

Quantitative Input Feature Usage

Denis Mazzucato, Marco Campion, Caterina Urban

École Normale Supérieure, Inria

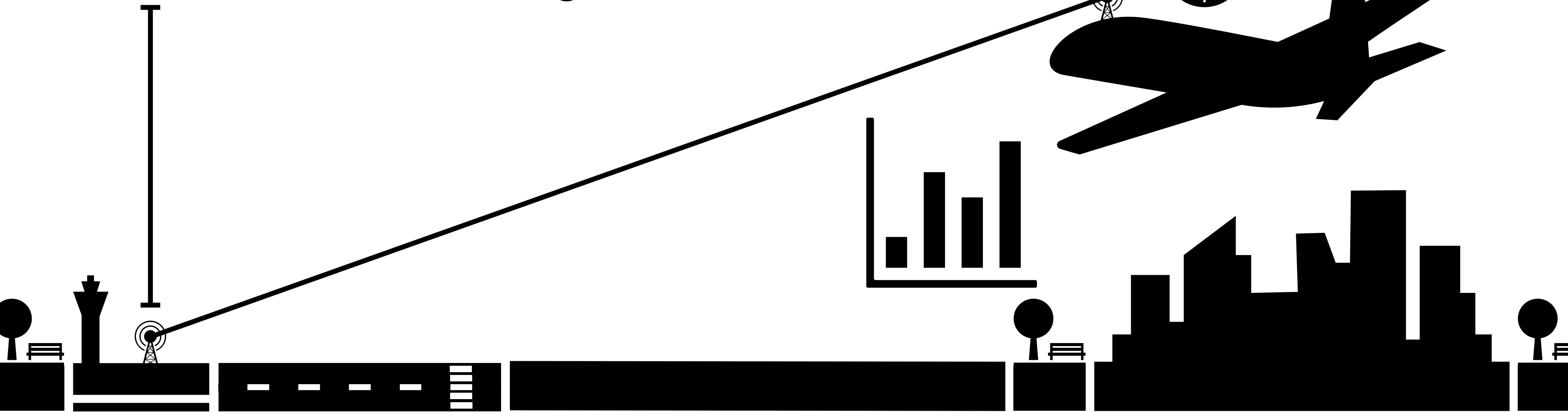
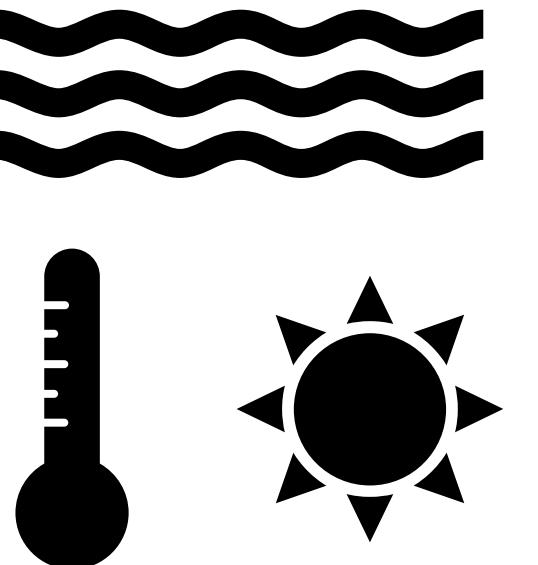
25th May 2023 — Challenges of Software Verification, Venice



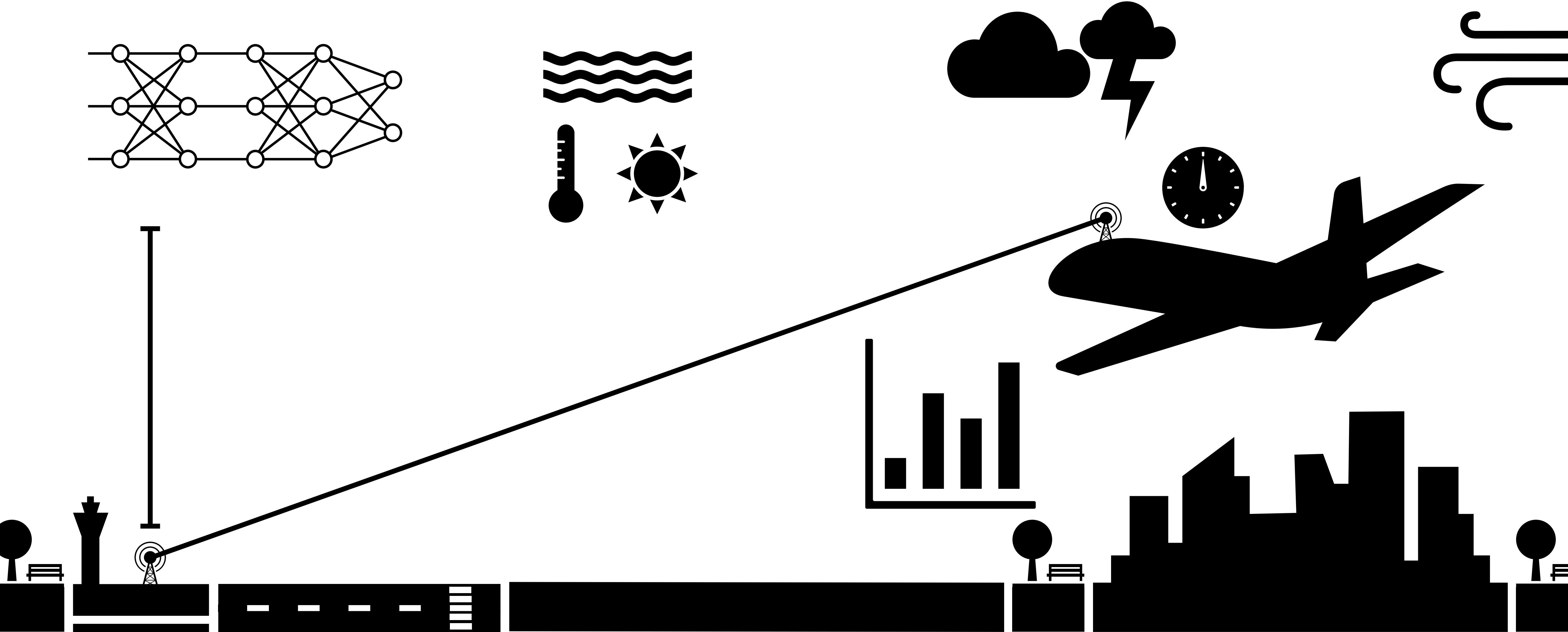
Landing alarm system



Landing alarm system



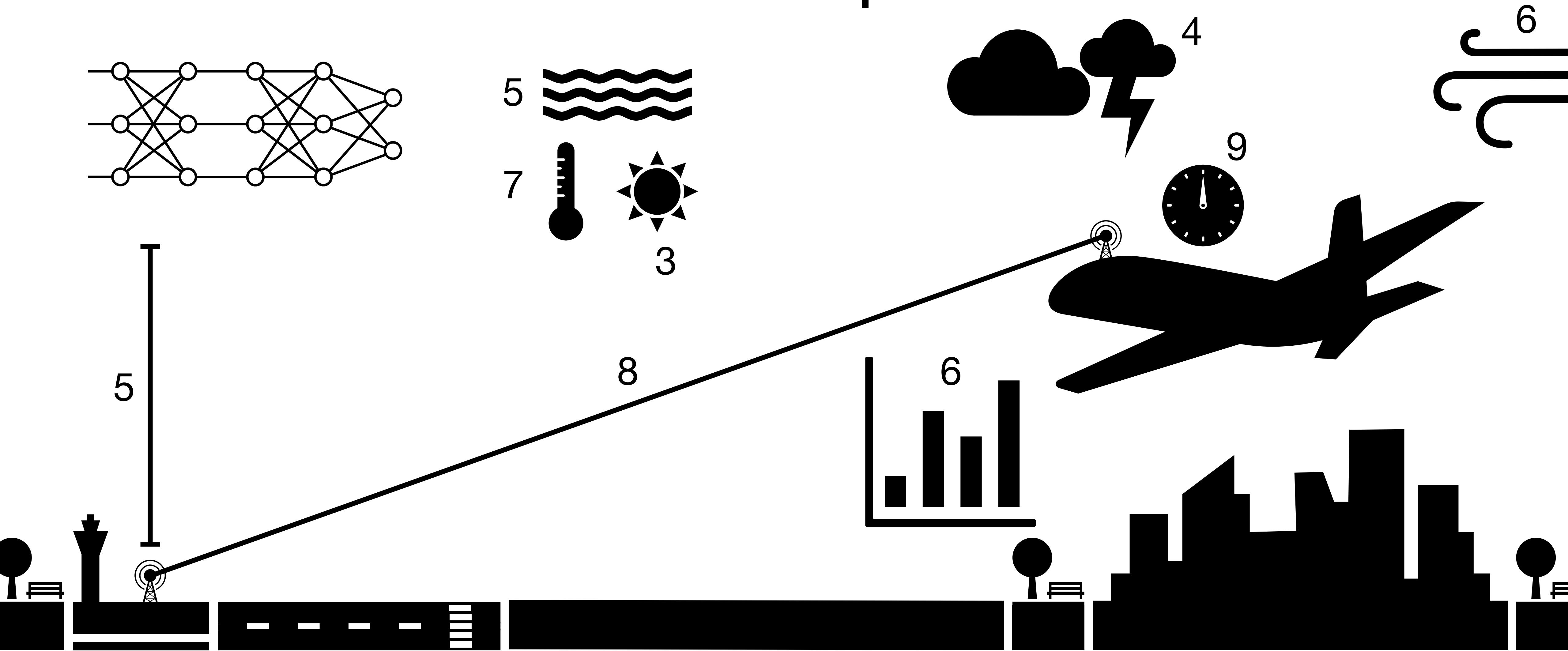
Landing alarm system



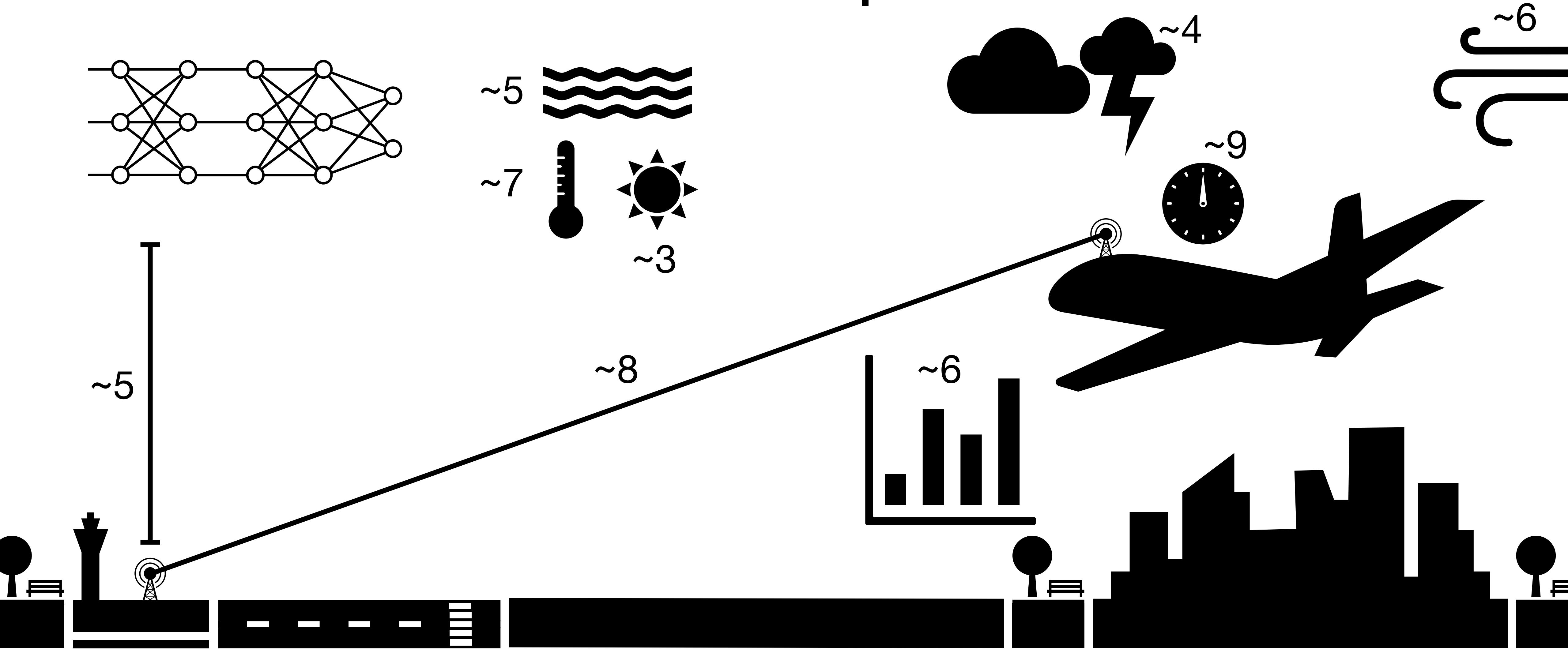
Landing alarm system



Stochastic methods: feature importance

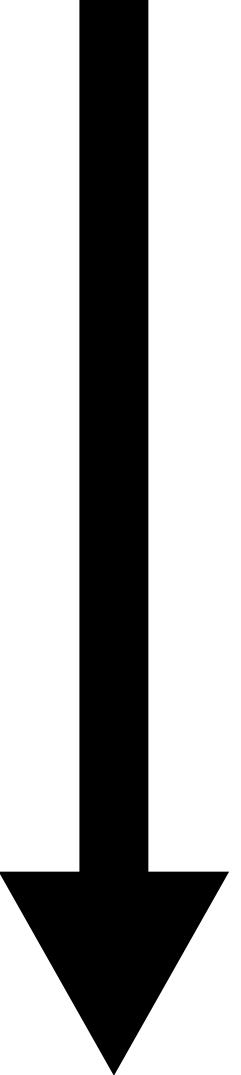


Stochastic methods: feature importance



Formal methods: quantitative analysis





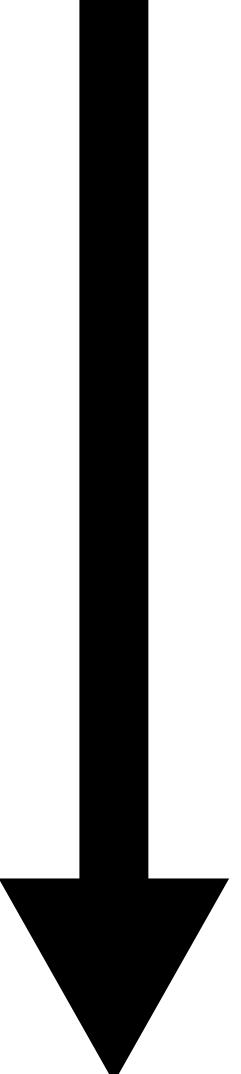
Qualitative

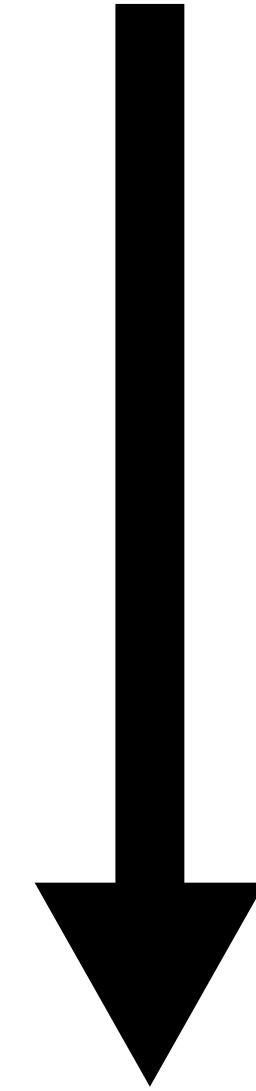
Is the i -th input feature of P unused?

} [Urban18]

Quantitative

How much is the i -th input feature of P used?

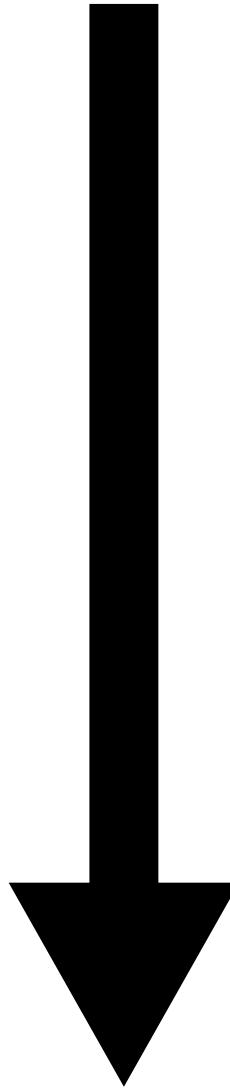




Qualitative
Is the i -th input feature of P unused?

} [Urban18]

$$P \models \mathbb{U}_i$$

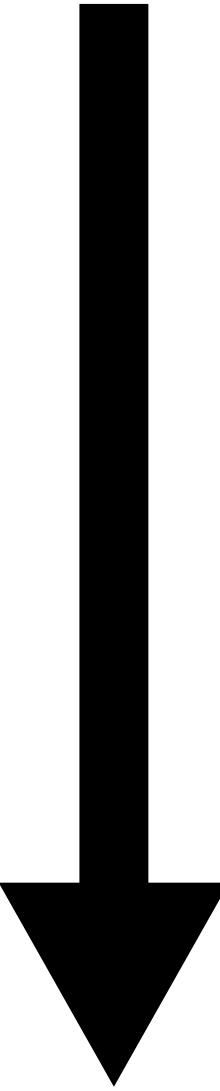


Qualitative
Is the i -th input feature of P unused?

} [Urban18]

$$P \models \mathbb{U}_i$$

$\mathbb{U}_i = \{ \text{program } S \mid S \text{ does not use the input feature } i \}$



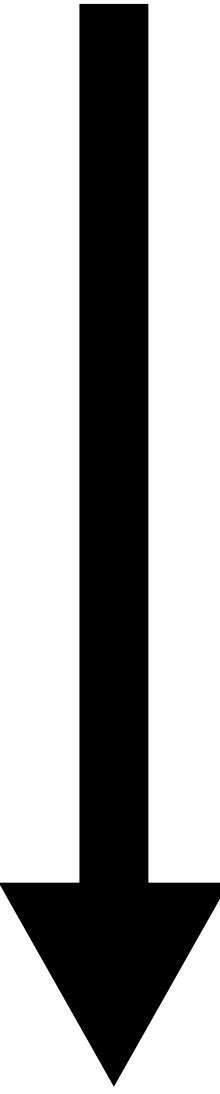
Qualitative

Is the i -th input feature of P unused?

} [Urban18]

$$P \models \mathbb{U}_i$$

$$\mathbb{U}_i = \{ \quad \llbracket S \rrbracket \quad | \quad S \text{ does not use the input feature } i \}$$



Qualitative

Is the i -th input feature of P unused?

} [Urban18]

$$P \models \mathbb{U}_i$$

$$\mathbb{U}_i = \{ \quad [\![S]\!] \quad | \quad \text{unused}_i([\![S]\!]) \}$$



Qualitative
Is the i -th input feature of P unused?

} [Urban18]

$$P \models \mathbb{U}_i$$

$$\mathbb{U}_i = \{ \llbracket S \rrbracket \mid \text{unused}_i(\llbracket S \rrbracket) \}$$



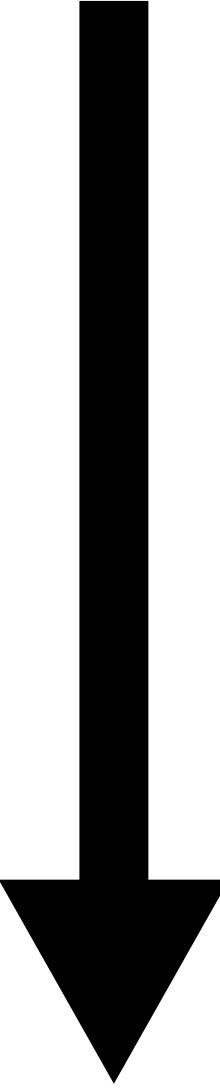
Qualitative
Is the i -th input feature of P unused?

} [Urban18]

$$P \models \mathbb{U}_i$$

$$\mathbb{U}_i = \{ \llbracket S \rrbracket \mid \text{unused}_i(\llbracket S \rrbracket) \}$$

$$P \models \mathbb{U}_i \iff \{\llbracket P \rrbracket\} \subseteq \mathbb{U}_i$$



Qualitative

Is the i -th input feature of P unused?

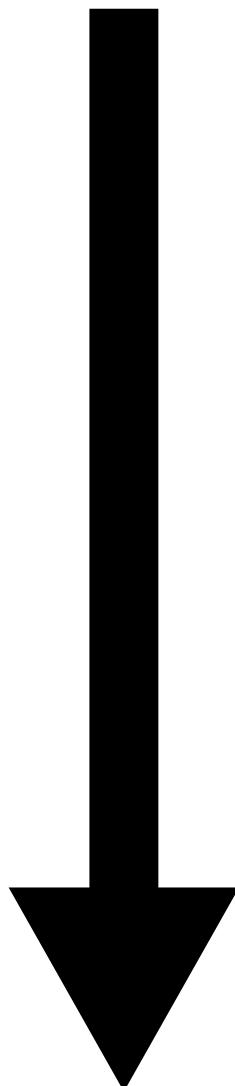
} [Urban18]

$$P \models \mathbb{U}_i$$

$$\mathbb{U}_i = \{ \llbracket S \rrbracket \mid \text{unused}_i(\llbracket S \rrbracket) \}$$

$$P \models \mathbb{U}_i \iff \{\llbracket P \rrbracket\} \subseteq \mathbb{U}_i \iff \gamma(\llbracket P \rrbracket^\natural) \subseteq \mathbb{U}_i$$

$$P \models Q_i$$

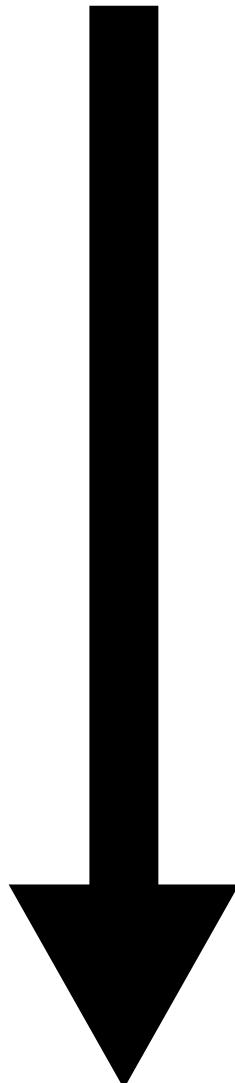


Quantitative

How much is the i -th input feature of P used?

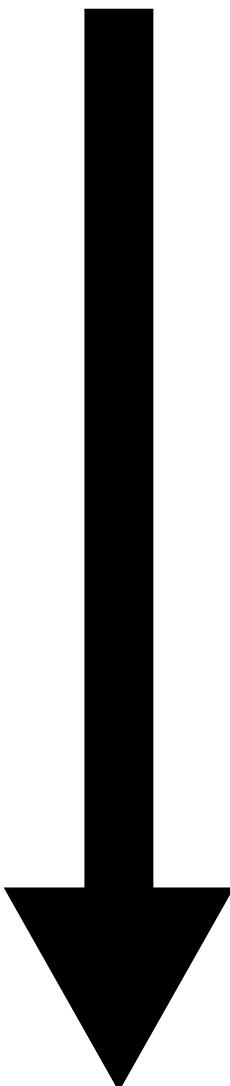
$$P \models \mathbb{Q}_i$$

$$\mathbb{Q}_i = \{ \llbracket S \rrbracket \mid \text{impact}_i(\llbracket S \rrbracket) \}$$



Quantitative

How much is the i -th input feature of P used?

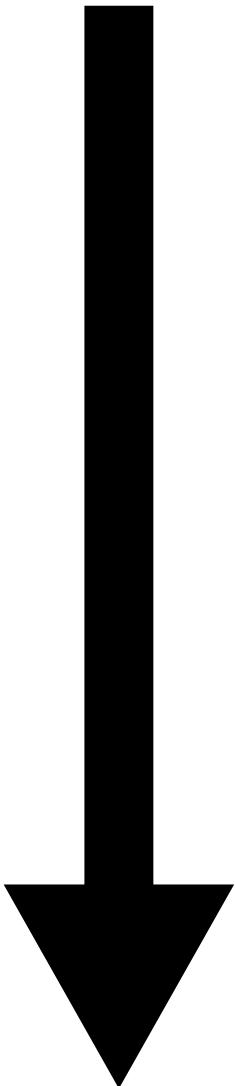
$P \models \mathbb{Q}_i$ $\text{impact}_i \in \text{Traces} \rightarrow \mathbb{R}$ $\mathbb{Q}_i = \{ \llbracket S \rrbracket \mid \text{impact}_i(\llbracket S \rrbracket) \}$ 

Quantitative

How much is the i -th input feature of P used?

$$P \models \mathbb{Q}_i^k \quad \text{impact}_i \in \text{Traces} \rightarrow \mathbb{R}$$

$$\mathbb{Q}_i^k = \{ \llbracket S \rrbracket \mid \text{impact}_i(\llbracket S \rrbracket) \leq k \}$$



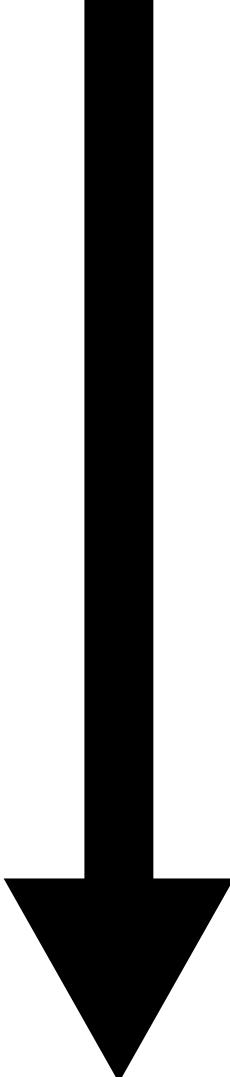
Quantitative

How much is the i -th input feature of P used?

$$P \models \mathbb{Q}_i^k \quad \text{impact}_i \in \text{Traces} \rightarrow \mathbb{R}$$

$$\mathbb{Q}_i^k = \{ \llbracket S \rrbracket \mid \text{impact}_i(\llbracket S \rrbracket) \leq k \}$$

$$P \models \mathbb{Q}_i^k \iff \{\llbracket P \rrbracket\} \subseteq \mathbb{Q}_i^k \iff \gamma(\llbracket P \rrbracket^\natural) \subseteq \mathbb{Q}_i^k$$



Quantitative

How much is the i -th input feature of P used?

$$P \models \mathbb{U}_i$$

$$\mathbb{U}_i = \{ \llbracket S \rrbracket \mid \text{unused}_i(\llbracket S \rrbracket) \}$$

$$P \models \mathbb{U}_i \iff \{\llbracket P \rrbracket\} \subseteq \mathbb{U}_i \iff \gamma(\llbracket P \rrbracket^\natural) \subseteq \mathbb{U}_i$$

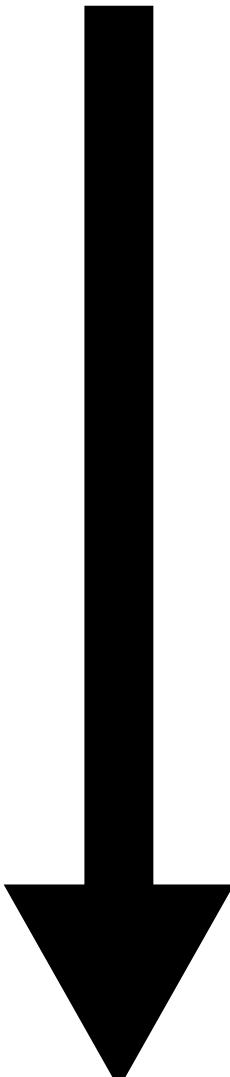
cato



Traces $\rightarrow \mathbb{R}$

$$\llbracket S \rrbracket) \leq k \}$$

$$P \models \mathbb{Q}_i^k \iff \{\llbracket P \rrbracket\} \subseteq \mathbb{Q}_i^k \iff \gamma(\llbracket P \rrbracket^\natural) \subseteq \mathbb{Q}_i^k$$



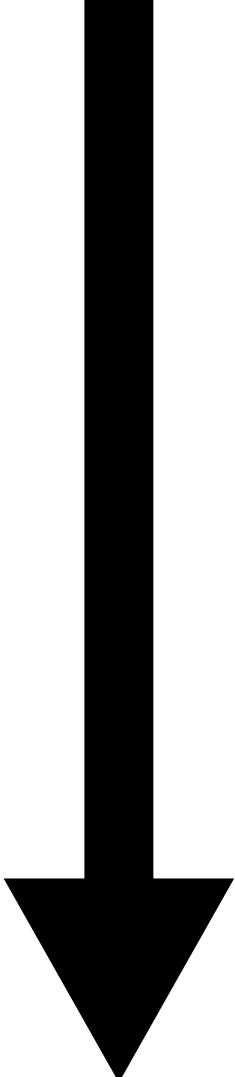
Quantitative

How much is the i -th input feature of P used?

$$P \models \mathbb{Q}_i^k \quad \text{impact}_i \in \text{Traces} \rightarrow \mathbb{R}$$

$$\mathbb{Q}_i^k = \{ \llbracket S \rrbracket \mid \text{impact}_i(\llbracket S \rrbracket) \leq k \}$$

$$P \models \mathbb{Q}_i^k \iff \{\llbracket P \rrbracket\} \subseteq \mathbb{Q}_i^k \iff \gamma(\llbracket P \rrbracket^\natural) \subseteq \mathbb{Q}_i^k$$



Quantitative

How much is the i -th input feature of P used?

$$P \models Q_i^k$$

$$\text{impact}_i \in \text{Traces} \rightarrow \mathbb{D}$$

$$Q_i^k = \{ \llbracket S \rrbracket \mid \text{impact}_i(\llbracket S \rrbracket) \leq k \}$$

Quantities
Domain

$$P \models Q_i^k \iff \{\llbracket P \rrbracket\} \subseteq Q_i^k \iff \gamma(\llbracket P \rrbracket^\natural) \subseteq Q_i^k$$

Quantitative

How much is the i -th input feature of P used?



