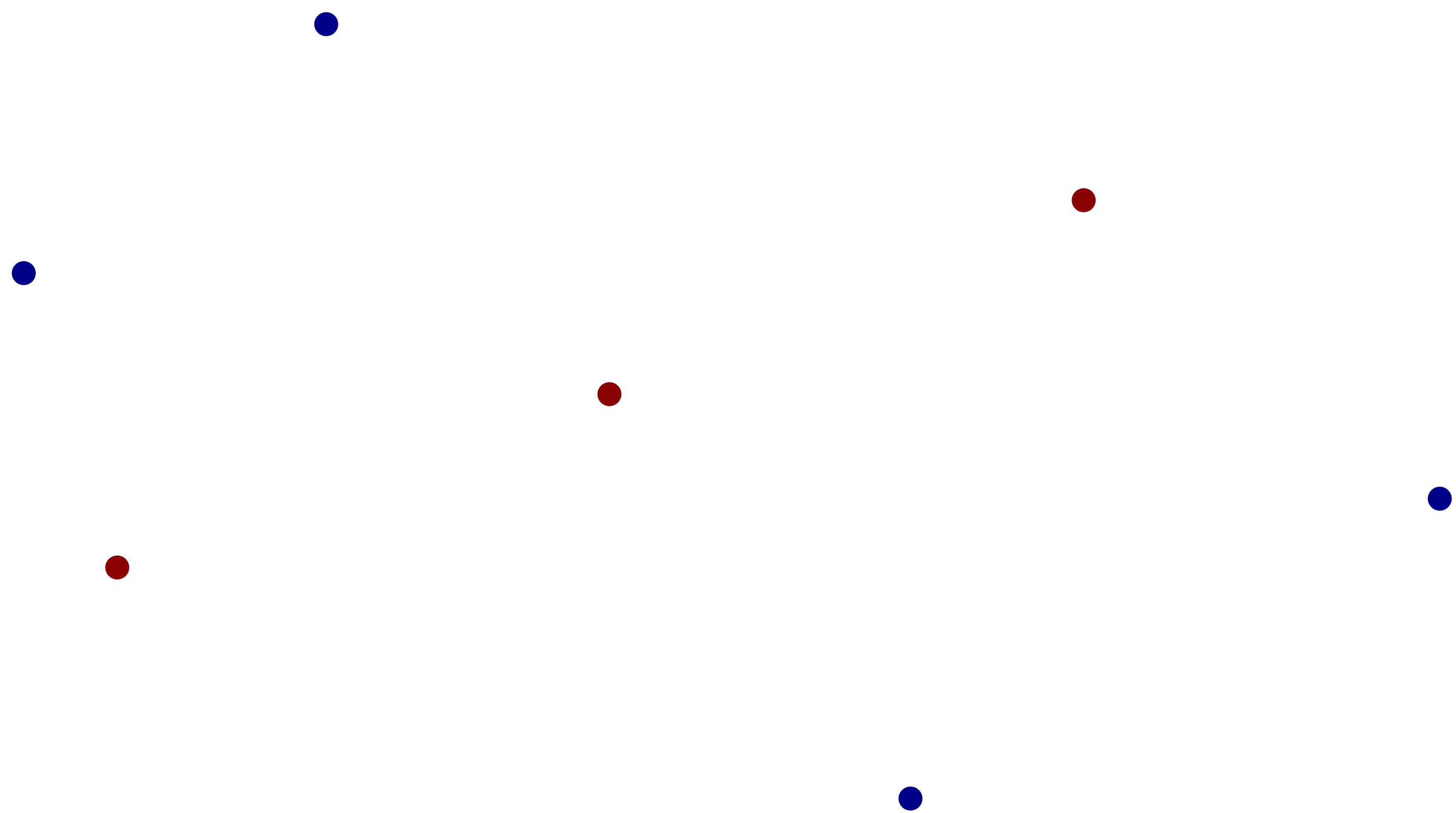
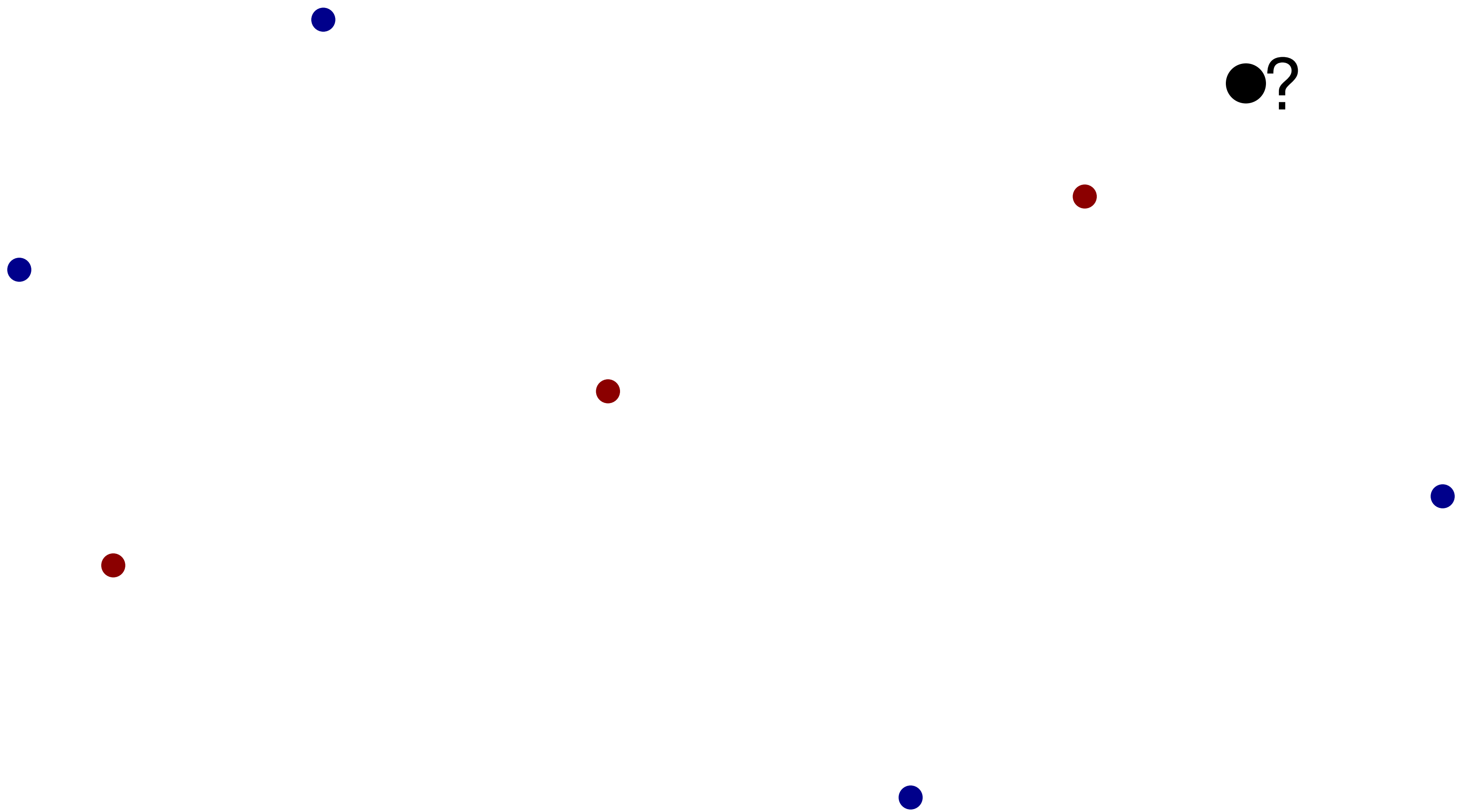
Reducing Nearest Neighbor Training Sets Optimally and Exactly

