \cap N \mid C \in \mathcal{C}\}$
  define reduced range space $(X, \mathcal{C}_{\Delta N}) = \{X \cap \psi_{\mathcal{C}}(Y) \mid Y \in (N, \mathcal{C}_{|N})\}$.
  Now for all $C \in (X, \mathcal{C})$ there exists $C' \in (S, \mathcal{C}_{\Delta N})$ so
  $|r(C) - r_S(C')| \leq \varepsilon \Rightarrow |\Phi(C) - \Phi_S(C')| \leq \varepsilon$.



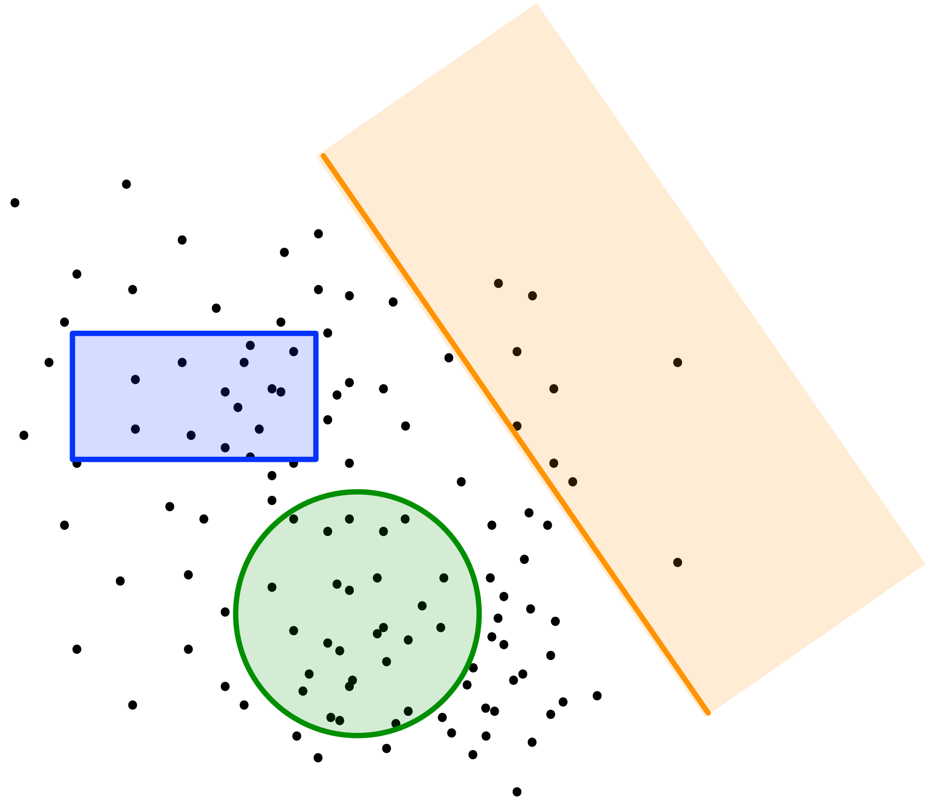
Two-Level Sample-then-Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  1. create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$.
  2. create an ε-net $N \subset X$; $|N| = n = (1/\varepsilon)\log(1/\varepsilon)$.

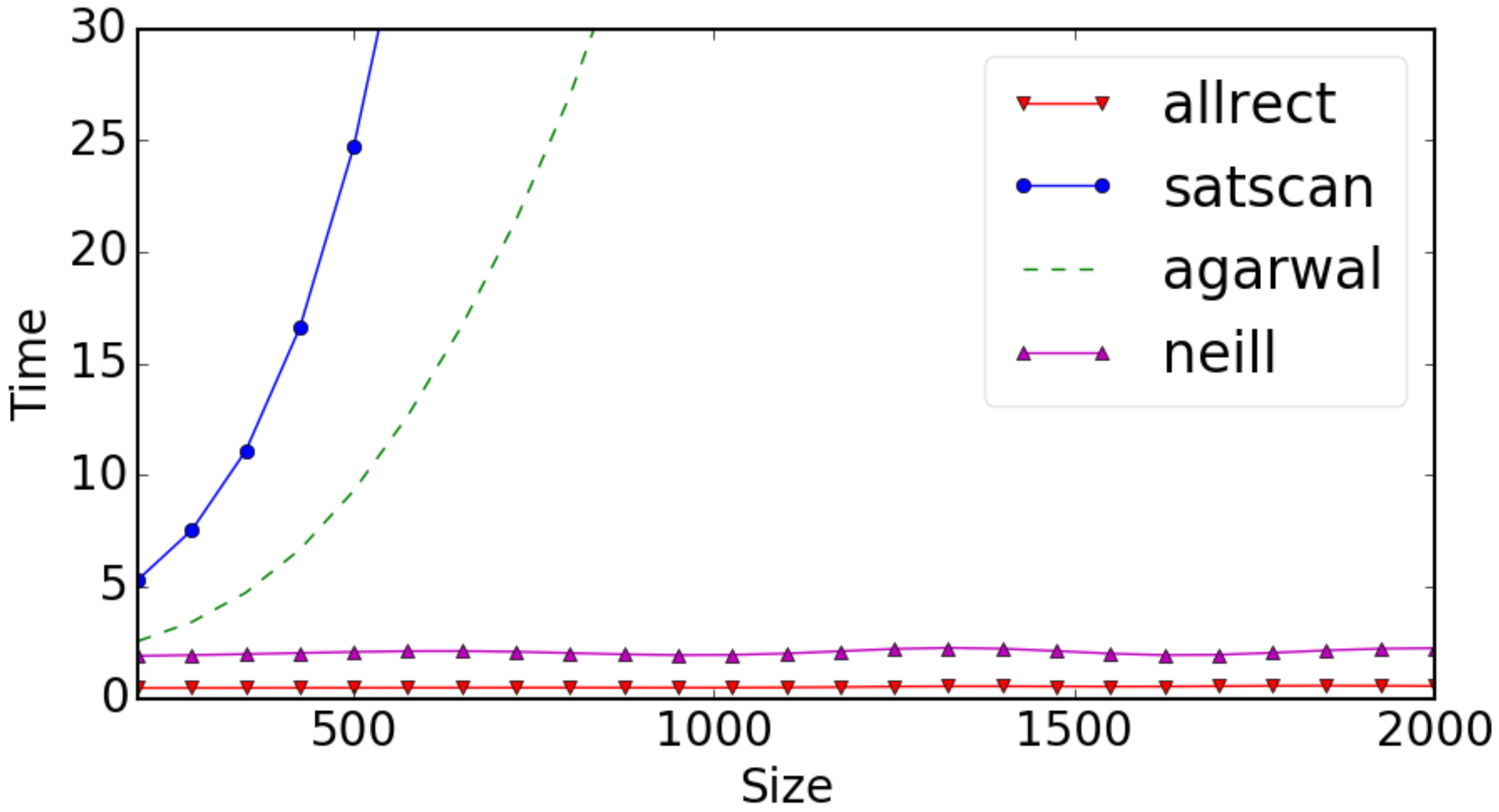
  (do same for $R \to S_R$, $R \to N_R$ independently)

- For induced range subsets $\mathcal{C}_{|N} = \{C \cap N \mid C \in \mathcal{C}\}$
  define reduced range space $(X, \mathcal{C}_{\Delta N}) = \{X \cap \psi_{\mathcal{C}}(Y) \mid Y \in (N, \mathcal{C}_{|N})\}$.
  Now for all $C \in (X, \mathcal{C})$ there exists $C' \in (S, \mathcal{C}_{\Delta N})$ so
  $|r(C) - r_S(C')| \leq \varepsilon \Rightarrow |\Phi(C) - \Phi_S(C')| \leq \varepsilon$.

- Enumerate all n^ν ranges $C' \in (S, \mathcal{C}_{\Delta N})$
  for each evaluate $\Phi_S(C')$ in s time.
  Total runtime $n^\nu s = (1/\varepsilon^{\nu+2})\log^\nu \frac{1}{\varepsilon}$ time.
  $\to$ disks $O(sn^2) = (1/\varepsilon^4)\log^2 \frac{1}{\varepsilon}$
  $\to$ halfspaces $O(sn) = (1/\varepsilon^3)\log \frac{1}{\varepsilon}$
  $\to$ rectangles $O(n^4 + s\log n) = (1/\varepsilon^4)\log^4(1/\varepsilon)$



Scalable Scanning and with Guarantees

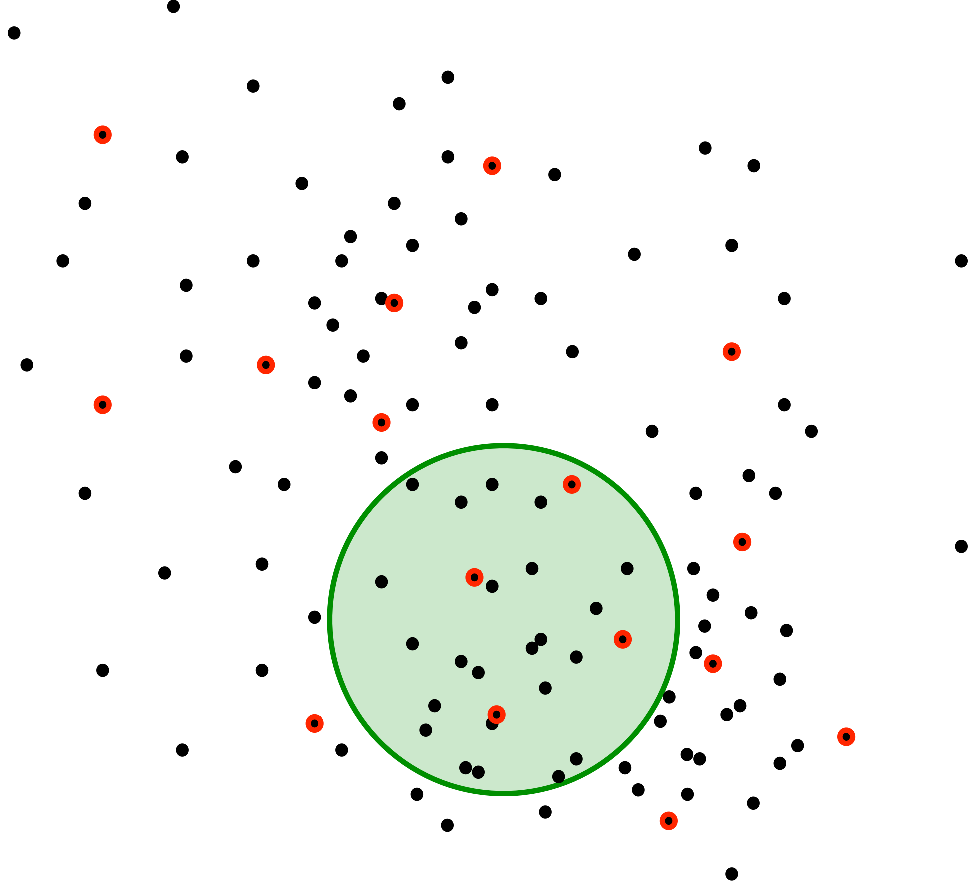


Statistical Power

Is an ε-approx $\hat{C} \in \mathcal{C}$ acceptable?

Statistical Power

Is an ε-approx $\hat{C} \in \mathcal{C}$ acceptable?



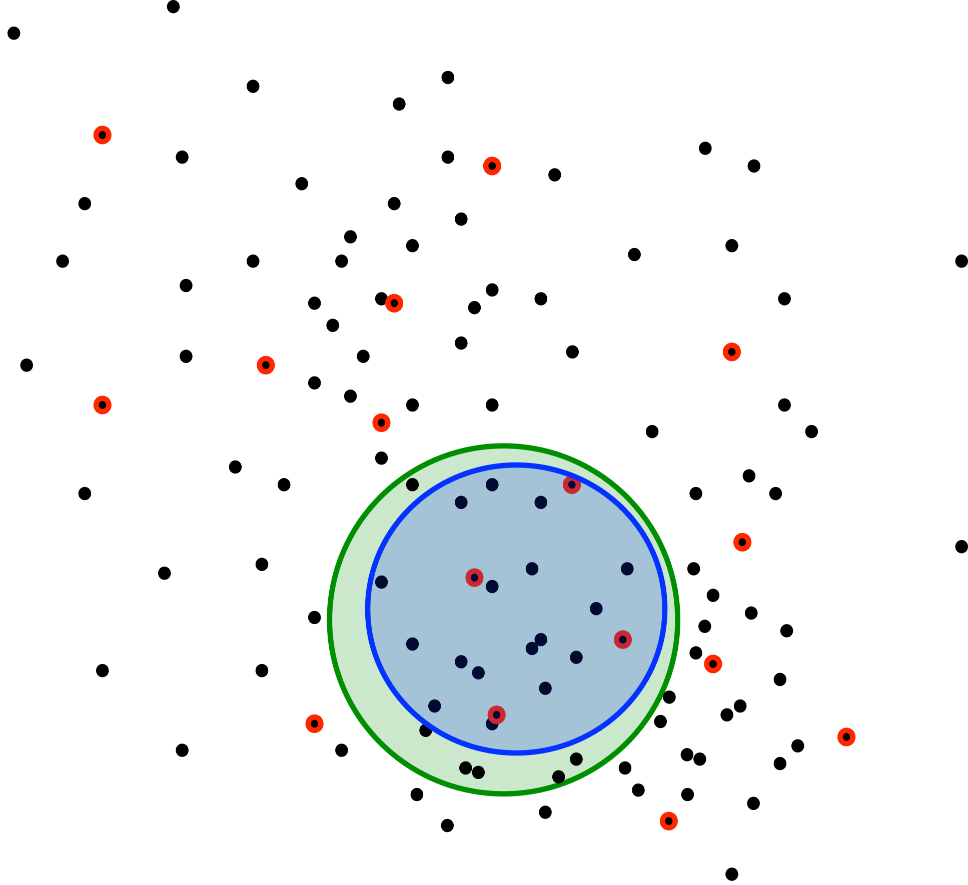
- Plant high-discrepancy region.
- Run algorithm to see if you find it.

Repeat many times,

what fraction find planted region?

Statistical Power

Is an ε-approx $\hat{C} \in \mathcal{C}$ acceptable?



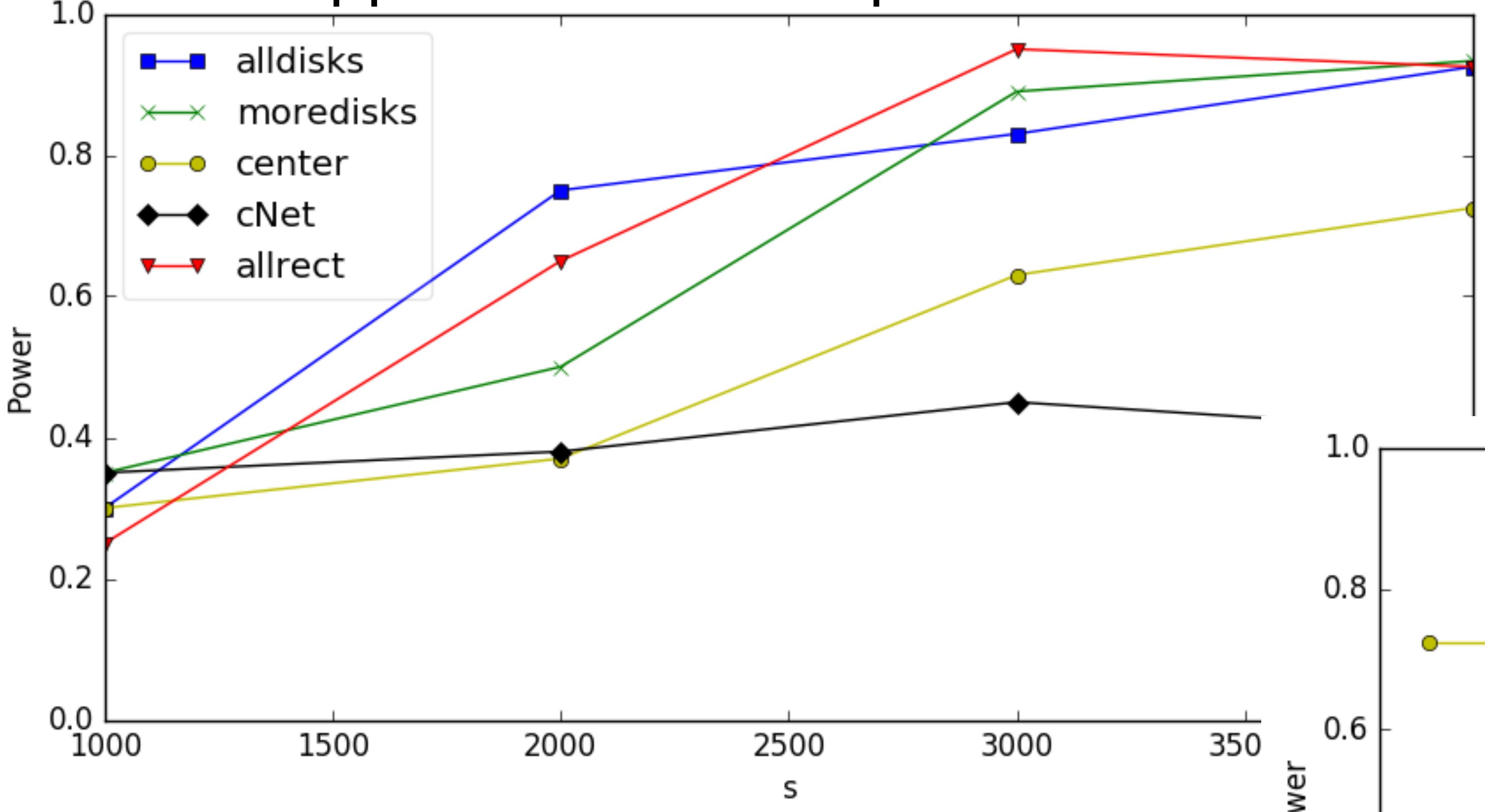
- Plant high-discrepancy region.
- Run algorithm to see if you find it.

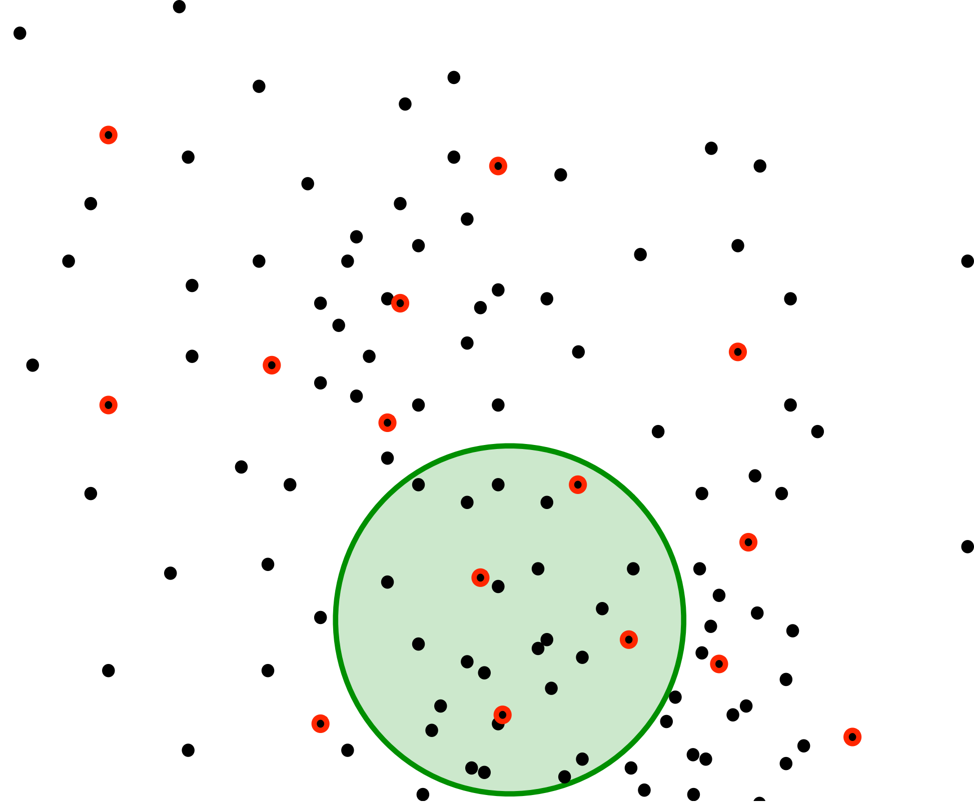
Repeat many times,

what fraction find planted region?

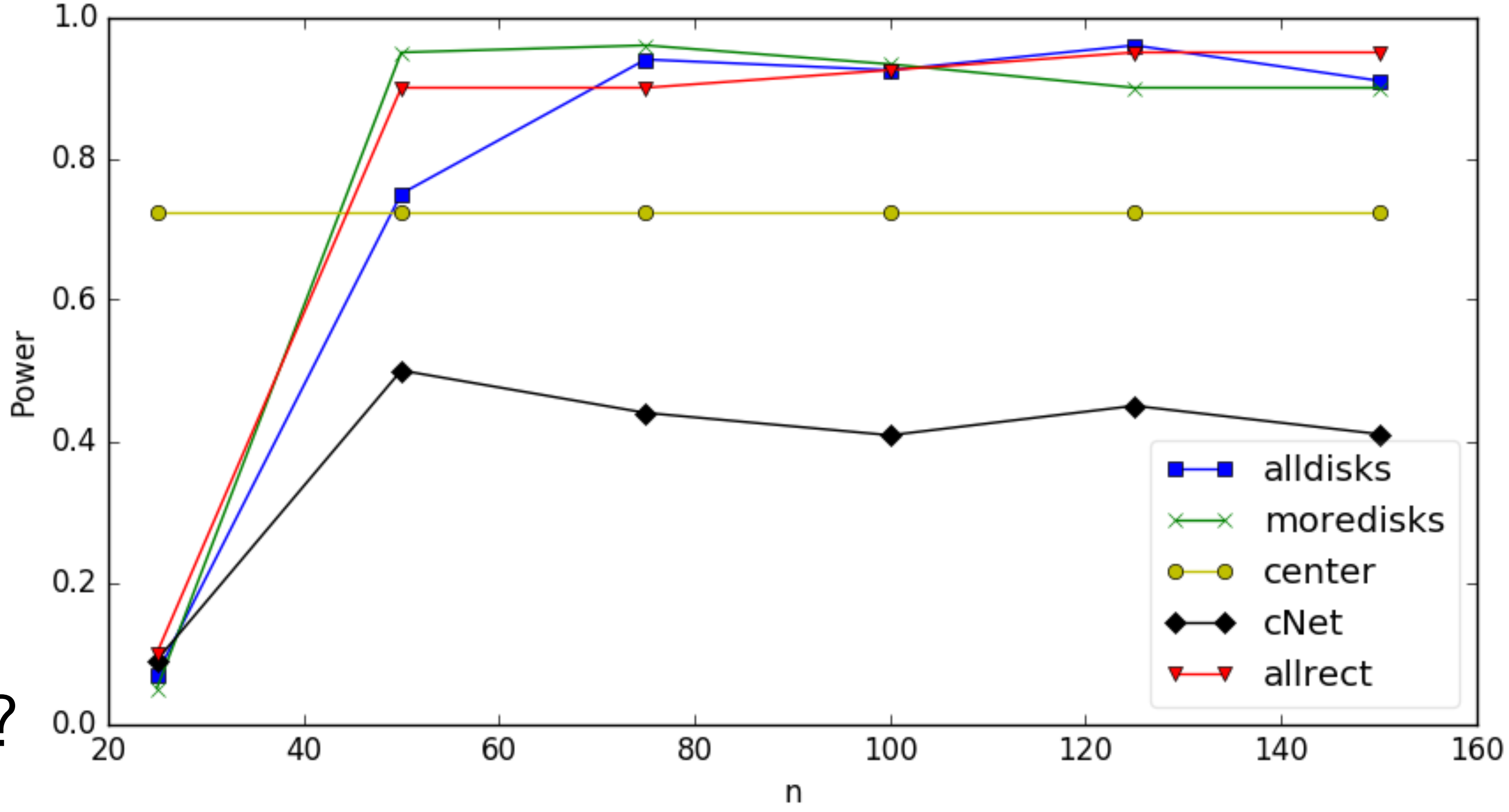
Statistical Power

Is an ε-approx $\hat{C} \in \mathcal{C}$ acceptable?