How to define “ $\text{impact}_i \in \text{Traces} \rightarrow \mathbb{D}$ ”?

How to define “ $\text{impact}_i \in \text{Traces} \rightarrow \mathbb{D}$ ”?

The number of output changes
(with repetitions)

result from perturbations on the i -th input feature
for any input X

How to define “ $\text{impact}_i \in \text{Traces} \rightarrow \mathbb{D}$ ”?

The number of **output changes**
(with repetitions)

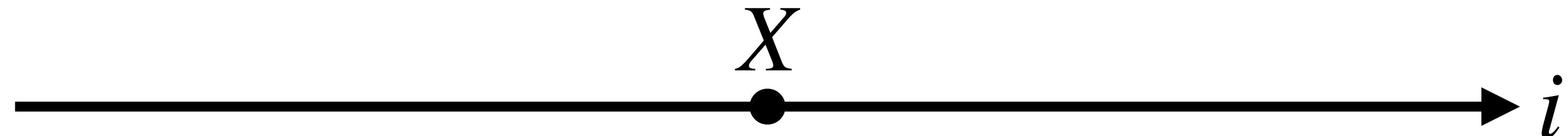
result from **perturbations** on the i -th input feature
for any input X

$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$

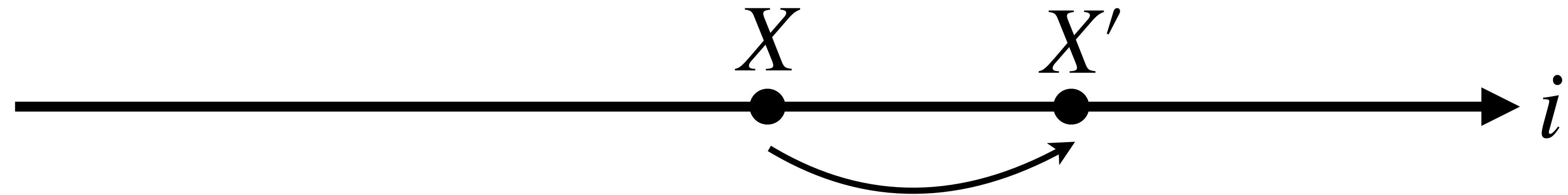
$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$



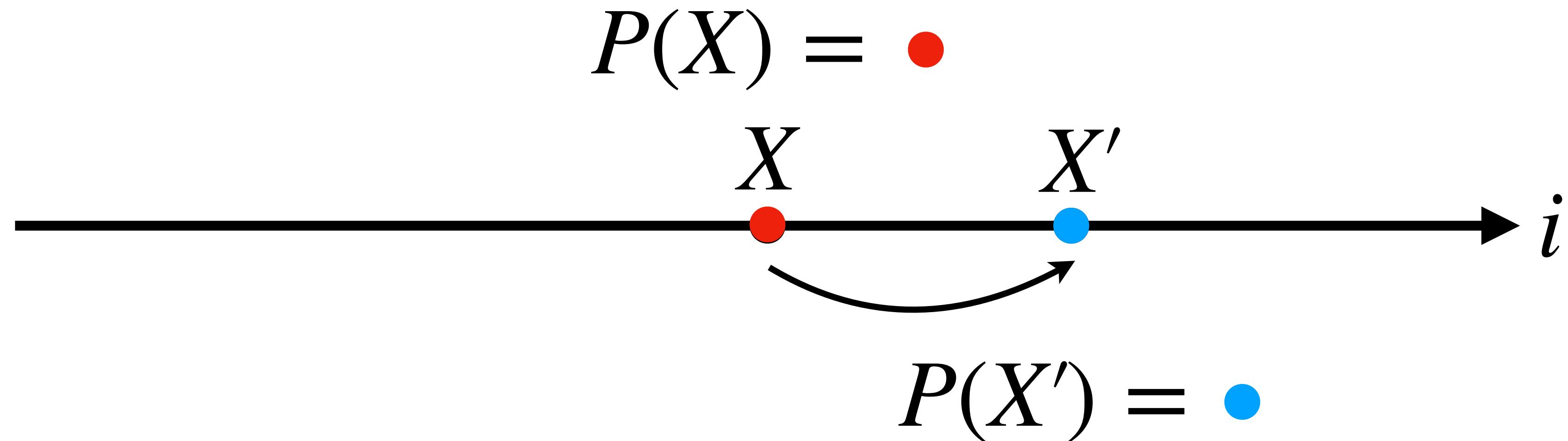
$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$



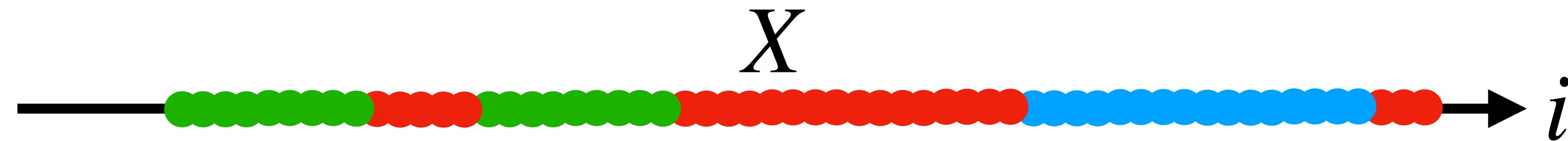
$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$



$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$

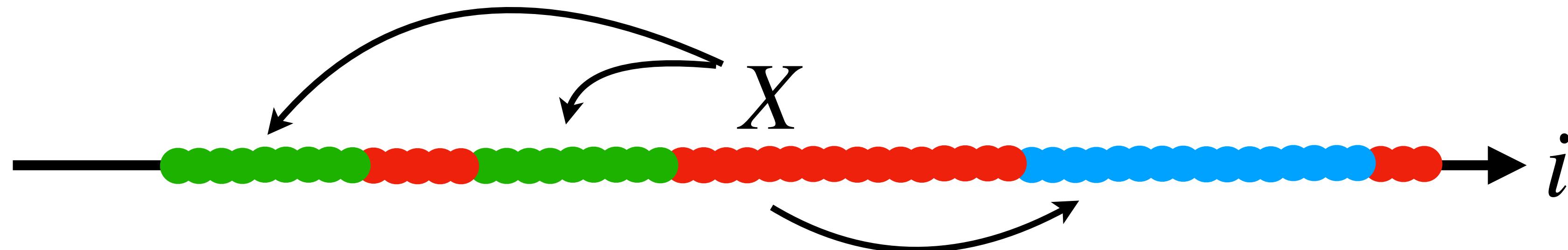


$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$



$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$

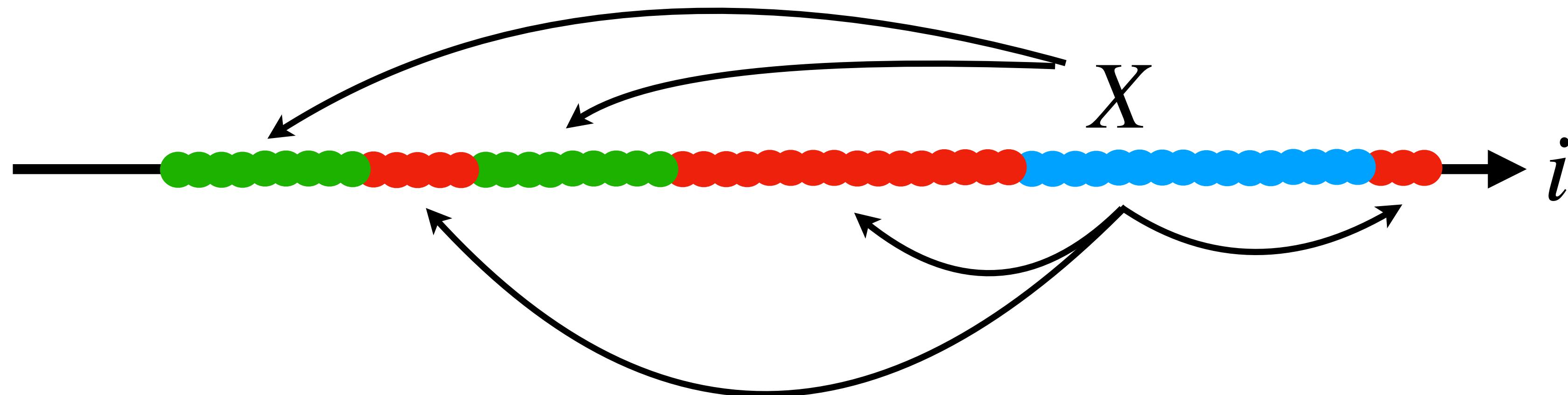
starting from X leading to \bullet $\implies 3$ changes



$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$

starting from X leading to \bullet $\Rightarrow 3$ changes

starting from X leading to \bullet $\Rightarrow 5$ changes

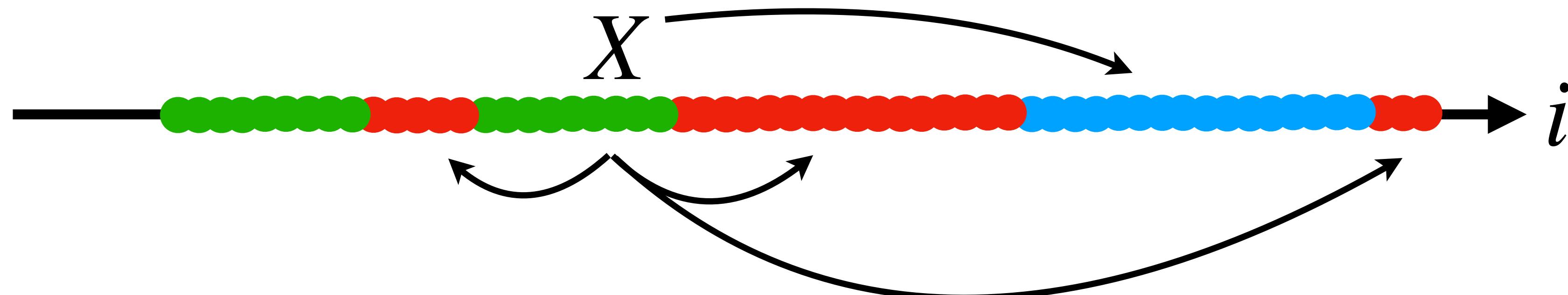


$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$

starting from X leading to \bullet $\Rightarrow 3$ changes

starting from X leading to \circ $\Rightarrow 5$ changes

starting from X leading to \bullet $\Rightarrow 4$ changes



$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$

starting from X leading to \bullet $\Rightarrow 3$ changes

starting from X leading to \bullet $\Rightarrow \underline{5 \text{ changes}}$

starting from X leading to \bullet $\Rightarrow 4$ changes



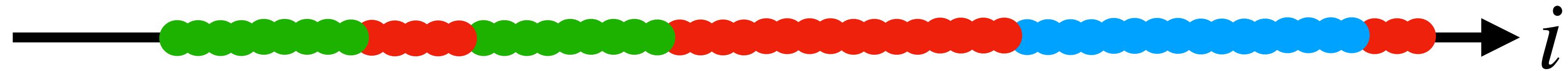
$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$

$\text{CountChanges}_i(P) = 5$



$\text{CountChanges}_i \in \text{Traces} \rightarrow \mathbb{N}^\infty$

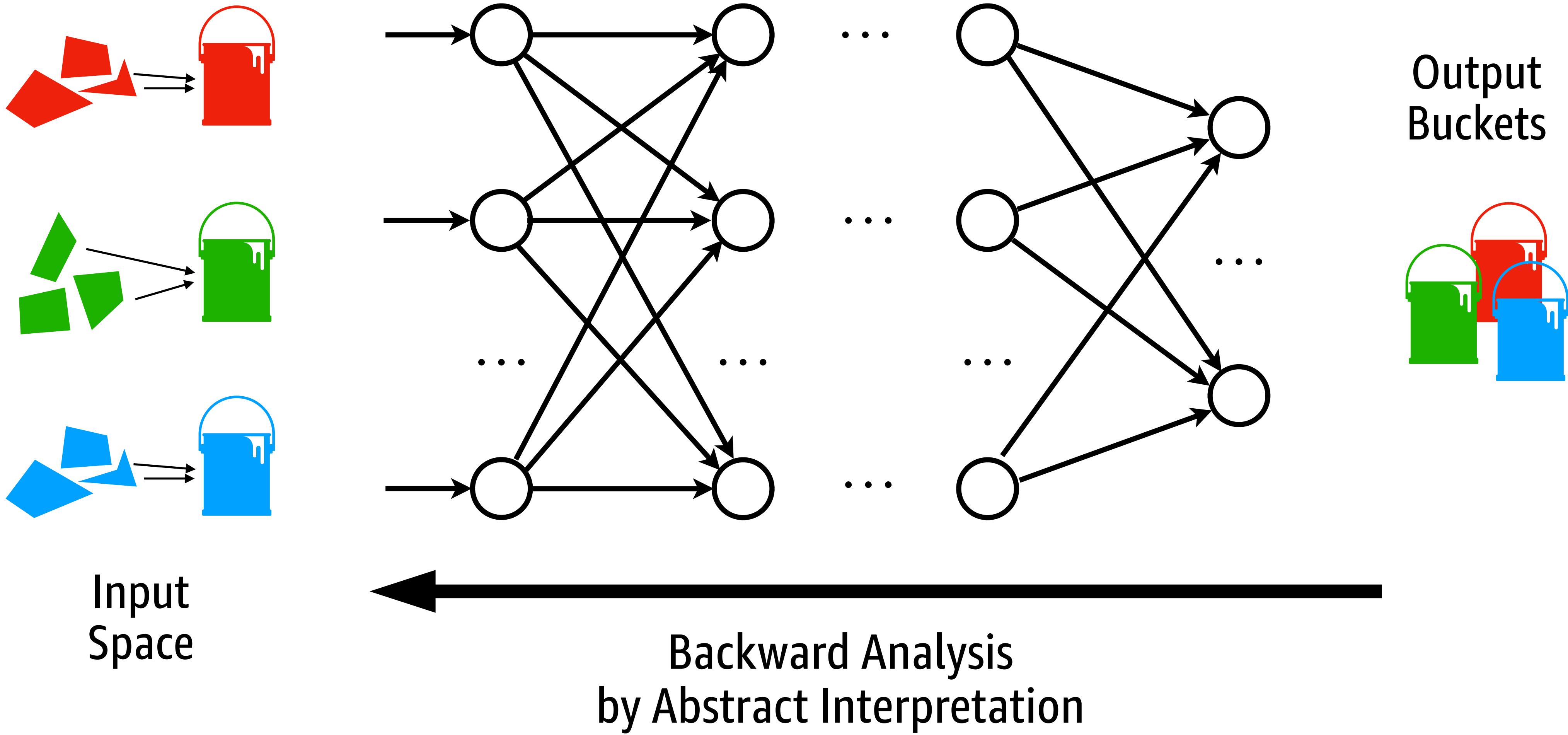
$\text{CountChanges}_i(P) = 5$



Continuous input space

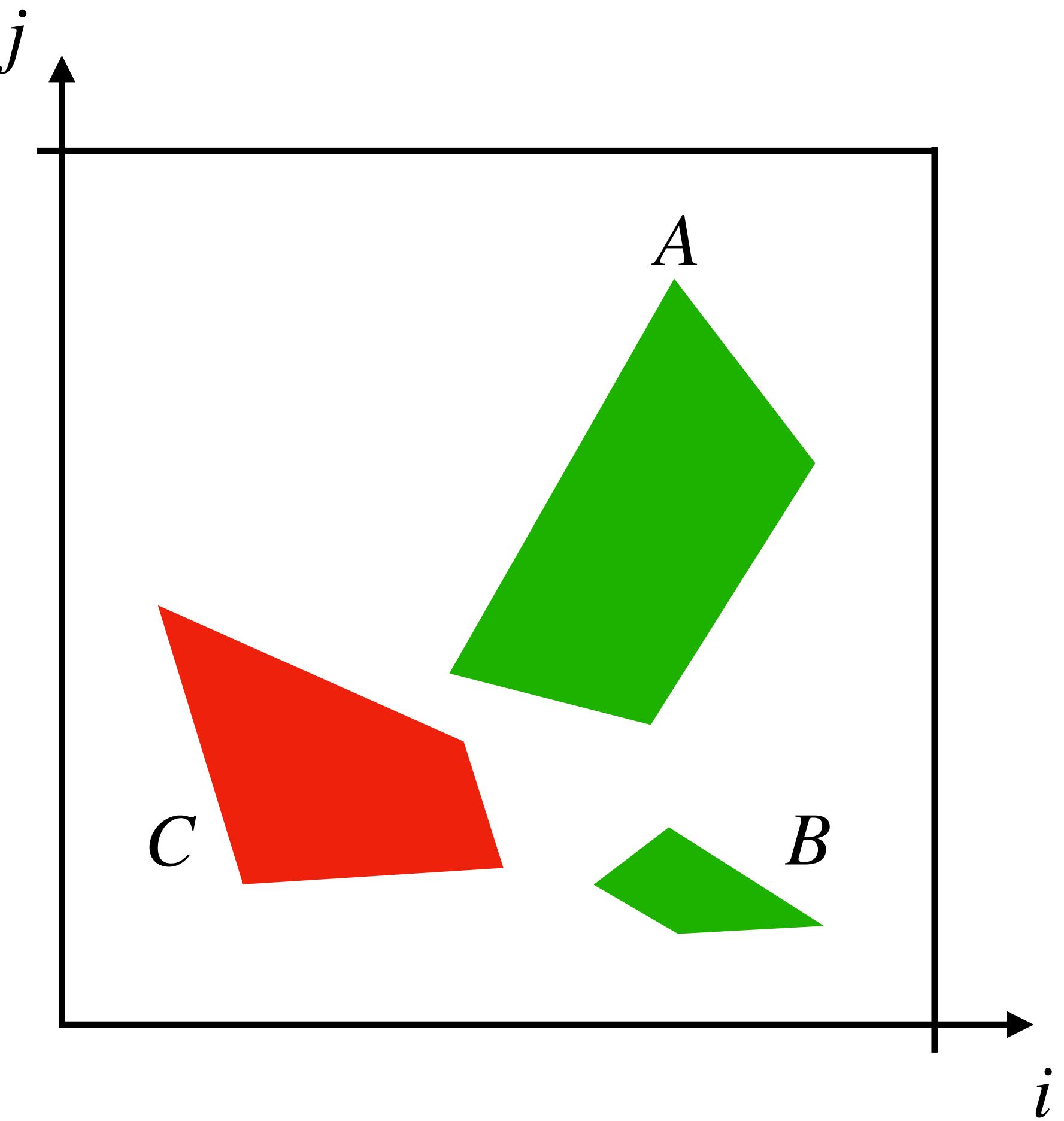
Discrete output space

ImpactAnalysis ^{\natural} _{i}

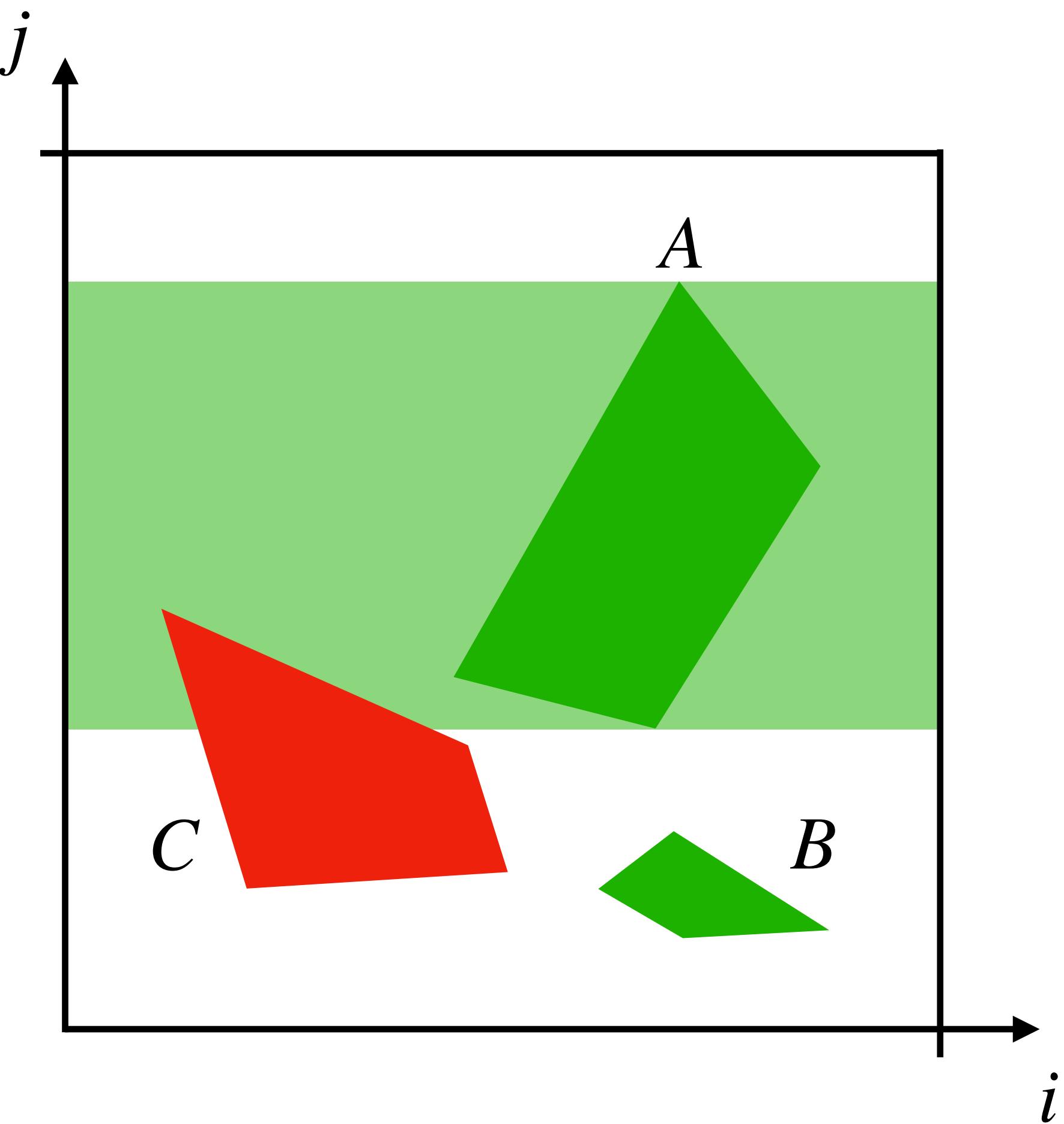


ImpactAnalysis_i[¶]