Josiah Rohrer and Simon Weber

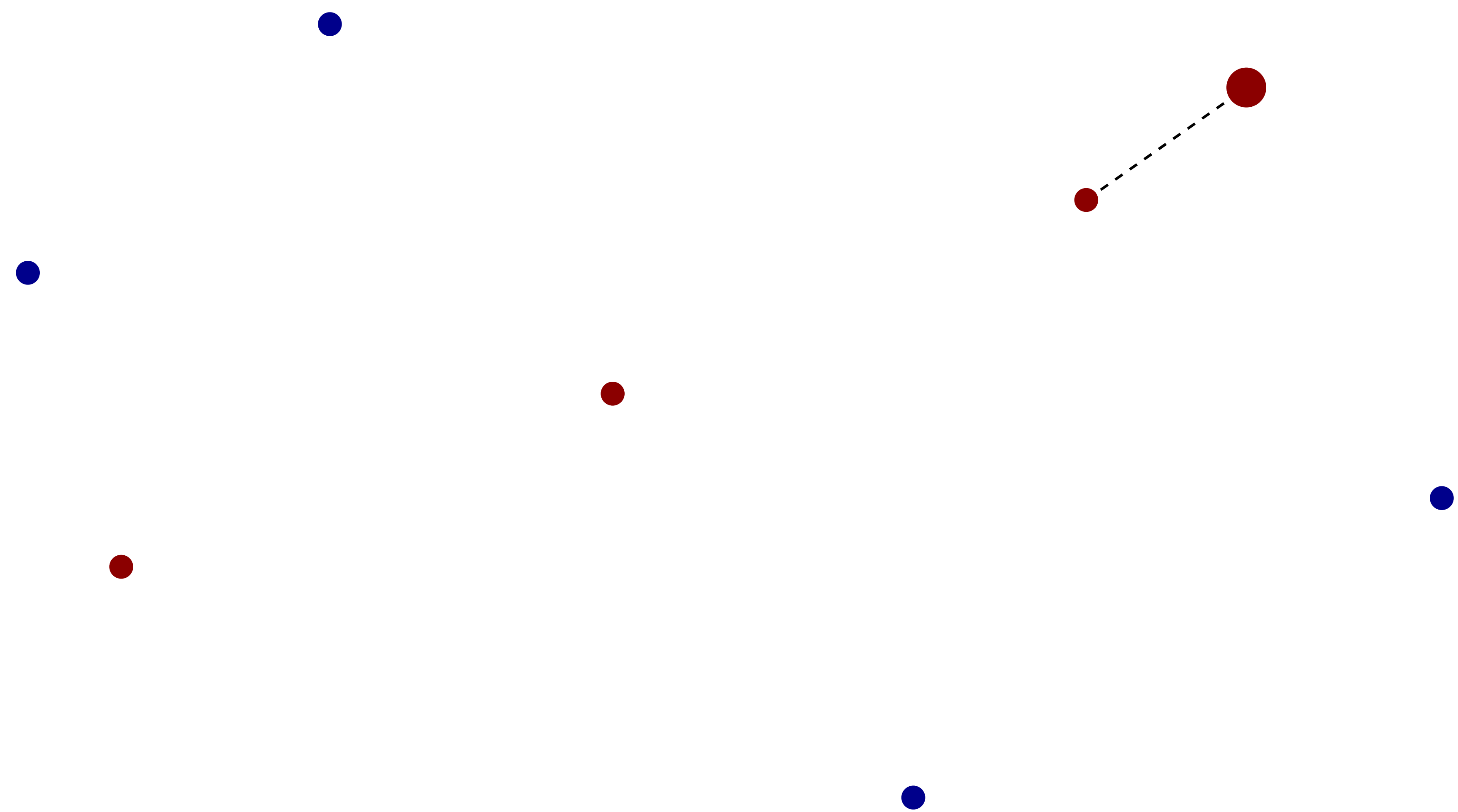
Nearest-Neighbor Classification



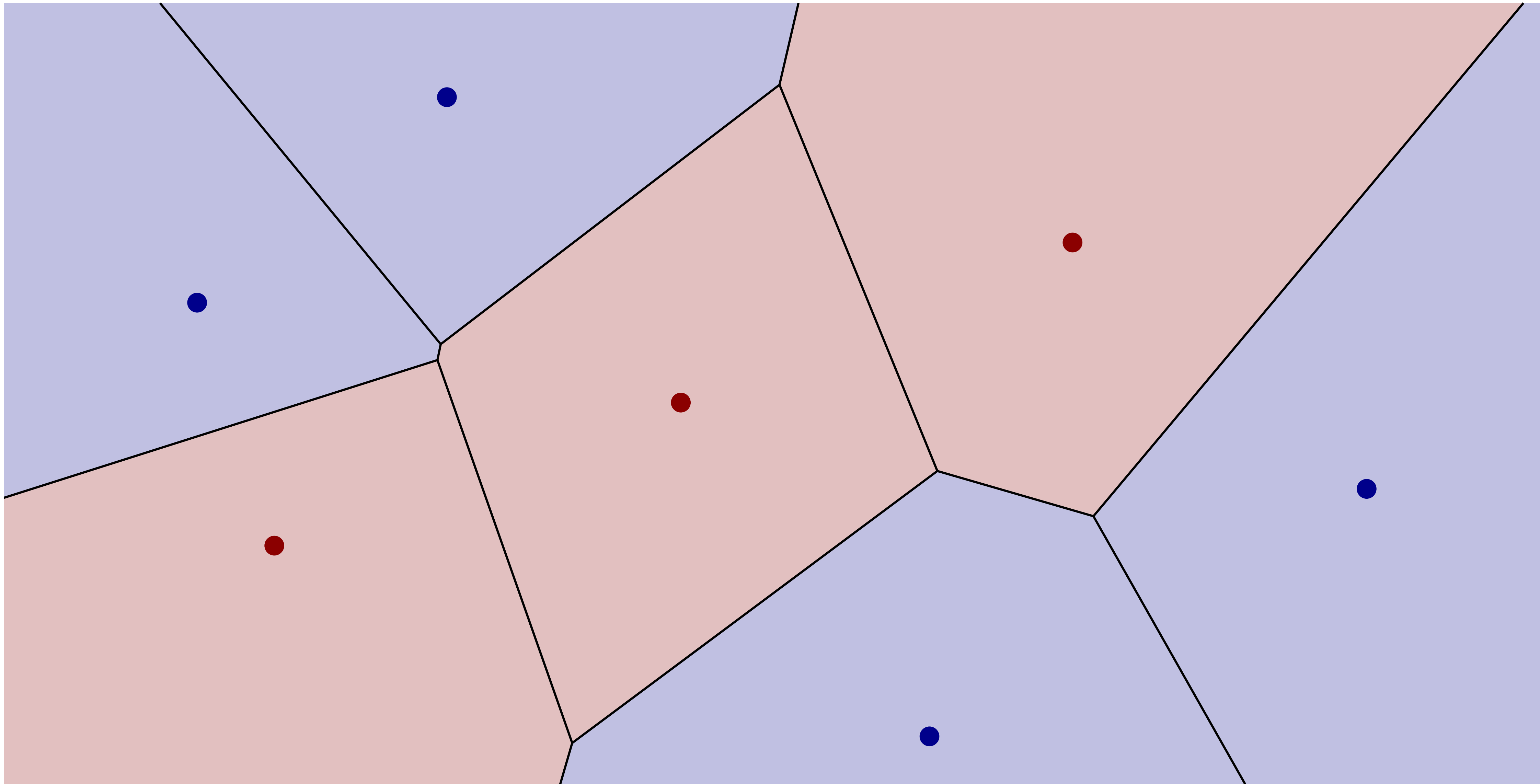
Nearest-Neighbor Classification



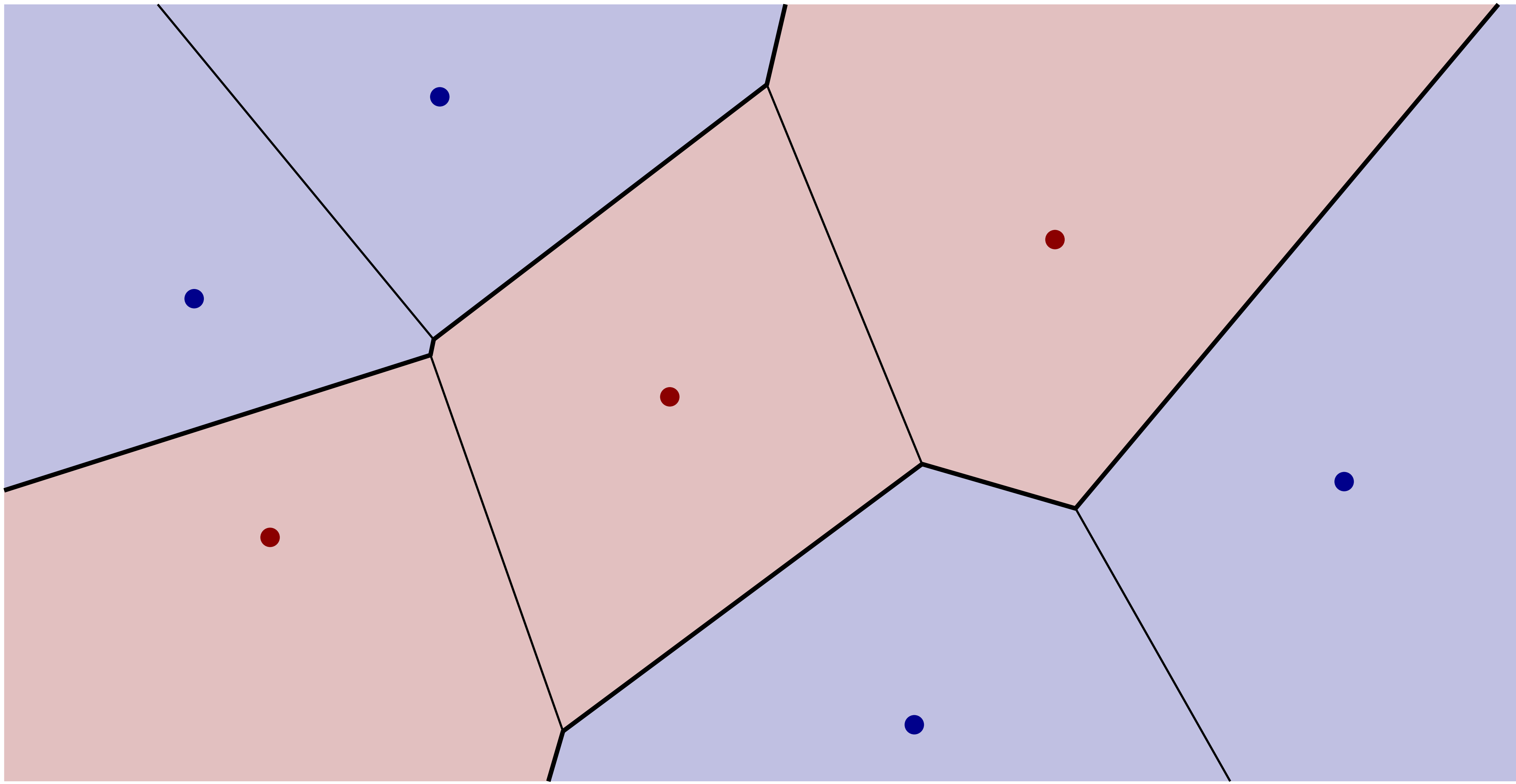
Nearest-Neighbor Classification



Nearest-Neighbor Classification



Nearest-Neighbor Classification



Efficient Nearest-Neighbor Classification

Nearest-Neighbor data structures and algorithms
scale badly in high dimensions!

Efficient Nearest-Neighbor Classification

Nearest-Neighbor data structures and algorithms scale badly in high dimensions!

What if we reduced the size of the training data set?

Efficient Nearest-Neighbor Classification

Nearest-Neighbor data structures and algorithms
scale badly in high dimensions!