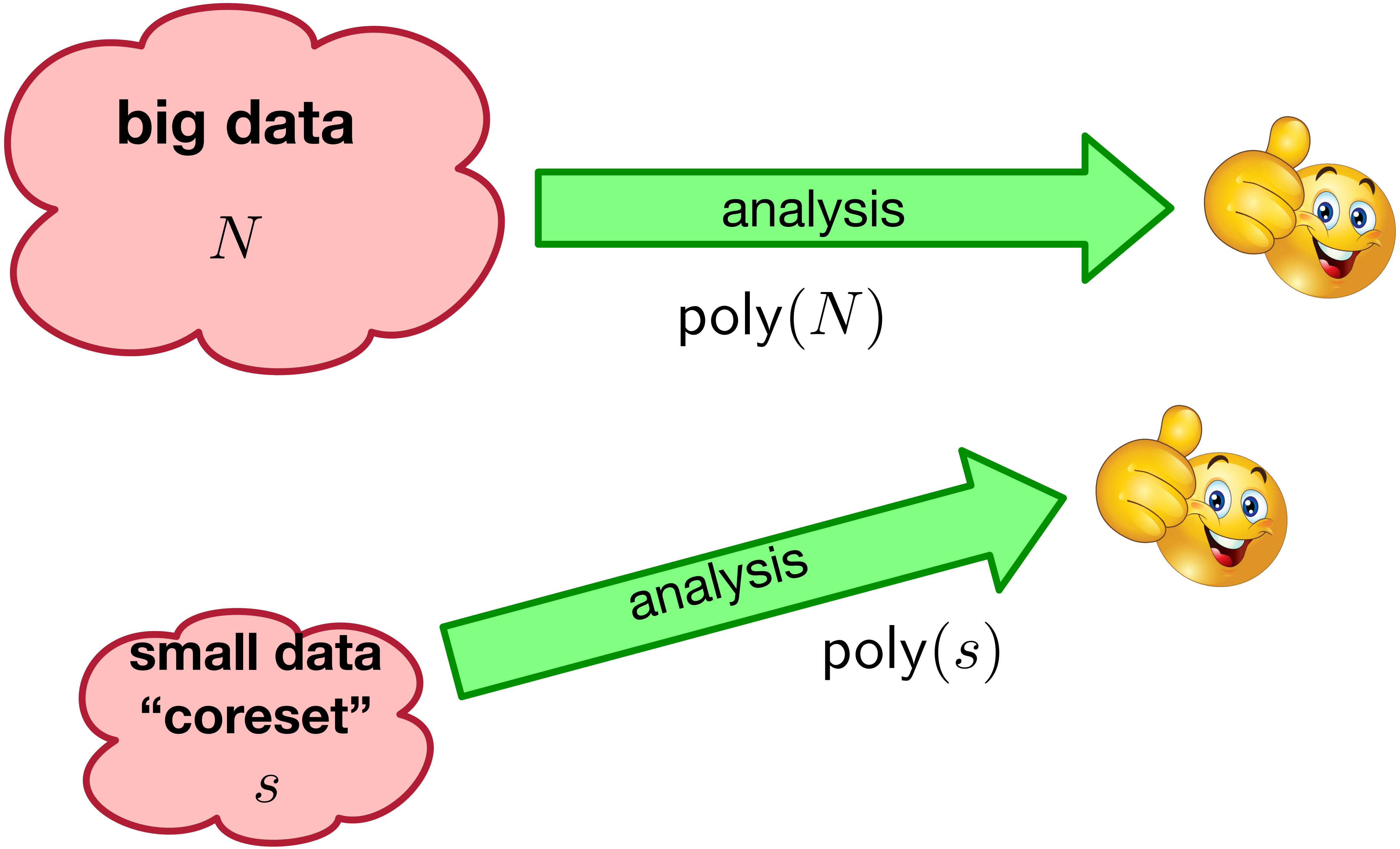
- Plant high-discrepancy region.
- Run algorithm to see if you find it.

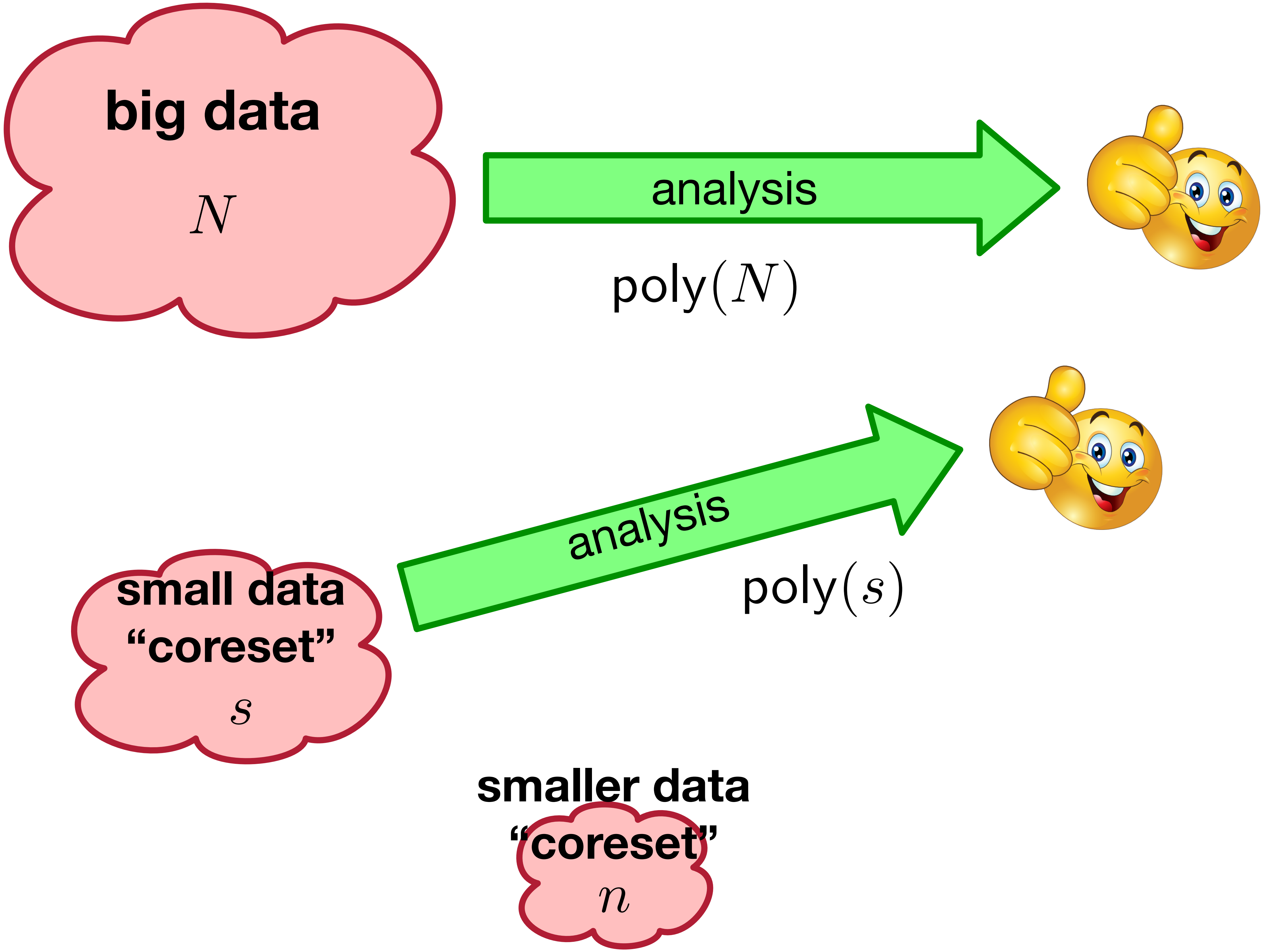


Repeat many times,

what fraction find planted region?

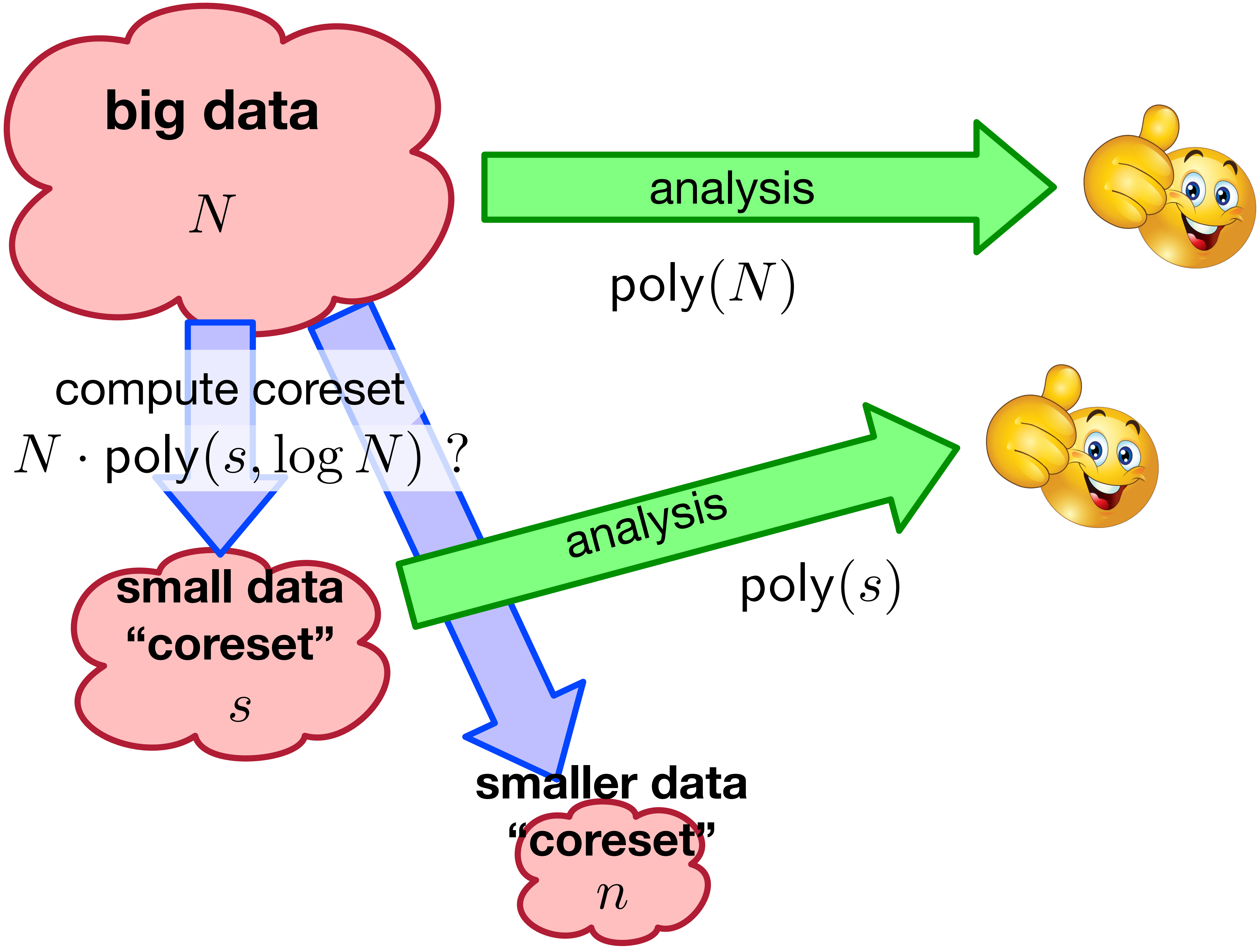
Can we do better?


big data
N
analysis
poly(N)
analysis
poly(s)
small data
“coreset”
s



Improvements?

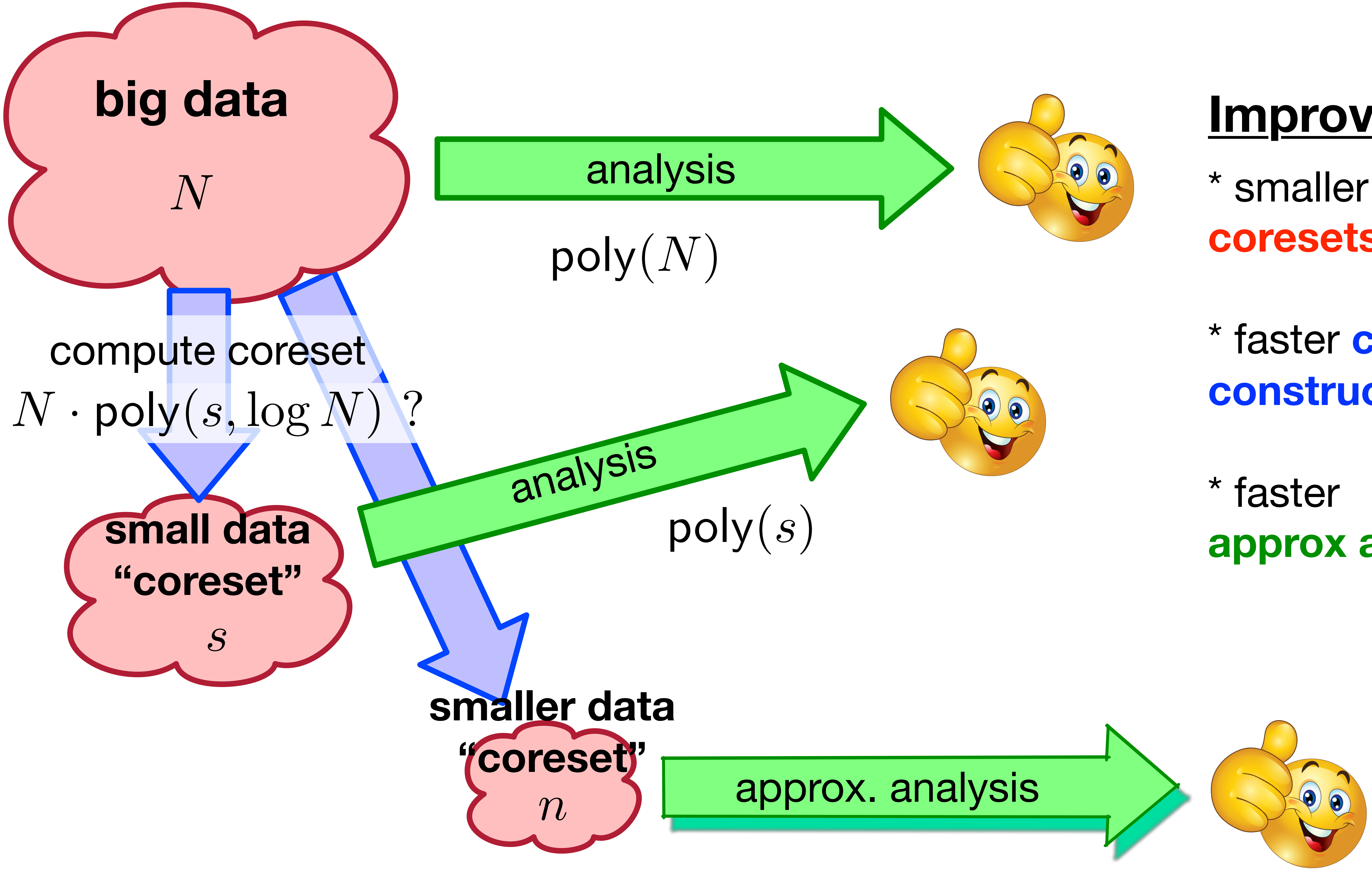
* smaller **coresets**?



Improvements?

* smaller **coresets**?

* faster **coreset construction**?



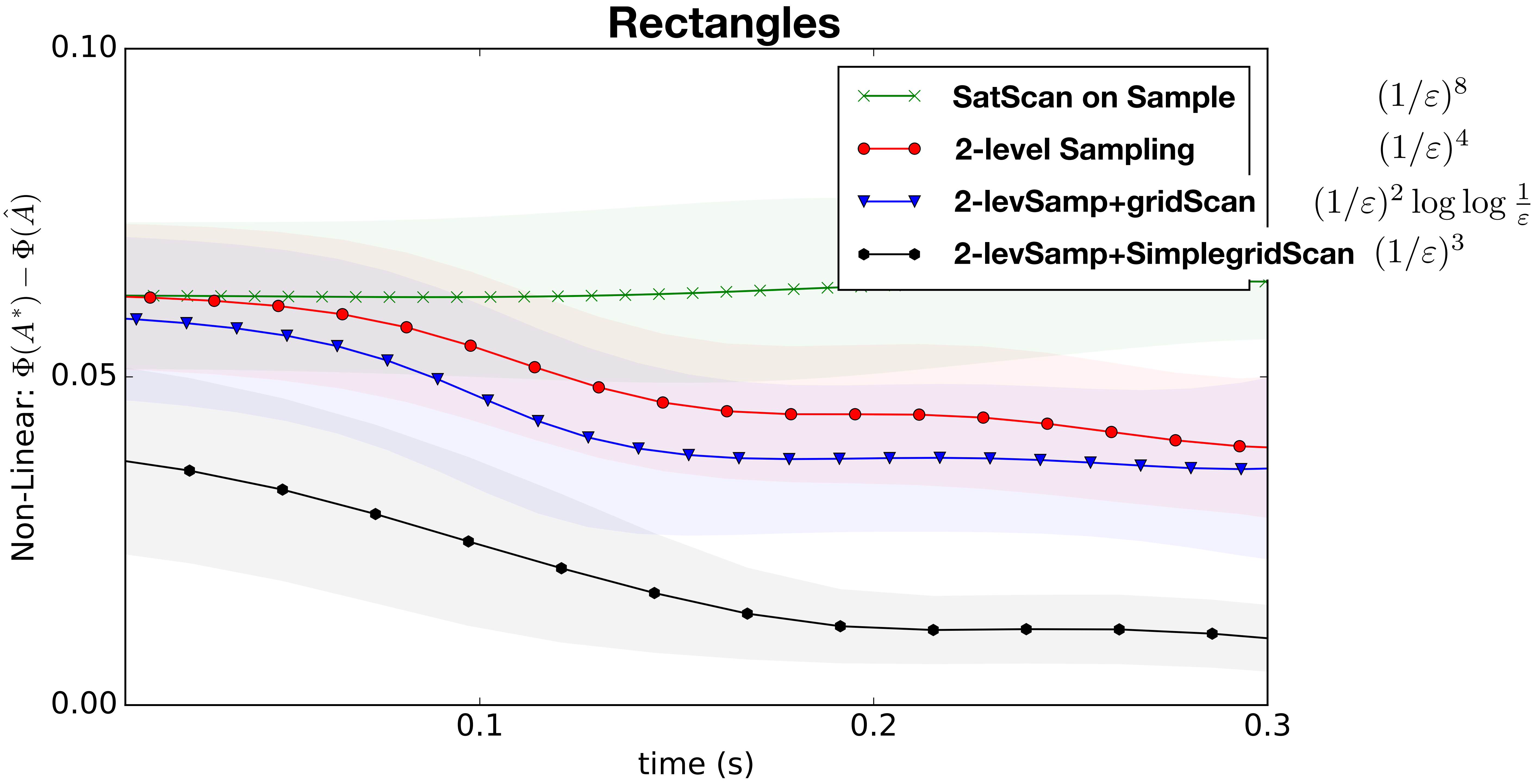
Improvements?

* smaller **coresets**?