2 input features

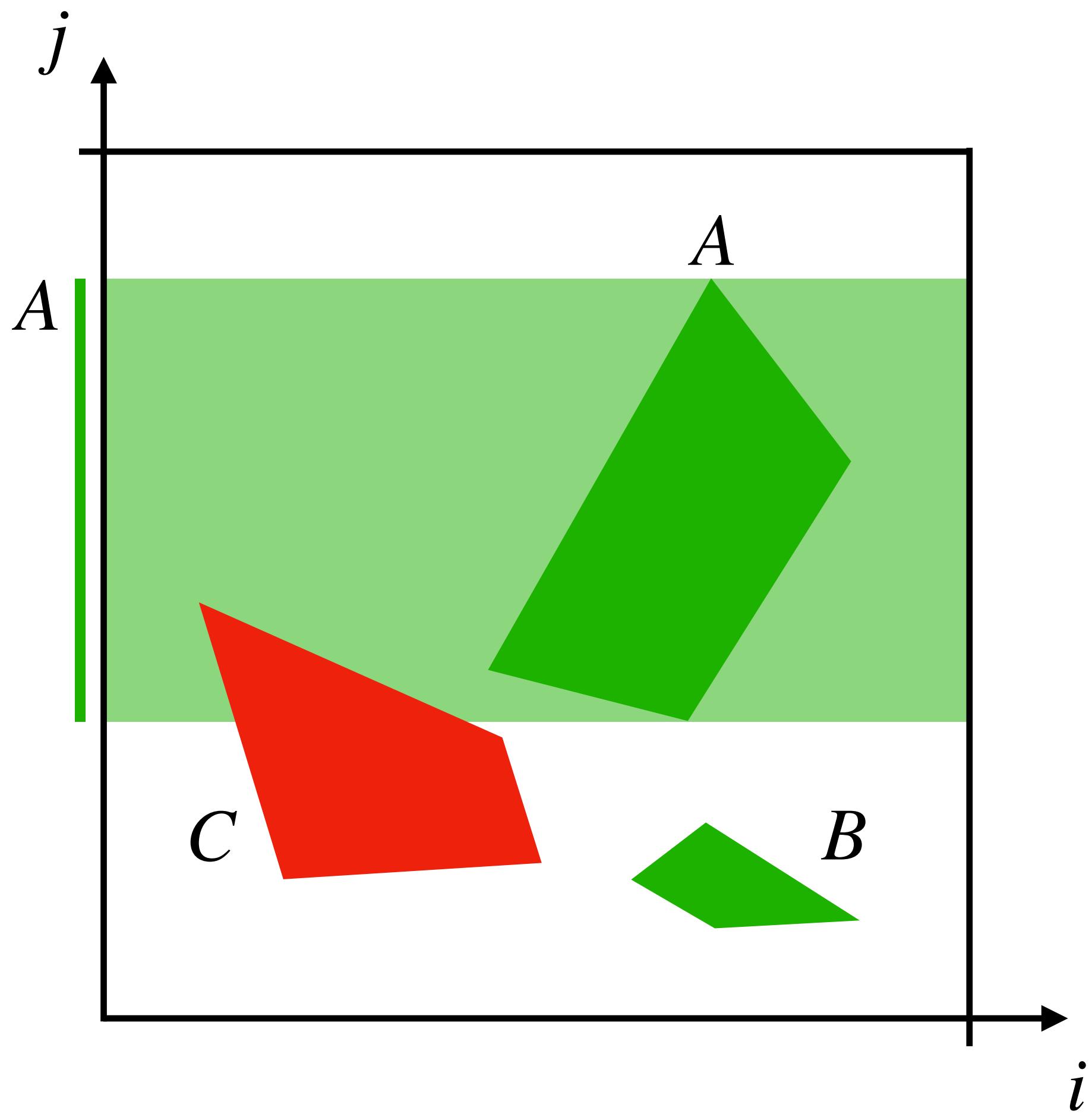


Perturbations of i



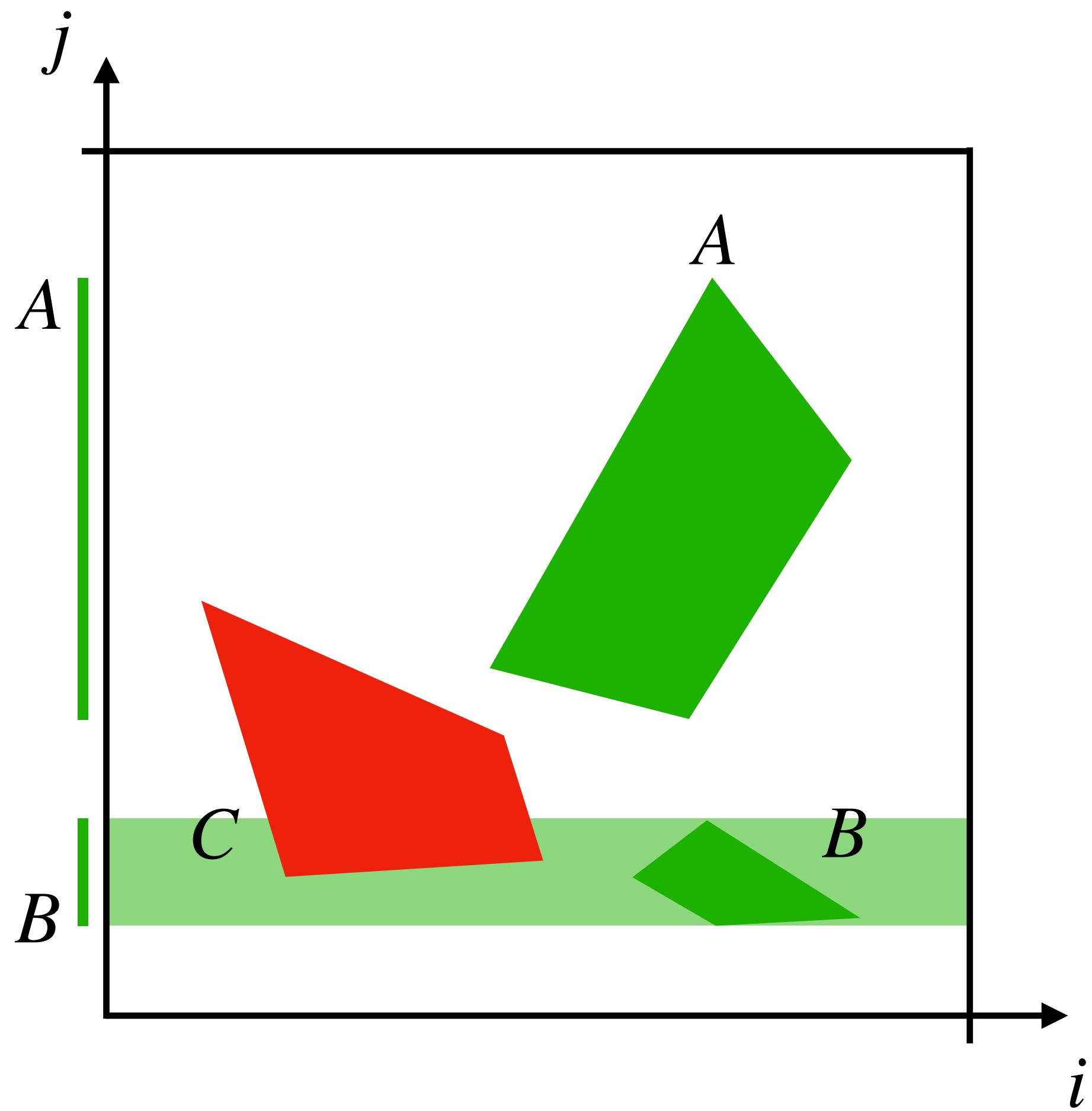
Projecting away
the feature i

Perturbations of i



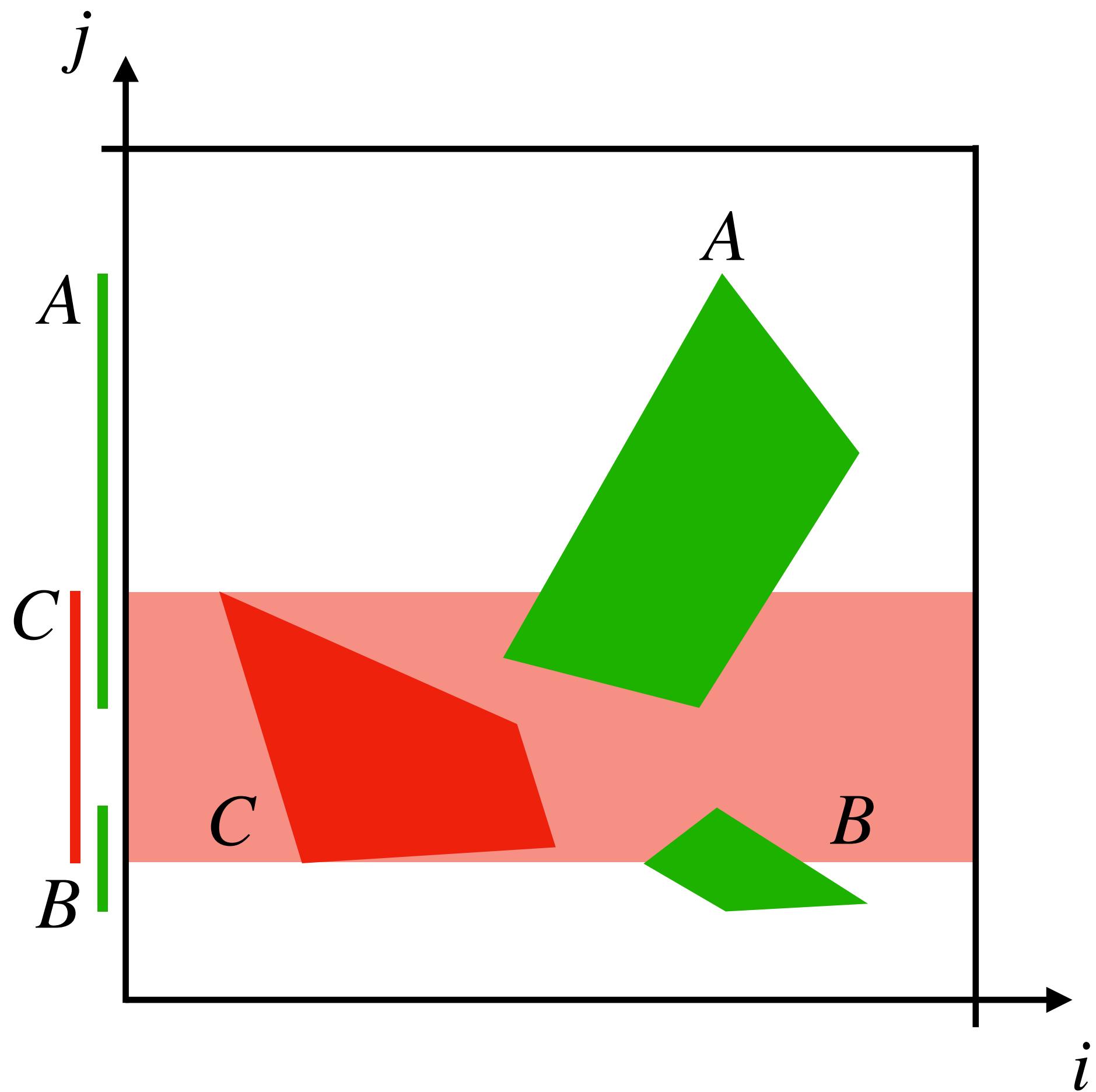
Projecting away
the feature i

Perturbations of i



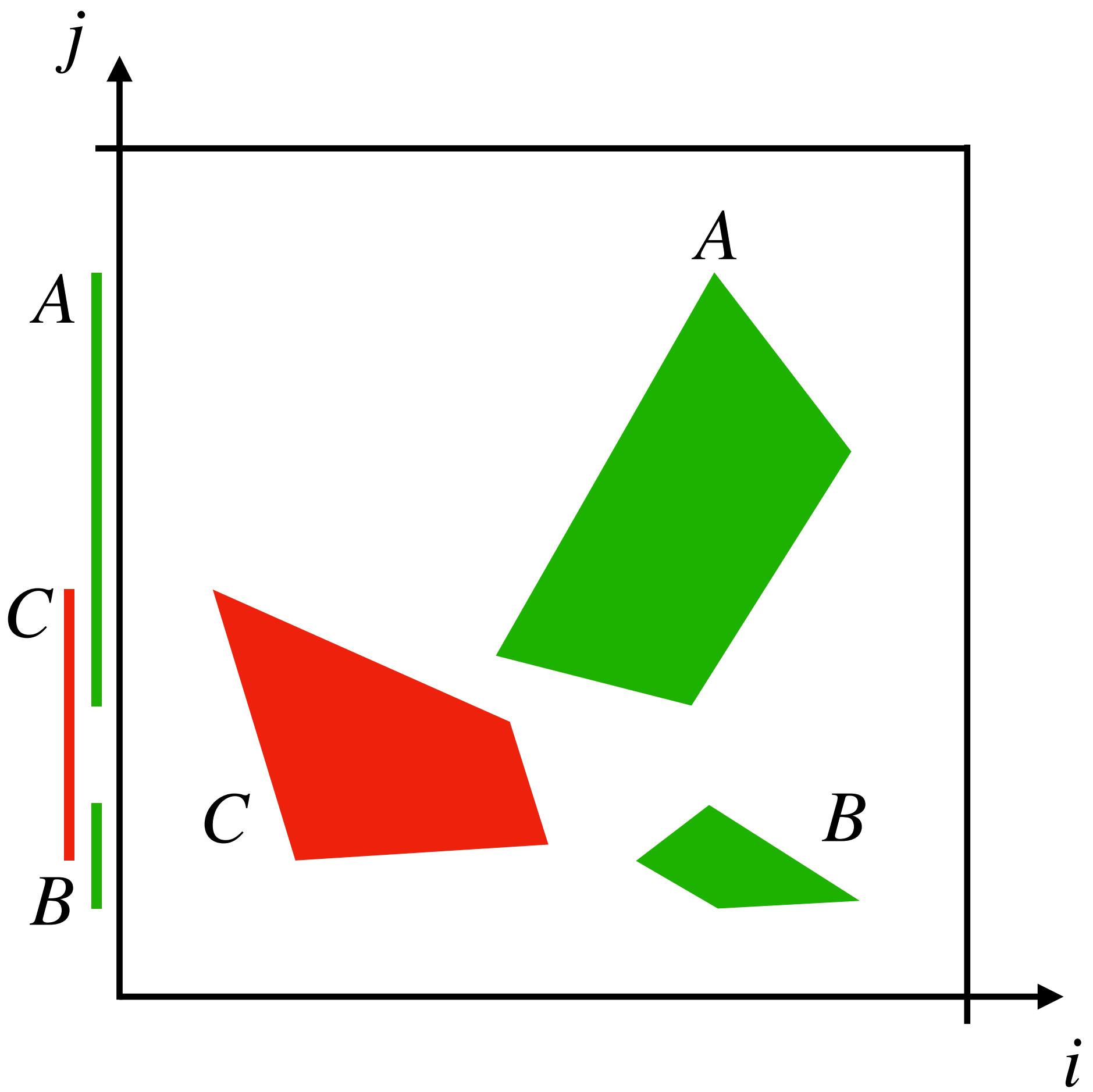
Projecting away
the feature i

Perturbations of i



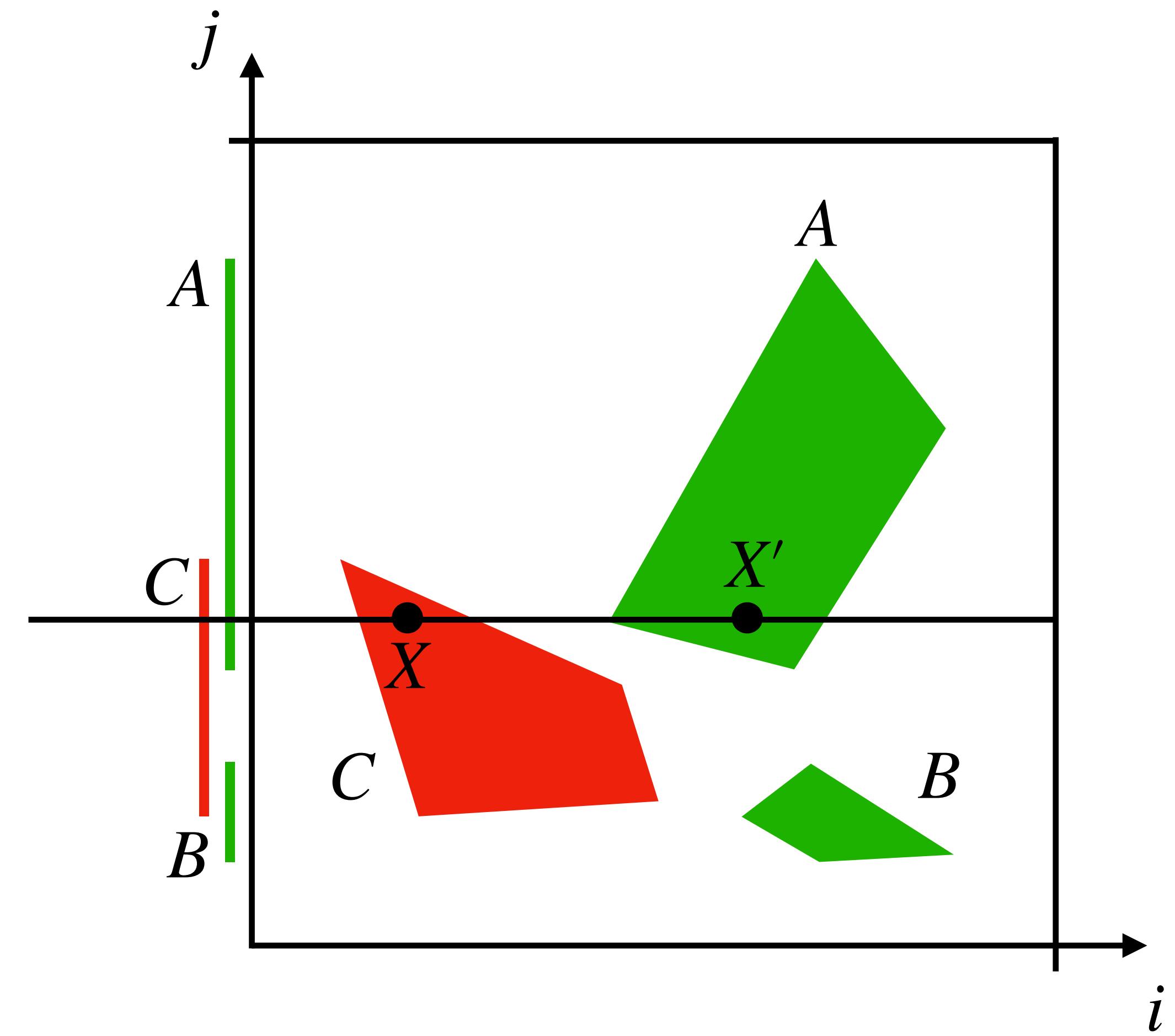
Projecting away
the feature i

Meaning of intersections

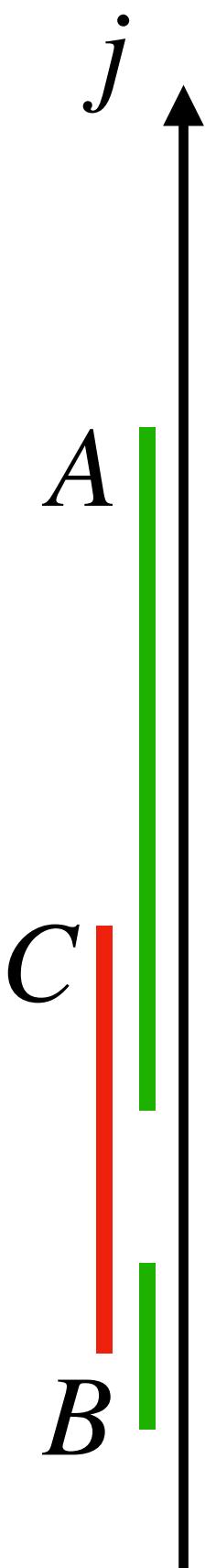


Meaning of intersections

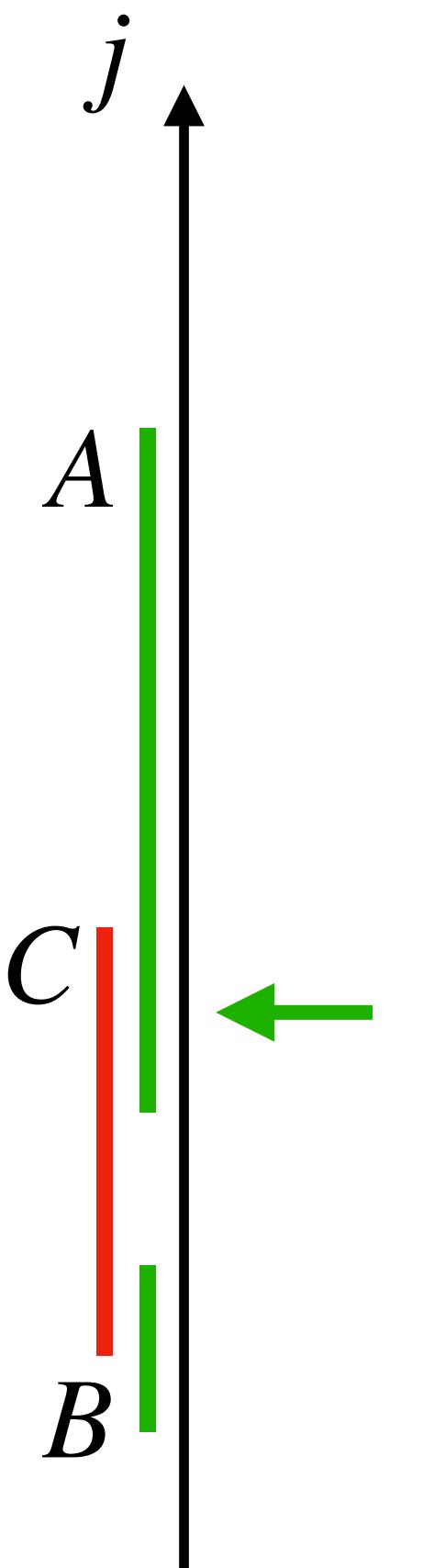
Intersection
between A and C



Count the intersections

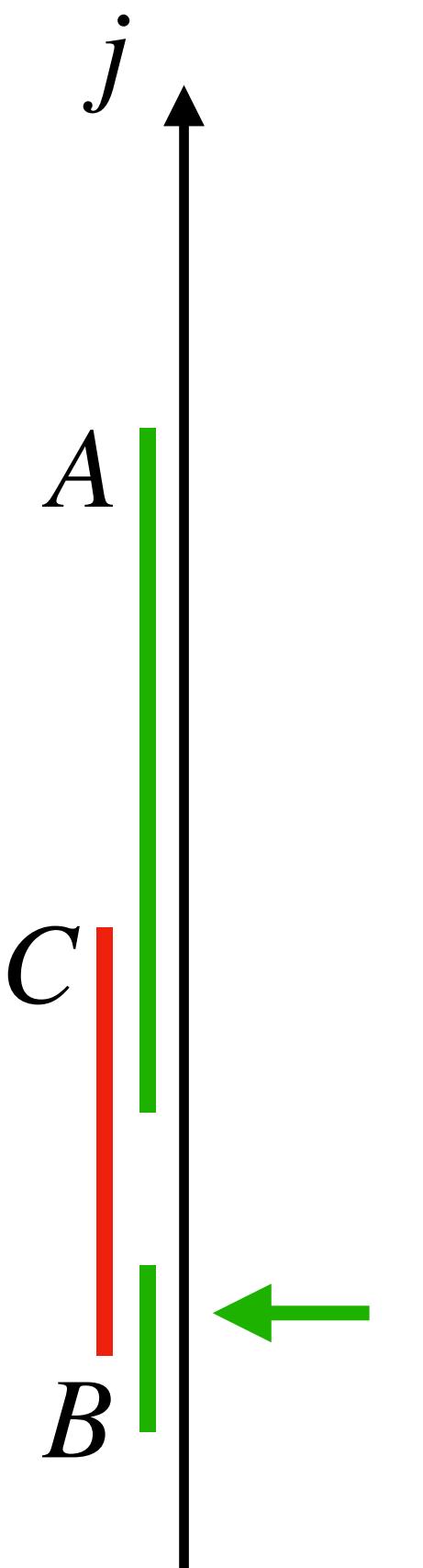


Count the intersections



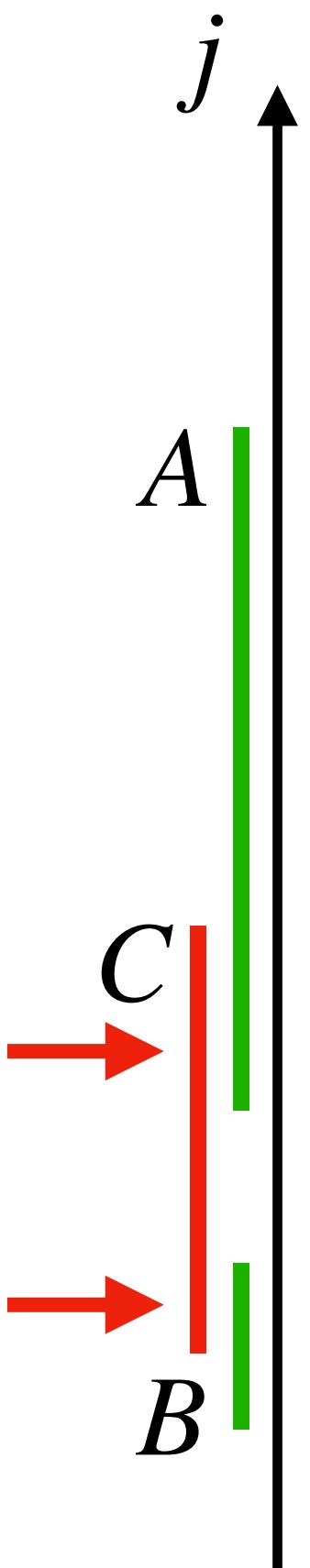
starting from • \Rightarrow 1 intersection

Count the intersections



starting from  \Rightarrow 1 intersection

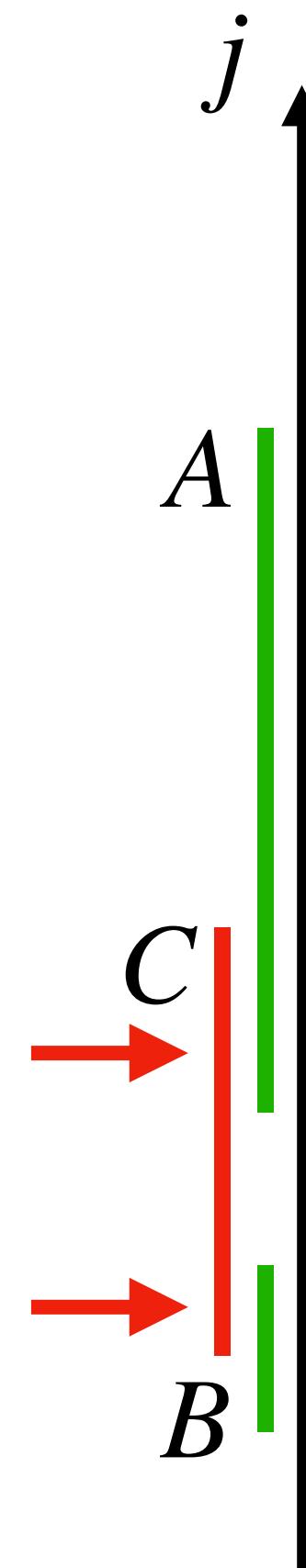
Count the intersections



starting from \Rightarrow 1 intersection

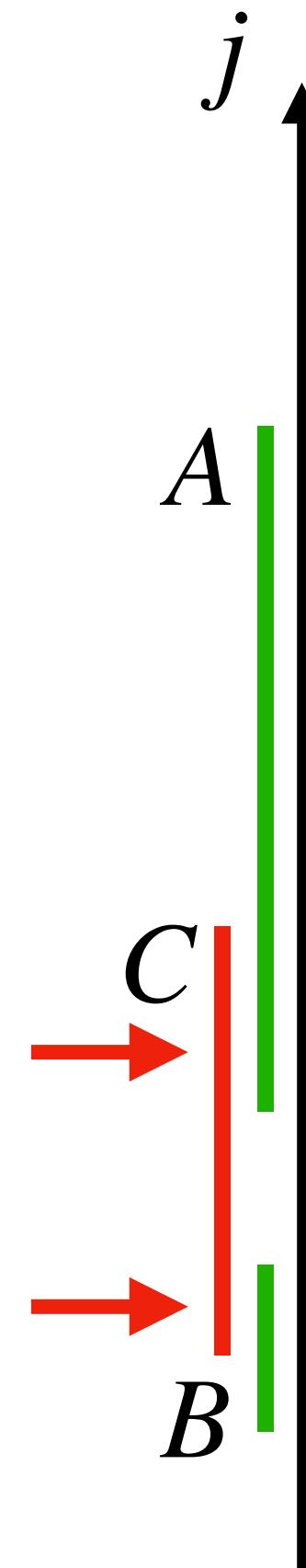
starting from \Rightarrow 2 intersections

Count the intersections



$\text{ImpactAnalysis}_i^\natural(P) = 2$

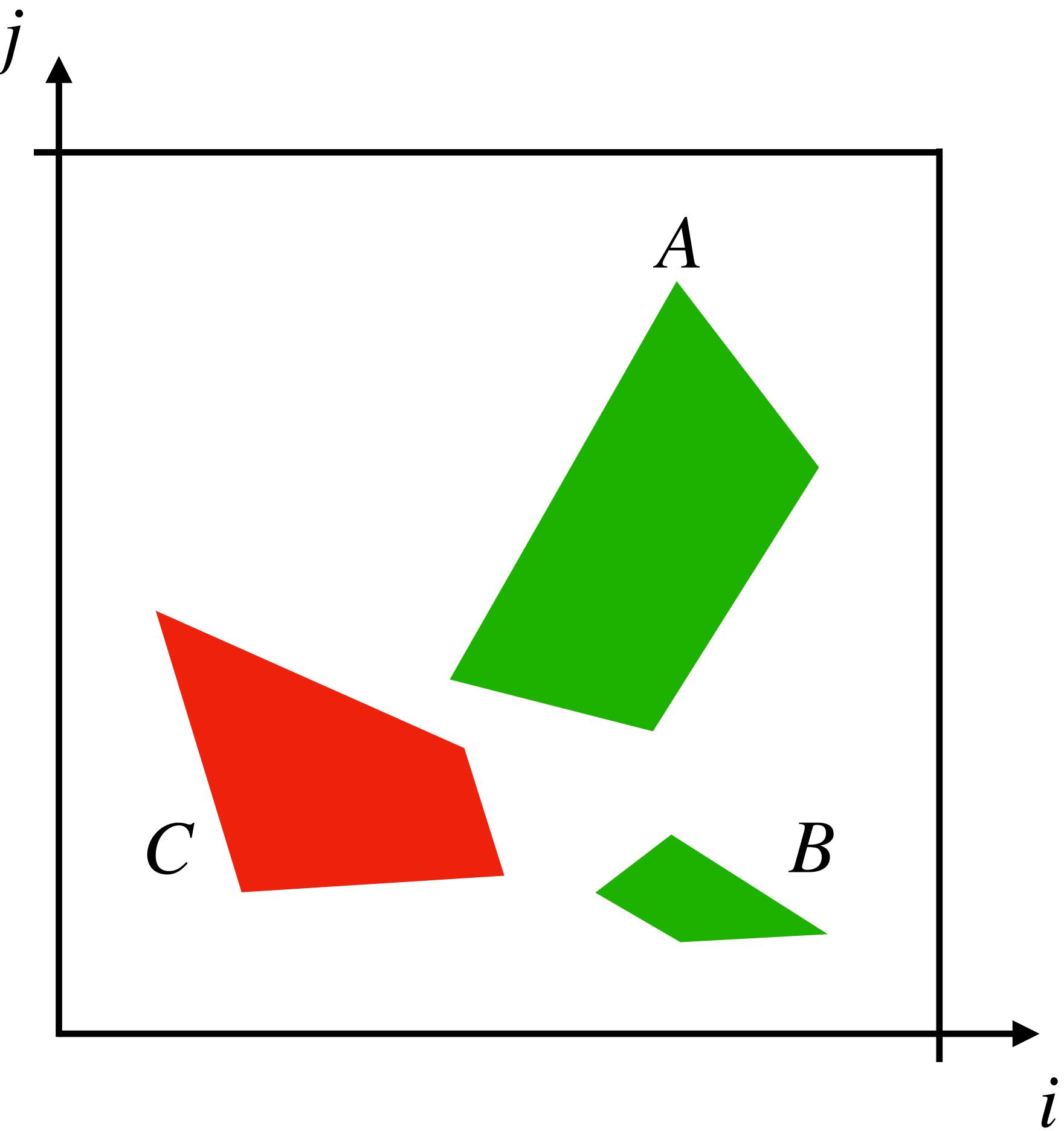
Count the intersections



concrete
CountChanges_{i.}(P)
 \leq
ImpactAnalysis_i[⊤](P) = 2
abstract

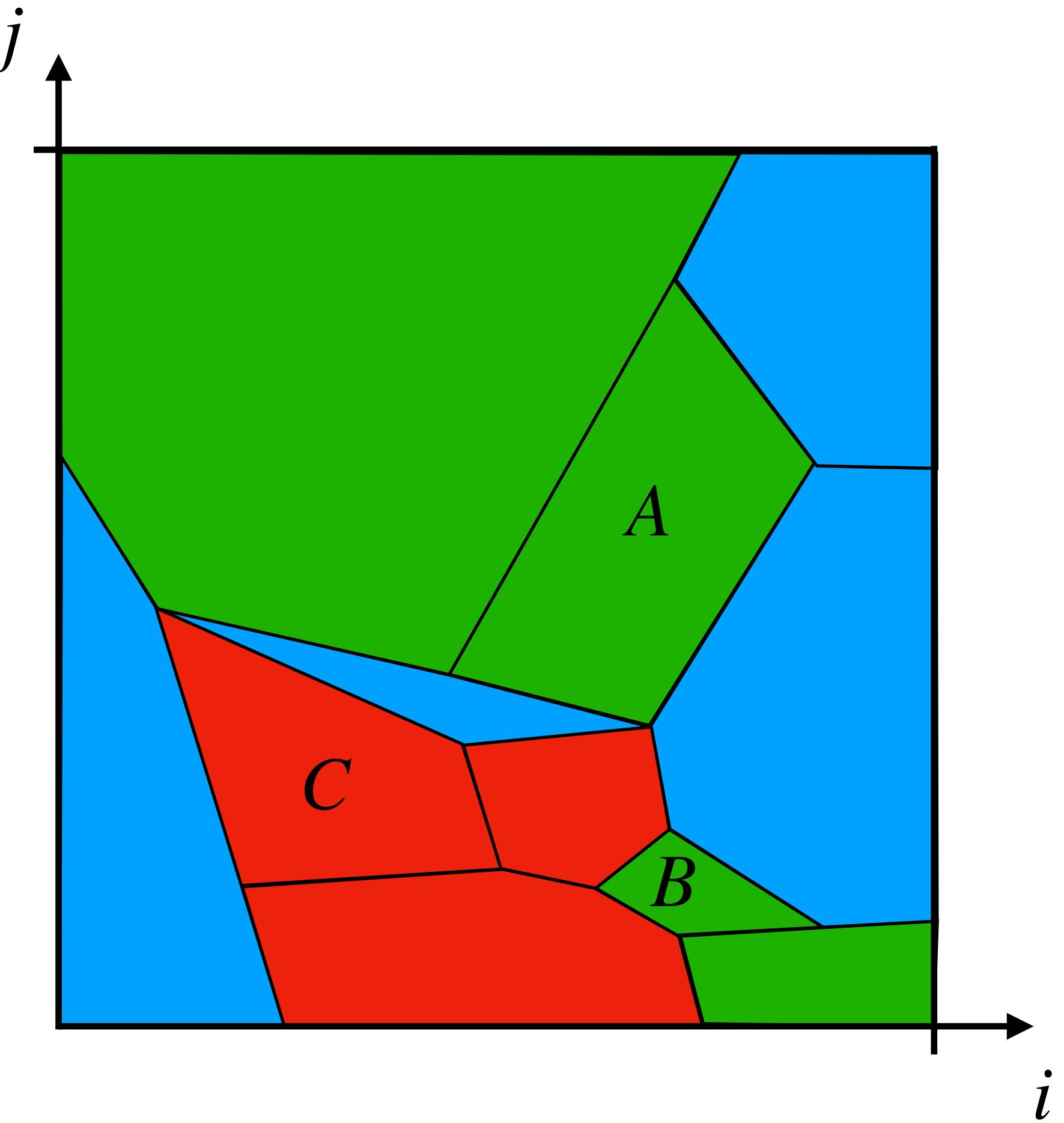
Sound?

2 input features

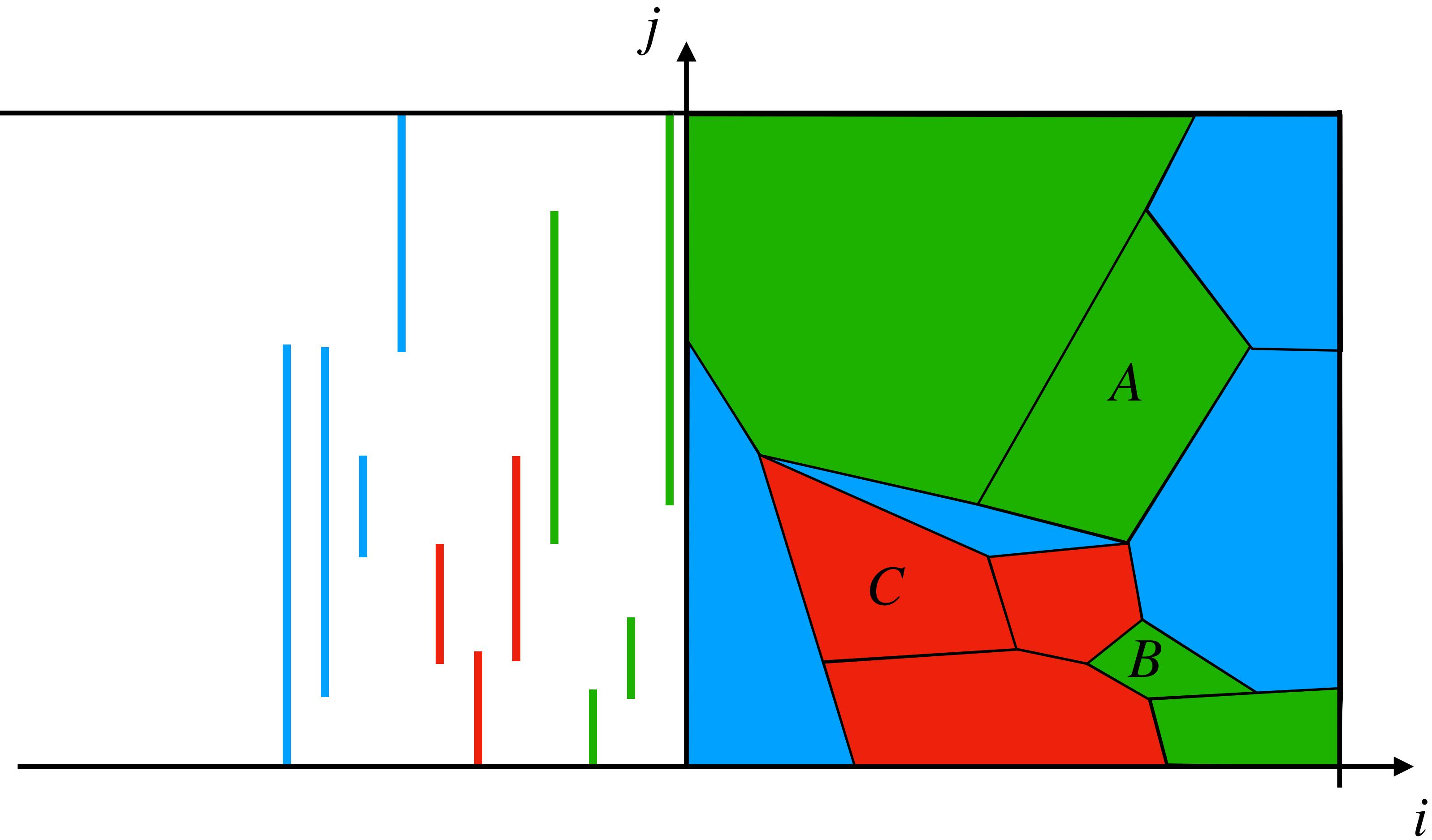


Sound?

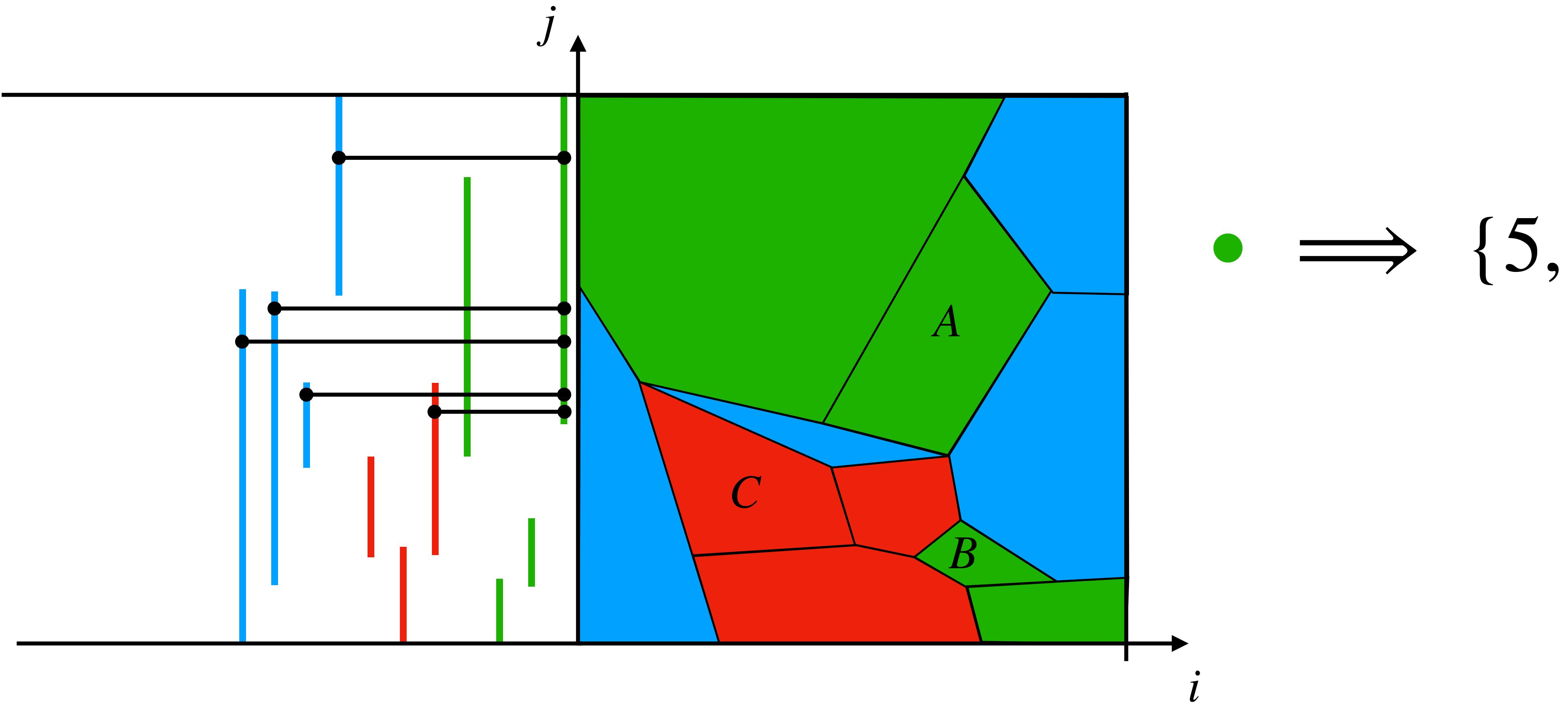
2 input features
3 output buckets



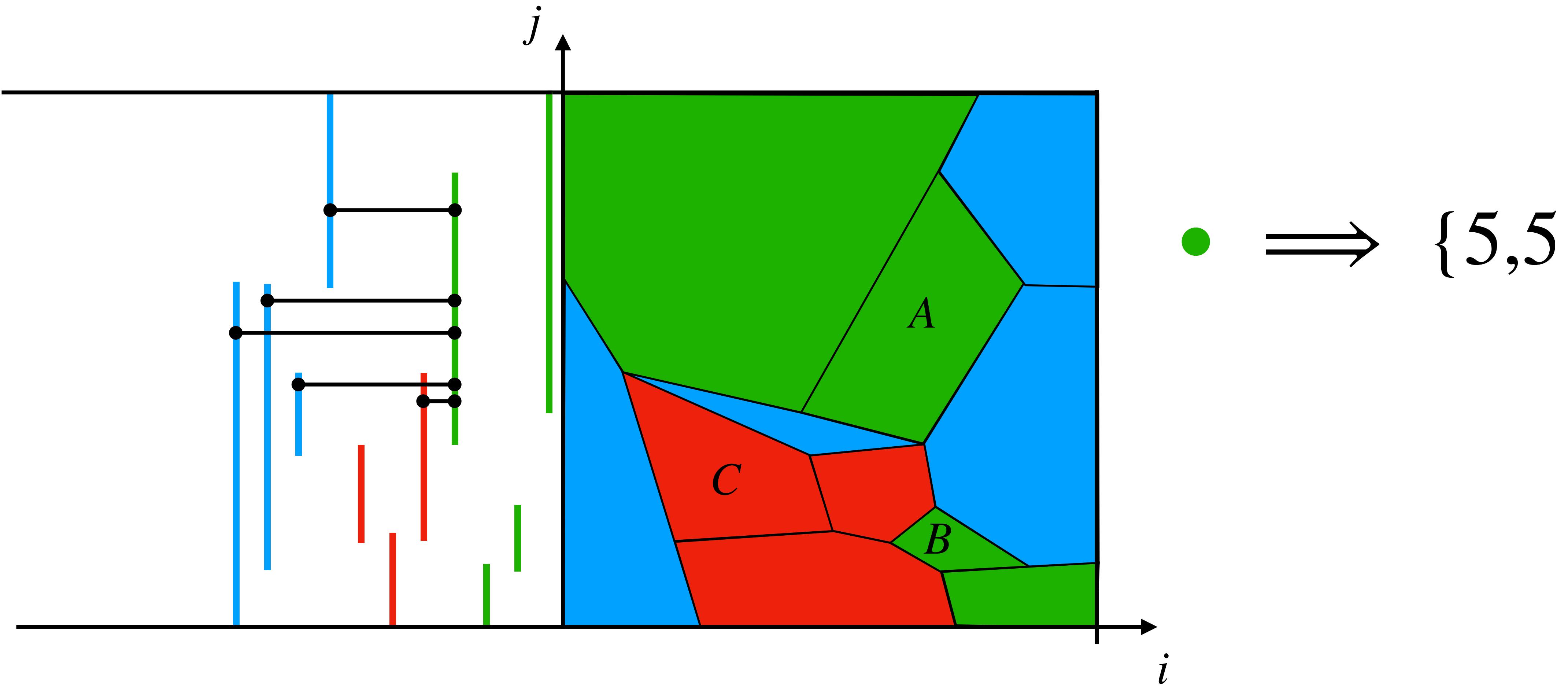
Sound?



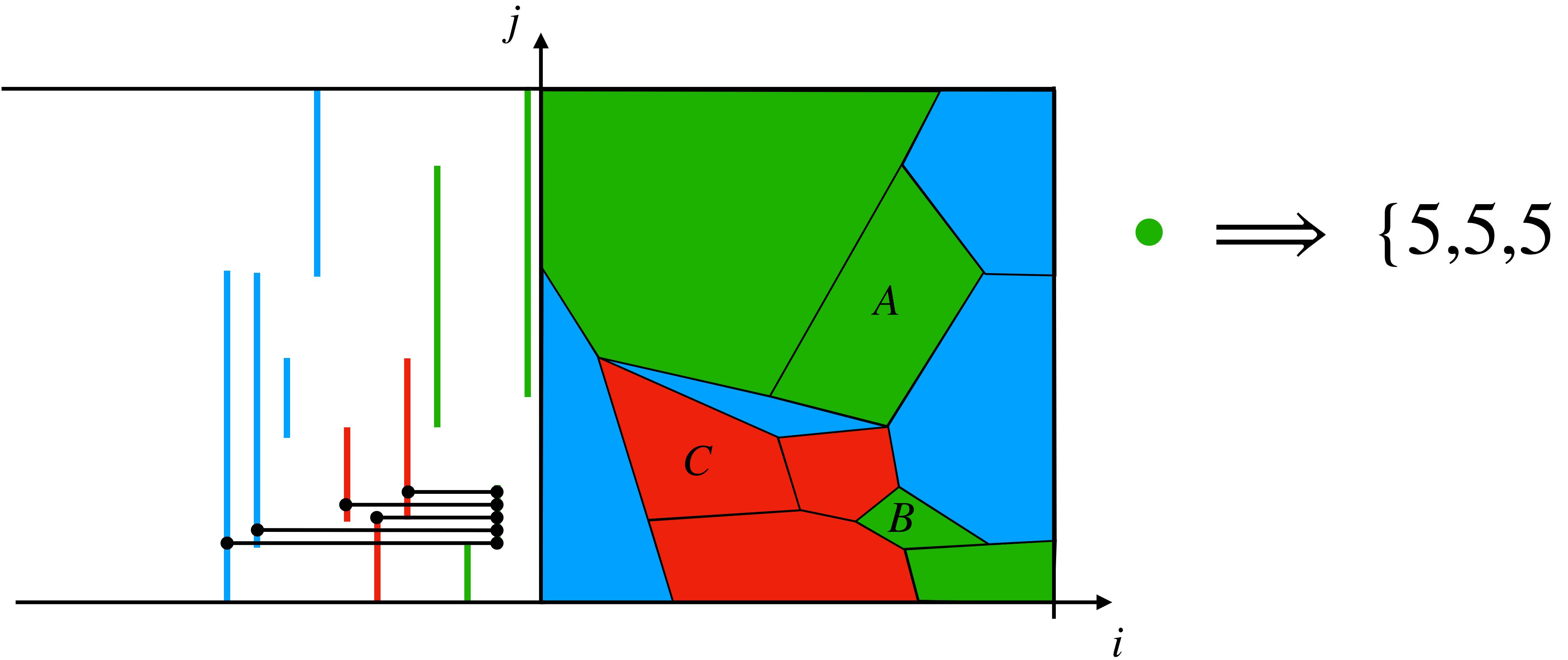
Sound?



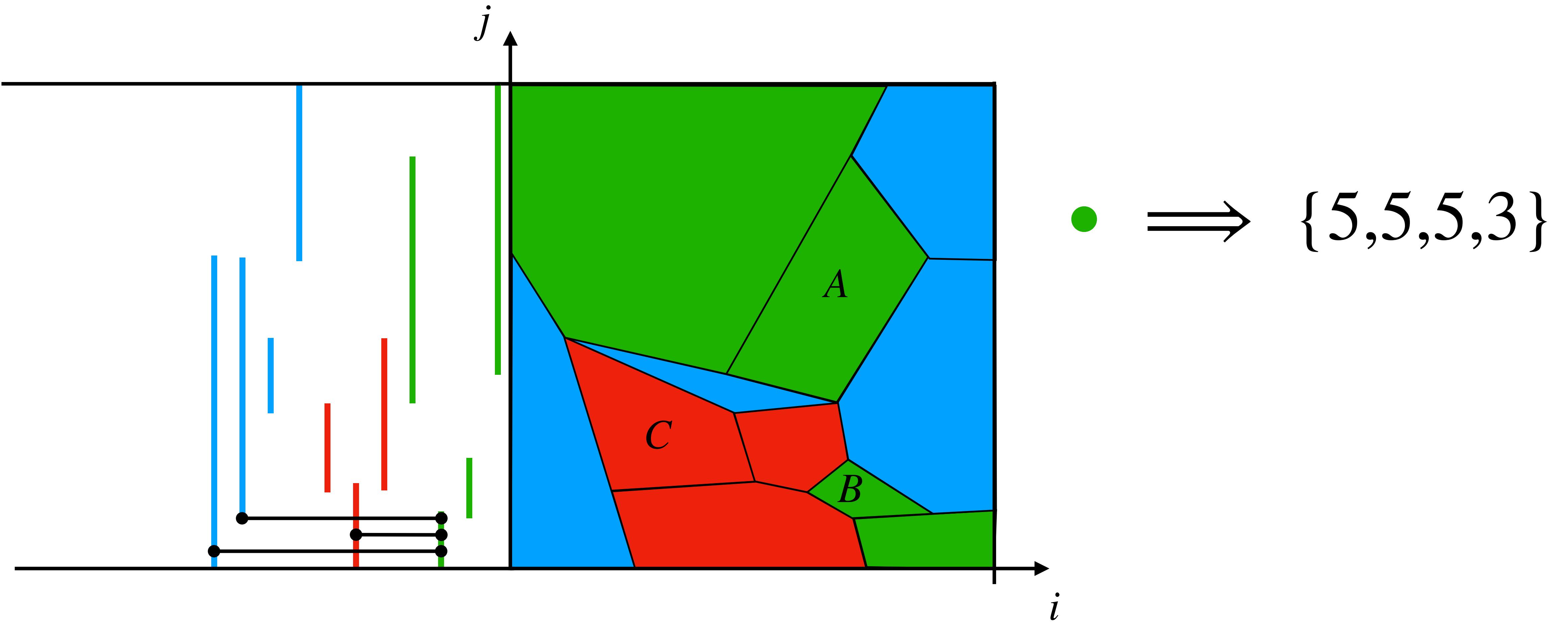
Sound?



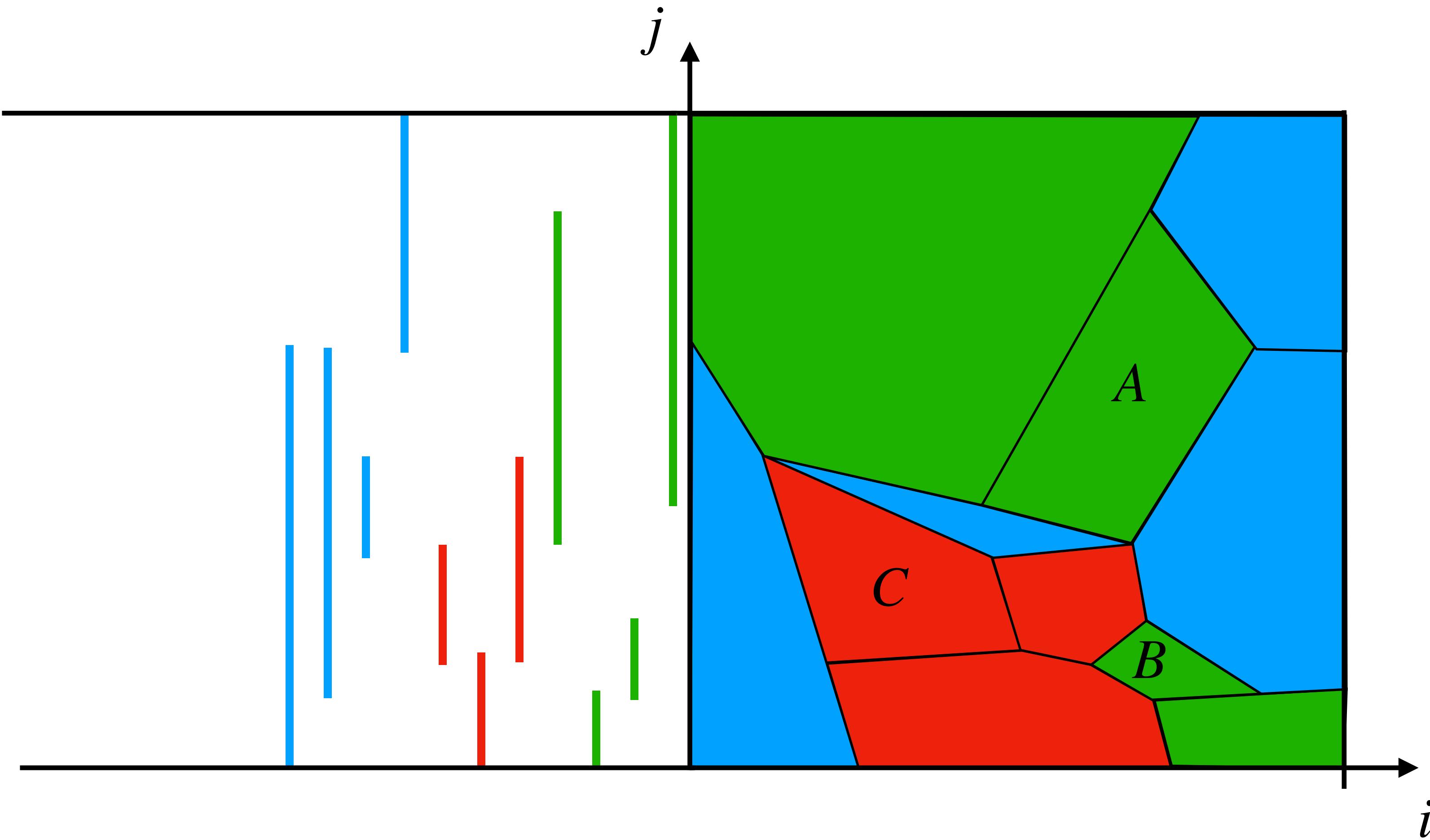
Sound?



Sound?

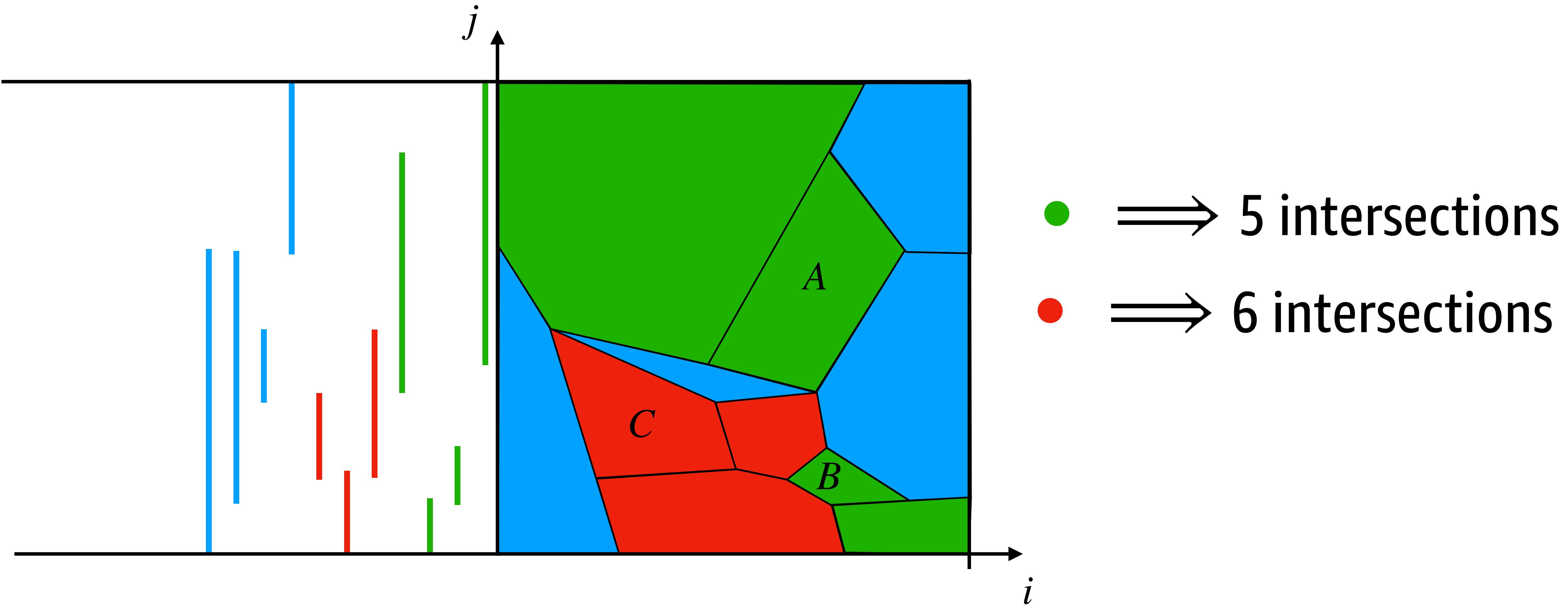


Sound?

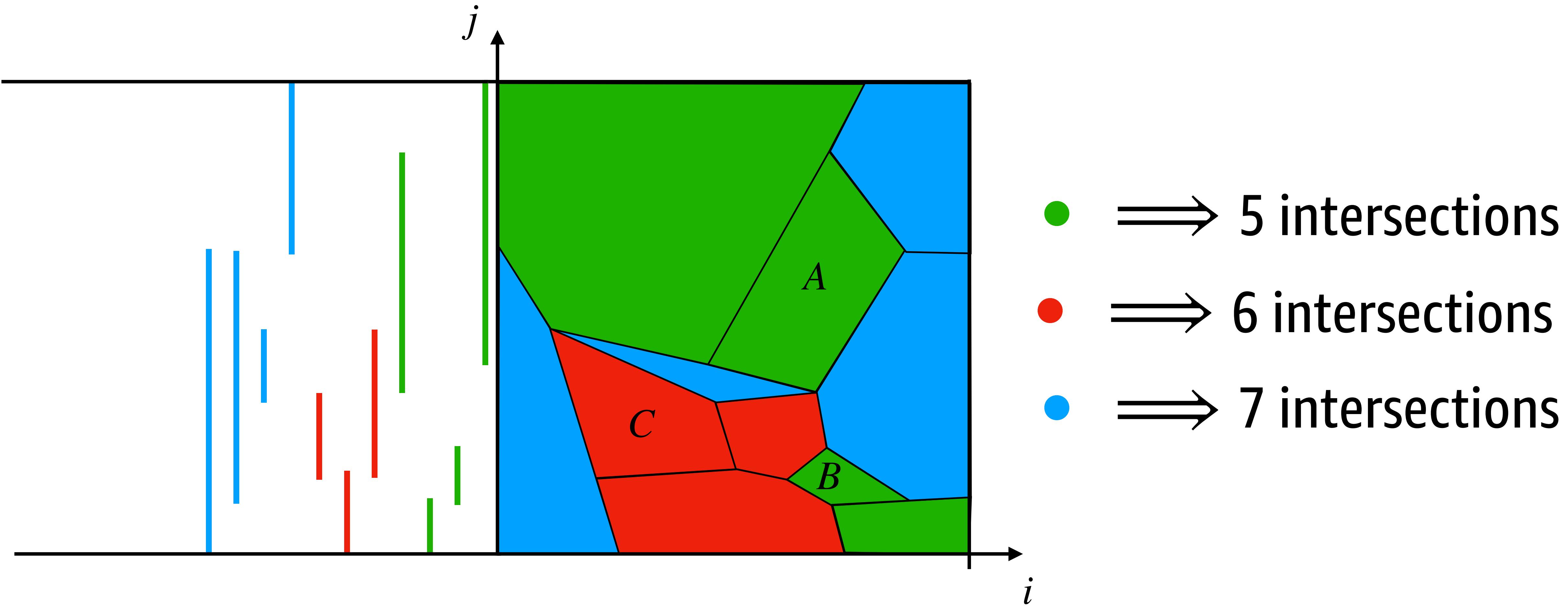


• \Rightarrow 5 intersections

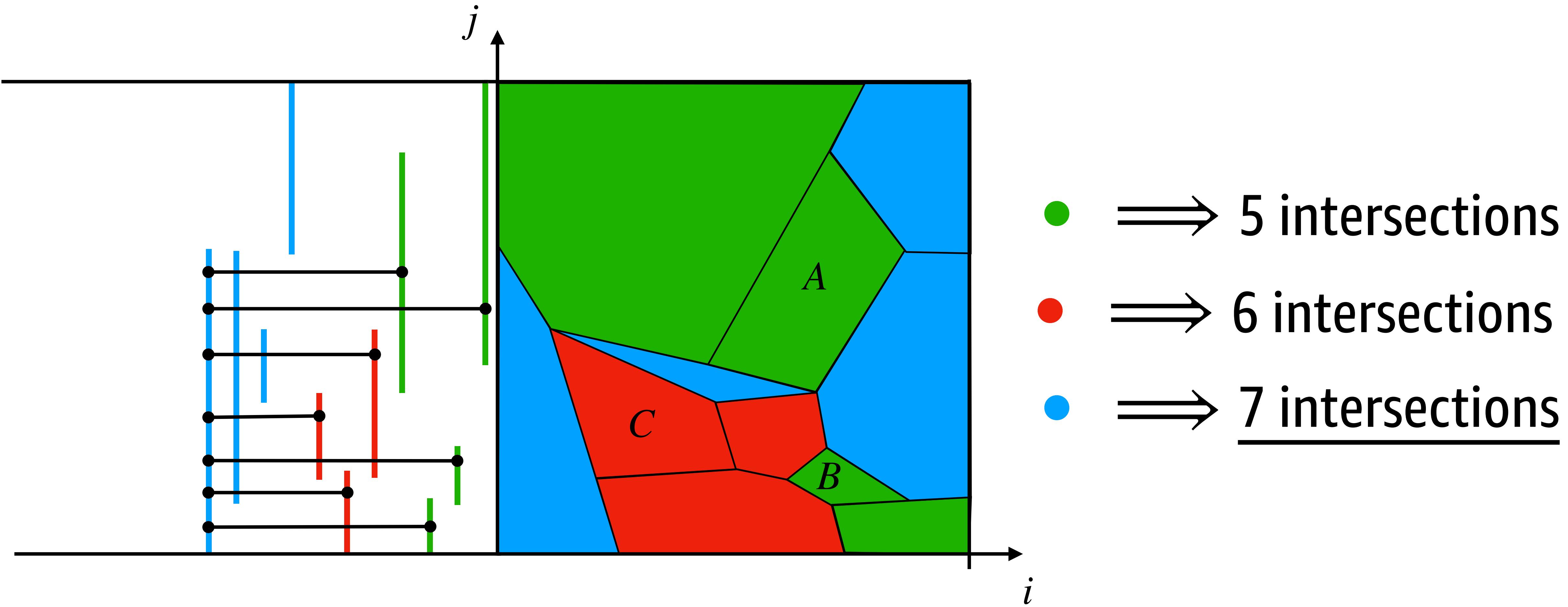
Sound?