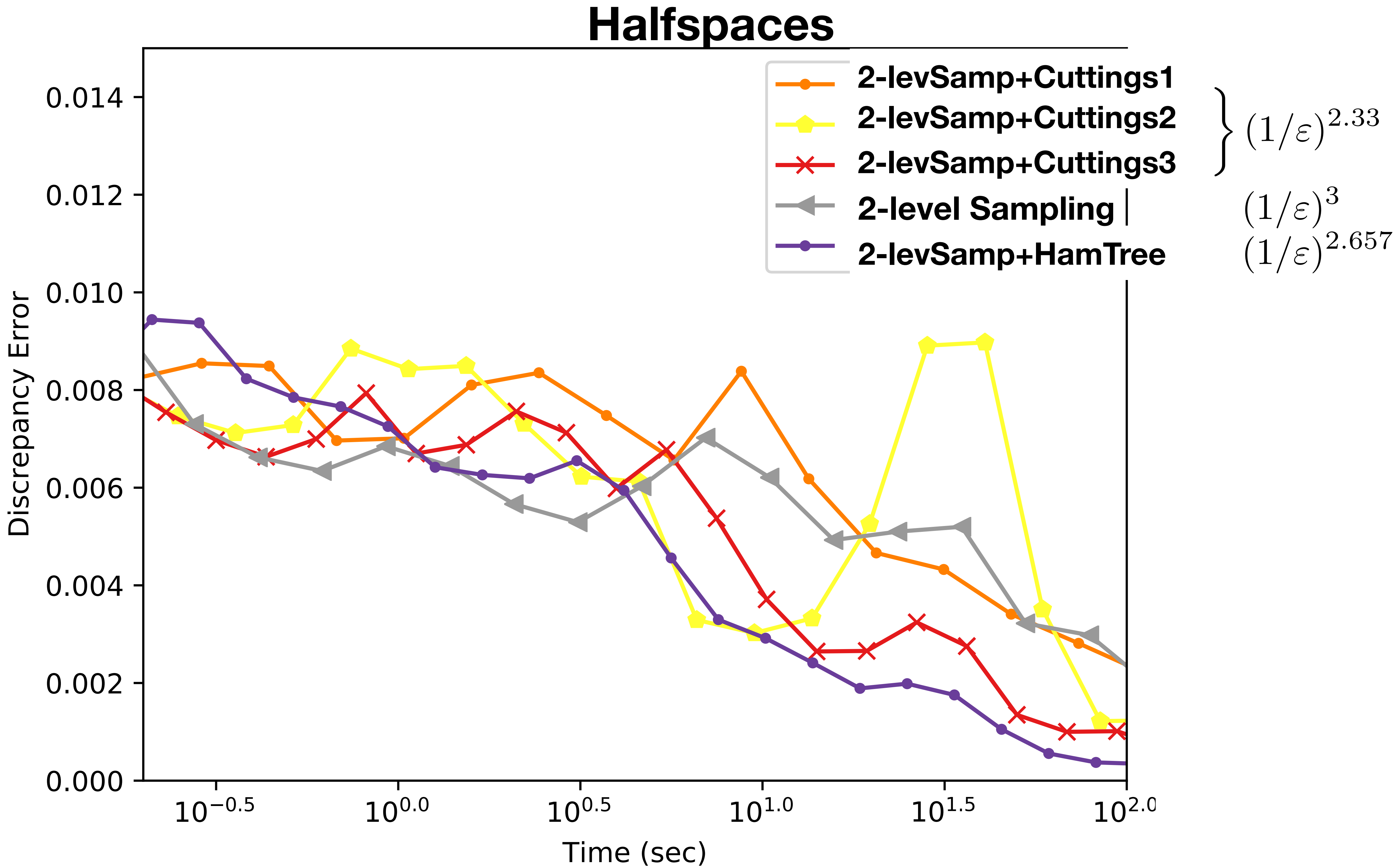
* faster **coreset construction**?

* faster **approx analysis**?

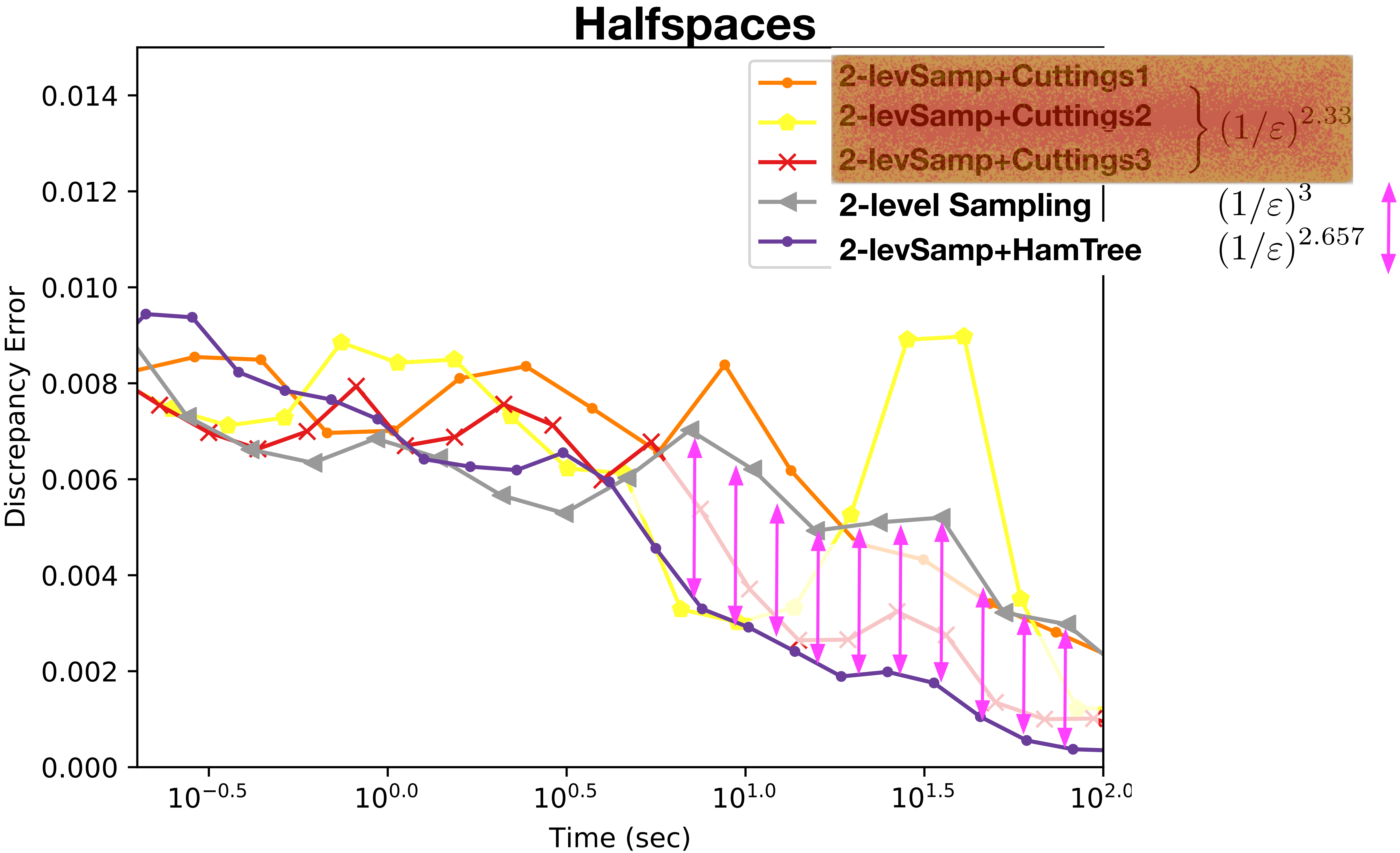
Can we do better?



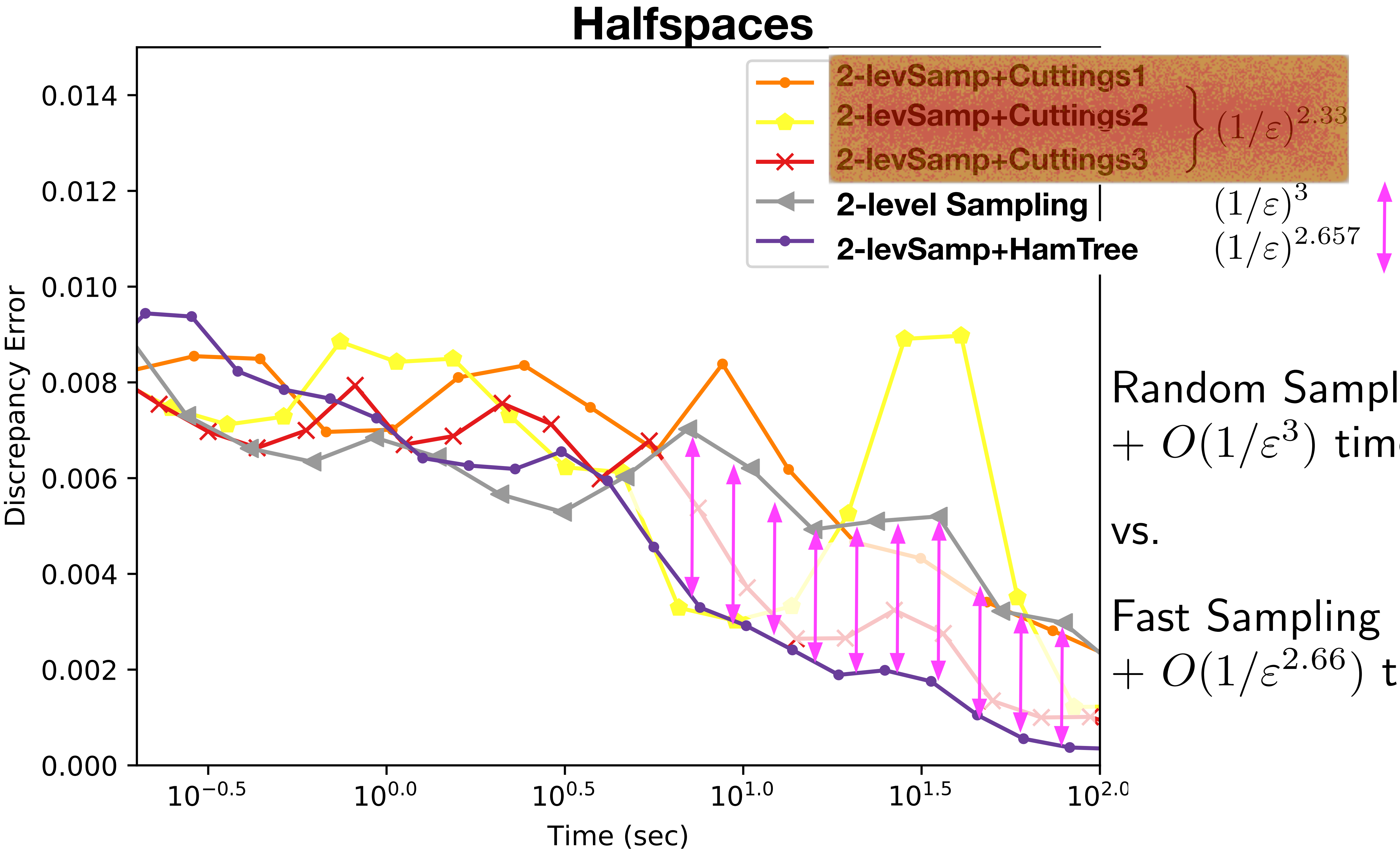
Can we do better?



Can we do better?



Can we do better?



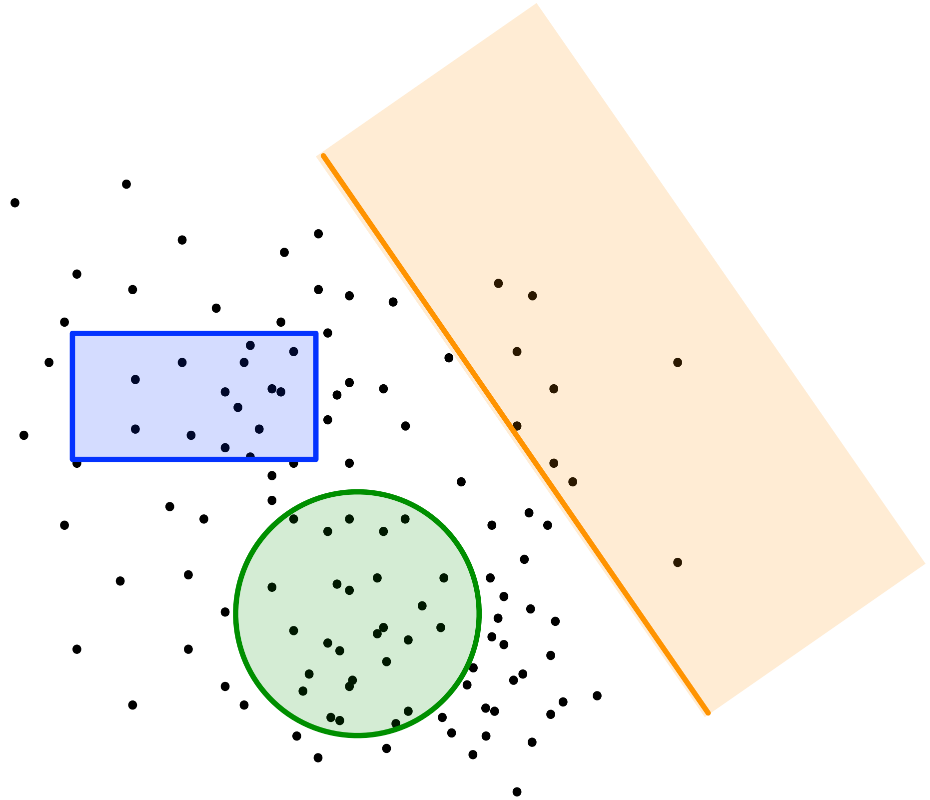
Random Sampling (very fast!)
$+\ O(1/\varepsilon^3)$ time scanning

vs.

Fast Sampling
$+\ O(1/\varepsilon^{2.66})$ time scanning

Improved Two-Level Sample-then-Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  1. create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$
  2. create an ε-net $N \subset X$; $|N| = n = (1/\varepsilon)\log(1/\varepsilon)$
  3. Scan all ranges $C \in (S, \mathcal{C}_{\Delta N})$ and evaluate $\Phi_S(C)$



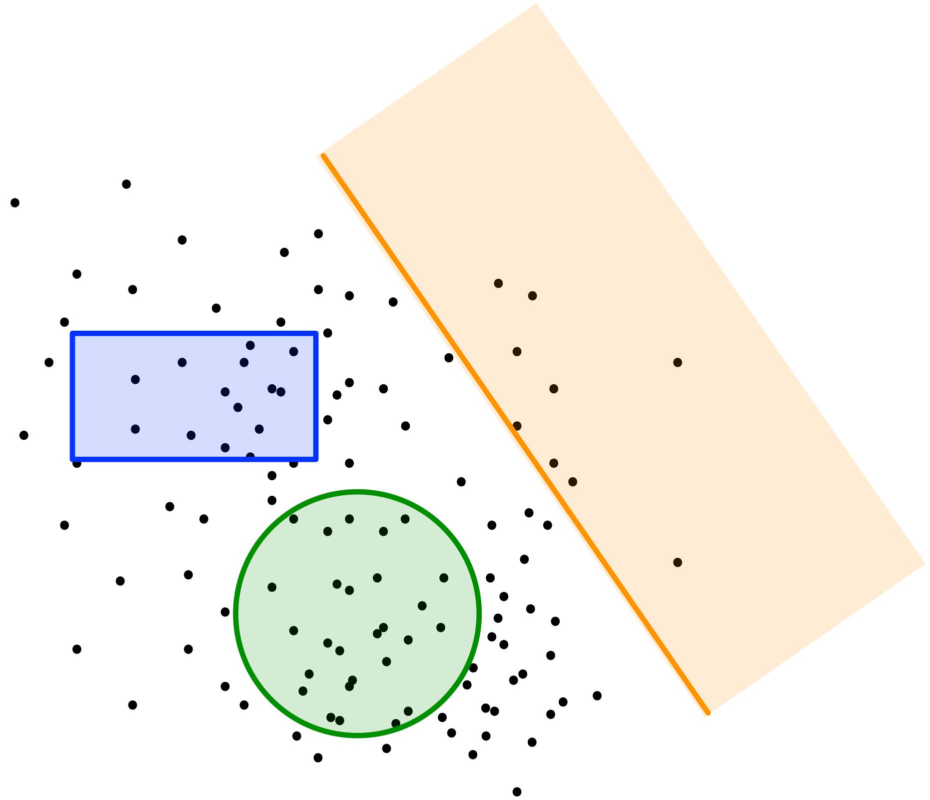
Improved Two-Level Sample-then-Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
    1. create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$
    2. create an ε-net $N \subset X$; $|N| = n = (1/\varepsilon)\log(1/\varepsilon)$
    3. Scan all ranges $C \in (S, \mathcal{C}_{\Delta N})$ and evaluate $\Phi_S(C)$

- For **halfspaces** in $\mathbb{R}^2$, there exist ε-samples of size $1/\varepsilon^{4/3}$.

  $\Rightarrow$ **NEW** construct S with $s = O(\frac{1}{\varepsilon^{4/3}} \log^{2/3})$ in $O(|X| \log |X|)$ time.

  $\Rightarrow$ $O(ns)$ time from $1/\varepsilon^3 \to (1/\varepsilon^{7/3}) \log^{2/3} \frac{1}{\varepsilon}$

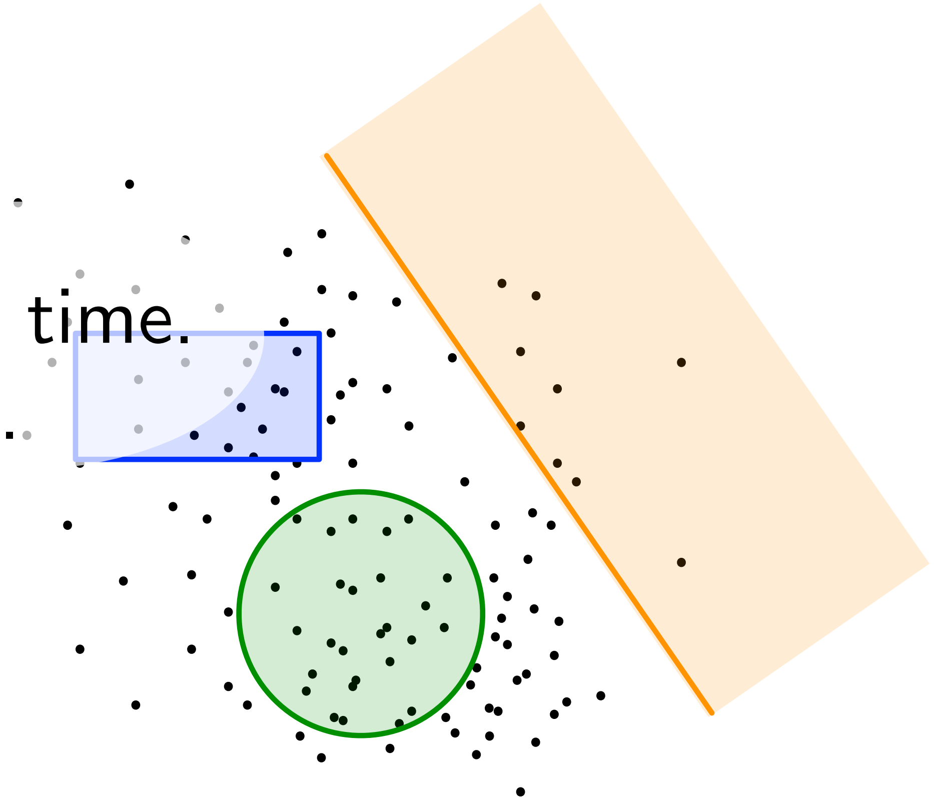


Improved Two-Level Sample-then-Scan

- Consider large conforming range space $(X, \mathcal{C})$ with map $\psi_{\mathcal{C}}$ and VC-dim ν
  1. create an ε-sample $S \subset X$; $|S| = s = 1/\varepsilon^2$
  2. create an ε-net $N \subset X$; $|N| = n = (1/\varepsilon)\log(1/\varepsilon)$
  3. Scan all ranges $C \in (S, \mathcal{C}_{\Delta N})$ and evaluate $\Phi_S(C)$

- For **halfspaces** in $\mathbb{R}^2$, there exist ε-samples of size $1/\varepsilon^{4/3}$.
  - $\Rightarrow$ **NEW** construct S with $s = O(\frac{1}{\varepsilon^{4/3}} \log^{2/3})$ in $O(|X| \log |X|)$ time.
  - $\Rightarrow$ $O(ns)$ time from $1/\varepsilon^3 \to (1/\varepsilon^{7/3}) \log^{2/3} \frac{1}{\varepsilon}$

- For **rectangles** in $\mathbb{R}^2$, only approximately scan $(S, \mathcal{C}_{\Delta N})$.
  - $\Rightarrow$ **NEW** ε-apx-scanning in $O(n^2 \log\log n + s \log\log n)$ time.
  - $\Rightarrow$ **NEW** simple ε-apx-scanning in $O(n^3 + s \log n)$ time.
  - $\Rightarrow$ from $n^4 + s \approx 1/\varepsilon^4$ to $n^2 + s \approx 1/\varepsilon^2$ time.



Fast Halfspace ε-Samples

- There exists ε-sample for $(X, \mathcal{H}_2)$ of size $\Theta(1/\varepsilon^{4/3})$.

  Alexander (Combinatorica '90), Matousek (DCG '95)

Fast Halfspace ε-Samples

- There exists ε-sample for $(X, \mathcal{H}_2)$ of size $\Theta(1/\varepsilon^{4/3})$.

  Alexander (Combinatorica '90), Matousek (DCG '95)

The high computational complexity of the currently known algorithms for these subroutines may be prohibitive for data stream applications. It is a long standing open problem to find efficient exact or approximation algorithms for either of them. – Suri,Toth,&Zhou (SoCG'04)

Fast Halfspace ε-Samples

- There exists ε-sample for $(X, \mathcal{H}_2)$ of size $\Theta(1/\varepsilon^{4/3})$.

  Alexander (Combinatorica '90), Matousek (DCG '95)

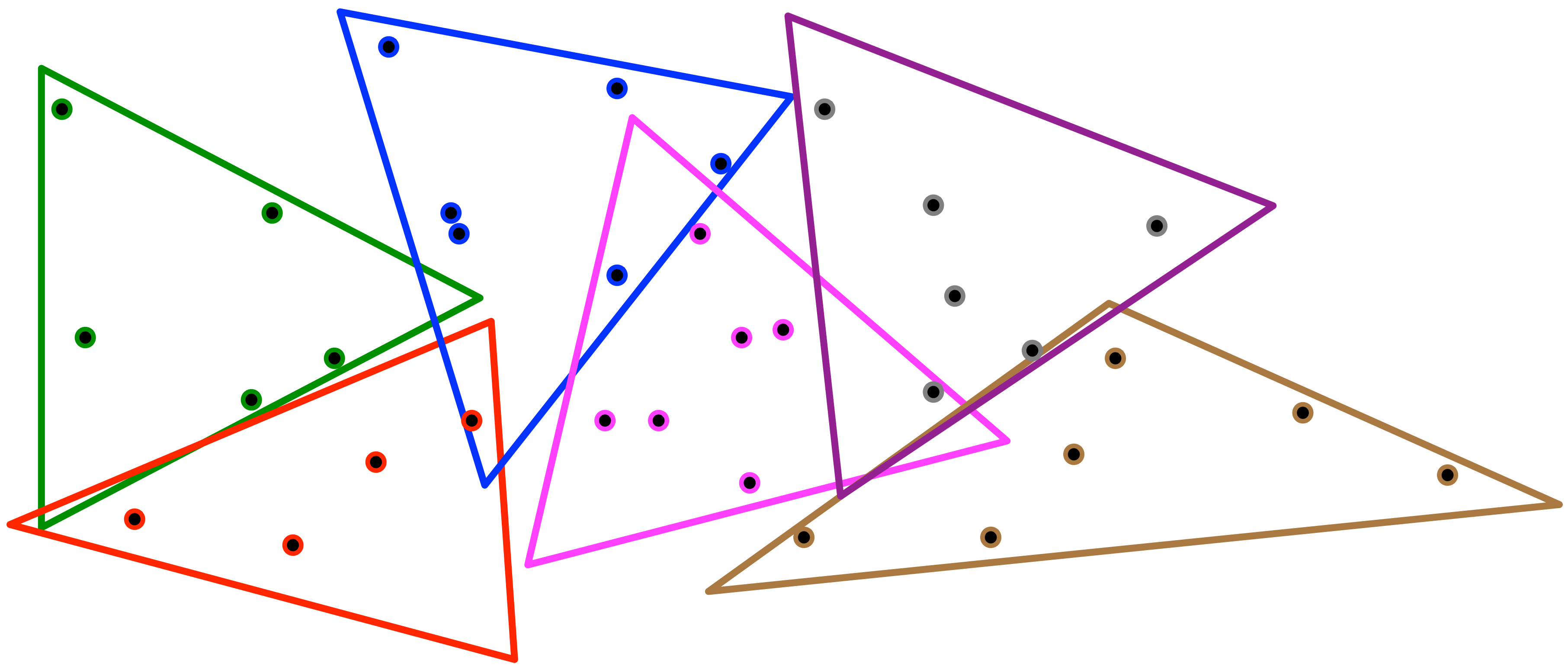
The high computational complexity of the currently known algorithms for these subroutines may be prohibitive for data stream applications. It is a long standing open problem to find efficient exact or approximation algorithms for either of them. – Suri,Toth,&Zhou (SoCG'04)

- Can be made constructed in polynomial time, about $O(|X|/\varepsilon^4)$. Uses SDP.