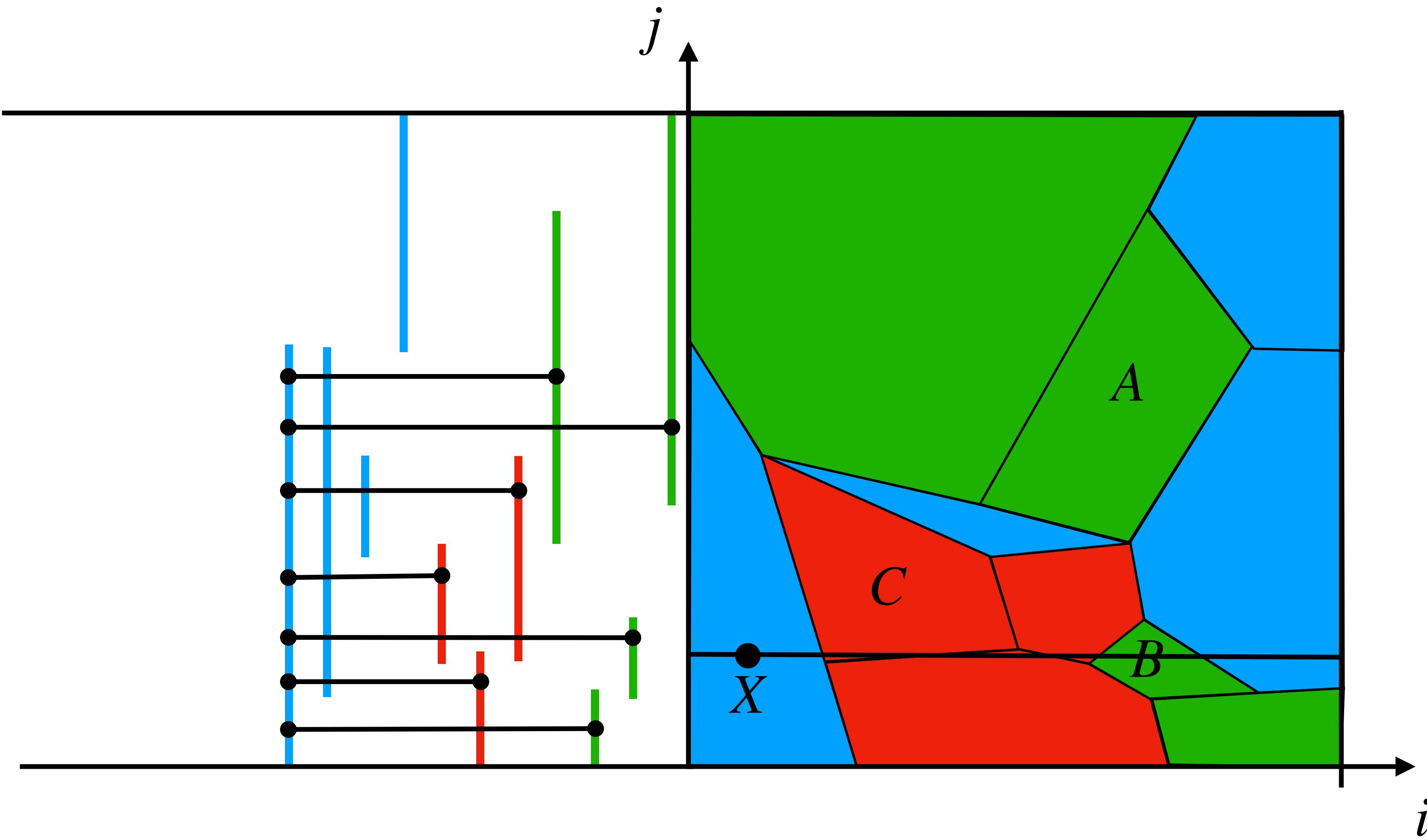
Sound?



Sound?

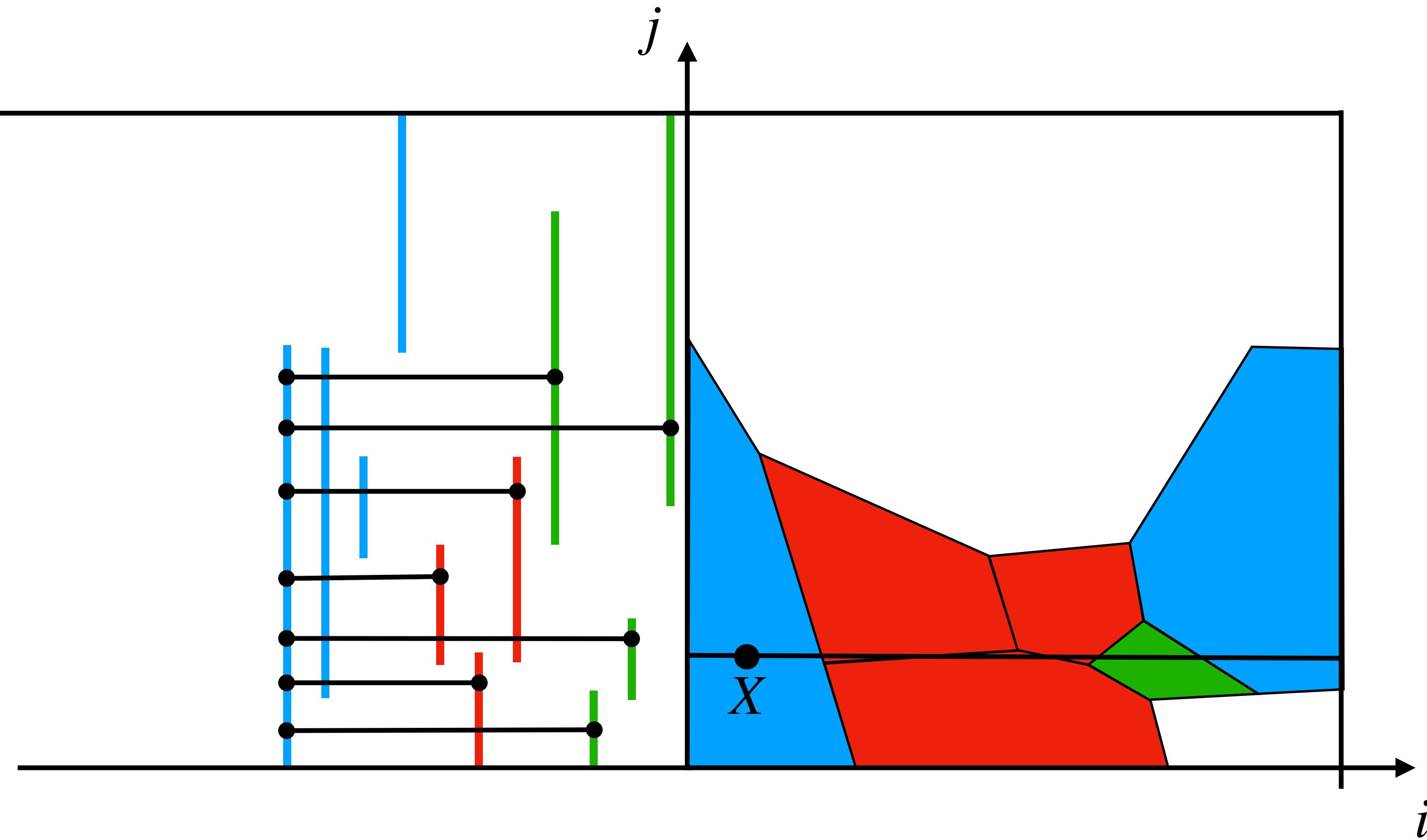


Sound?



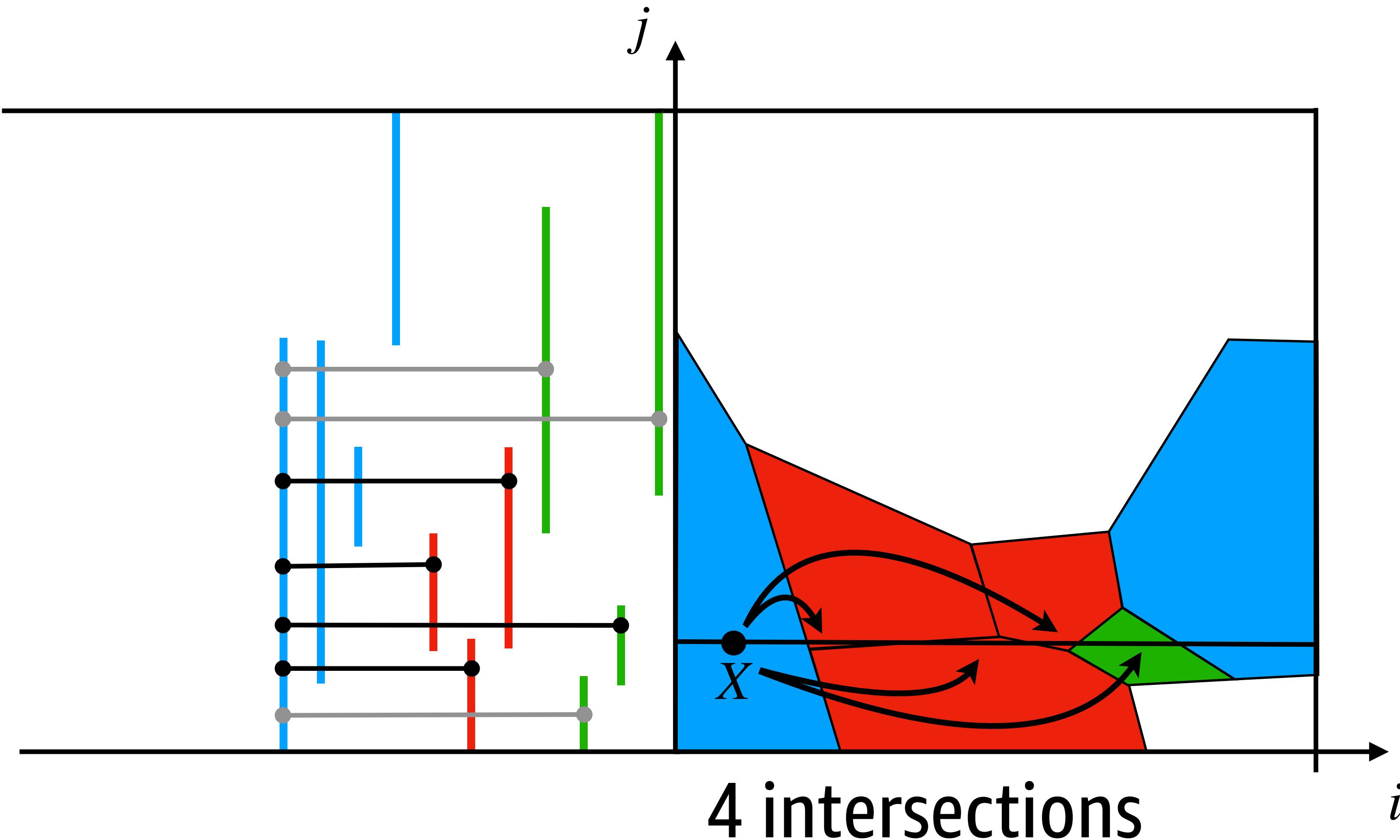
7 intersections

Sound?

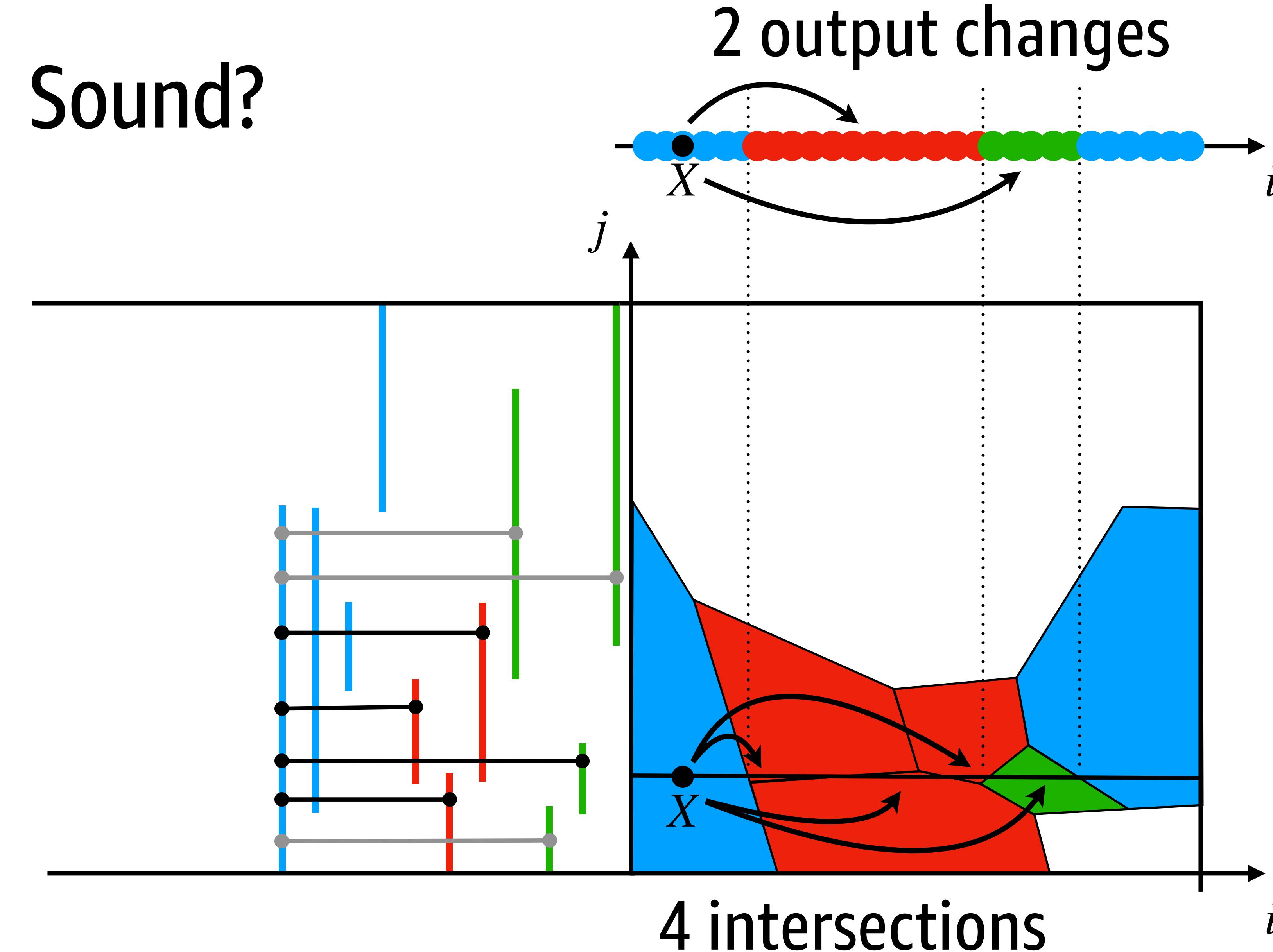


7 intersections

Sound?



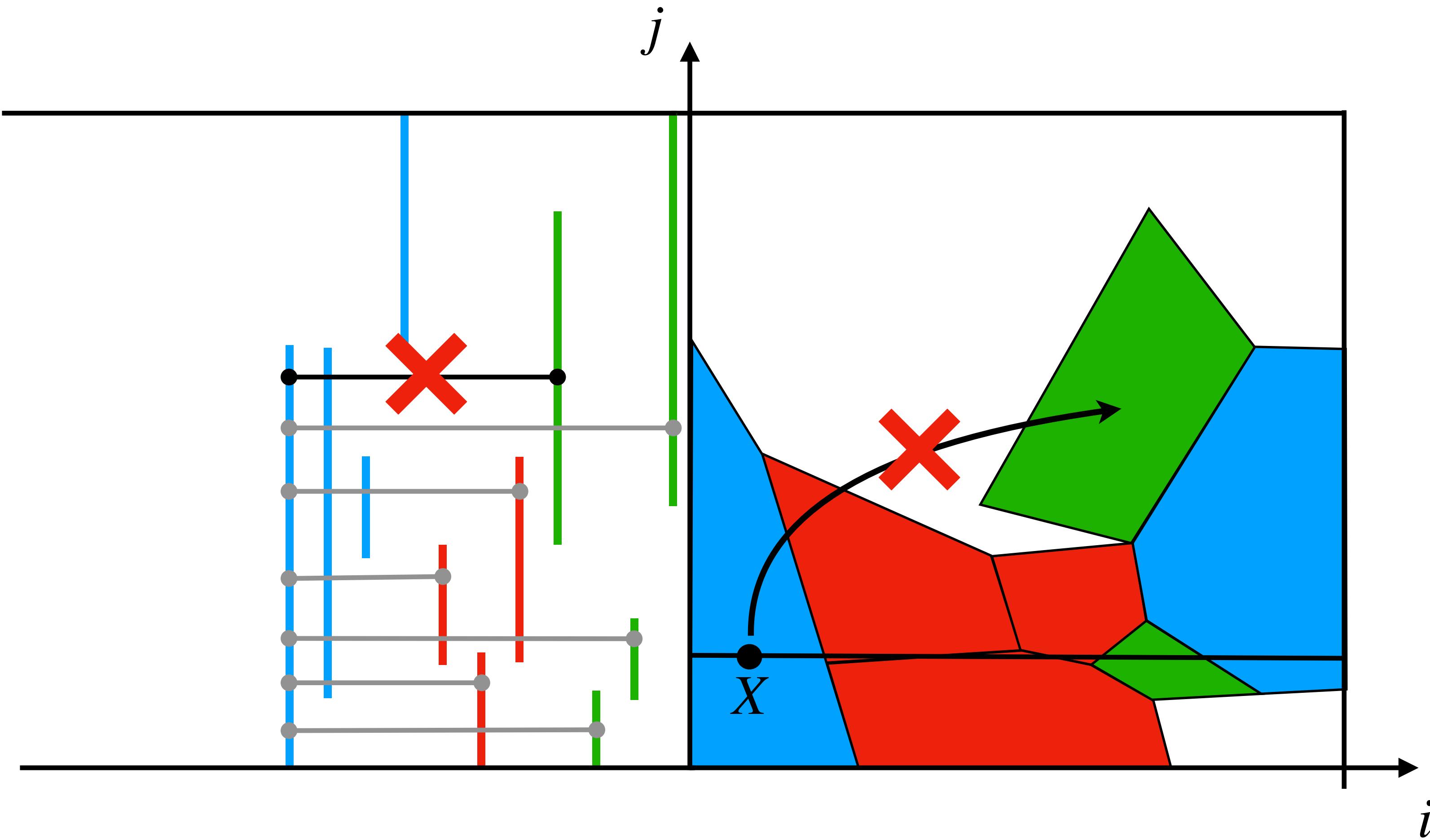
Sound?



The abstraction counts
more changes than real

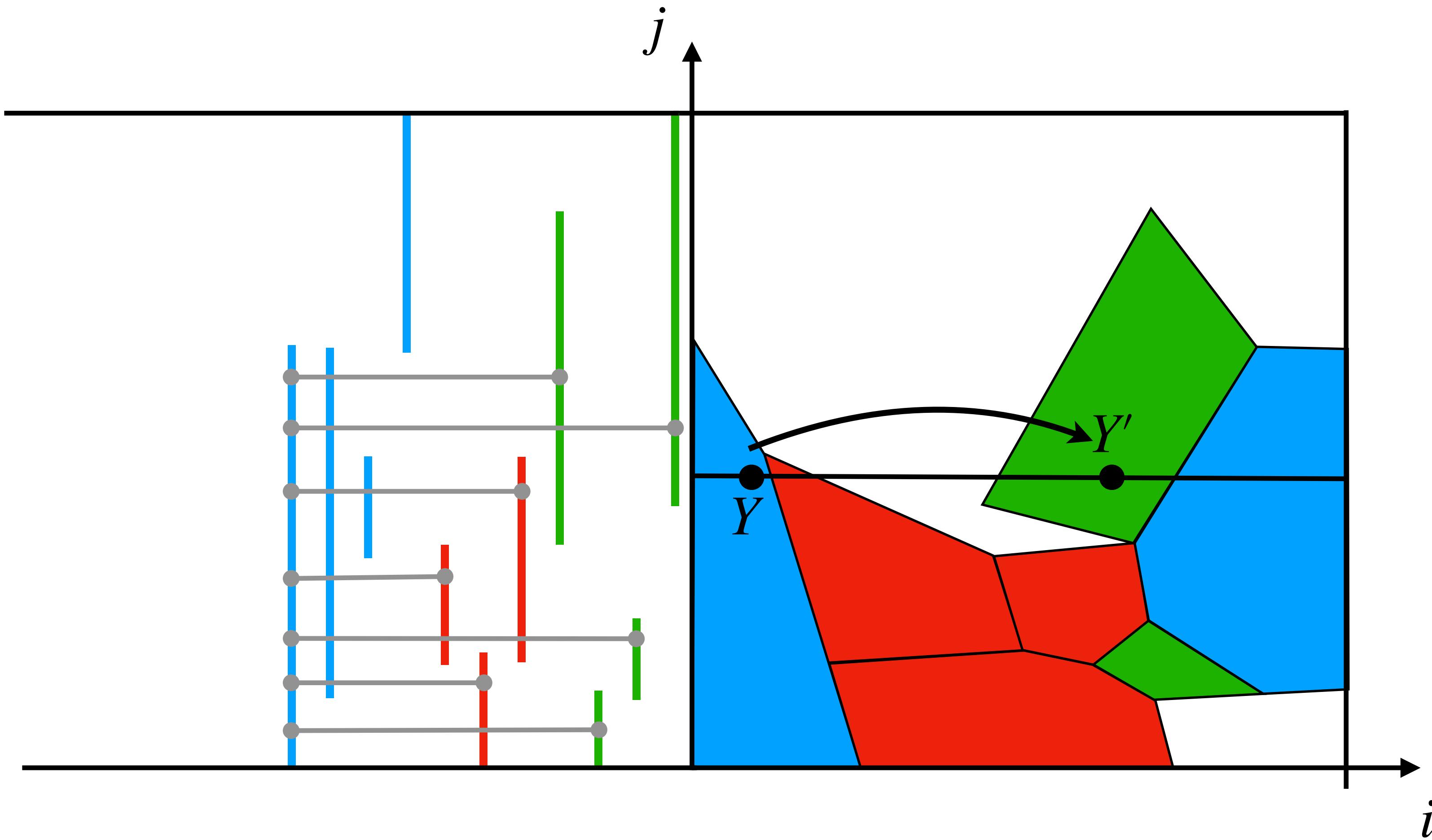
7 intersections

Sound?



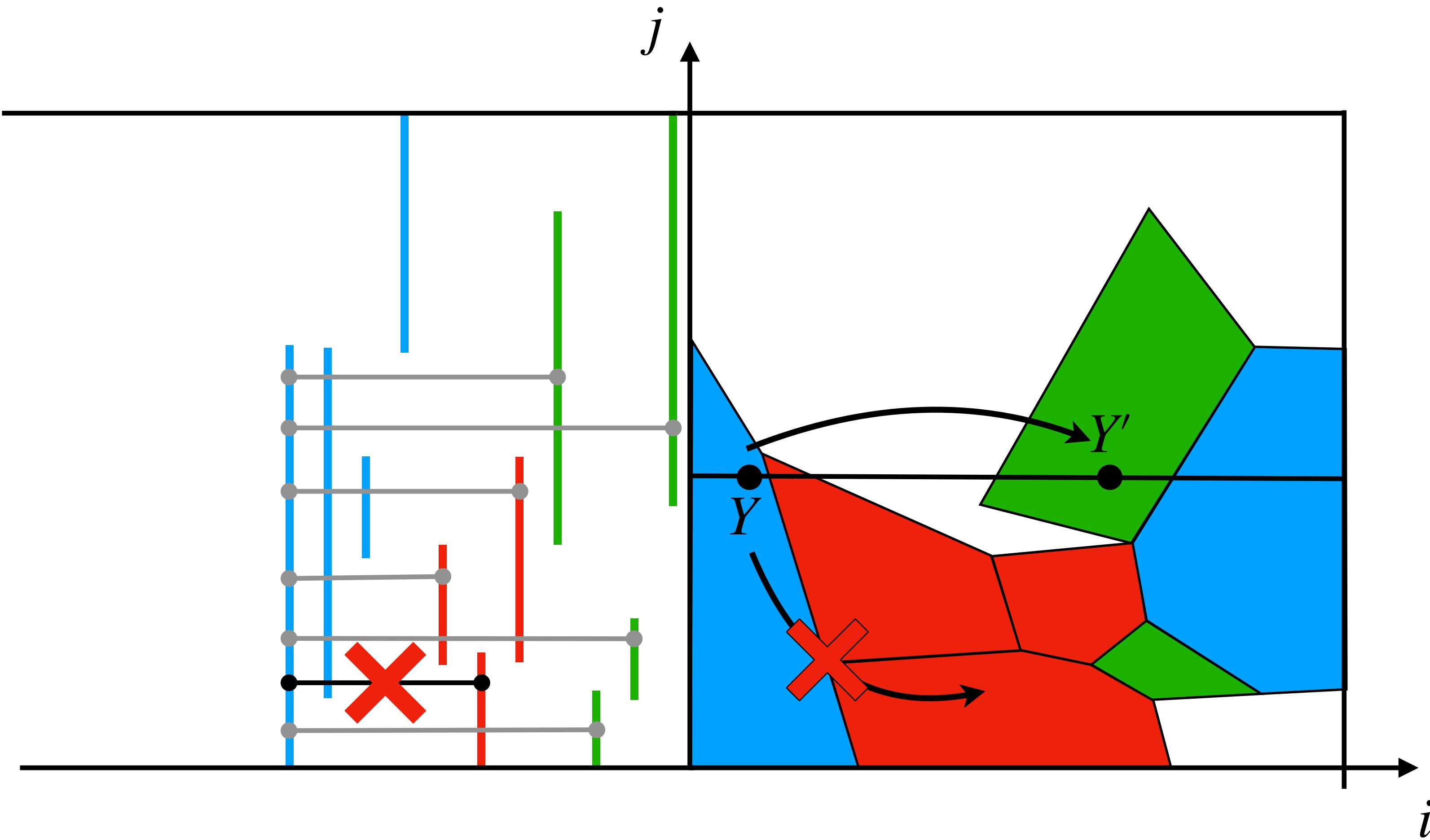
The abstraction adds
spurious perturbations

Sound?