  Bansal (FOCS '10)

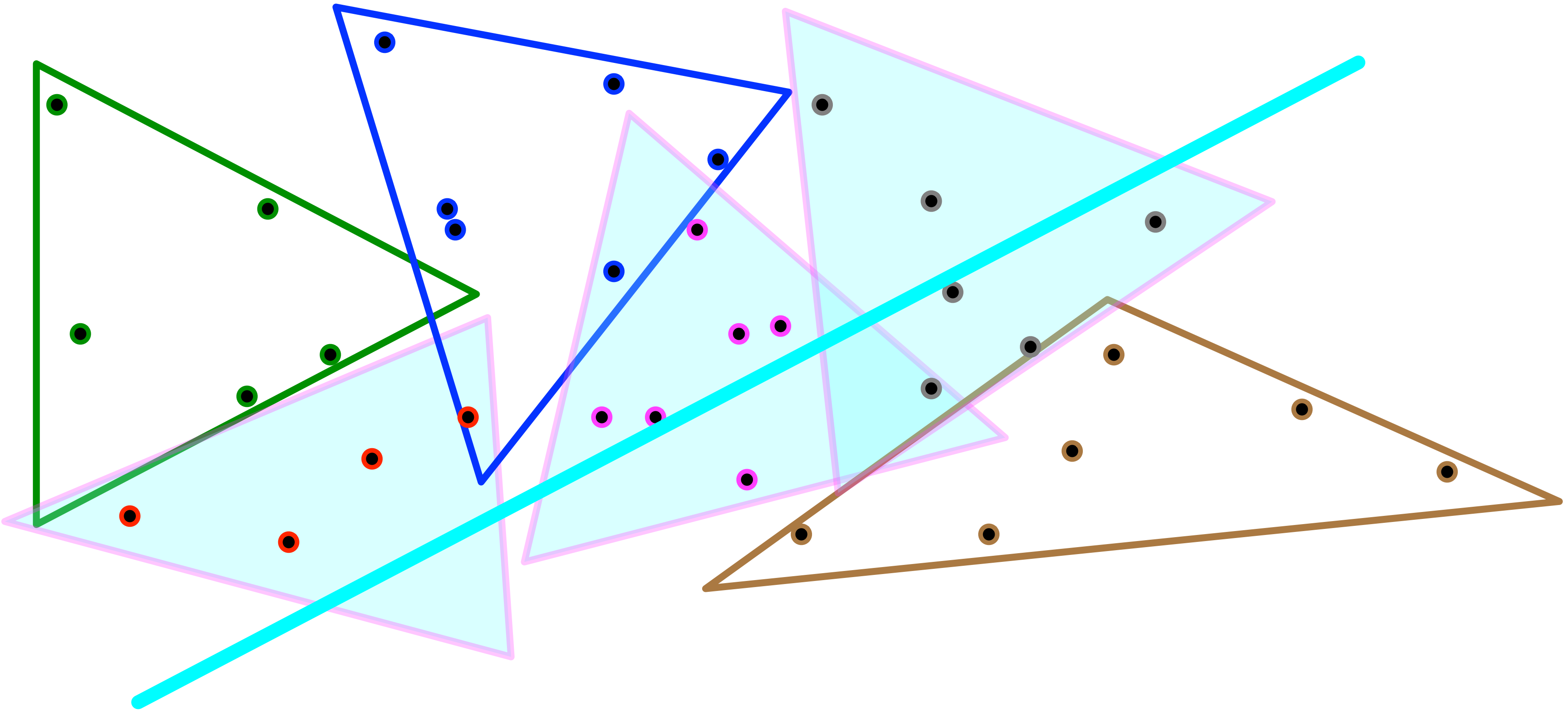
Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.



Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.



Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.

- For any large enough $t > 0$, a $(t, 1/2)$-partition can be found in $O(|X| \log t)$ time.

  Matousek (DCG '92), Chan (SoCG, '10)

Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.

- For any large enough $t > 0$, a $(t, 1/2)$-partition can be found in $O(|X| \log t)$ time.
  Matousek (DCG '92), Chan (SoCG, '10)

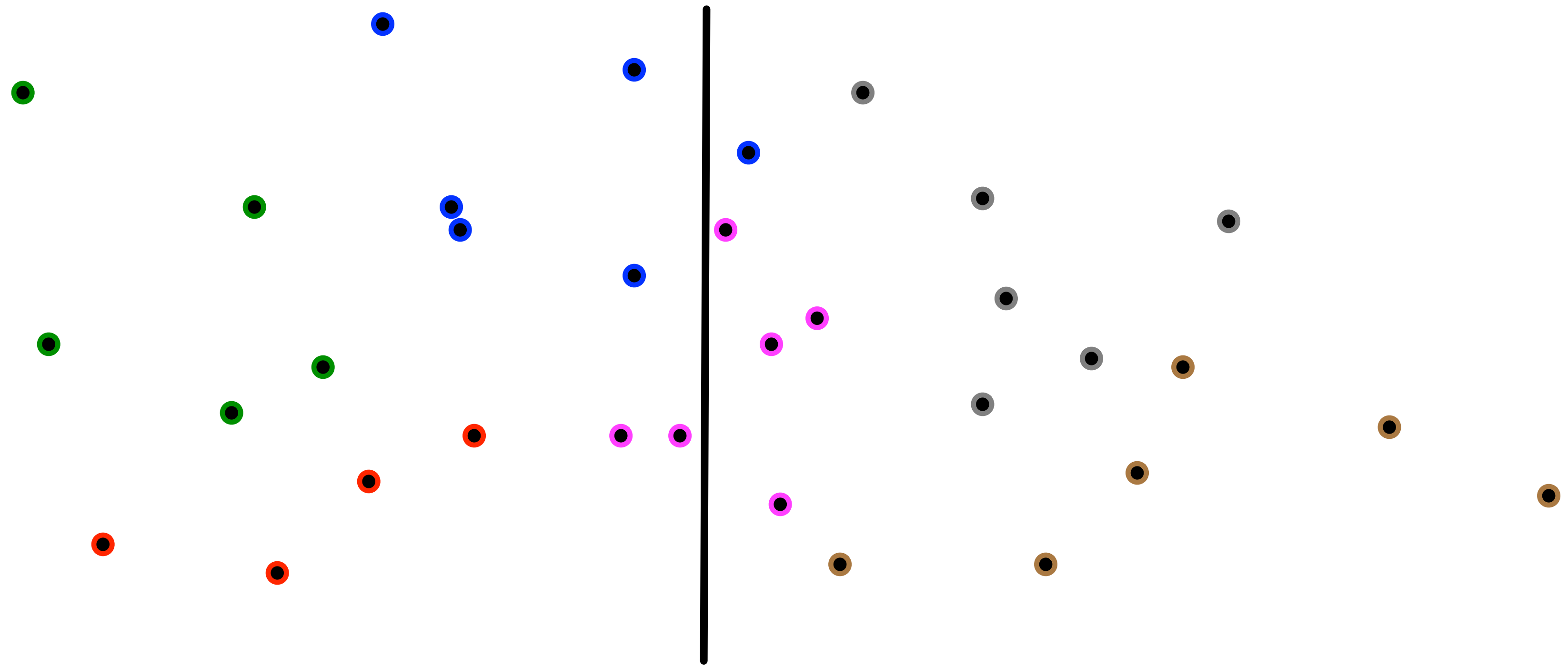
- A $(t, 0.7925)$-partition can be found in $O(|X| \log \frac{|X|}{t})$ time; $z = \log_4(3)$.
  Willard (SICOMP '82), Edelsbrunner+Welzl (IPL, '86)

Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.

- A $(t, 0.7925)$-partition can be found in $O(|X| \log \frac{|X|}{t})$ time; $z = \log_4(3)$.

  Willard (SICOMP '82), Edelsbrunner+Welzl (IPL, '86)

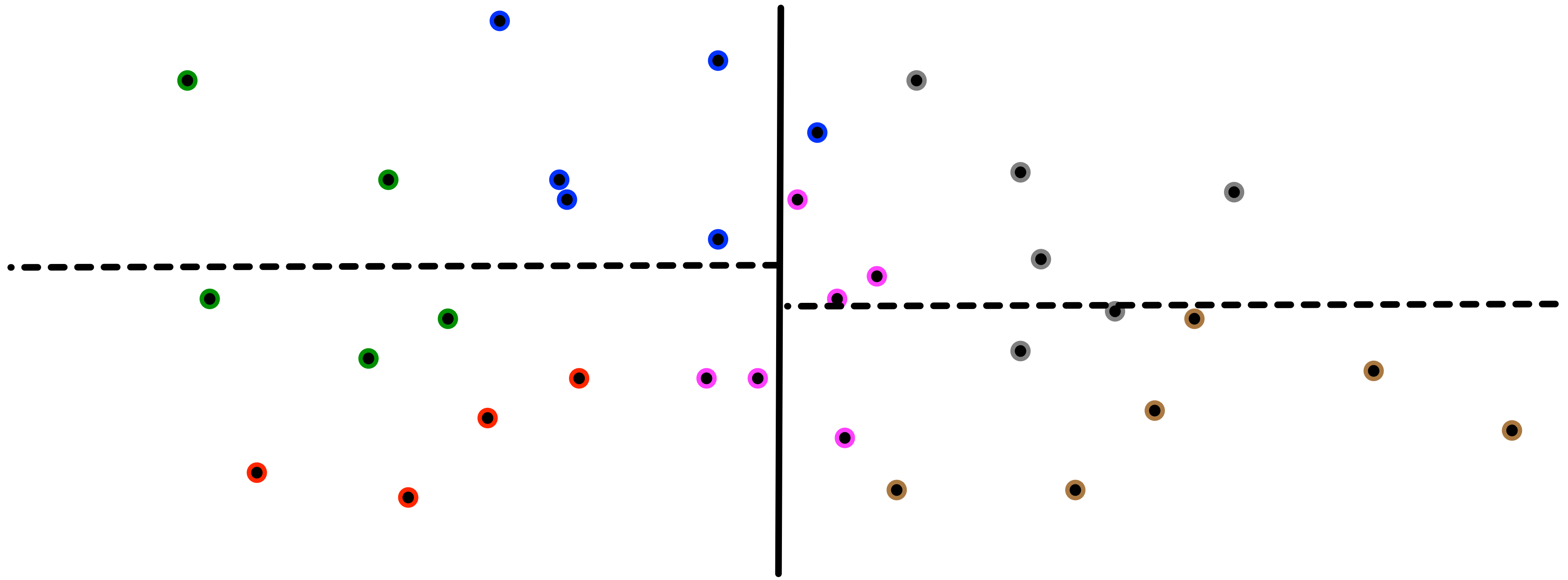


Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.

- A $(t, 0.7925)$-partition can be found in $O(|X| \log \frac{|X|}{t})$ time; $z = \log_4(3)$.

  Willard (SICOMP '82), Edelsbrunner+Welzl (IPL, '86)

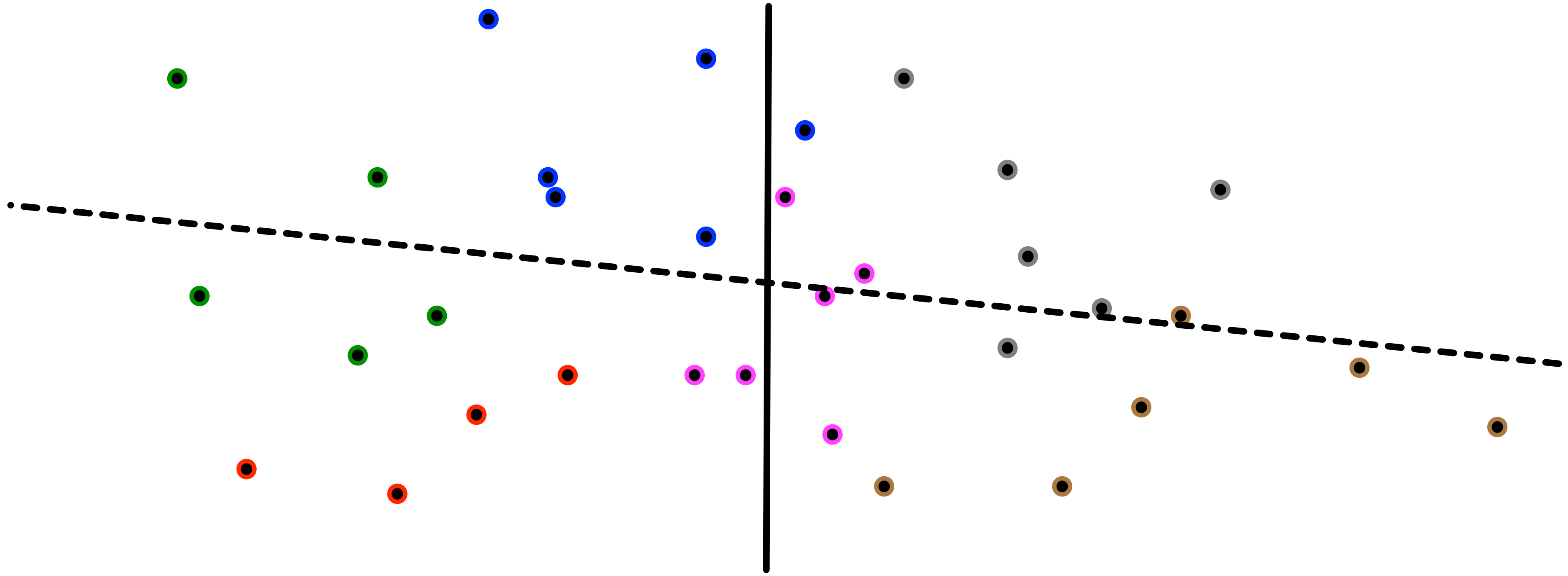


Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.

- A $(t, 0.7925)$-partition can be found in $O(|X| \log \frac{|X|}{t})$ time; $z = \log_4(3)$.

  Willard (SICOMP '82), Edelsbrunner+Welzl (IPL, '86)

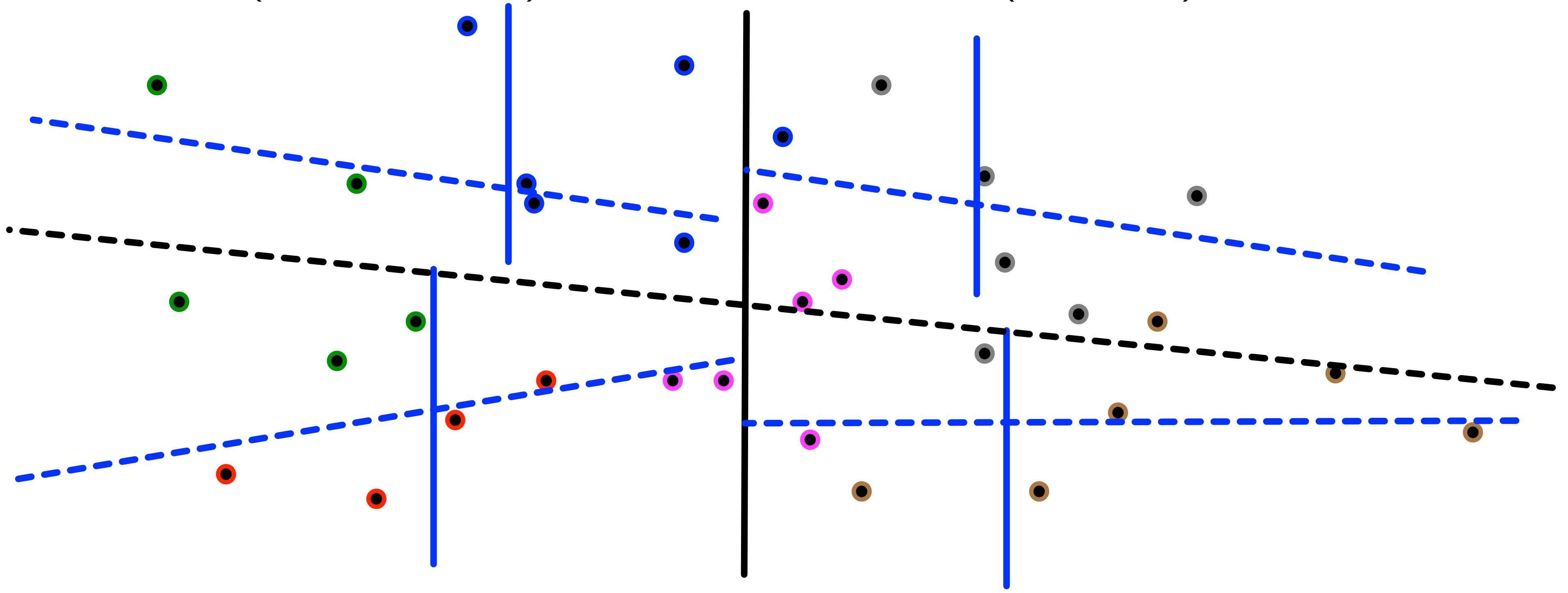


Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.

- A $(t, 0.7925)$-partition can be found in $O(|X| \log \frac{|X|}{t})$ time; $z = \log_4(3)$.

  Willard (SICOMP '82), Edelsbrunner+Welzl (IPL, '86)



Fast Halfspace ε-Samples

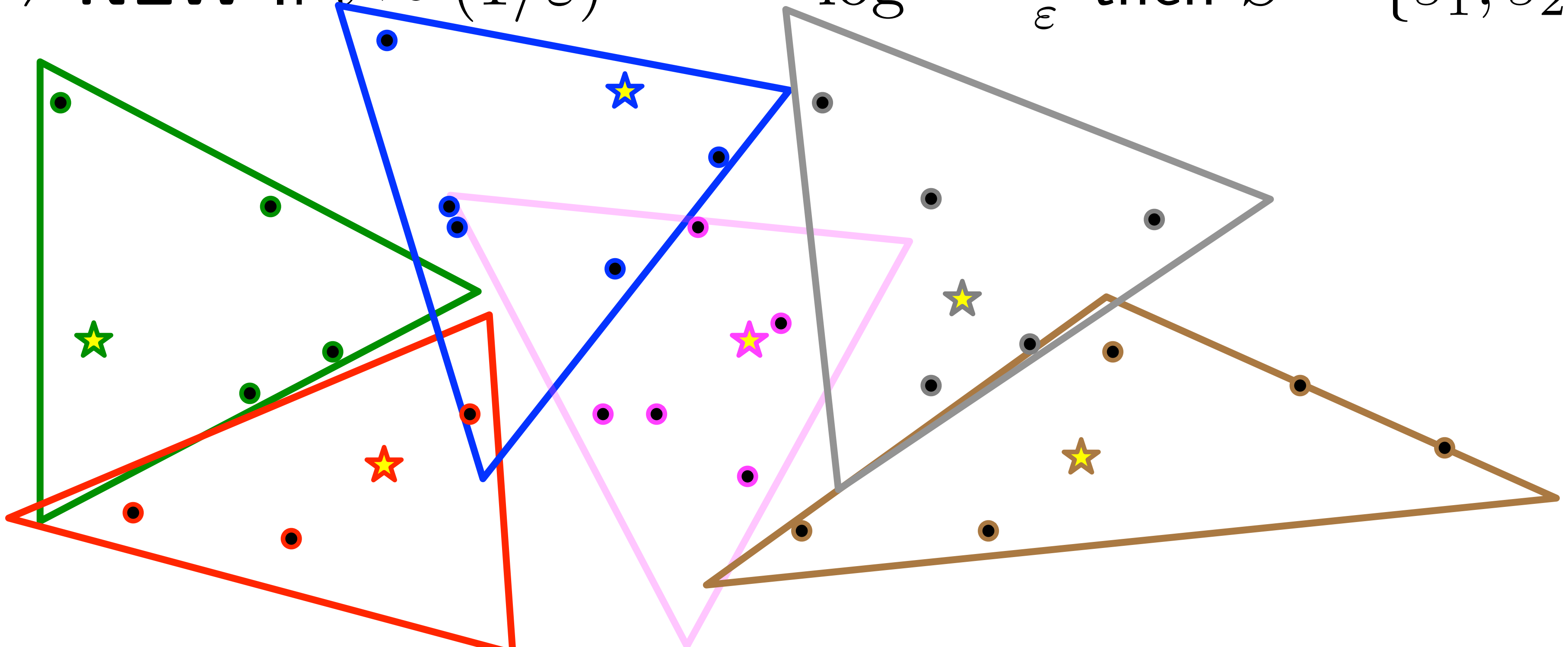
- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.

- Choose one $s_i \in X_i$ randomly for each pair (Δ_i, X_i).
  - $\Rightarrow$ **NEW** If $t \approx (1/\varepsilon)^{2/(2-z)} \log^{\frac{1}{2-z}} \frac{1}{\varepsilon}$ then $S = \{s_1, s_2, \ldots\}$ is an ε-sample.

Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.