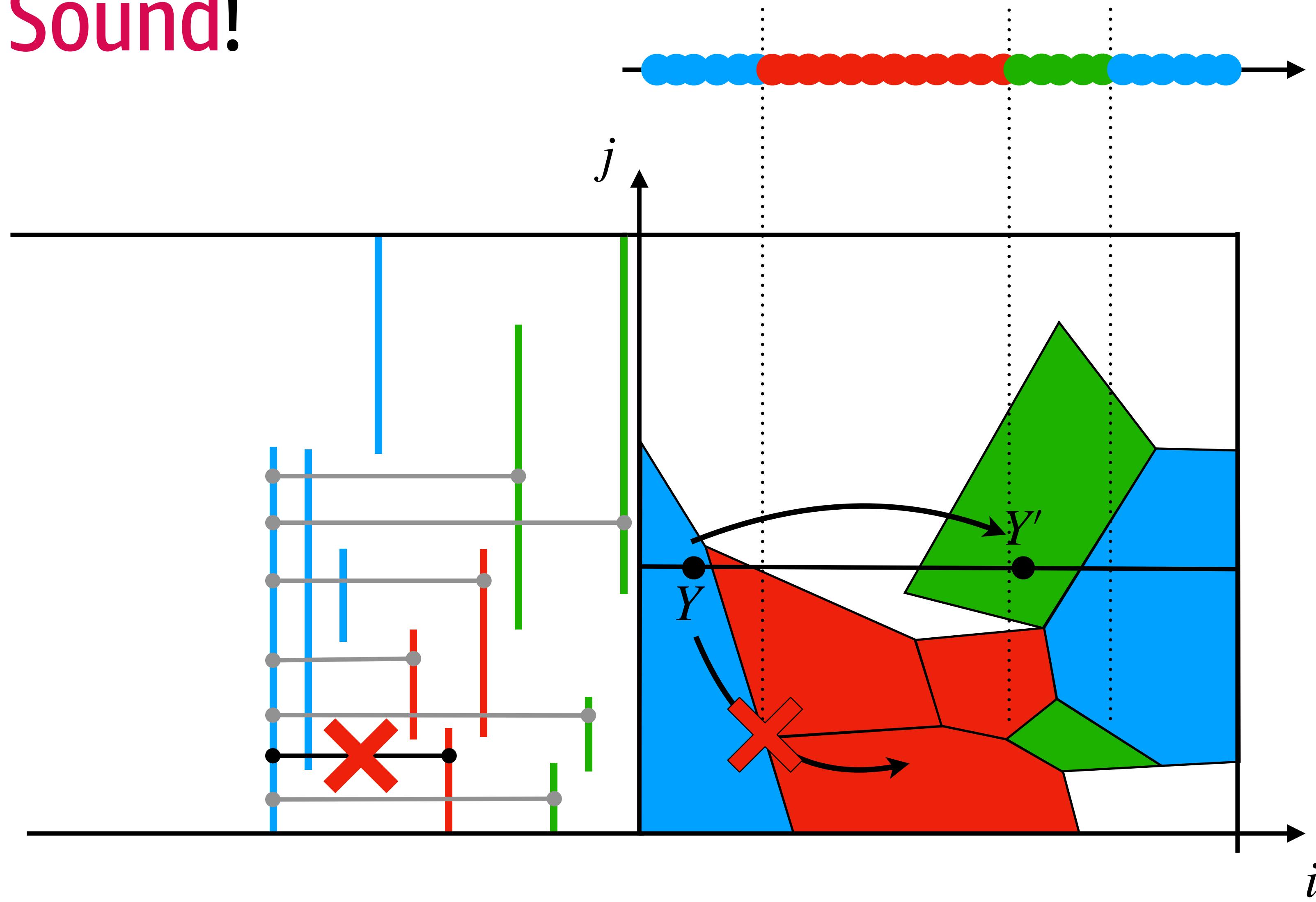
The abstraction adds
spurious perturbations

Sound?

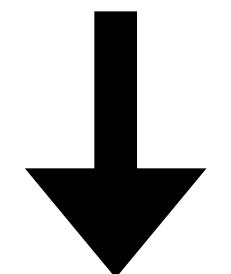


The abstraction adds
spurious perturbations

Sound!



The abstraction adds
spurious perturbations
and
counts more changes
than real



Sound!

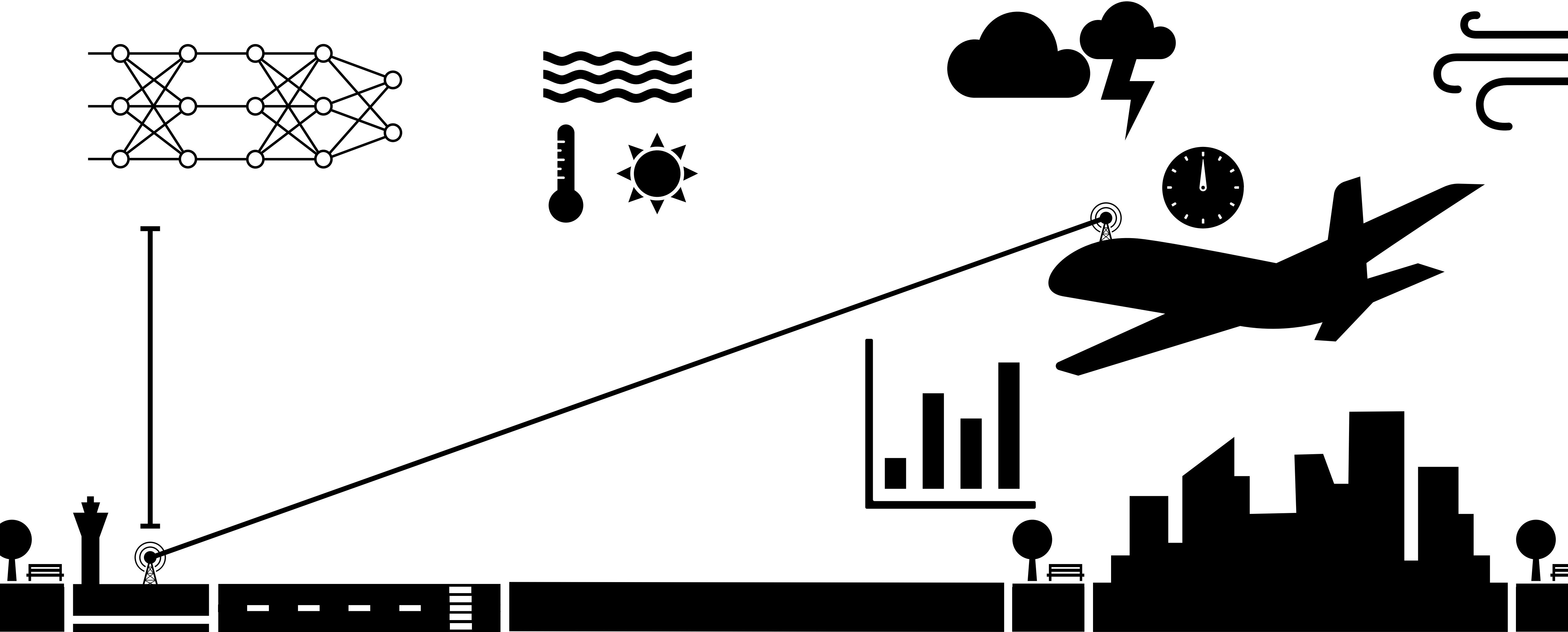
Sound!

concrete

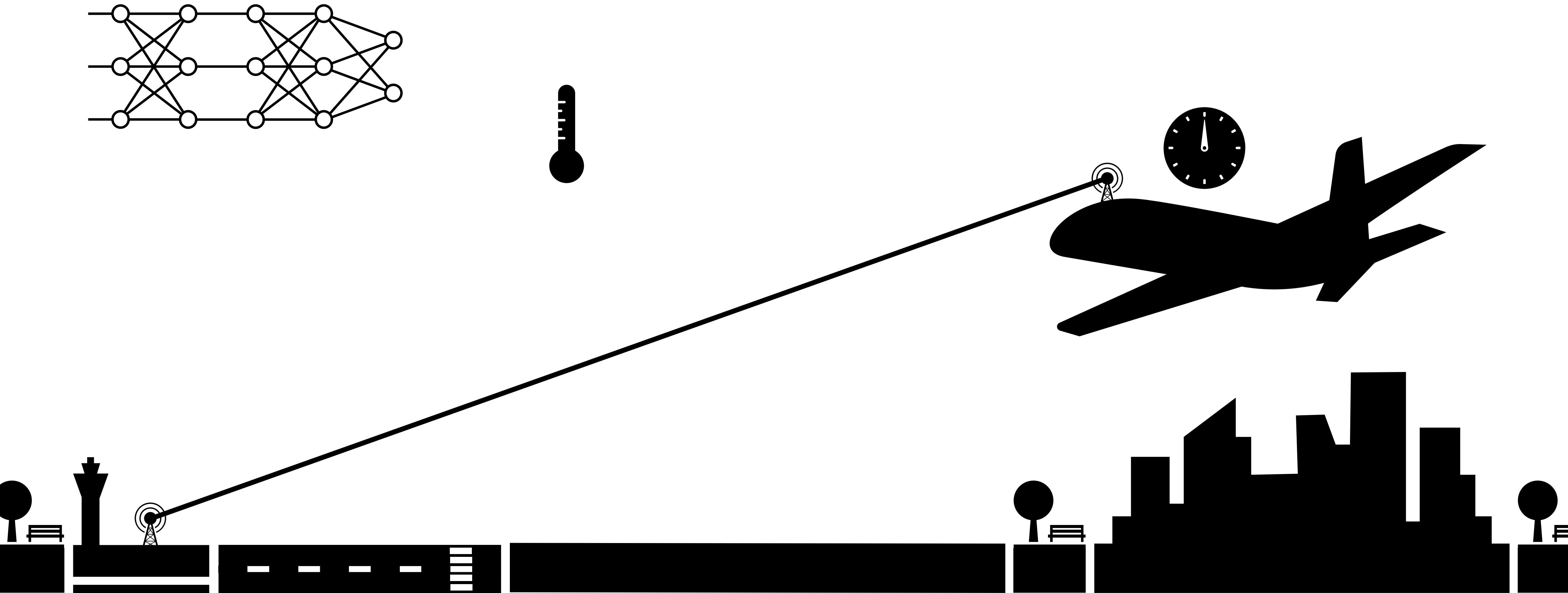
abstract

$$\text{CountChanges}_i(P) \leq \text{ImpactAnalysis}_i^\natural(P)$$

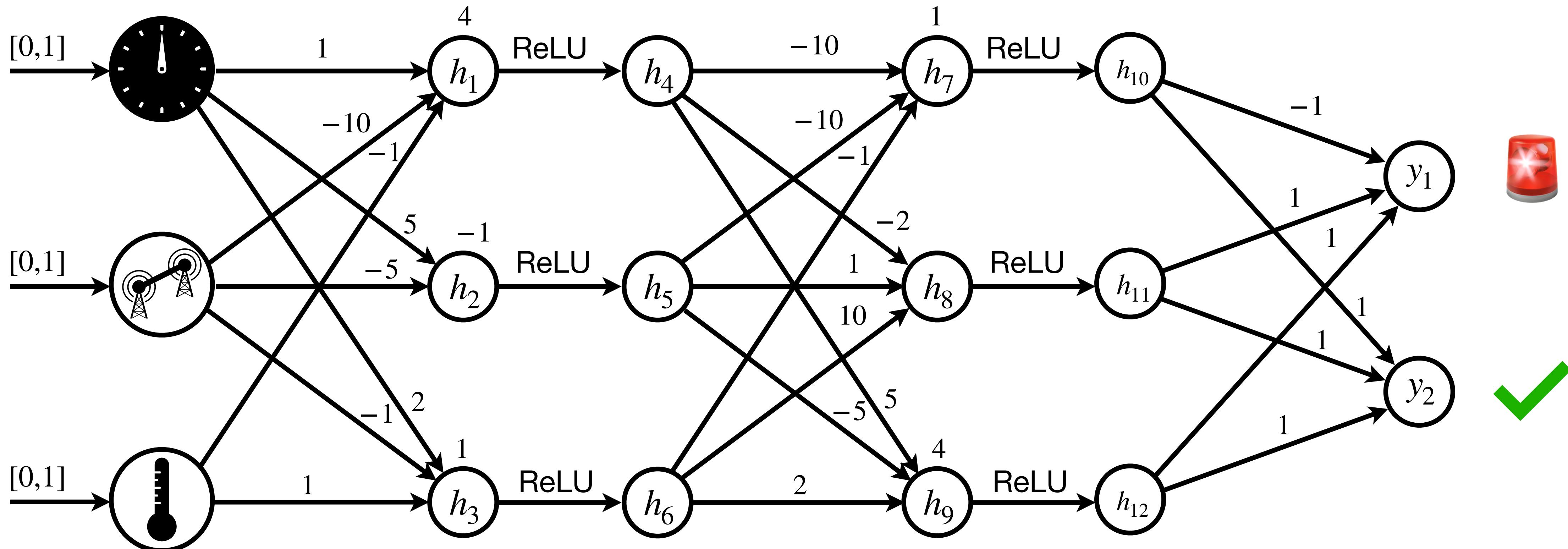
Landing alarm system



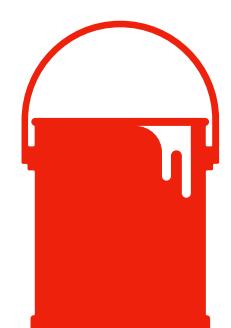
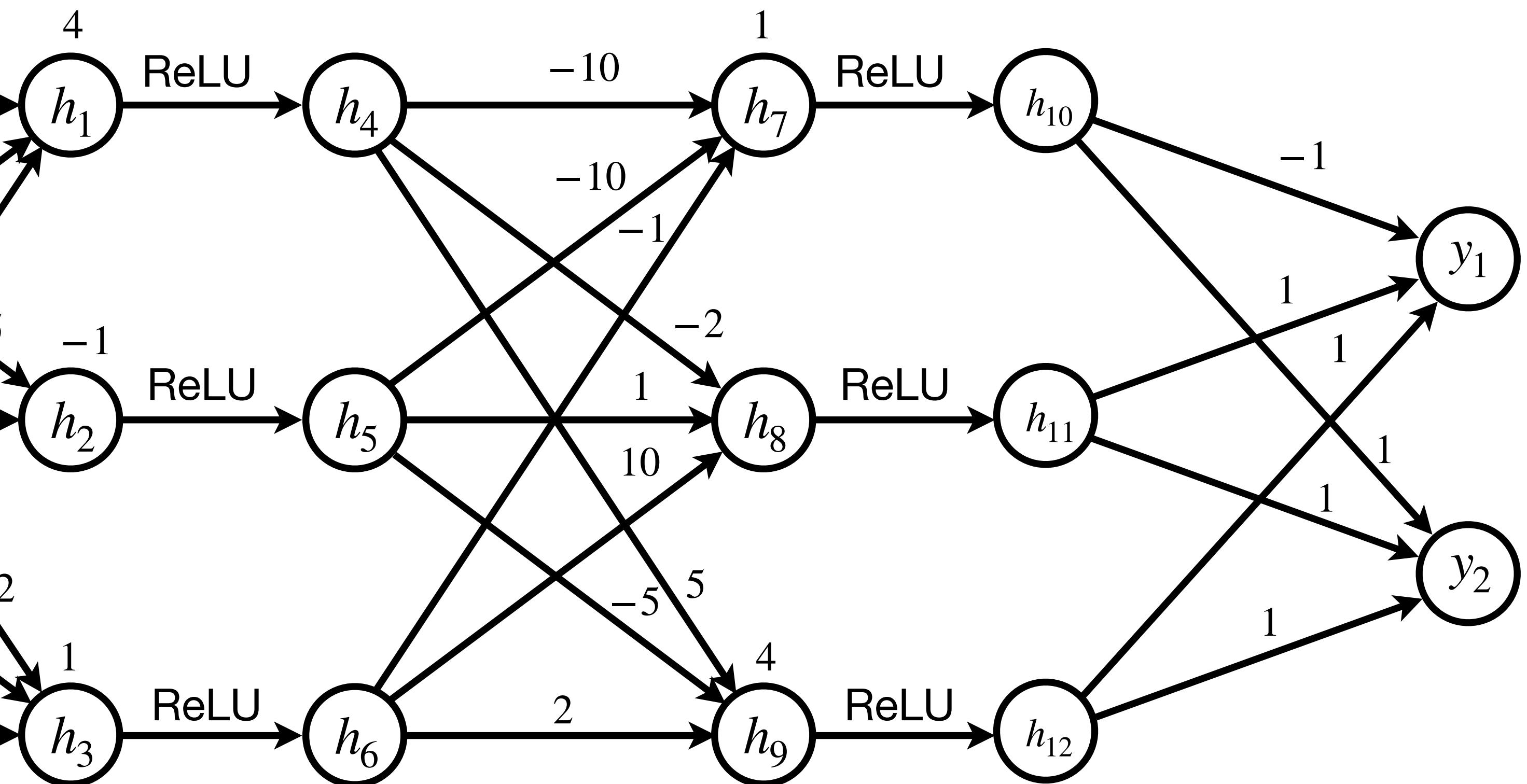
Simplified landing alarm system



Feed-forward neural network



Bucket abstraction



$y_2 \leq y_1$



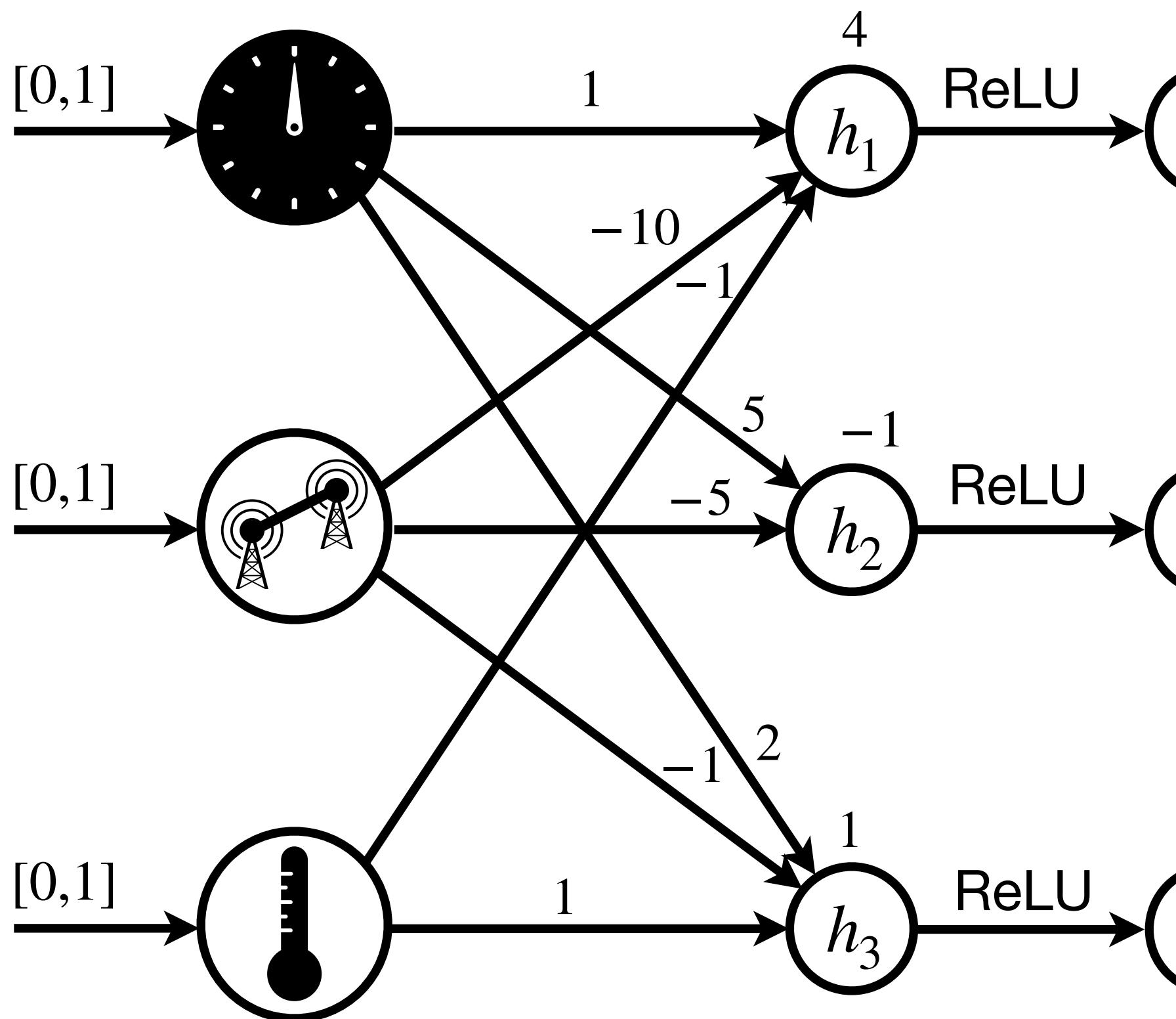
$y_1 \leq y_2$



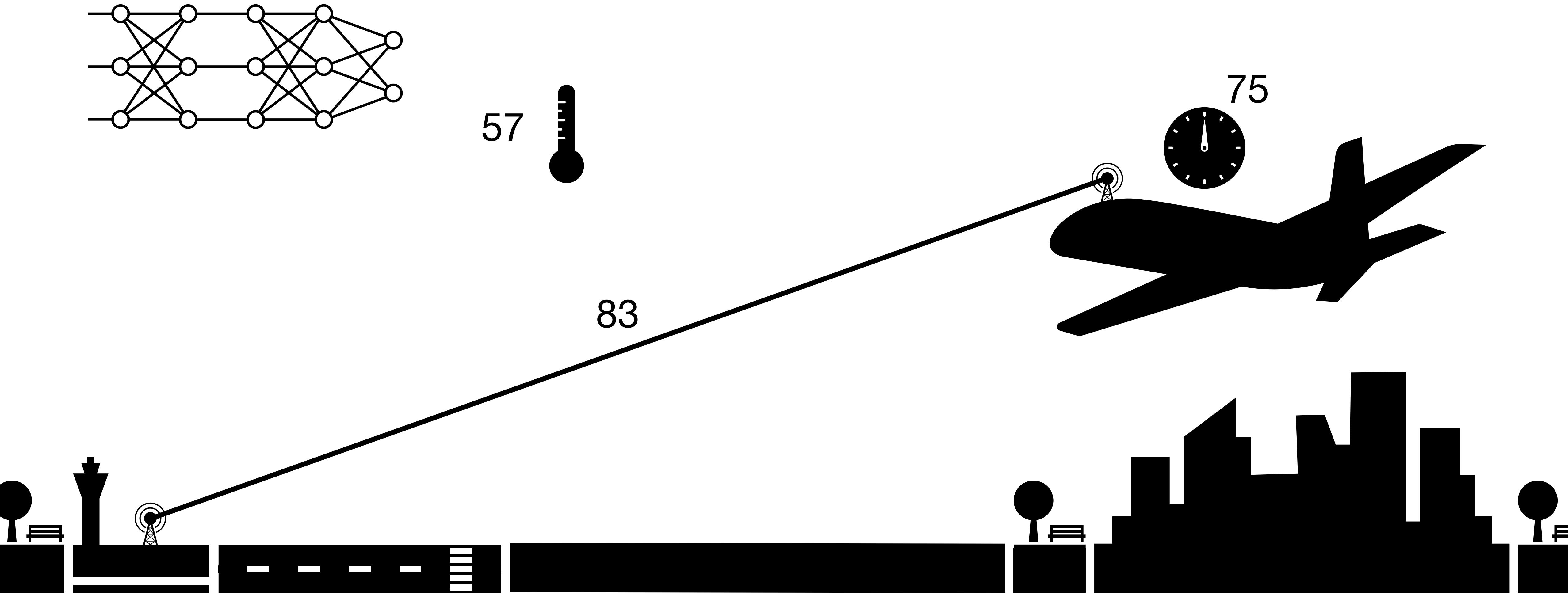
ImpactAnalysis  $(P) = 75$

ImpactAnalysis  $(P) = 83$

ImpactAnalysis  $(P) = 57$



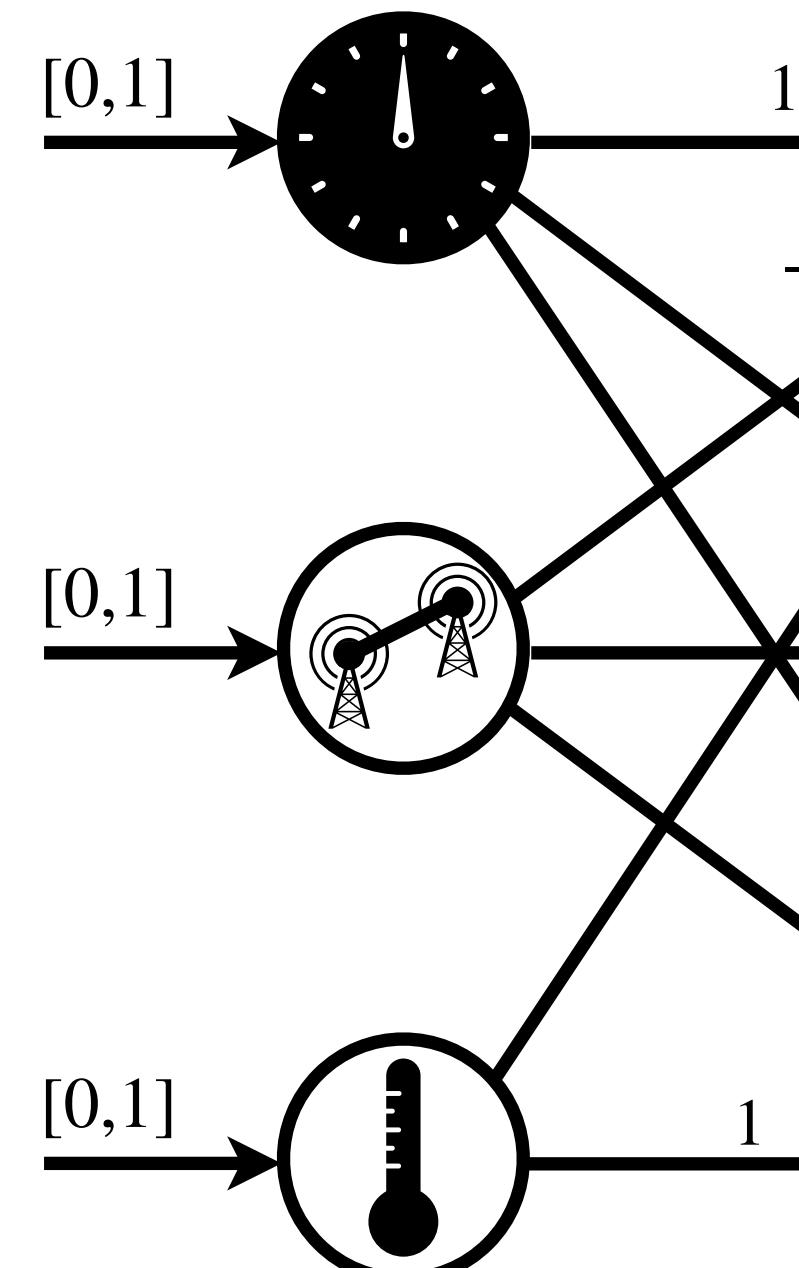
Simplified landing alarm system

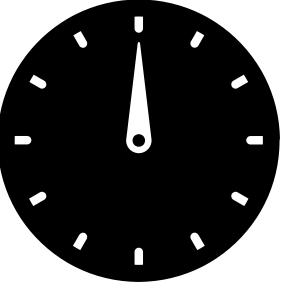
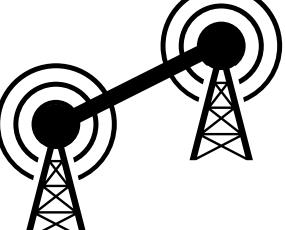


ImpactAnalysis_i

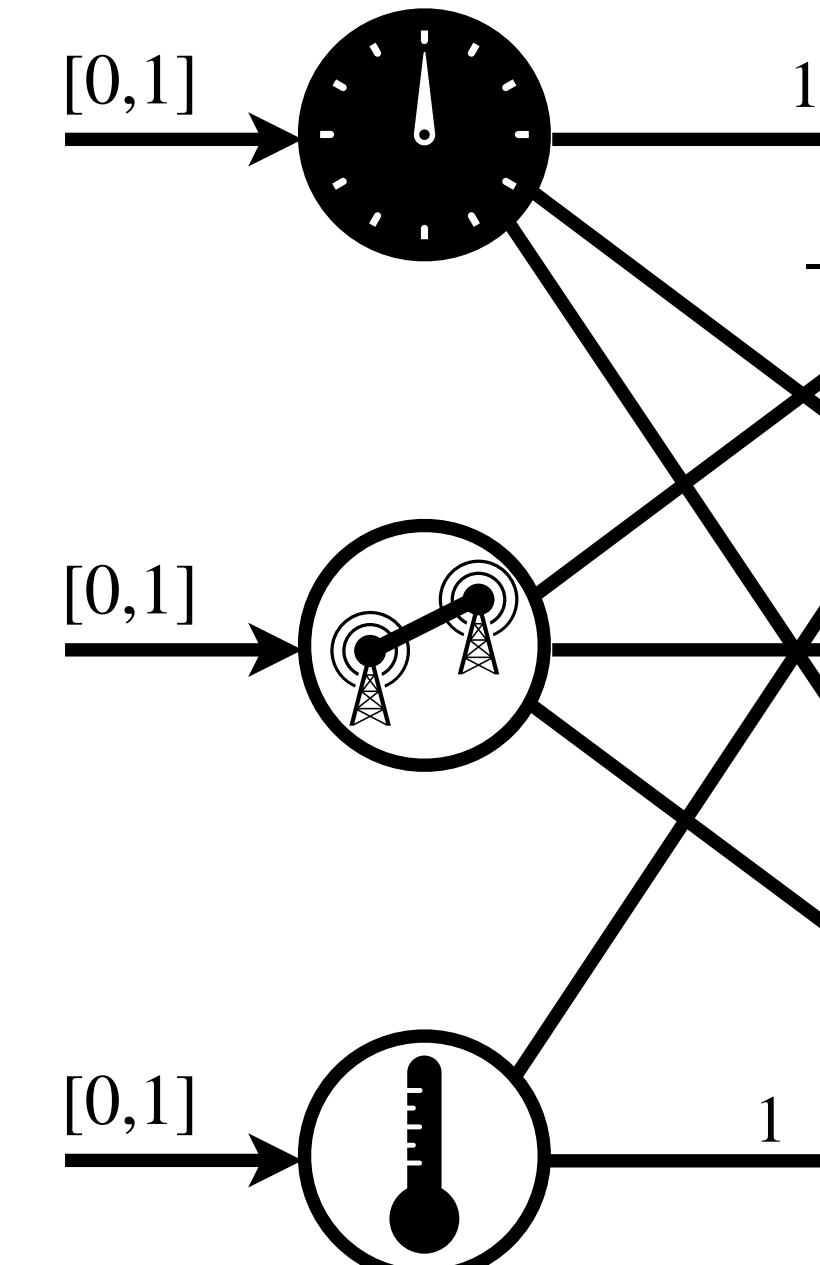
Baseline

	75	??
	83	??
	57	??



	ImpactAnalysis_i	Baseline	Stochastic
	75	??	??
	83	??	??
	57	??	??

Permutation
Feature
Importance



Experiments setup

- Diabetes
 - Wine quality
 - RPG Videogame
 - Rain Sidney
- 4 Databases

Experiments setup

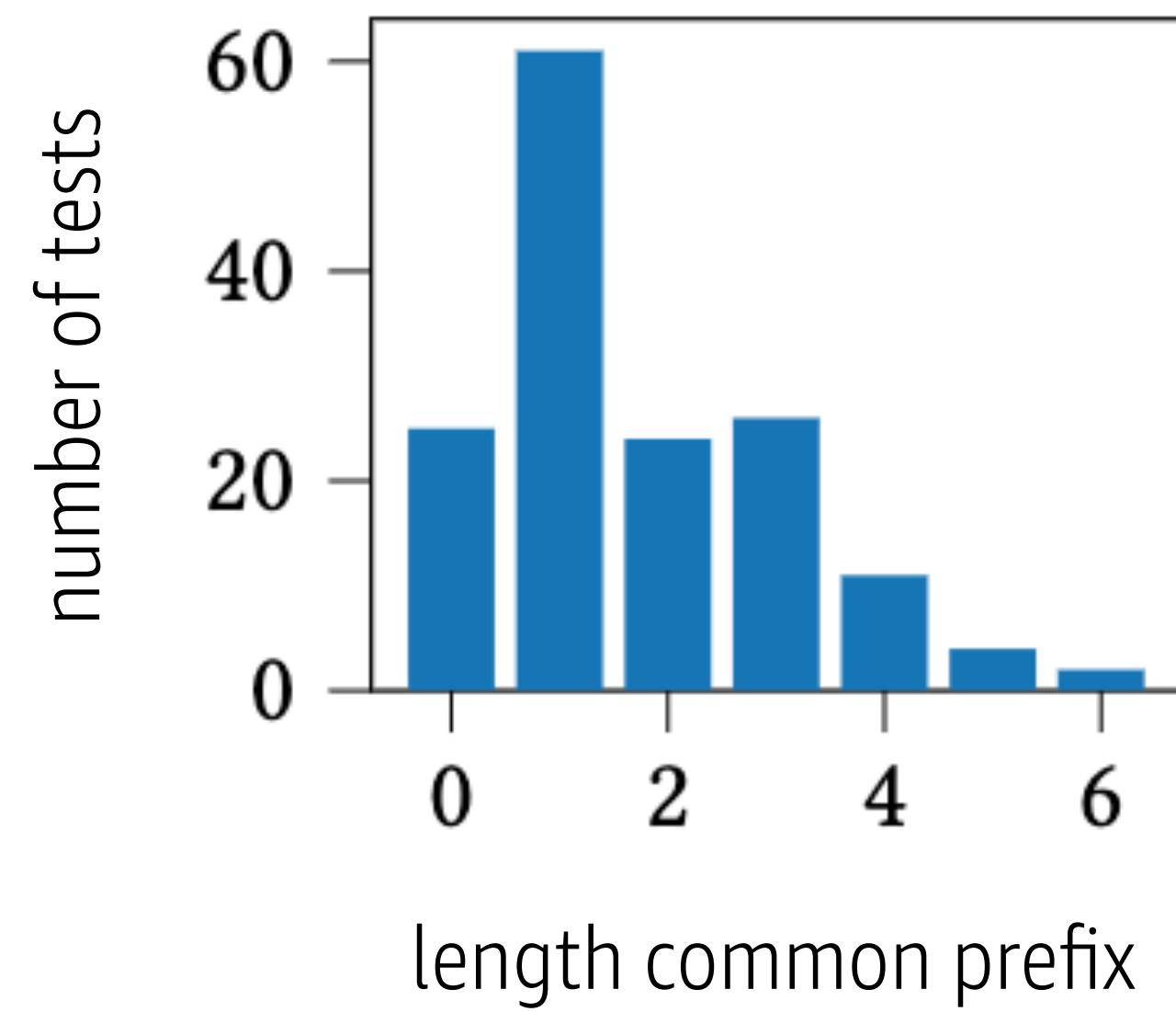
- Diabetes
- Wine quality
- RPG Videogame
- Rain Sidney

Common prefix
between ordering
of input features

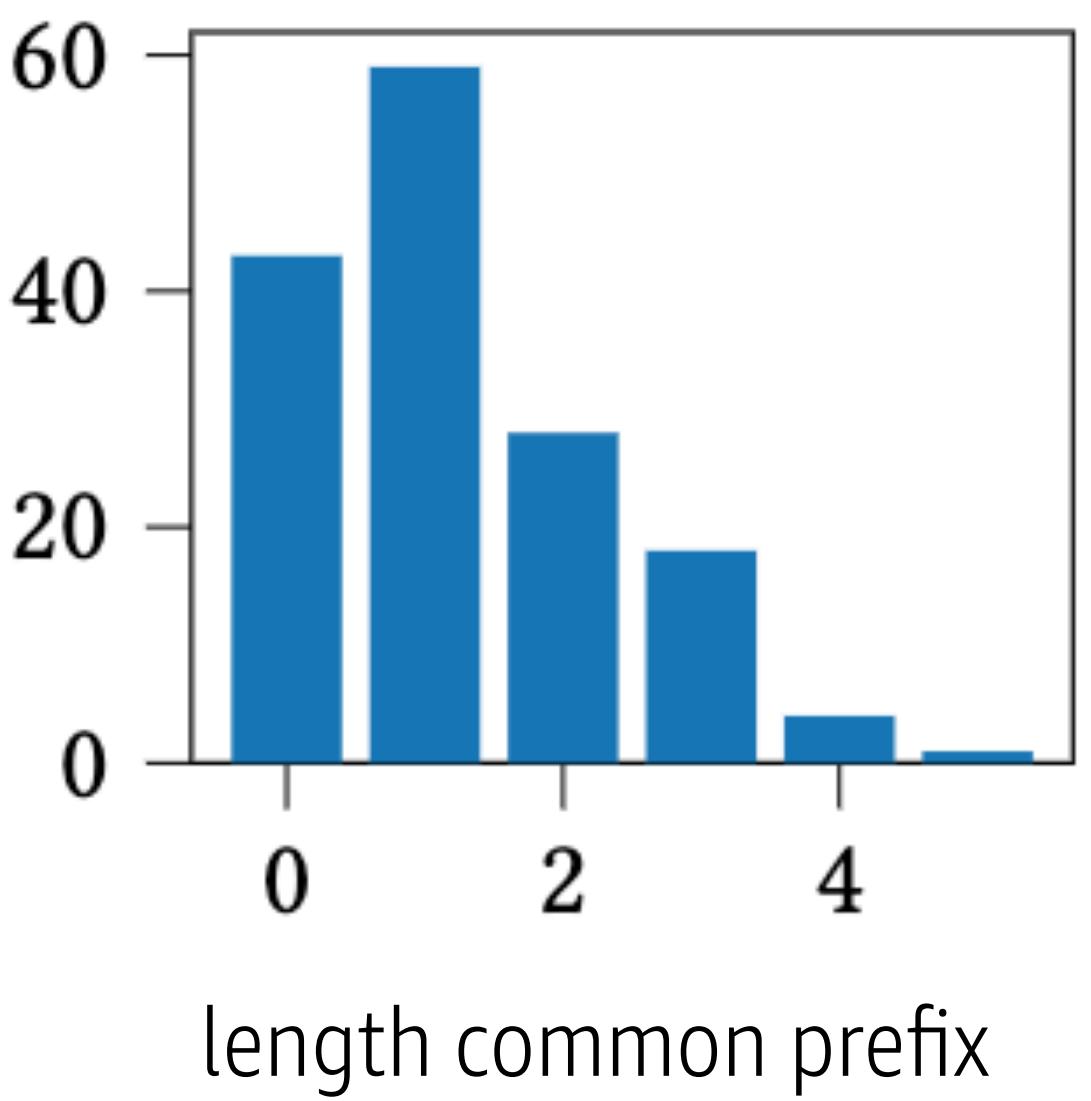
- 4 Databases
 - Baseline vs Permutation Feature Importance
 - Baseline vs retraining
 - Baseline vs our ImpactAnalysis $_i^{\bowtie}$

Baseline vs

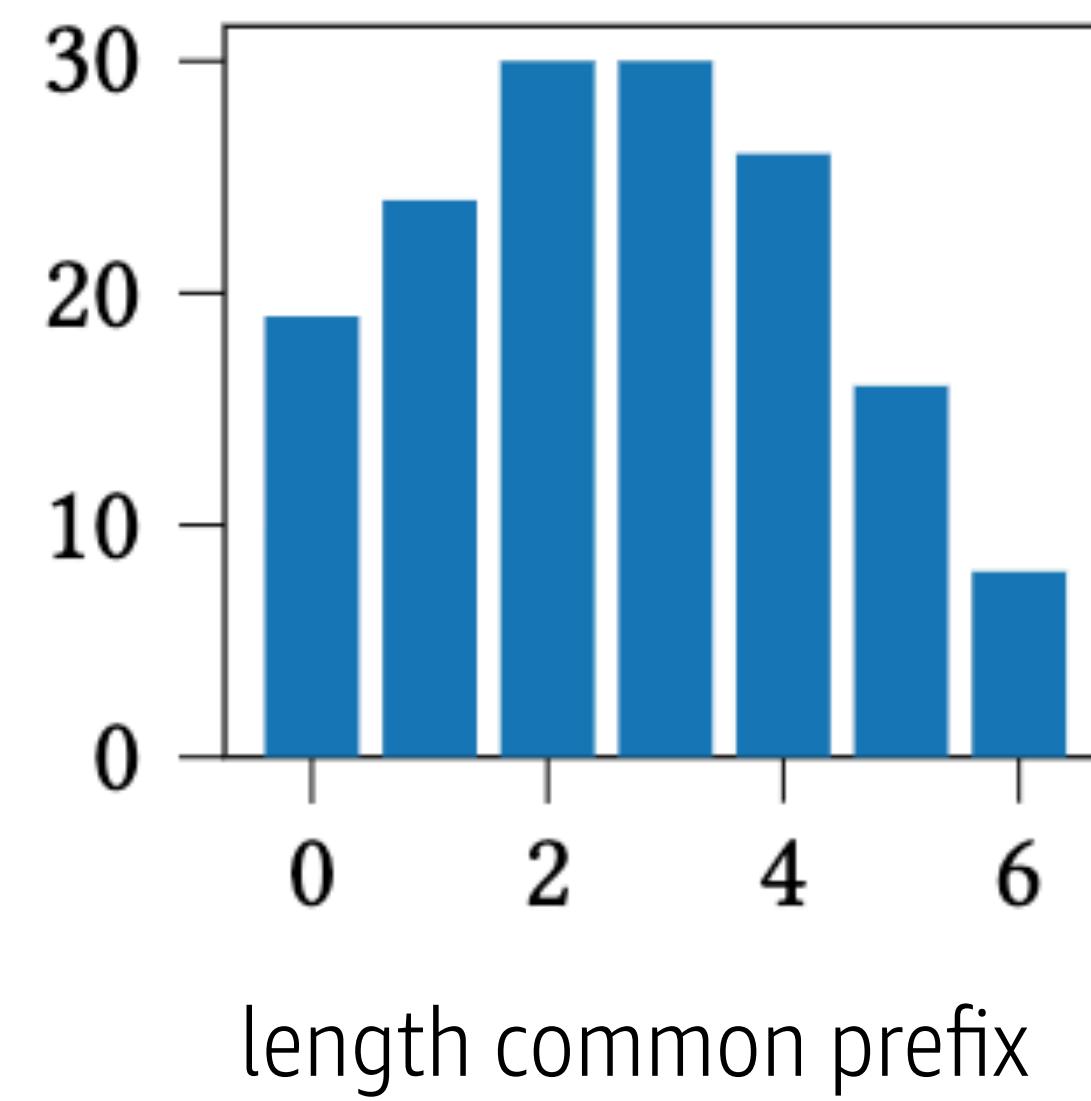
PFI



Retraining



ImpactAnalysis_i



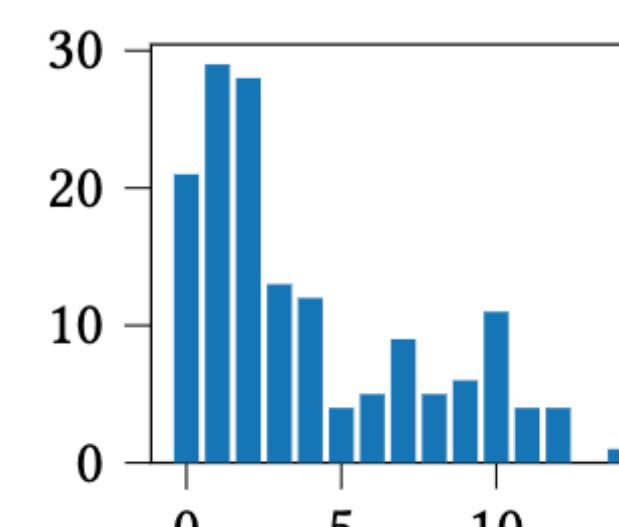
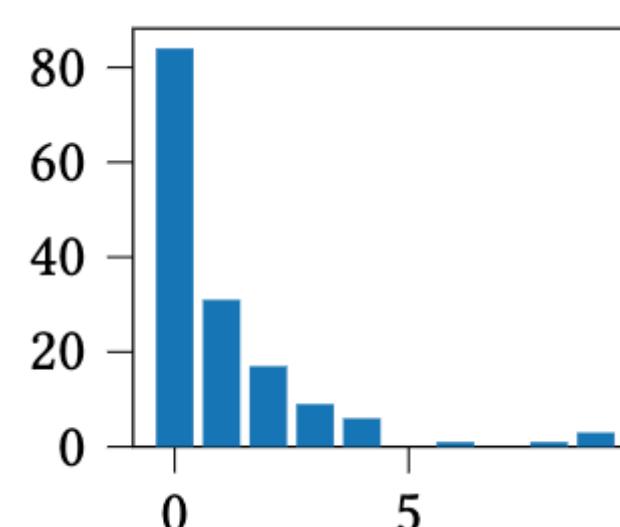
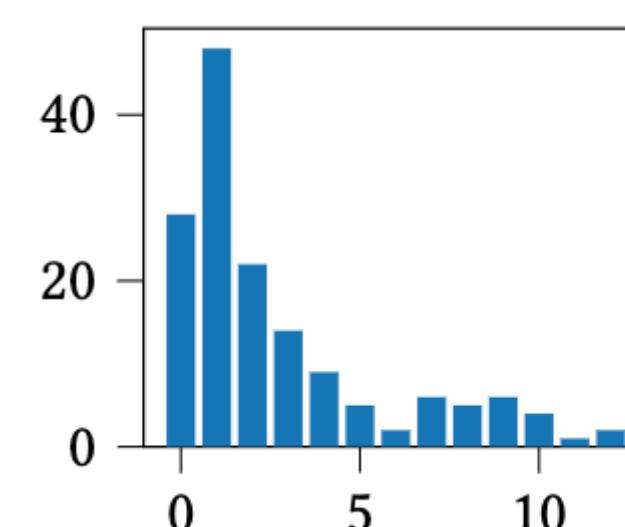
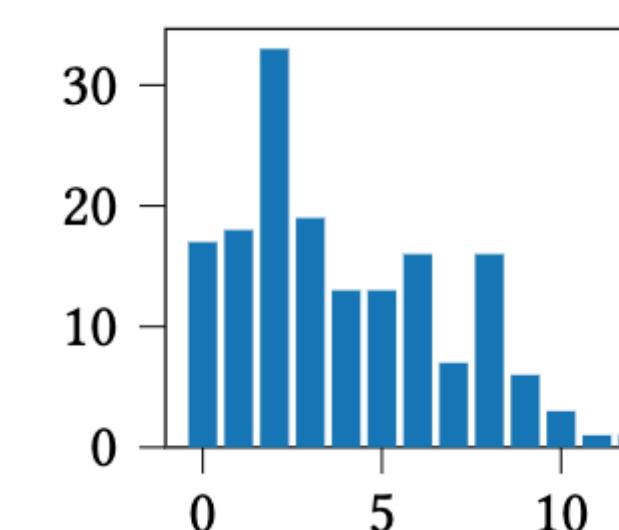
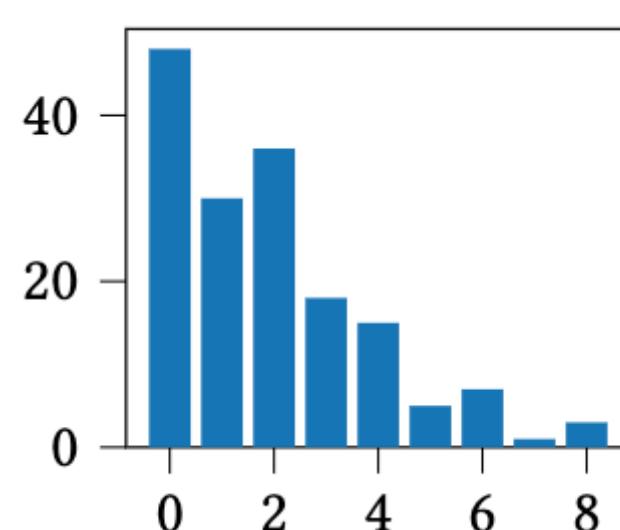
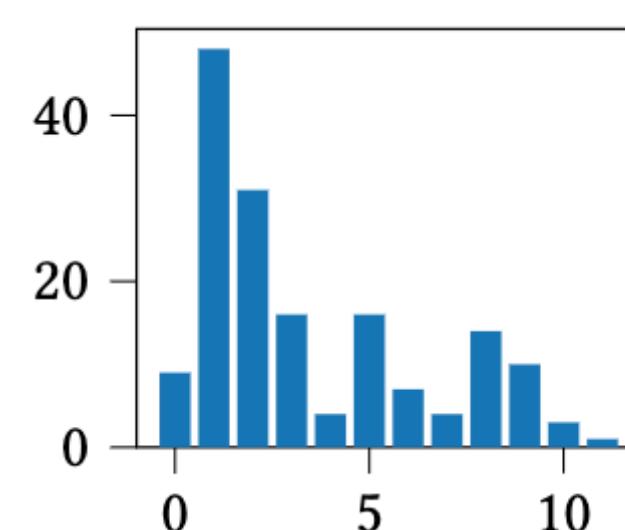
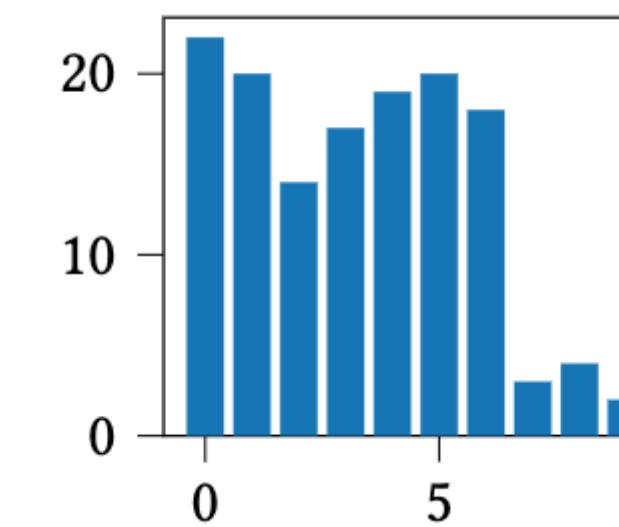
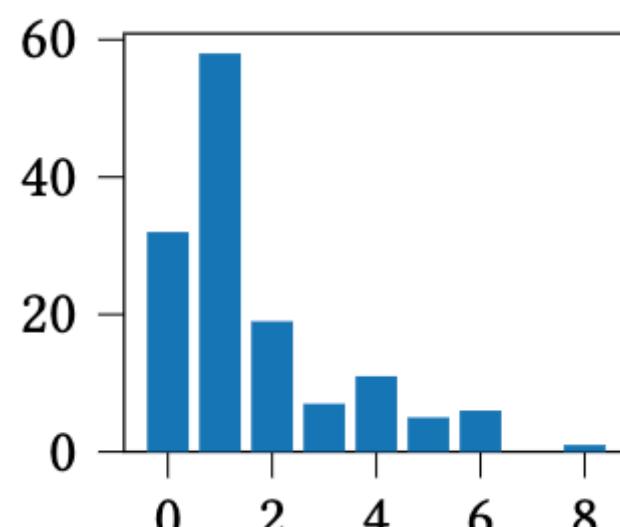
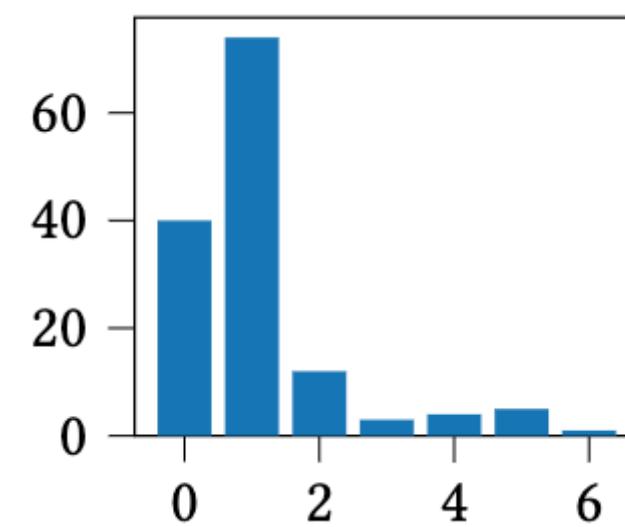
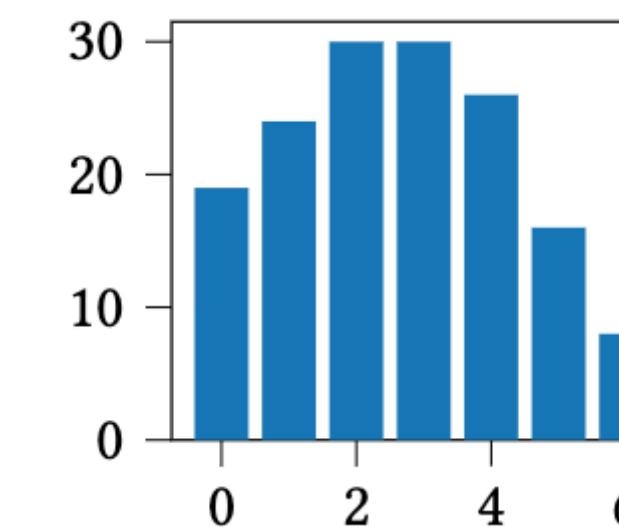
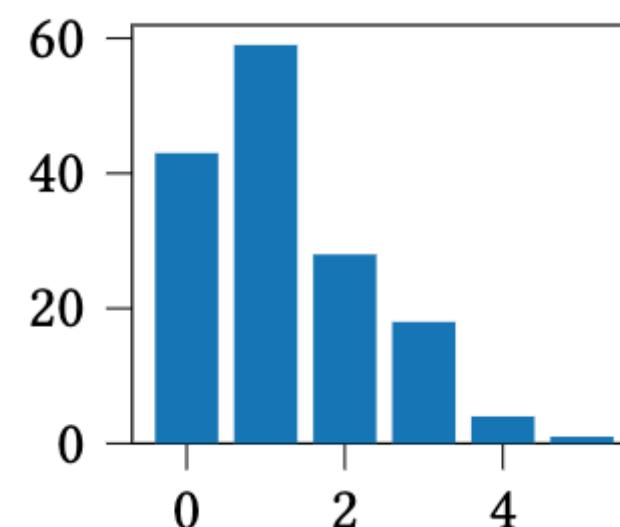
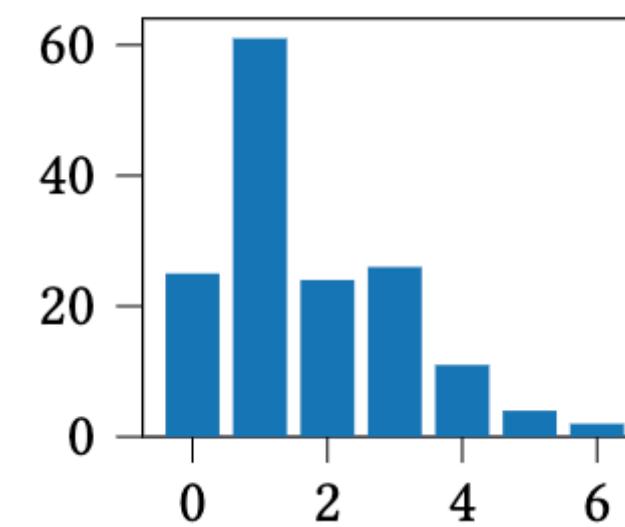
Diabetes