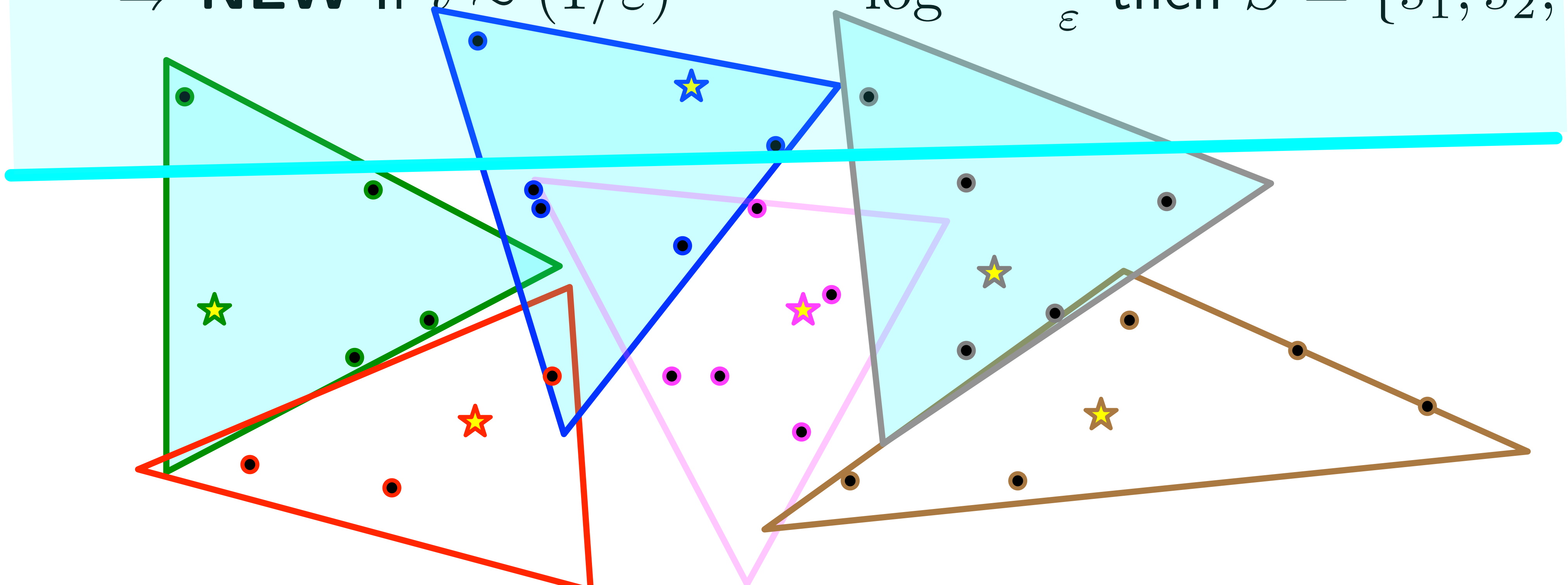
- Choose one $s_i \in X_i$ randomly for each pair (Δ_i, X_i).
  - $\Rightarrow$ **NEW** If $t \approx (1/\varepsilon)^{2/(2-z)} \log^{\frac{1}{2-z}} \frac{1}{\varepsilon}$ then $S = \{s_1, s_2, \ldots\}$ is an ε-sample.



Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.

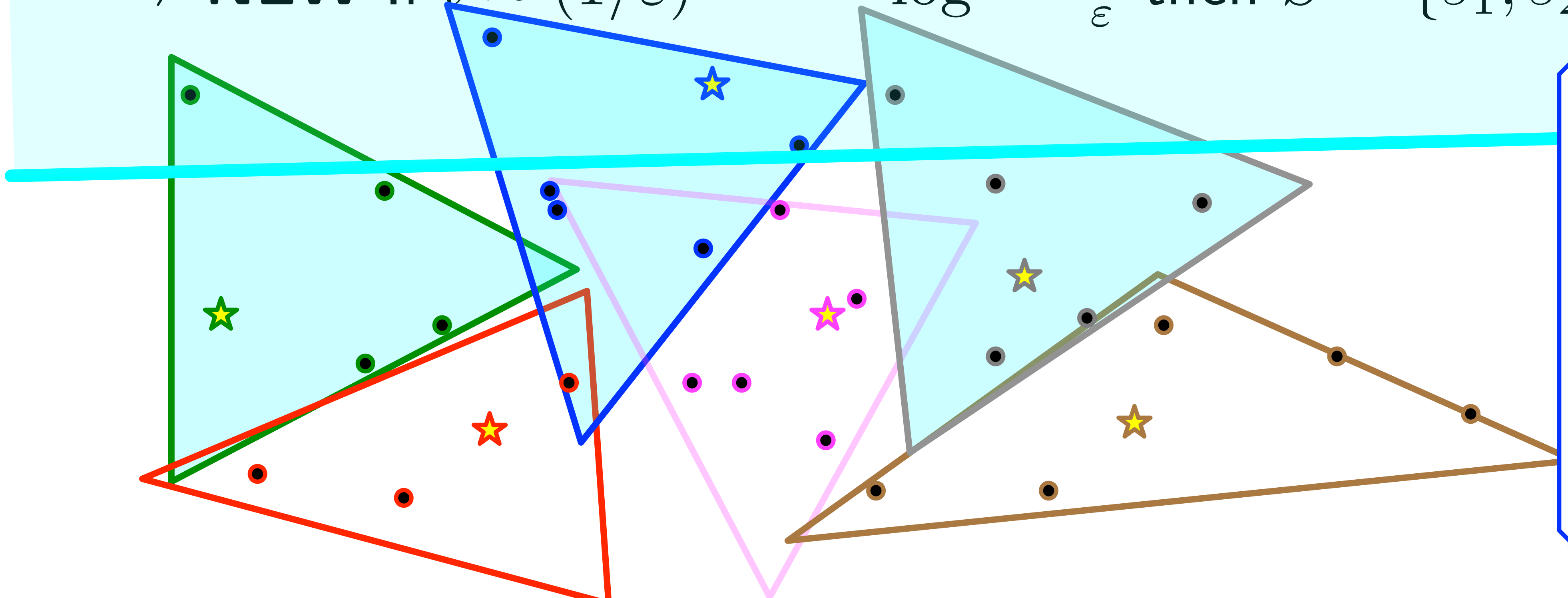
- Choose one $s_i \in X_i$ randomly for each pair (Δ_i, X_i).
  - $\Rightarrow$ **NEW** If $t \approx (1/\varepsilon)^{2/(2-z)} \log^{\frac{1}{2-z}} \frac{1}{\varepsilon}$ then $S = \{s_1, s_2, \ldots\}$ is an ε-sample.



Fast Halfspace ε-Samples

- A **(t,z)-partition** of $(X, \mathcal{H}_2)$ is a set of pairs $\{(\Delta_1, X_1), (\Delta_2, X_2), \ldots\}$ so
  - $\rightarrow$ each cell Δ_i is a simple region that contains X_i
  - $\rightarrow$ X is a disjoint union of $X_1 \cup X_2 \cup \ldots$
  - $\rightarrow$ there are $O(t)$ pairs, each with $|X_i| \leq 2|X|/t$,
  - $\rightarrow$ and each $h \in \mathcal{H}_2$ crosses $O(t^z)$ cells.

- Choose one $s_i \in X_i$ randomly for each pair (Δ_i, X_i).
  - $\Rightarrow$ **NEW** If $t \approx (1/\varepsilon)^{2/(2-z)} \log^{\frac{1}{2-z}} \frac{1}{\varepsilon}$ then $S = \{s_1, s_2, \ldots\}$ is an ε-sample.

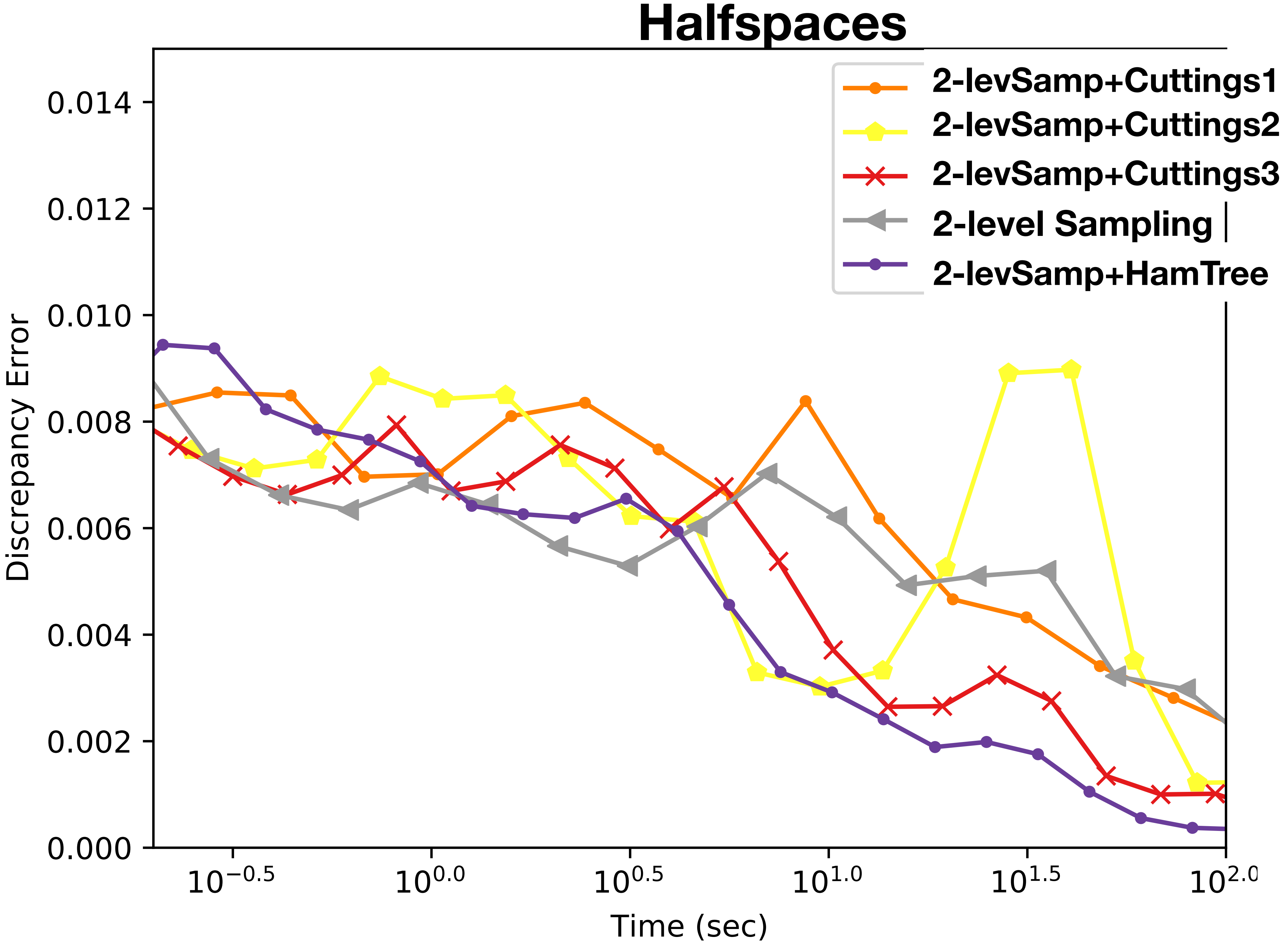


+ h hits $t^{1/2}$ cells
+ each $|X_i| \leq |X|/t$
+ $t \approx 1/\varepsilon^{4/3}$

proof in $\mathbb{R}^2$ $z = \frac{1}{2}$

Apply Hoeffding bound:
$\Pr[\text{err} > \varepsilon|X|] \leq$
$\exp(-\frac{(\varepsilon|X|)^2}{t^{1/2} \cdot (\frac{|X|}{t})^2}) = \exp(-\varepsilon^2 t^{3/2})$
$\approx$ const

Even Faster Halfspace Scanning



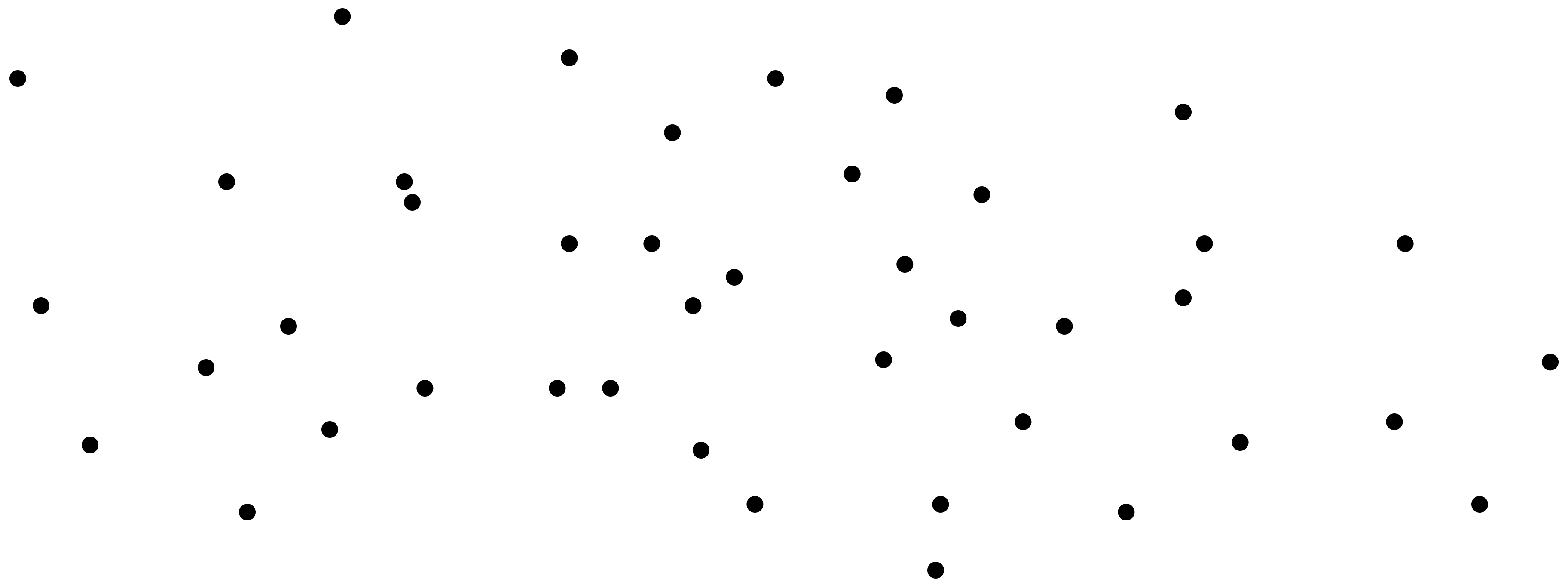
$\left.\right\}(1/\varepsilon)^{2.33}$

$(1/\varepsilon)^{3}$

$(1/\varepsilon)^{2.657}$

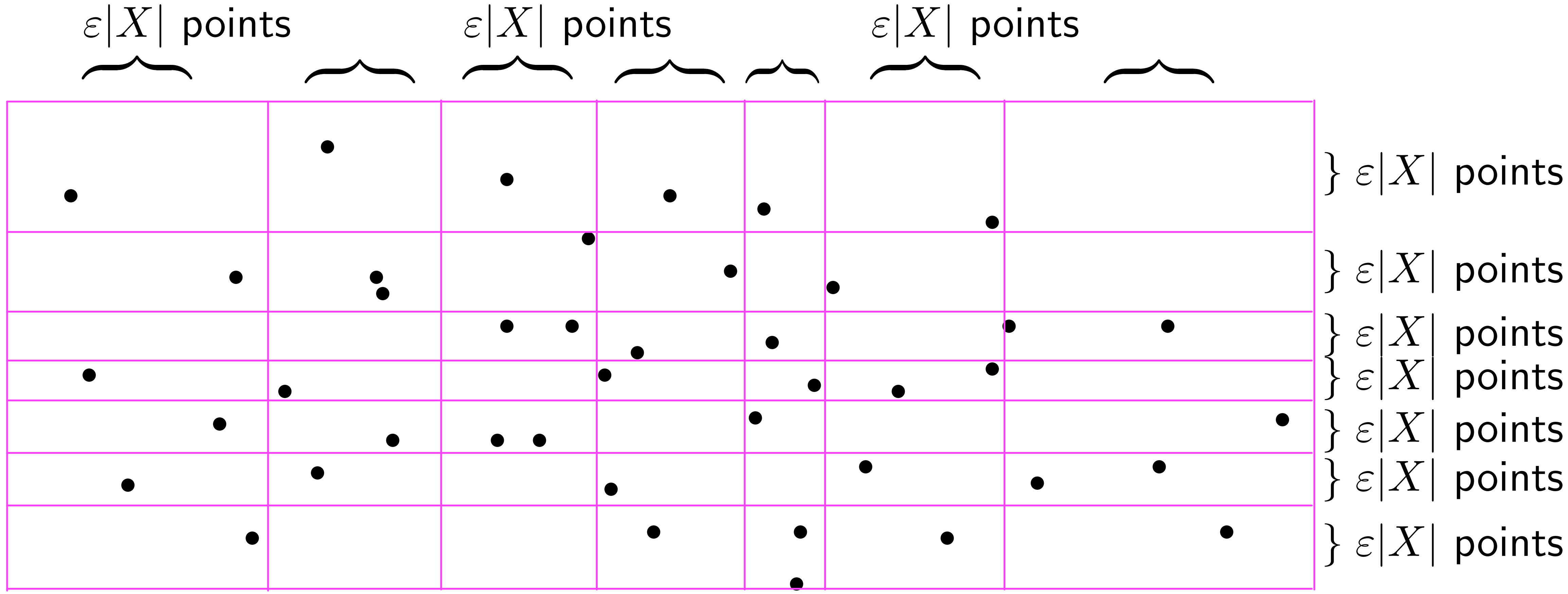
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.



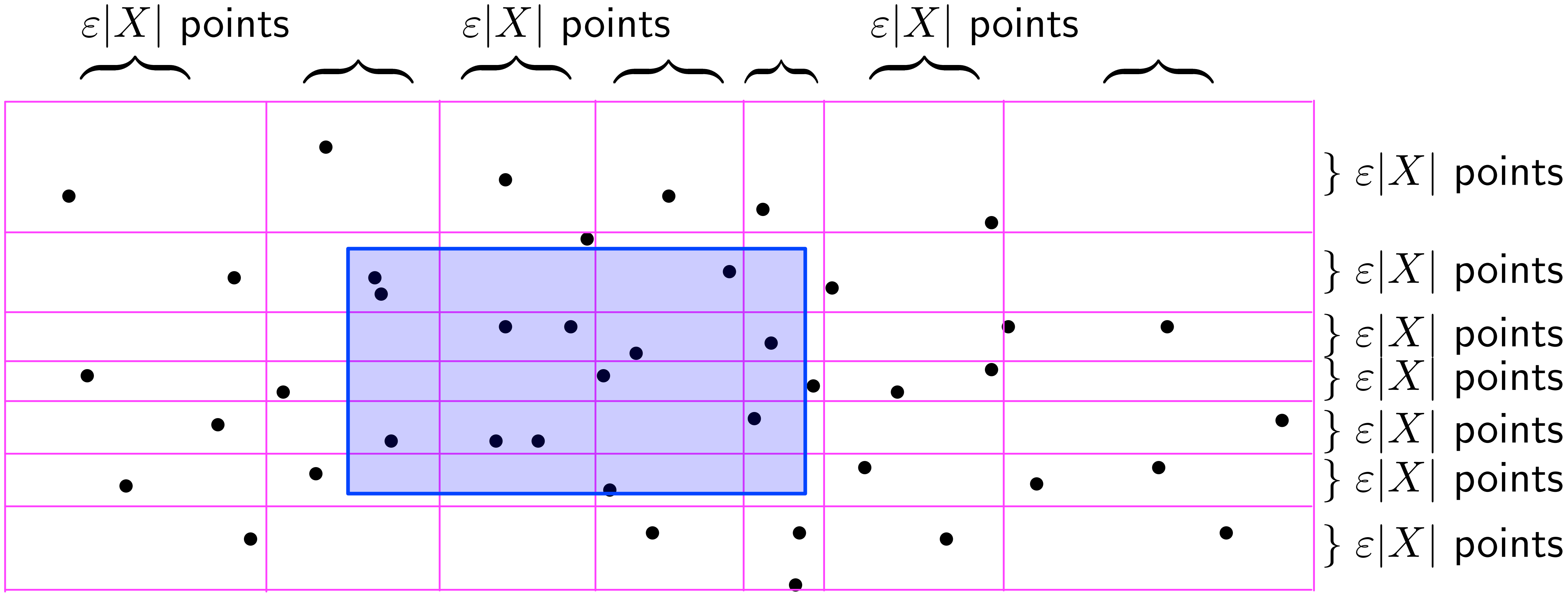
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.