Baseline vs

PFI

Retraining

ImpactAnalysis^h_i



Diabetes

Wine quality

Videogame

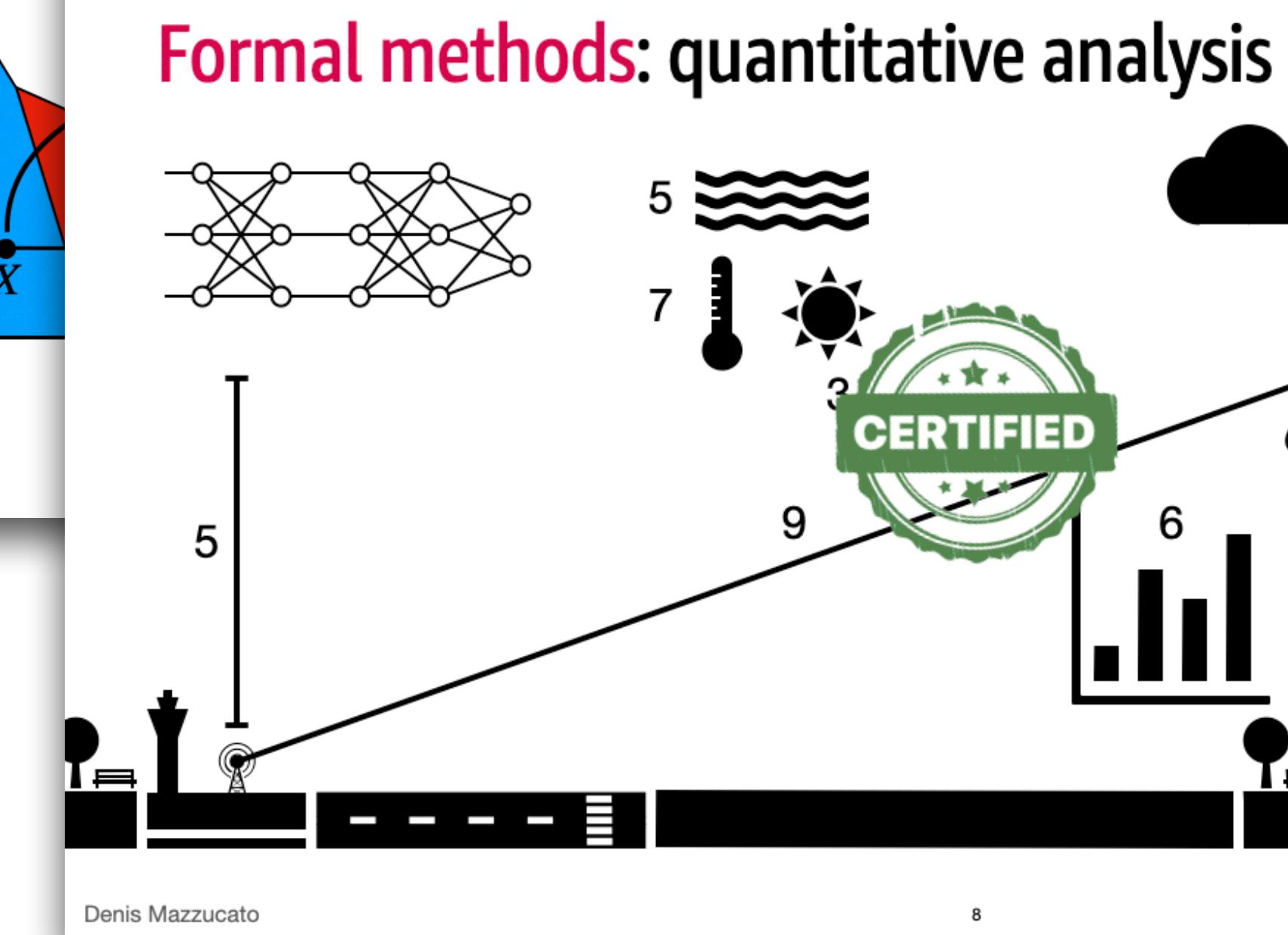
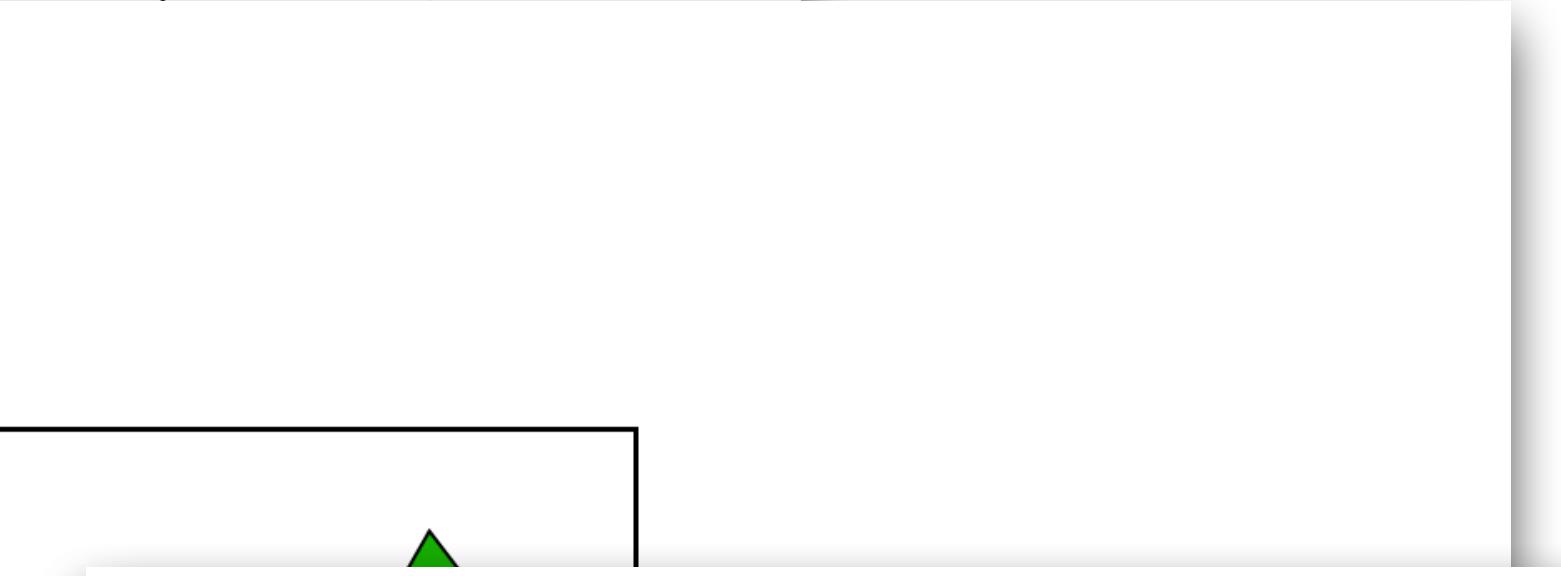
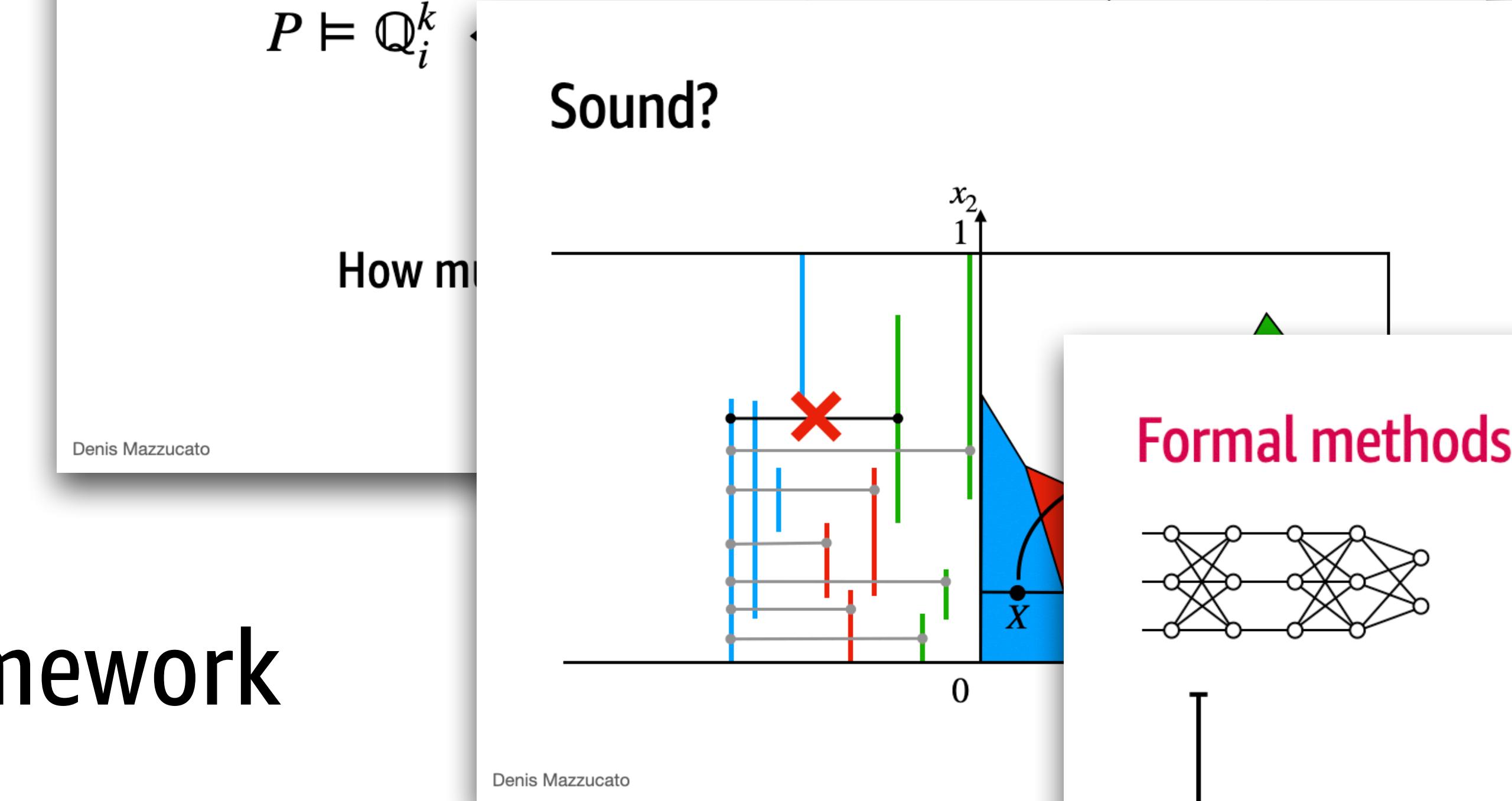
Rain Sidney

Conclusion

- Impact formal framework
- Concrete CountChanges_i
- Sound abstraction ImpactAnalysis_i

$$P \models Q_i^k \quad \text{impact}_i \in \text{Traces} \rightarrow \mathbb{D}$$
$$Q_i^k = \{ \llbracket X \rrbracket \mid \text{impact}_i(\llbracket X \rrbracket) \leq k \}$$

Quantities Domain



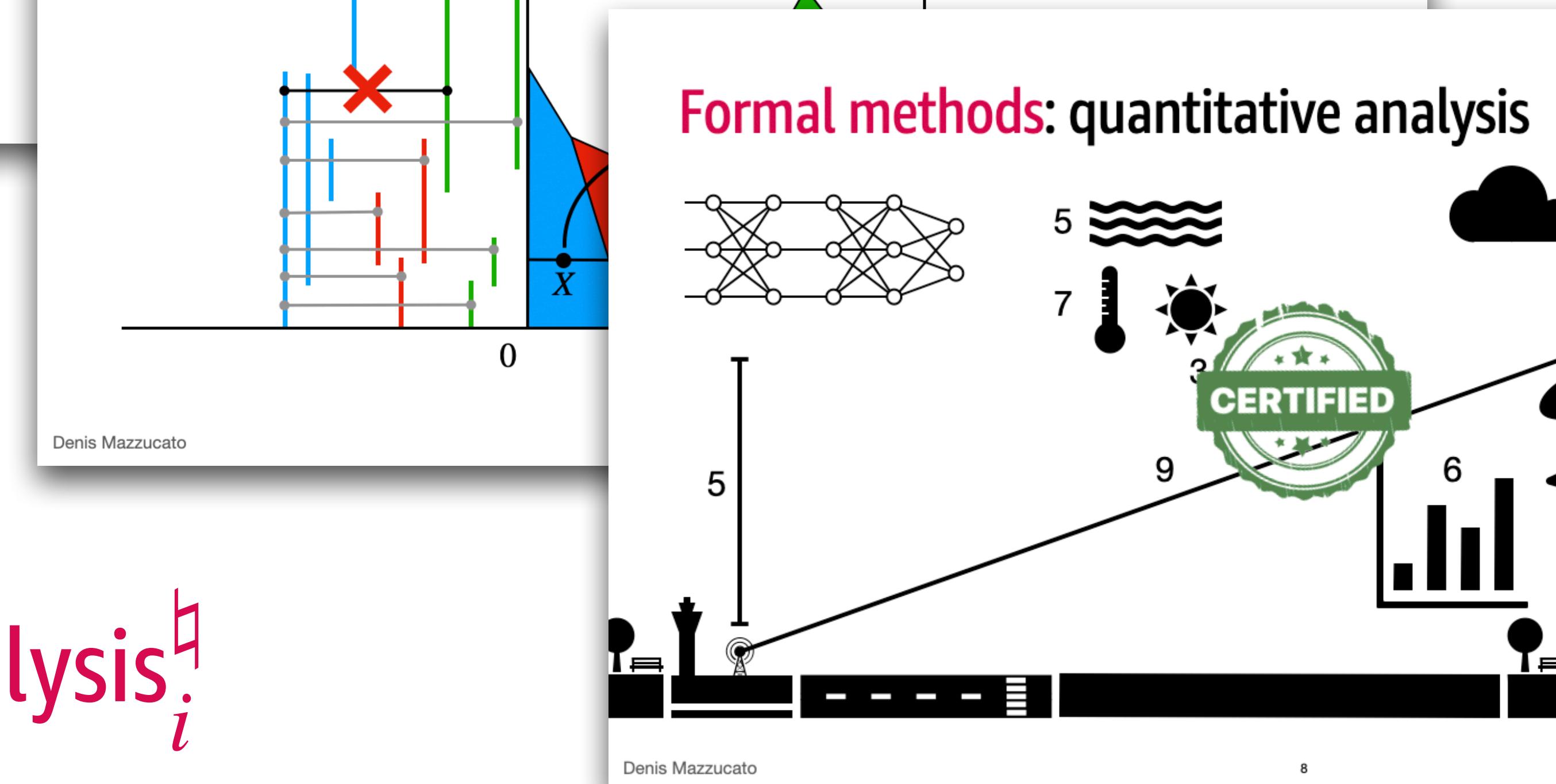
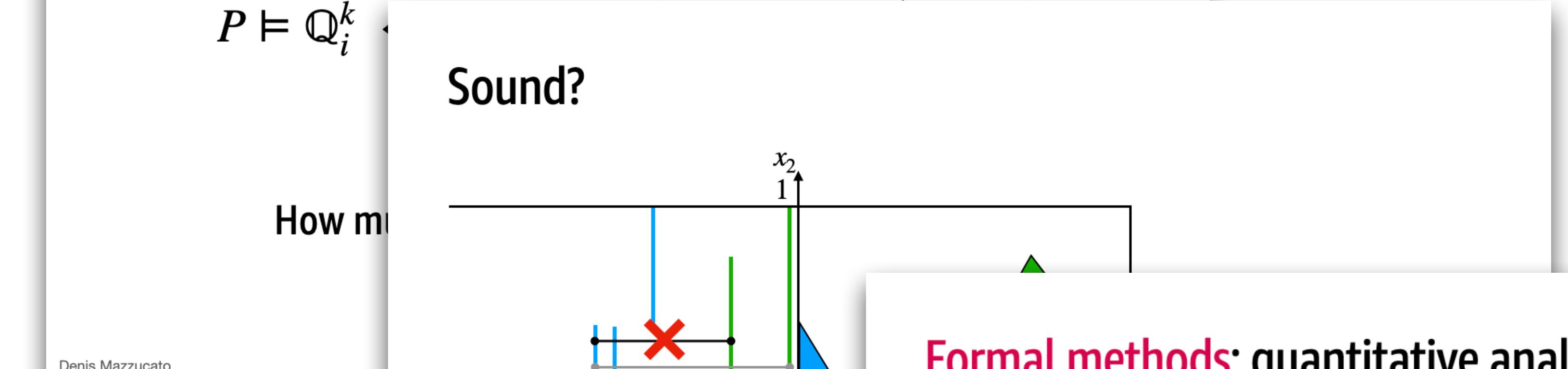
Conclusion

Sound by construction!

- Impact formal framework
- Concrete CountChanges_i
- Sound abstraction ImpactAnalysis_i

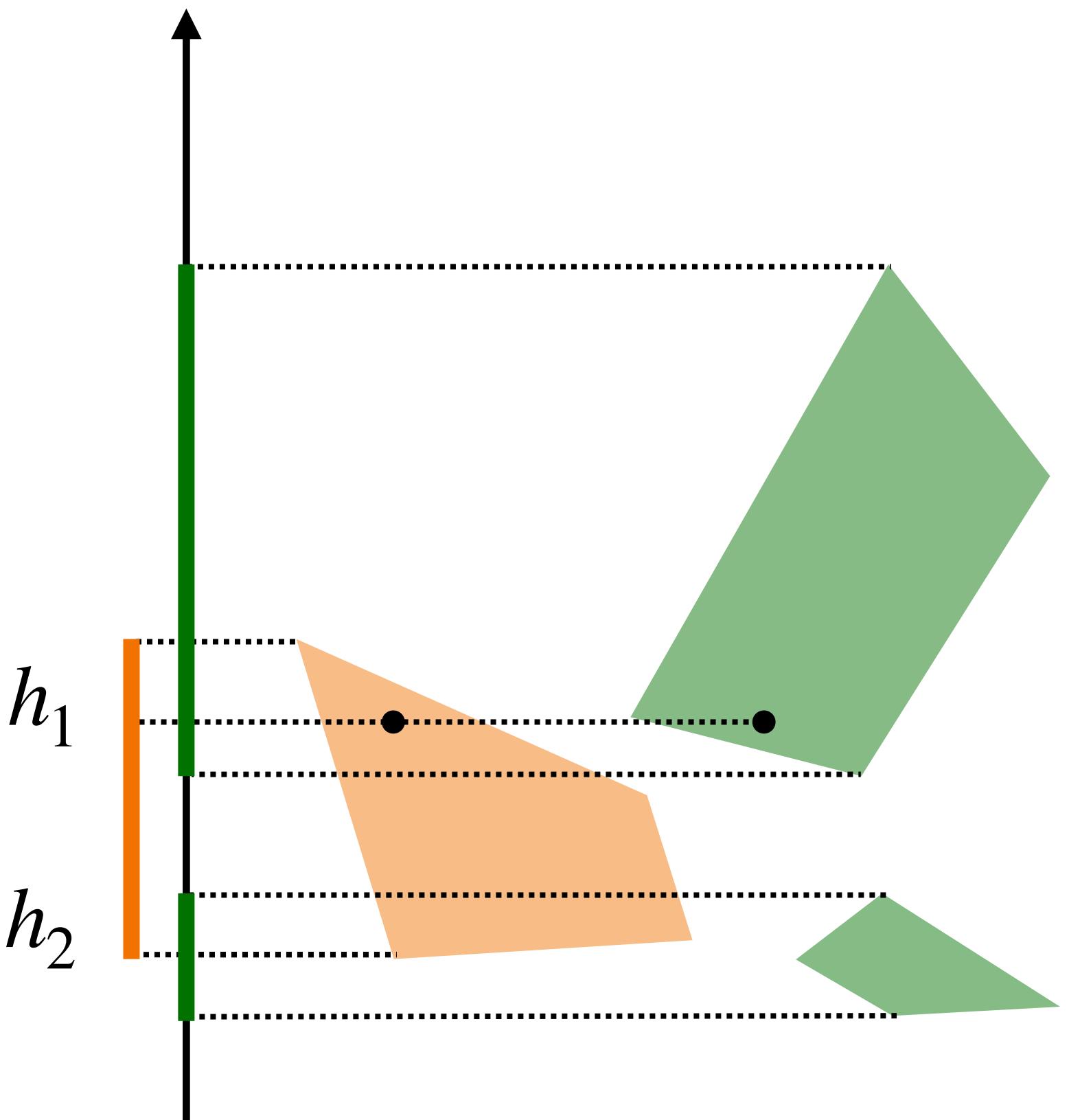
$$P \models Q_i^k \quad \text{impact}_i \in \text{Traces} \rightarrow \mathbb{D}$$
$$Q_i^k = \{ \llbracket X \rrbracket \mid \text{impact}_i(\llbracket X \rrbracket) \leq k \}$$

Quantities Domain



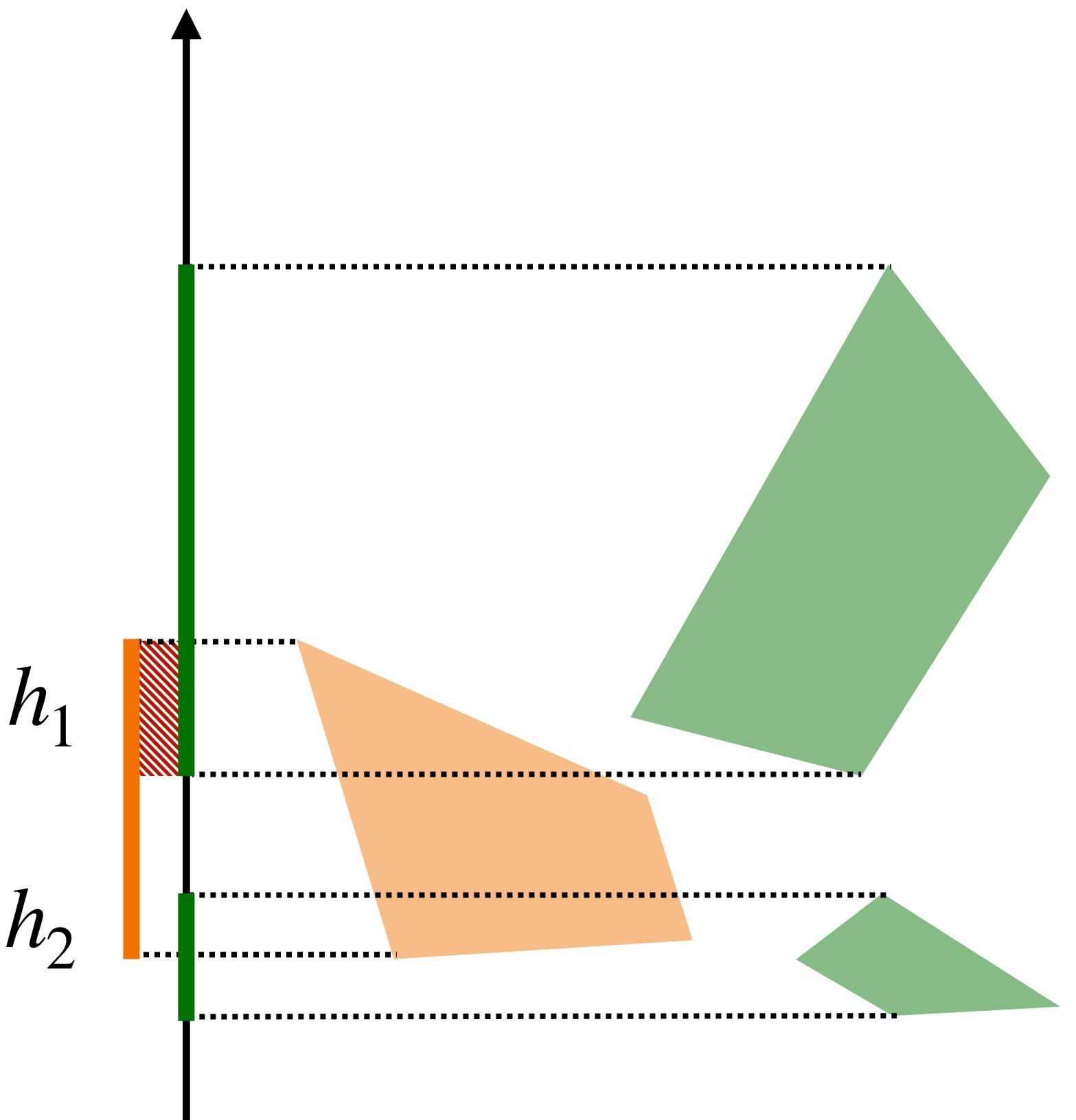
Future work

- Other **impact** instances
 - Checking Intersections



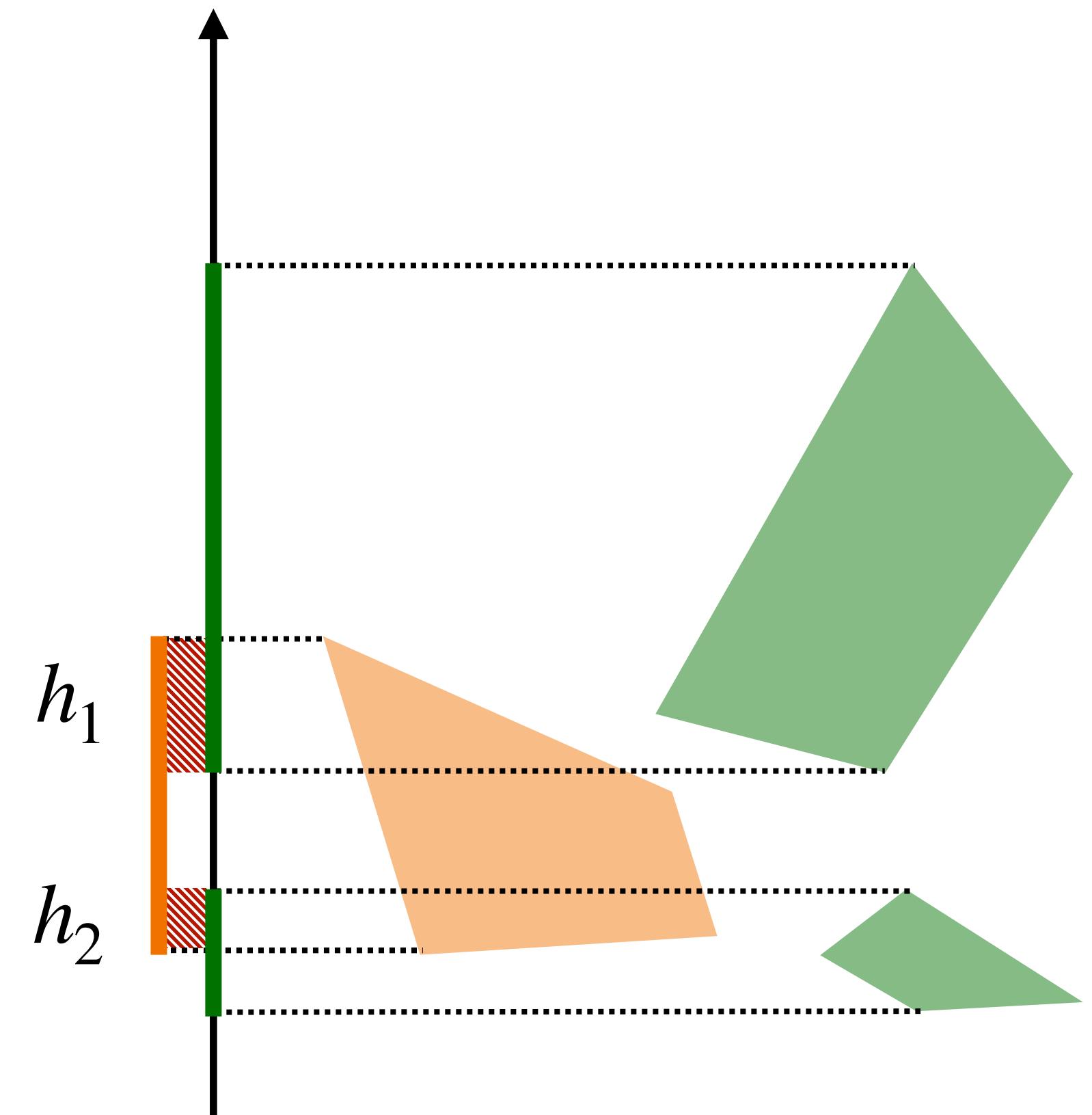
Future work

- Other **impact** instances
 - Checking Intersections
 - Maximum Volume



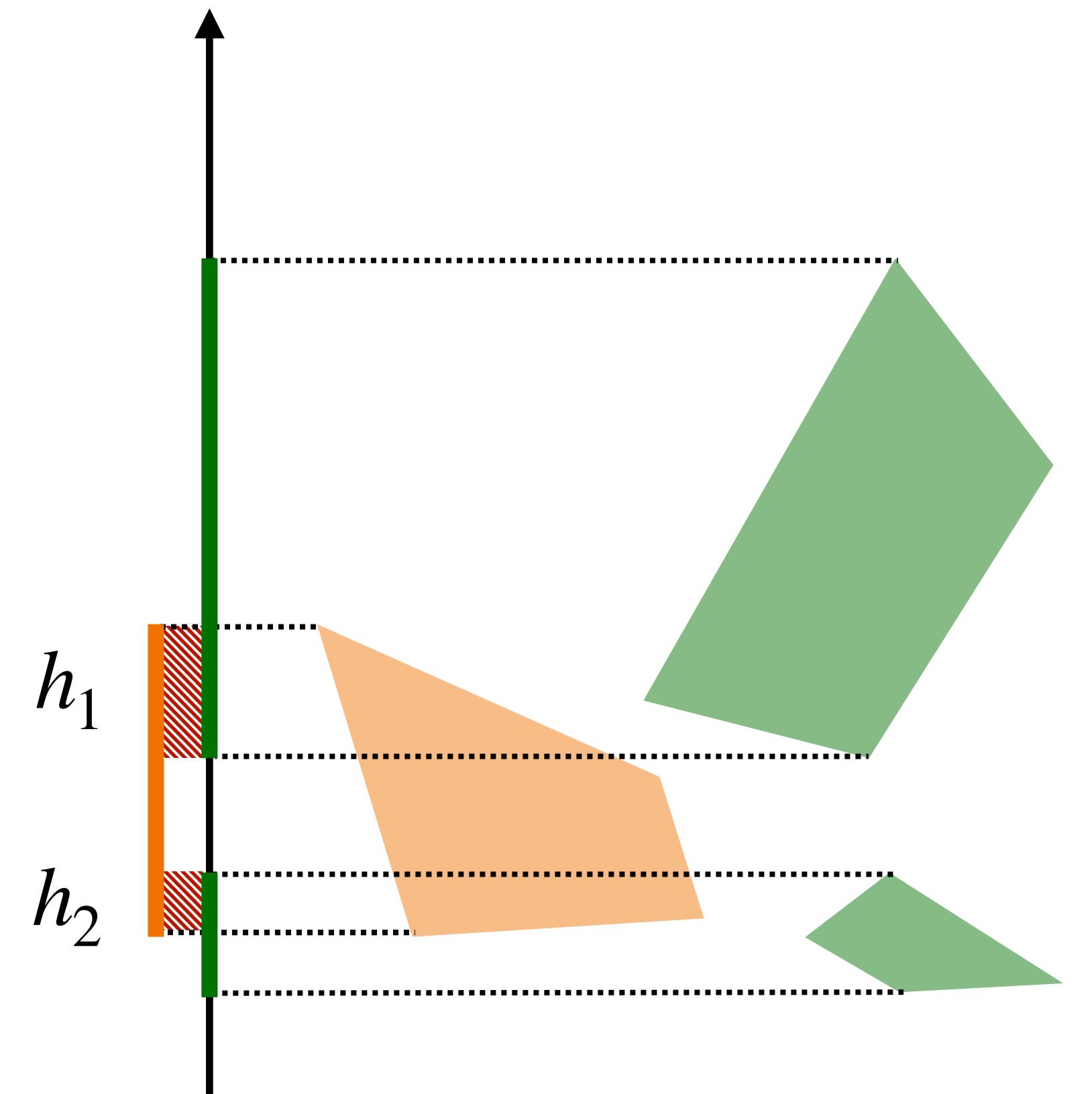
Future work

- Other **impact** instances
 - Checking Intersections
 - Maximum Volume
 - Disjoint Sum of Volumes



Future work

- Other **impact** instances
 - Checking Intersections
 - Maximum Volume
 - Disjoint Sum of Volumes
- Other **application contexts**
 - Not only neural networks
 - Jupyter notebooks



Future work

- Other **impact** instances
 - Checking Intersections
 - Maximum Volume
 - Disjoint Sum of Volumes
- Other application contexts
 - Not only neural networks
 - Jupyter notebooks

Our work

- Impact formal framework
- Concrete **CountChanges**.
- Abstract **ImpactAnalysis**.

Sound by
construction!