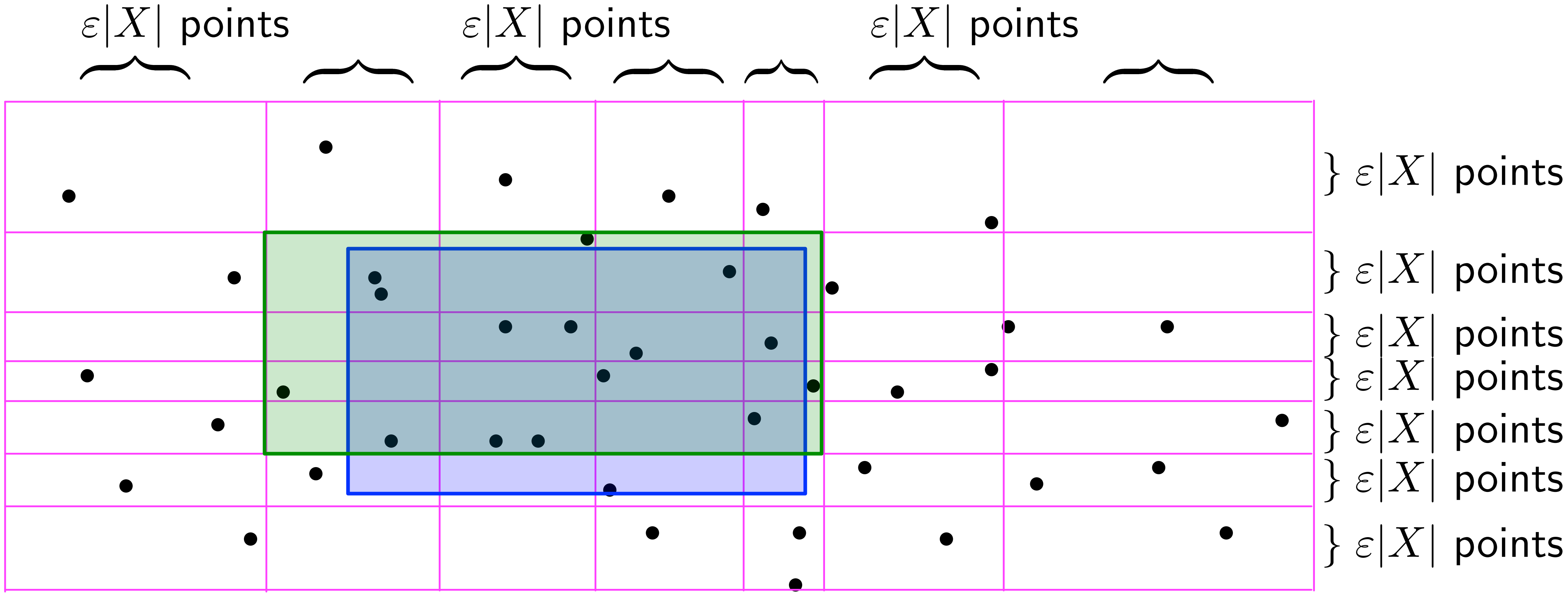
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Each rectangle $R \in (X, \mathcal{R}_2)$ is ε-approximated by on in $R_g \in G$



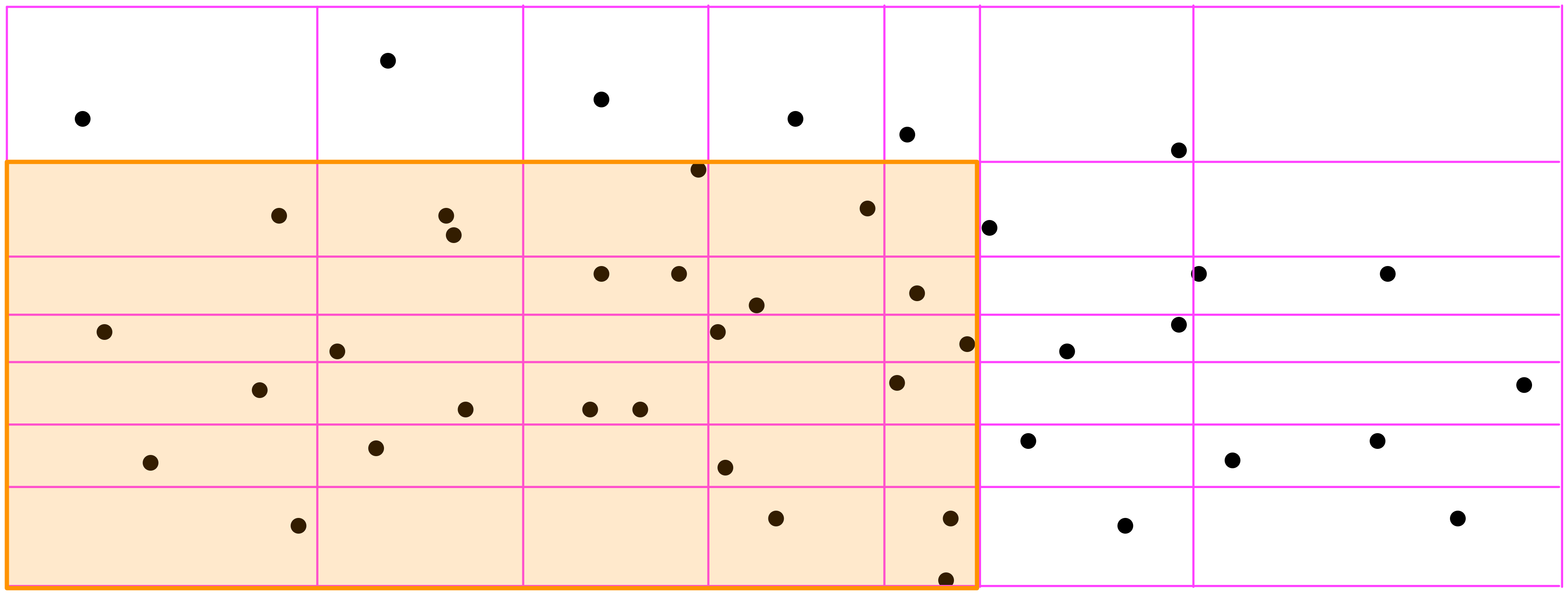
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Each rectangle $R \in (X, \mathcal{R}_2)$ is ε-approximated by on in $R_g \in G$



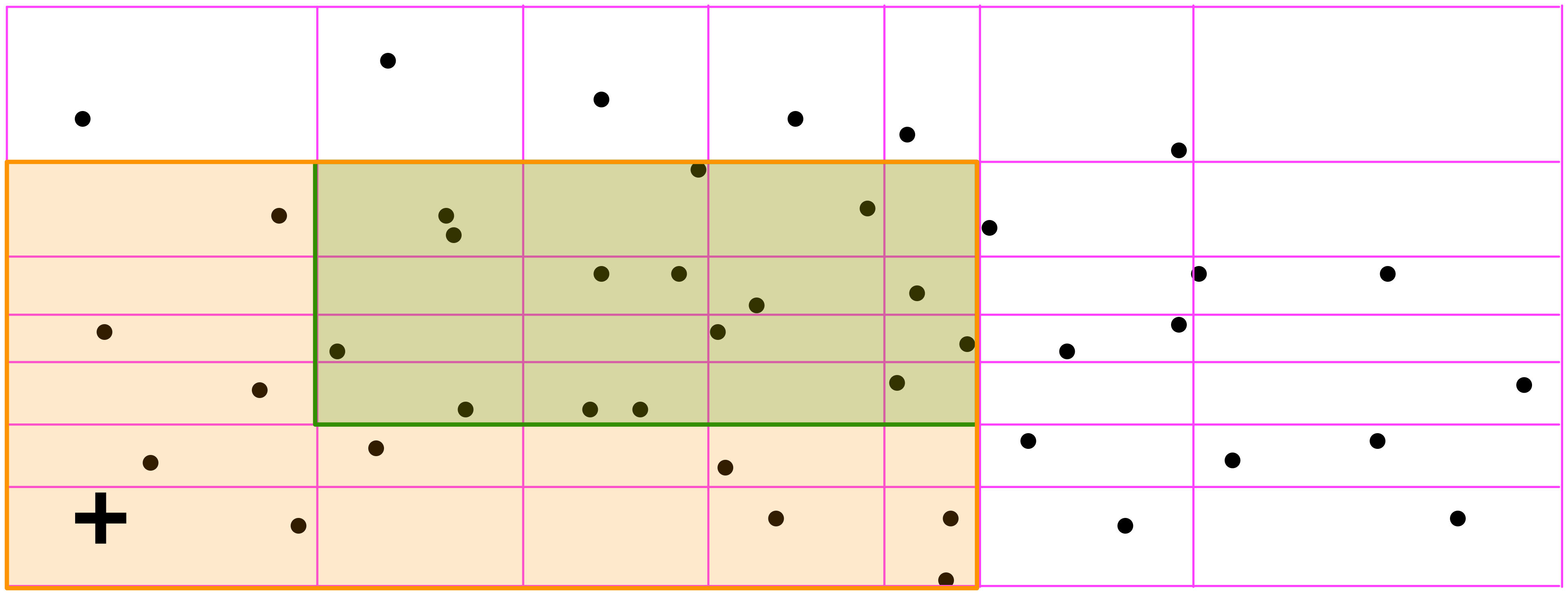
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Each rectangle $R \in (X, \mathcal{R}_2)$ is ε-approximated by on in $R_g \in G$
- Compute subset sum for all 2-sided rectangle in $1/\varepsilon^2$ time.
  Can now compute $\Phi(R_g)$ for each $R_g \in G$ in $O(1)$ time.



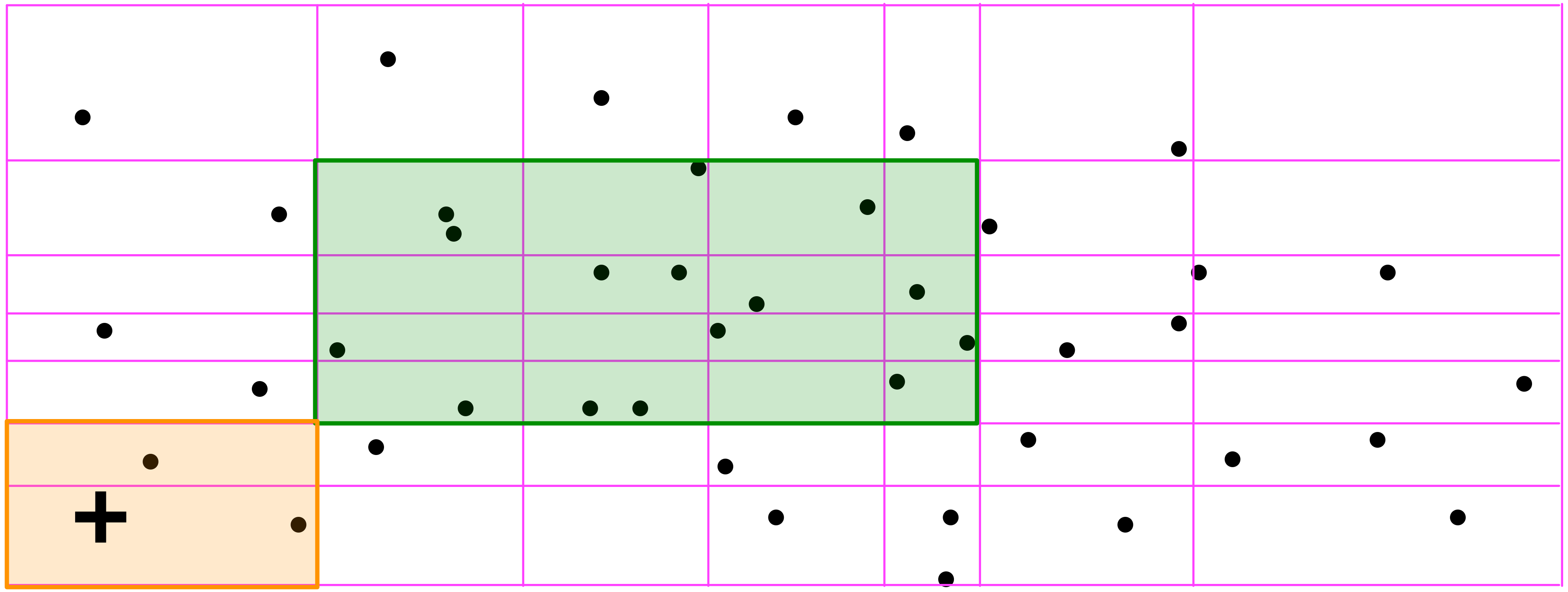
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Each rectangle $R \in (X, \mathcal{R}_2)$ is ε-approximated by on in $R_g \in G$
- Compute subset sum for all 2-sided rectangle in $1/\varepsilon^2$ time.
  Can now compute $\Phi(R_g)$ for each $R_g \in G$ in $O(1)$ time.



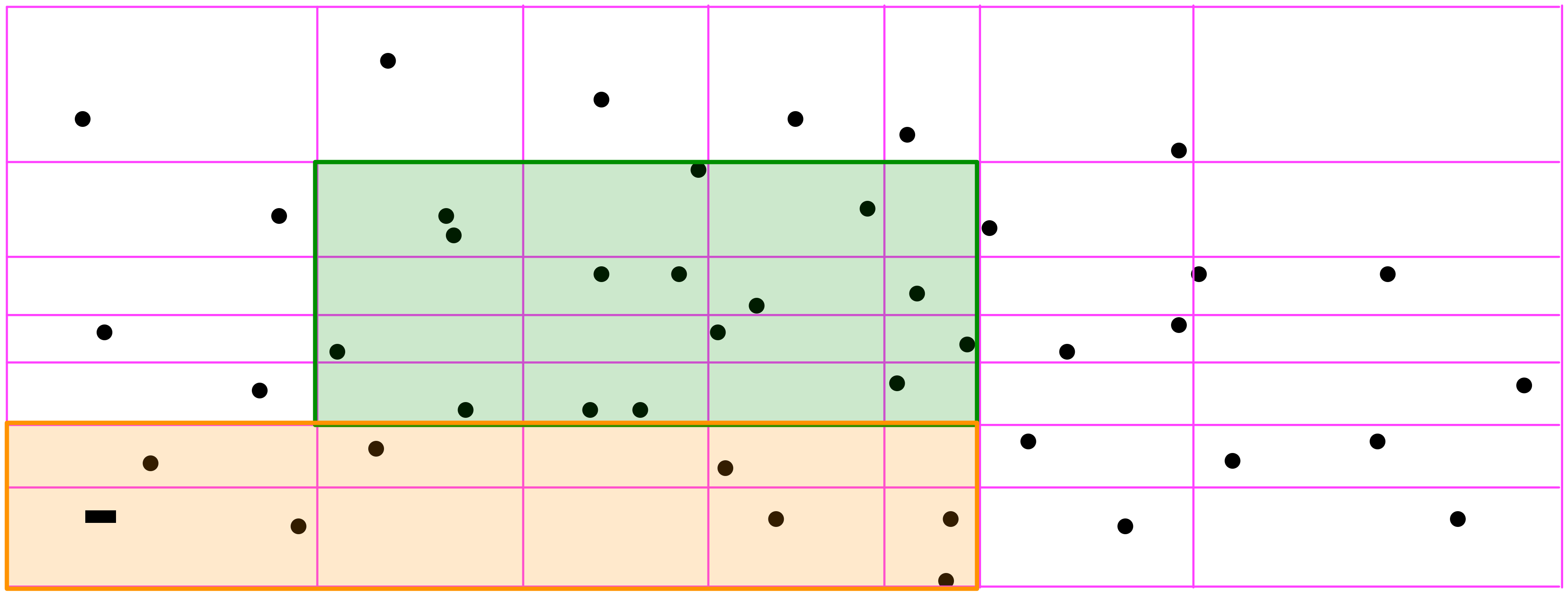
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Each rectangle $R \in (X, \mathcal{R}_2)$ is ε-approximated by on in $R_g \in G$
- Compute subset sum for all 2-sided rectangle in $1/\varepsilon^2$ time.
  Can now compute $\Phi(R_g)$ for each $R_g \in G$ in $O(1)$ time.



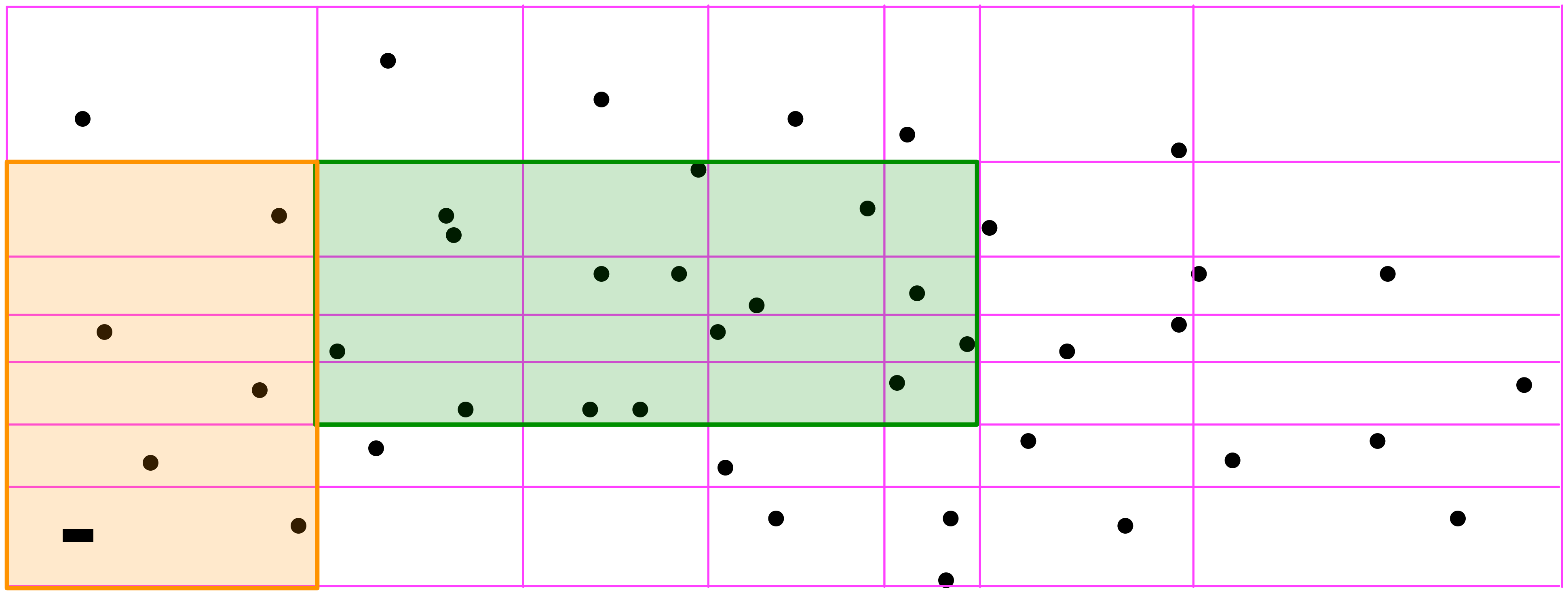
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Each rectangle $R \in (X, \mathcal{R}_2)$ is ε-approximated by on in $R_g \in G$
- Compute subset sum for all 2-sided rectangle in $1/\varepsilon^2$ time.
  Can now compute $\Phi(R_g)$ for each $R_g \in G$ in $O(1)$ time.



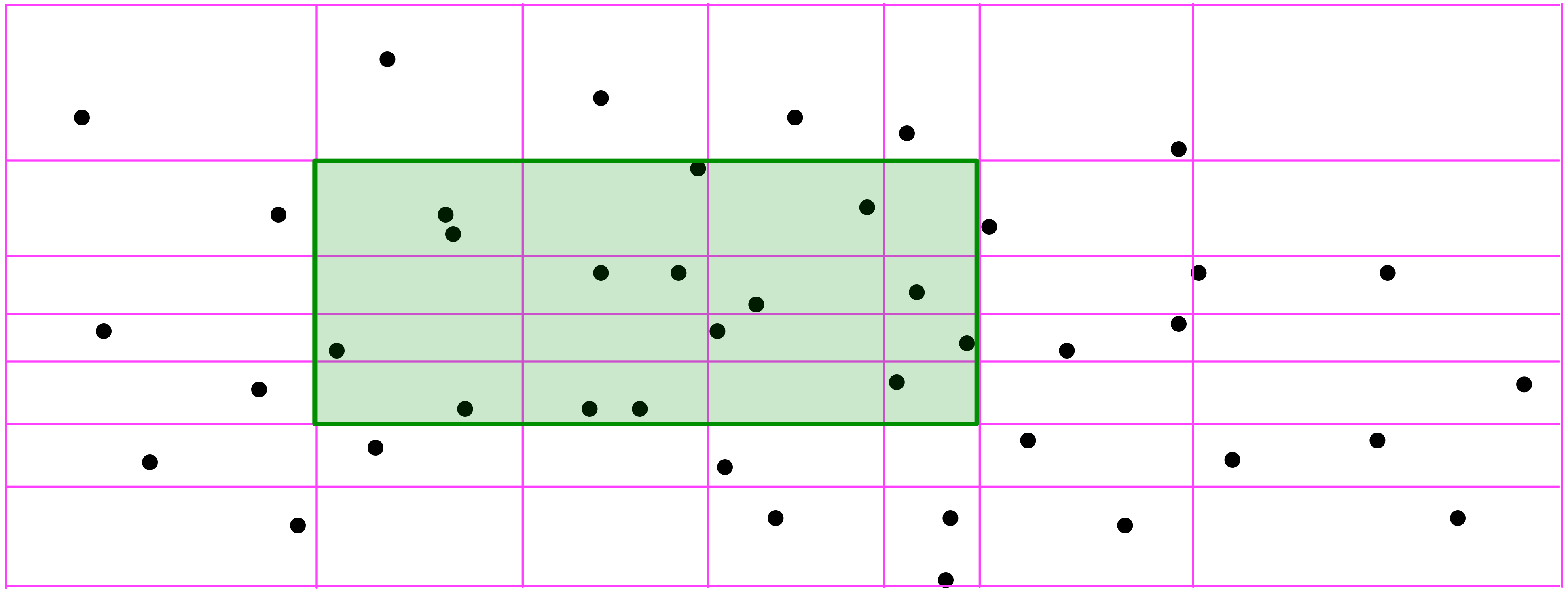
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Each rectangle $R \in (X, \mathcal{R}_2)$ is ε-approximated by on in $R_g \in G$
- Compute subset sum for all 2-sided rectangle in $1/\varepsilon^2$ time.
  Can now compute $\Phi(R_g)$ for each $R_g \in G$ in $O(1)$ time.



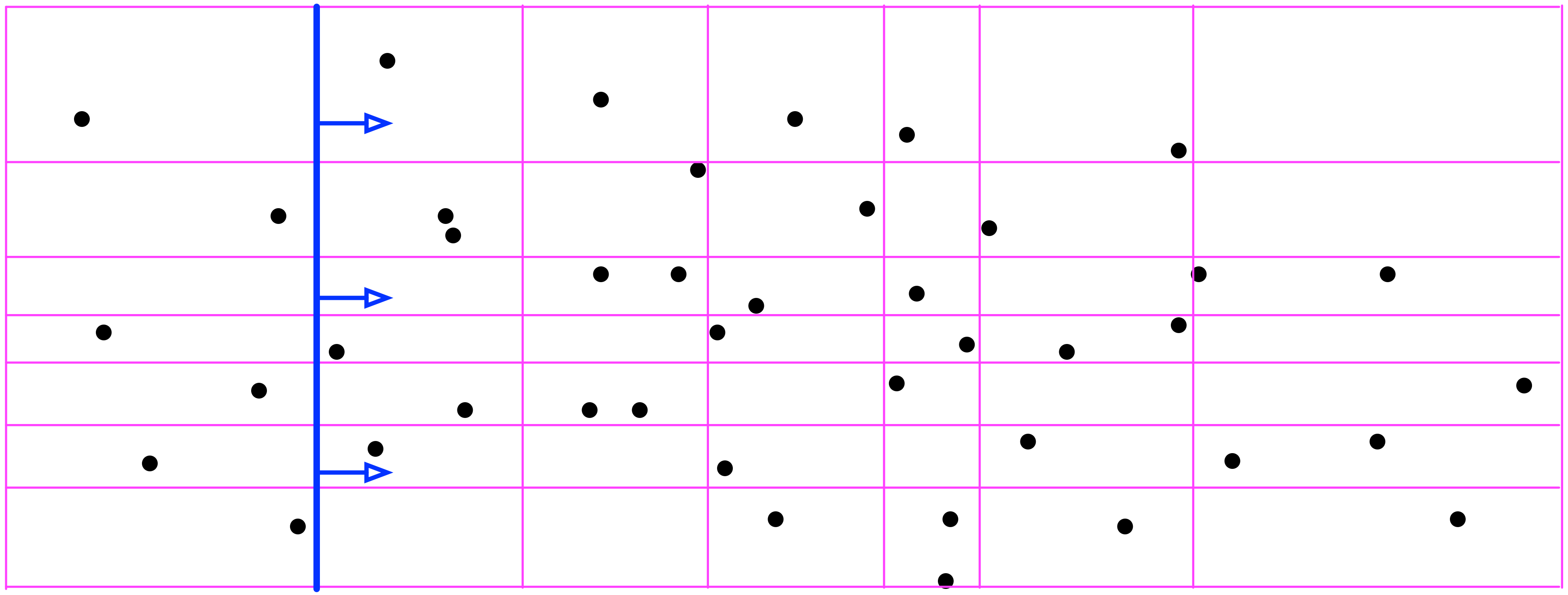
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Each rectangle $R \in (X, \mathcal{R}_2)$ is ε-approximated by on in $R_g \in G$
- Compute subset sum for all 2-sided rectangle in $1/\varepsilon^2$ time.
  Can now compute $\Phi(R_g)$ for each $R_g \in G$ in $O(1)$ time.
  Enumerate all in $O(1/\varepsilon^4)$ time.



Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$
  Consider all $1/\varepsilon$ left end points
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints
  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



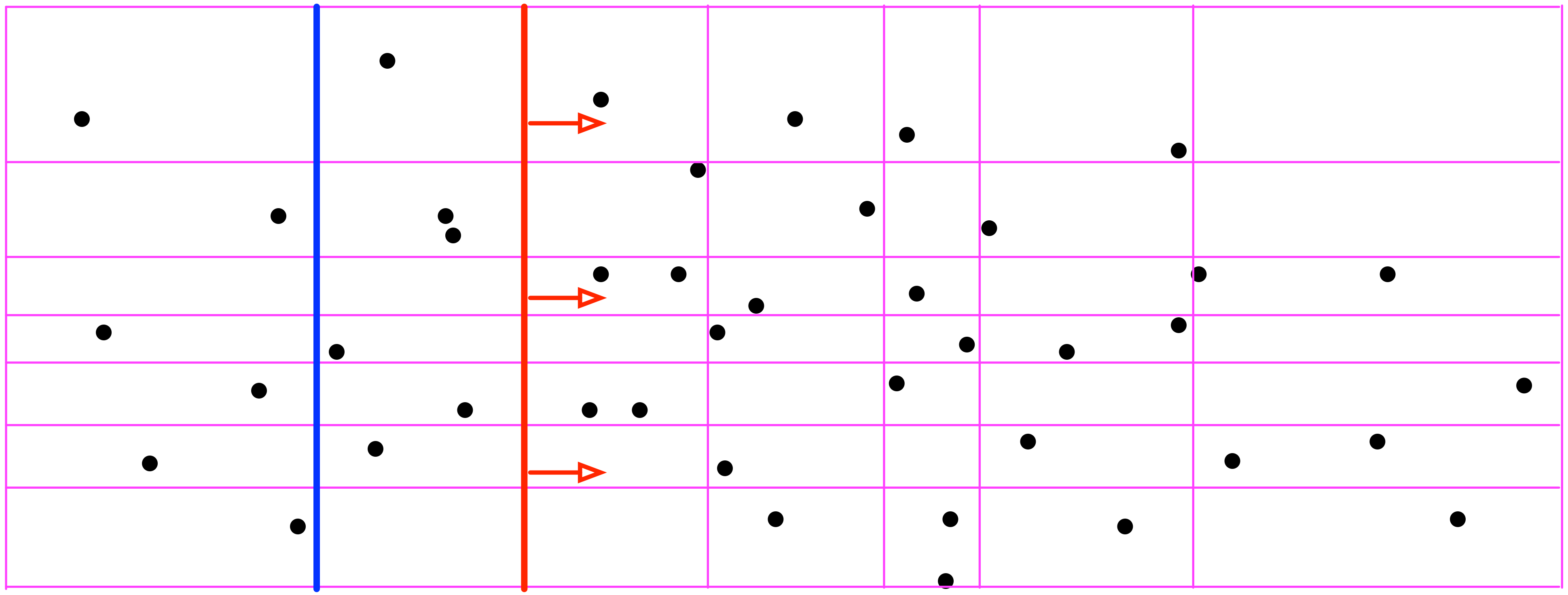
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$

  Consider all $1/\varepsilon$ left end points

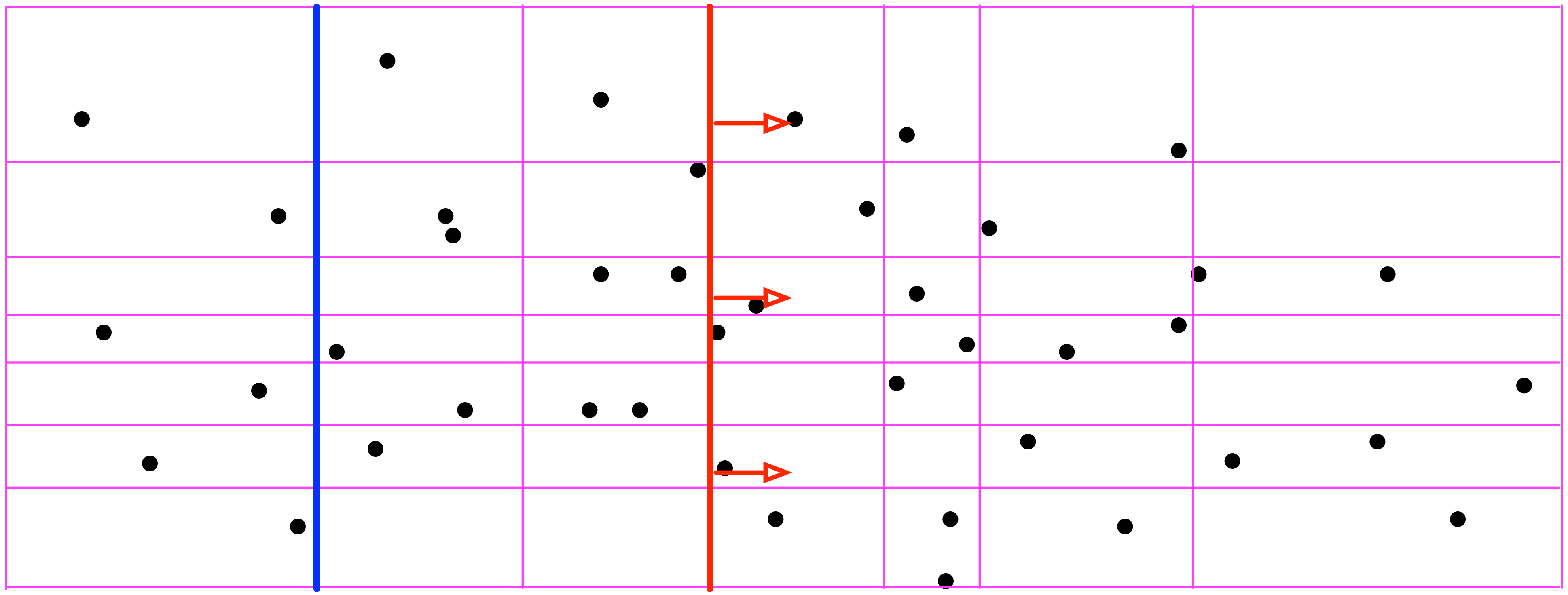
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints

  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$
  Consider all $1/\varepsilon$ left end points
  $\to$ Sweep $1/\varepsilon$ right endpoints
  $- \to$ Scan to calculate best vertical rect in $1/\varepsilon$ time



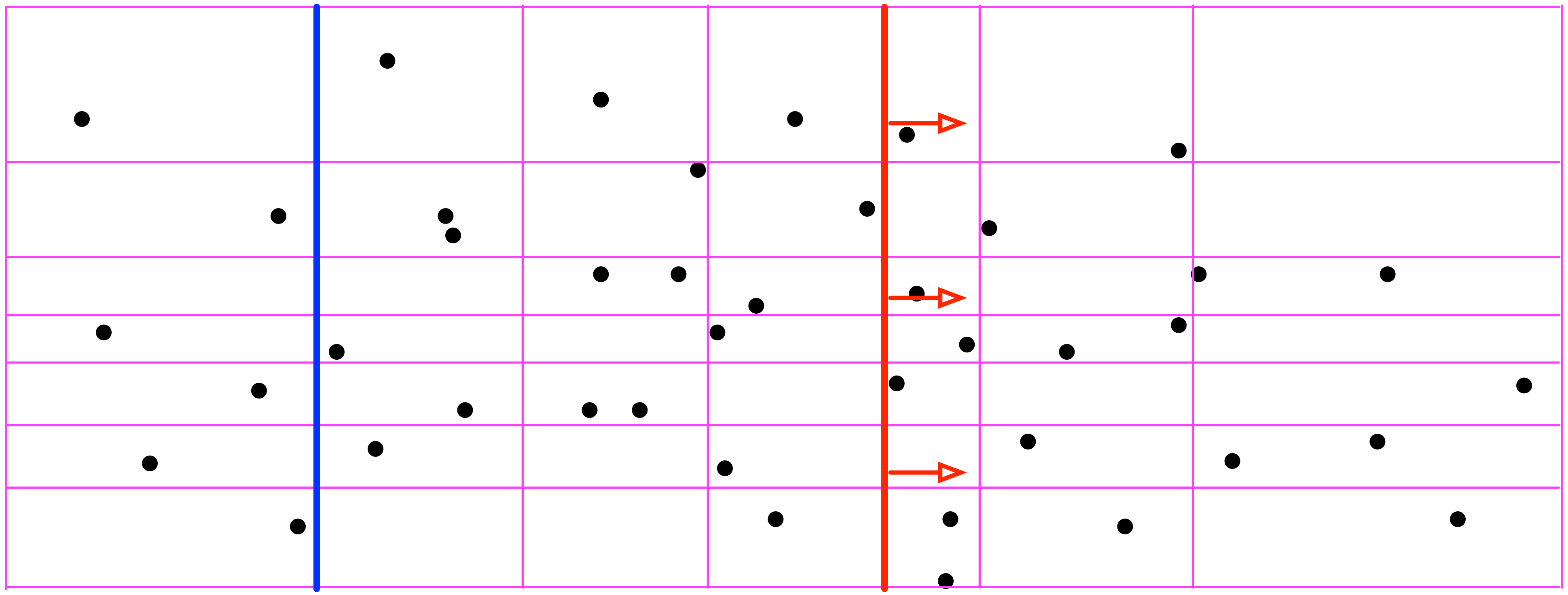
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$

  Consider all $1/\varepsilon$ left end points

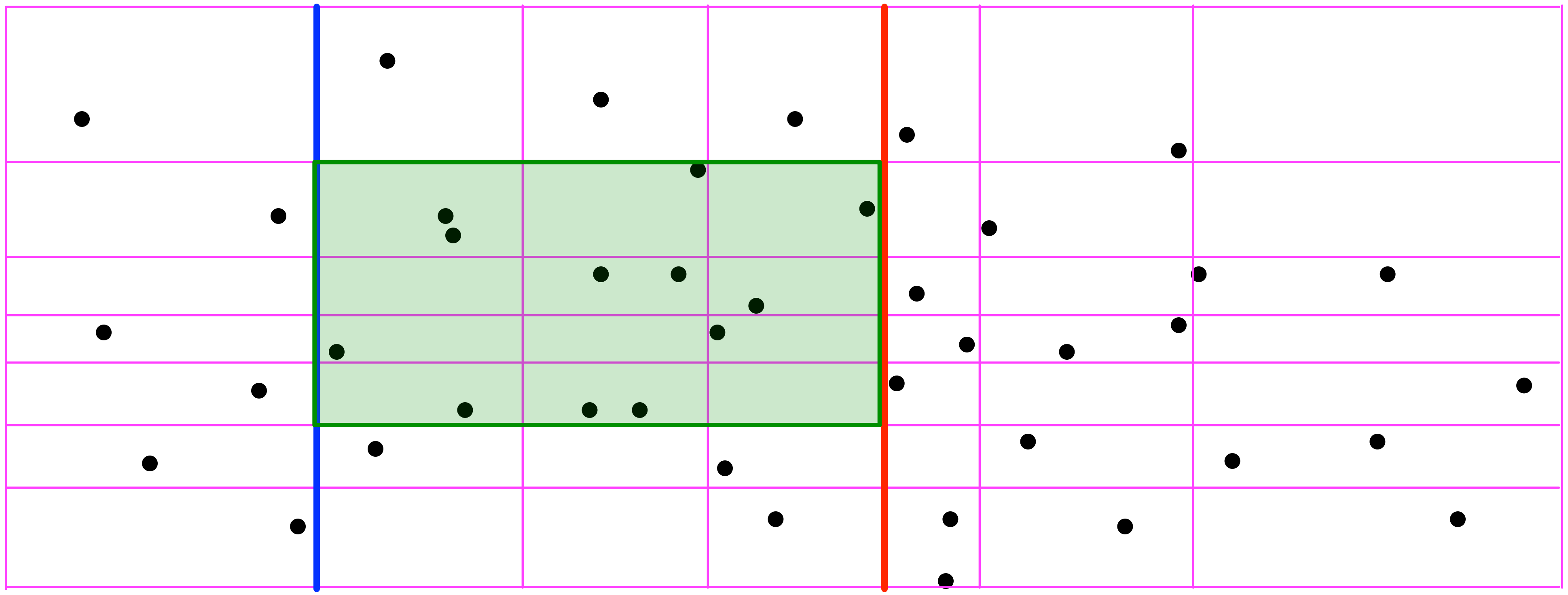
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints

  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



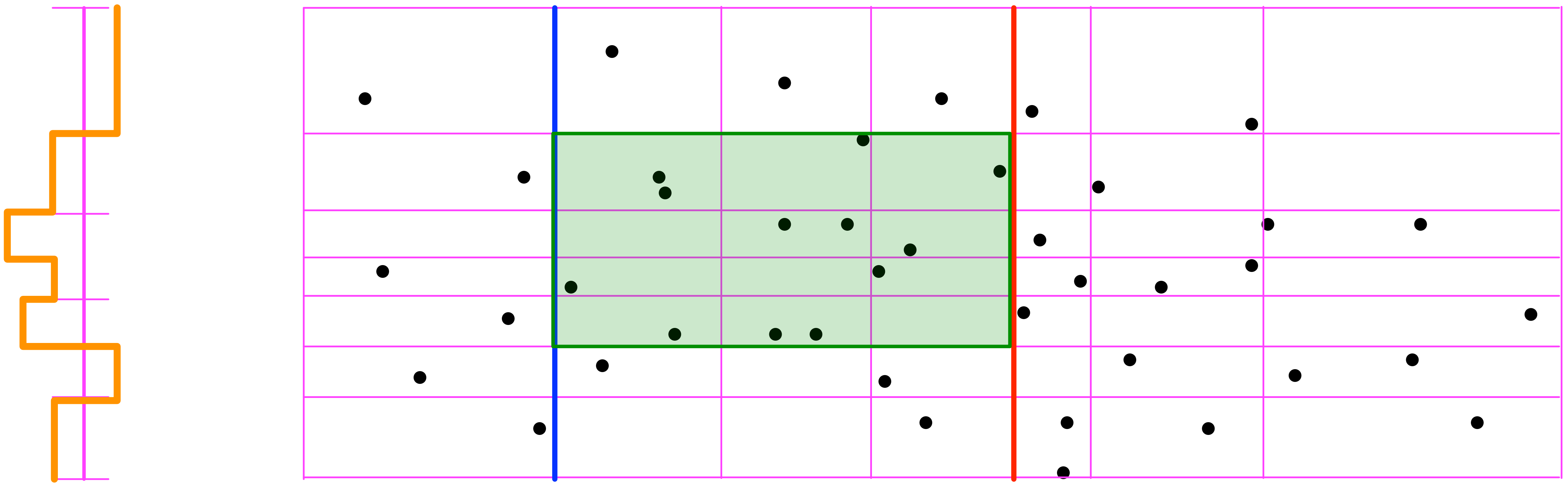
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$
  Consider all $1/\varepsilon$ left end points
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints
  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$
  Consider all $1/\varepsilon$ left end points
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints
  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



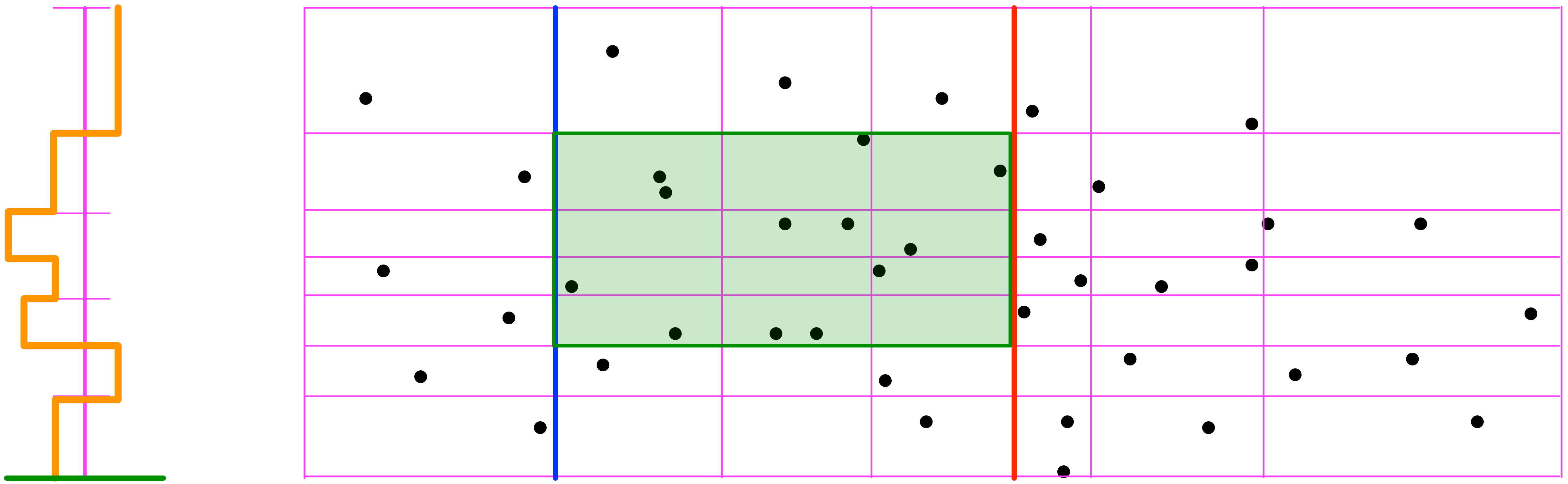
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$

  Consider all $1/\varepsilon$ left end points

  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints

  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



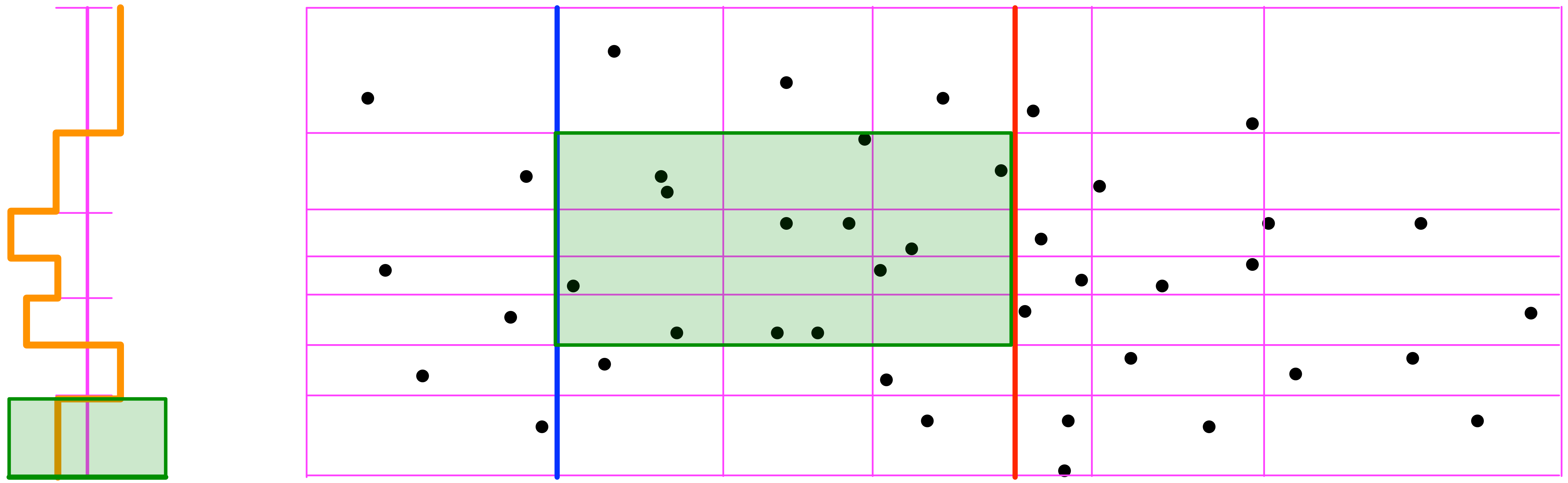
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$

  Consider all $1/\varepsilon$ left end points

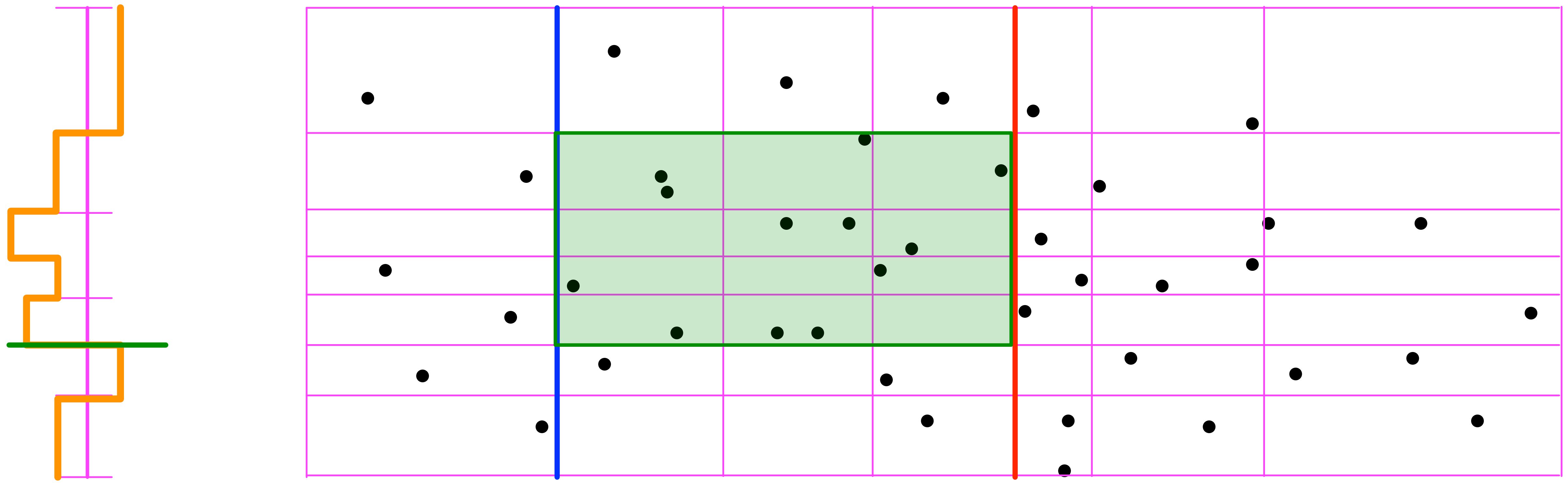
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints

  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$
  Consider all $1/\varepsilon$ left end points
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints
  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



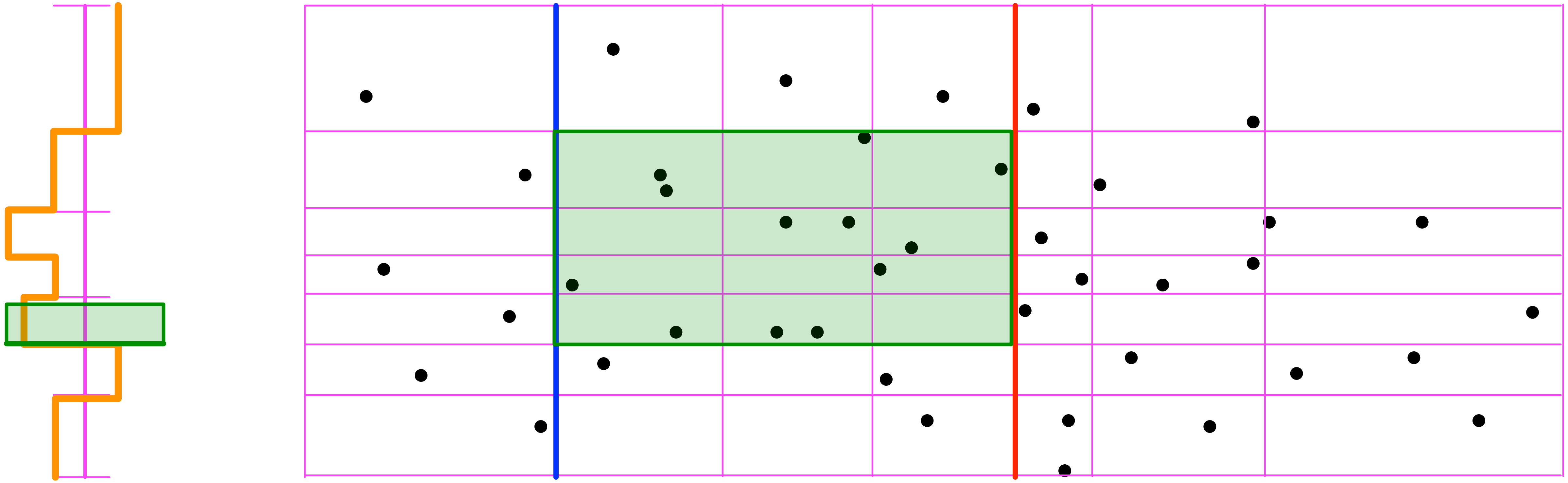
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$

  Consider all $1/\varepsilon$ left end points

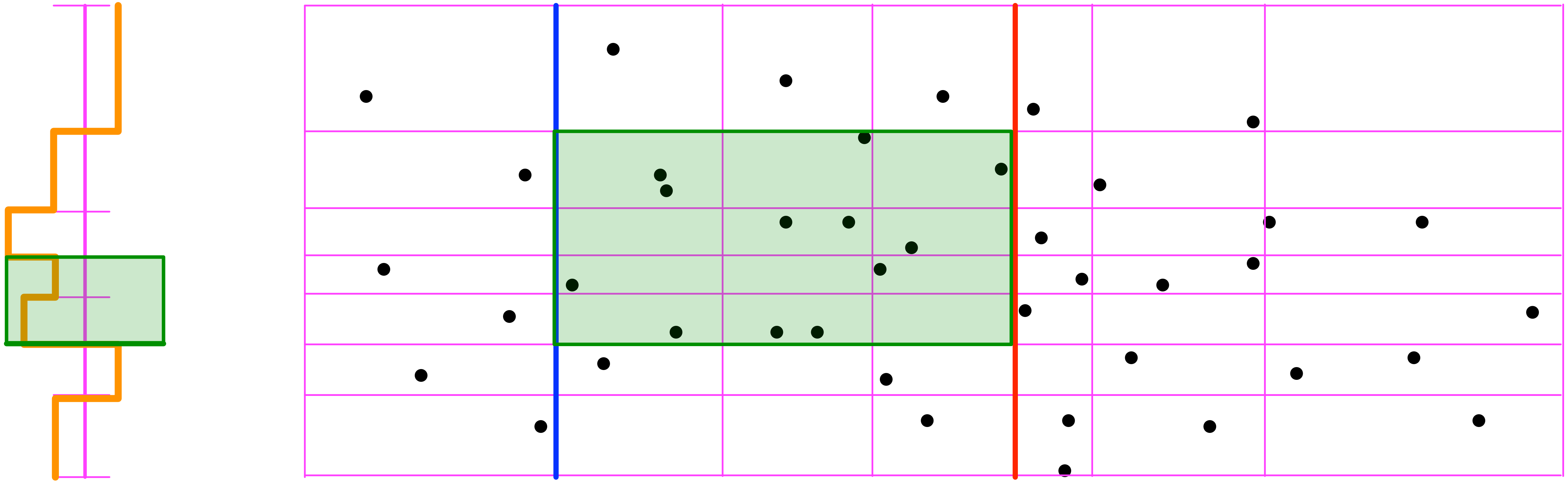
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints

  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$
  Consider all $1/\varepsilon$ left end points
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints
  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



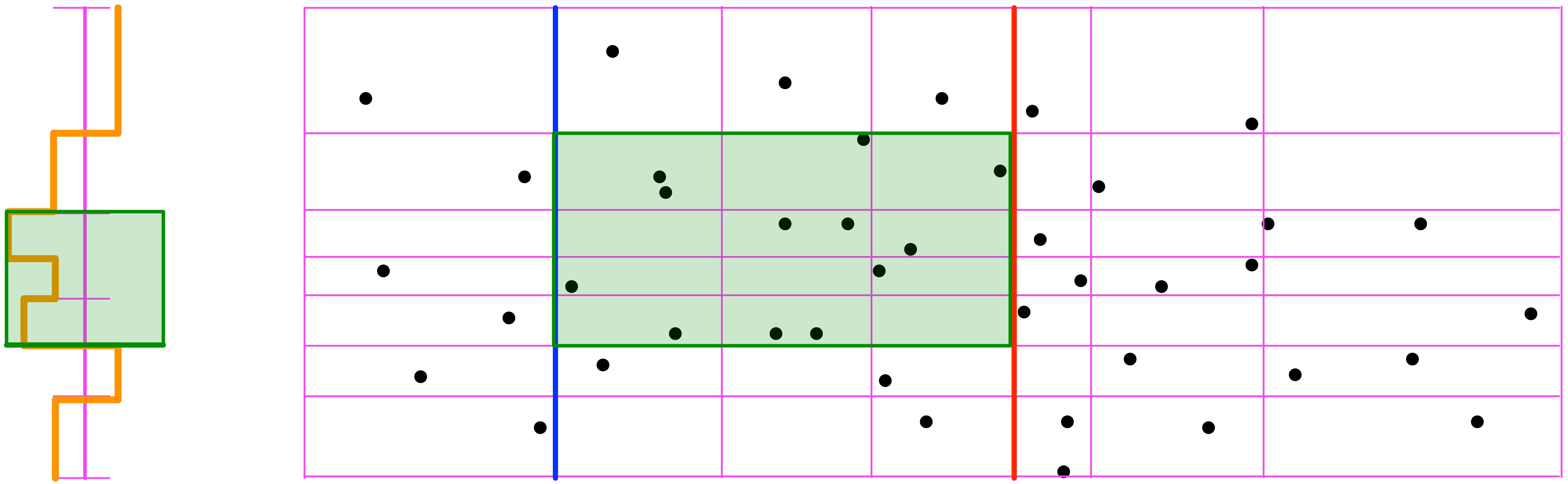
Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$

  Consider all $1/\varepsilon$ left end points

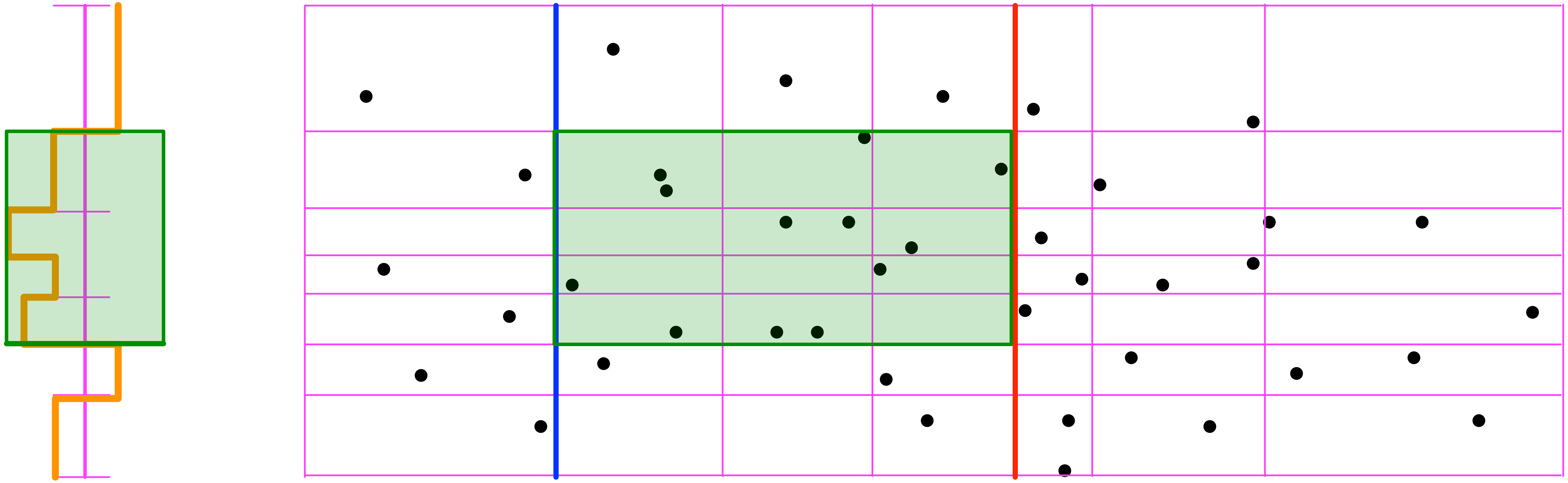
  $\to$ Sweep $1/\varepsilon$ right endpoints

  $- \to$ Scan to calculate best vertical rect in $1/\varepsilon$ time



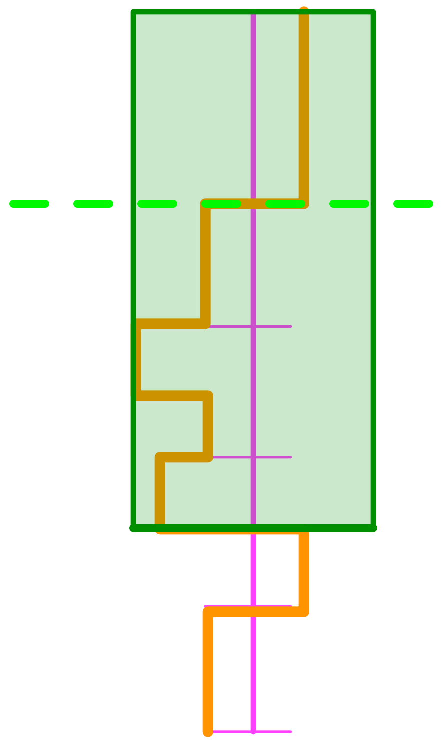
Approximate Rectangle Scanning

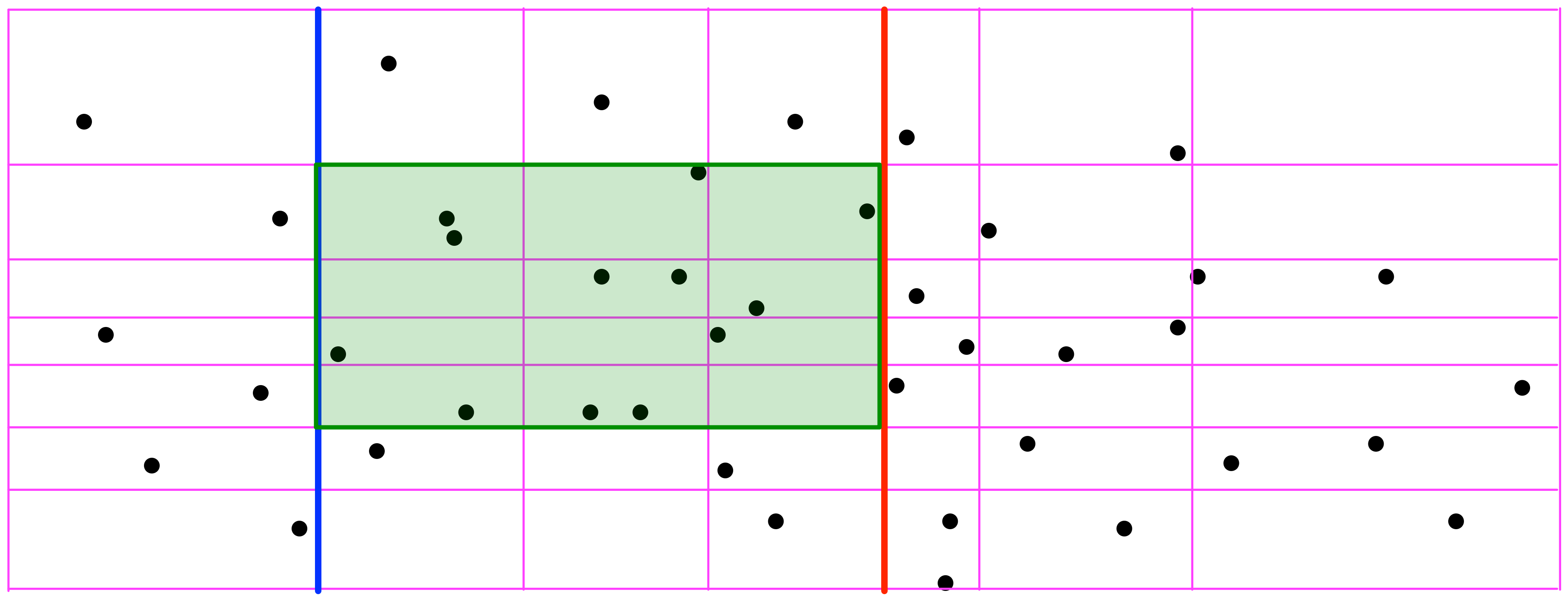
- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$
  Consider all $1/\varepsilon$ left end points
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints
  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



Approximate Rectangle Scanning

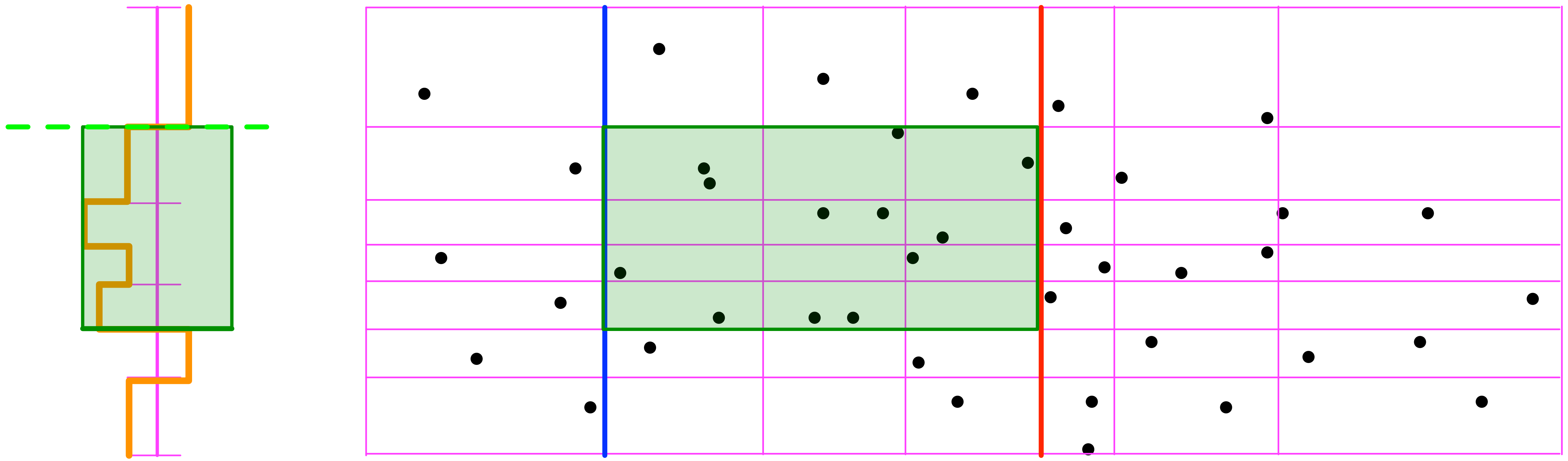
- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$
  Consider all $1/\varepsilon$ left end points
  $\rightarrow$ Sweep $1/\varepsilon$ right endpoints
  $- \rightarrow$ Scan to calculate best vertical rect in $1/\varepsilon$ time



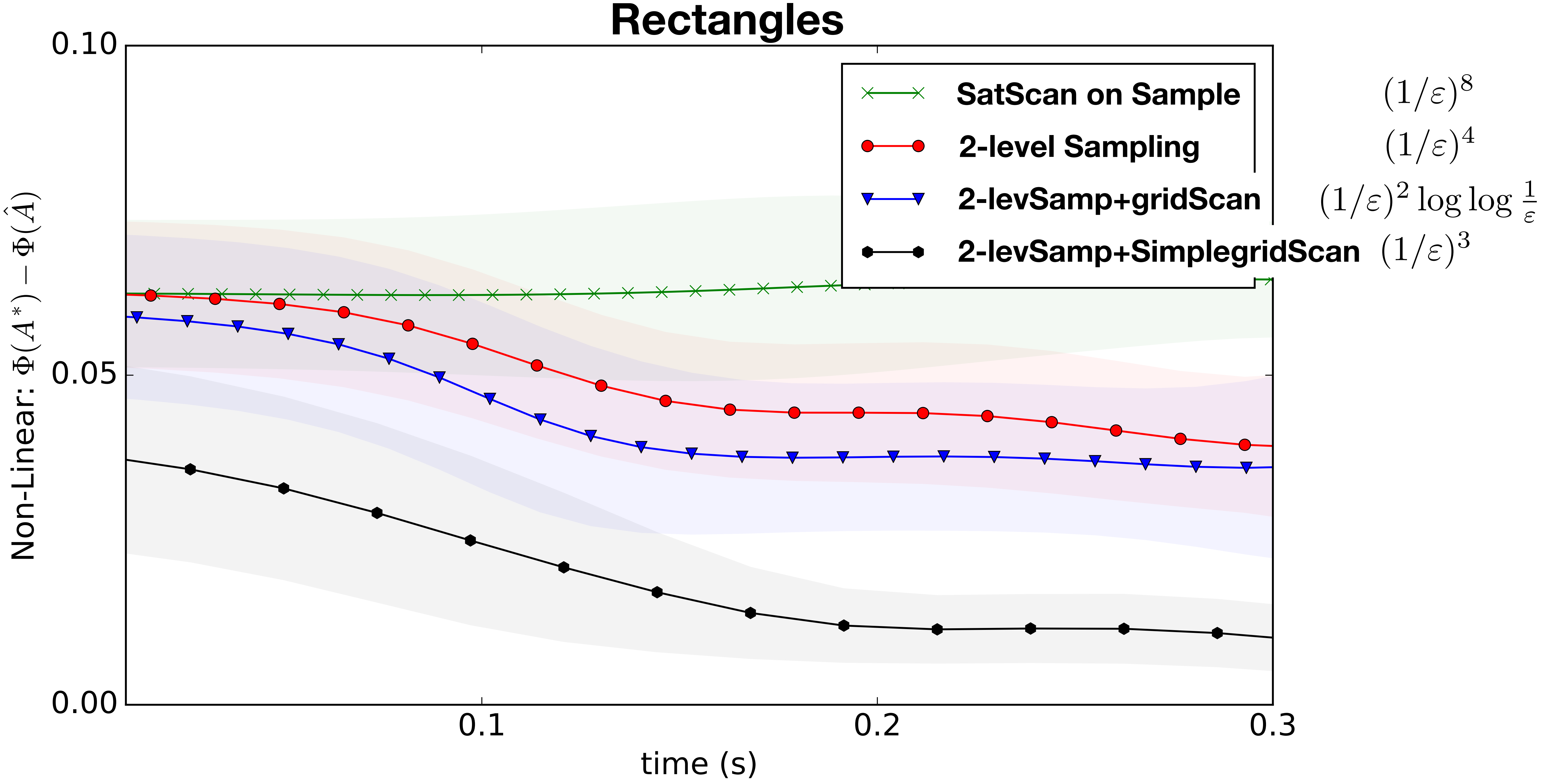


Approximate Rectangle Scanning

- ε-cover X with a grid G so each strip has $\approx \varepsilon|X|$ points.
- Run Kadane's Algorithm to find $R^* = \arg\max_{R_g \in G} \Phi(R_g)$
  Consider all $1/\varepsilon$ left end points
  $\to$ Sweep $1/\varepsilon$ right endpoints
  $- \to$ Scan to calculate best vertical rect in $1/\varepsilon$ time



Very Fast Rectangle Scanning



Conclusion

Michael Matheny

- $\mathbb{R}^2$: Halfspaces $O(N + 1/\varepsilon^{7/3})$,
  Rectangles in $O(N + 1/\varepsilon^2)$,
  Disks in $O(N + 1/\varepsilon^{10/3})$.
- Rectangles conditionally tight (APSP). Conjecture Halfspaces $\Theta(N + 1/\varepsilon^2)$.
- Code online: `pyscan` | https://github.com/michaelmathen/pyscan