What if we reduced the size of the training data set?
"Nearest Neighbor Condensation"

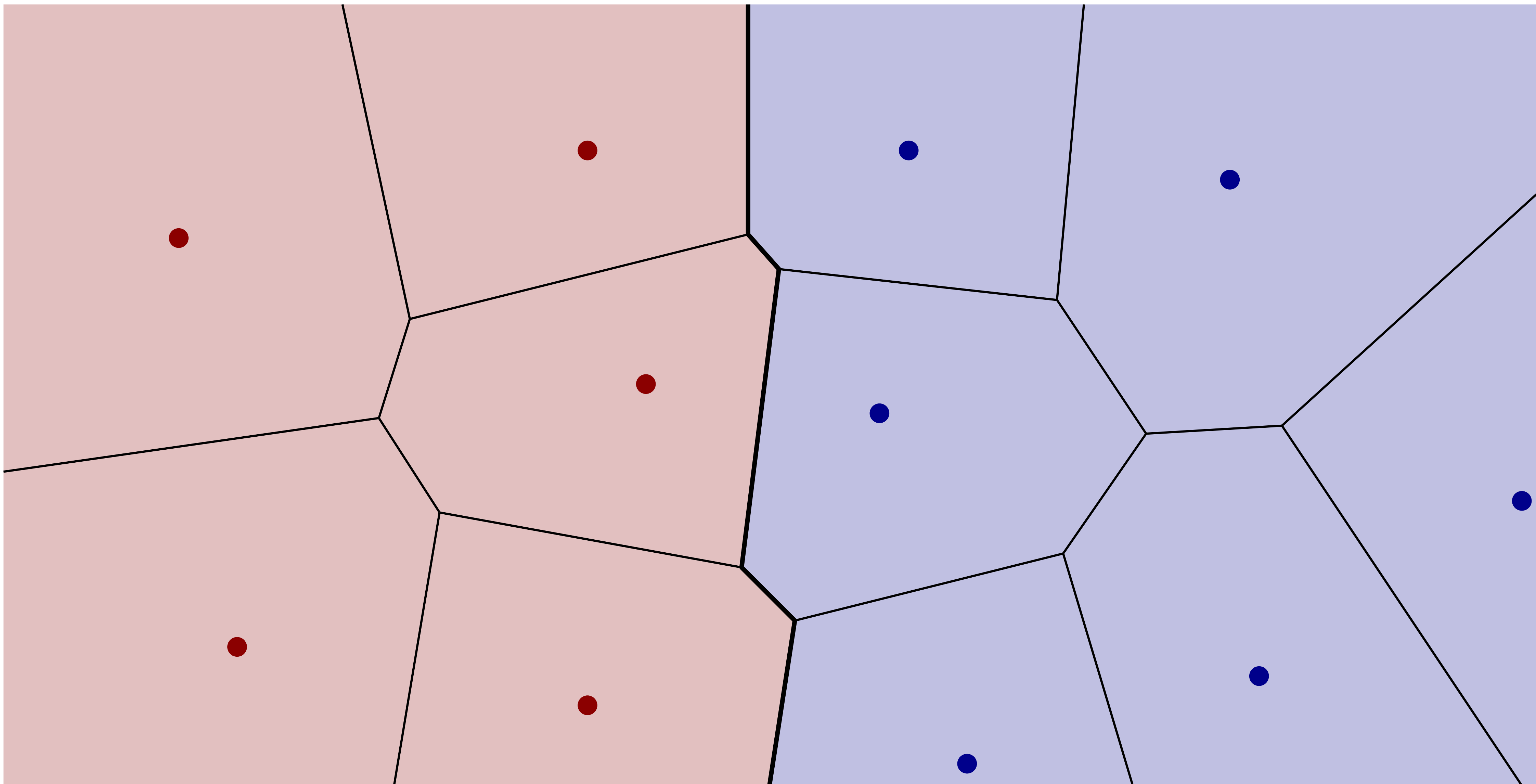
Efficient Nearest-Neighbor Classification

Nearest-Neighbor data structures and algorithms scale badly in high dimensions!

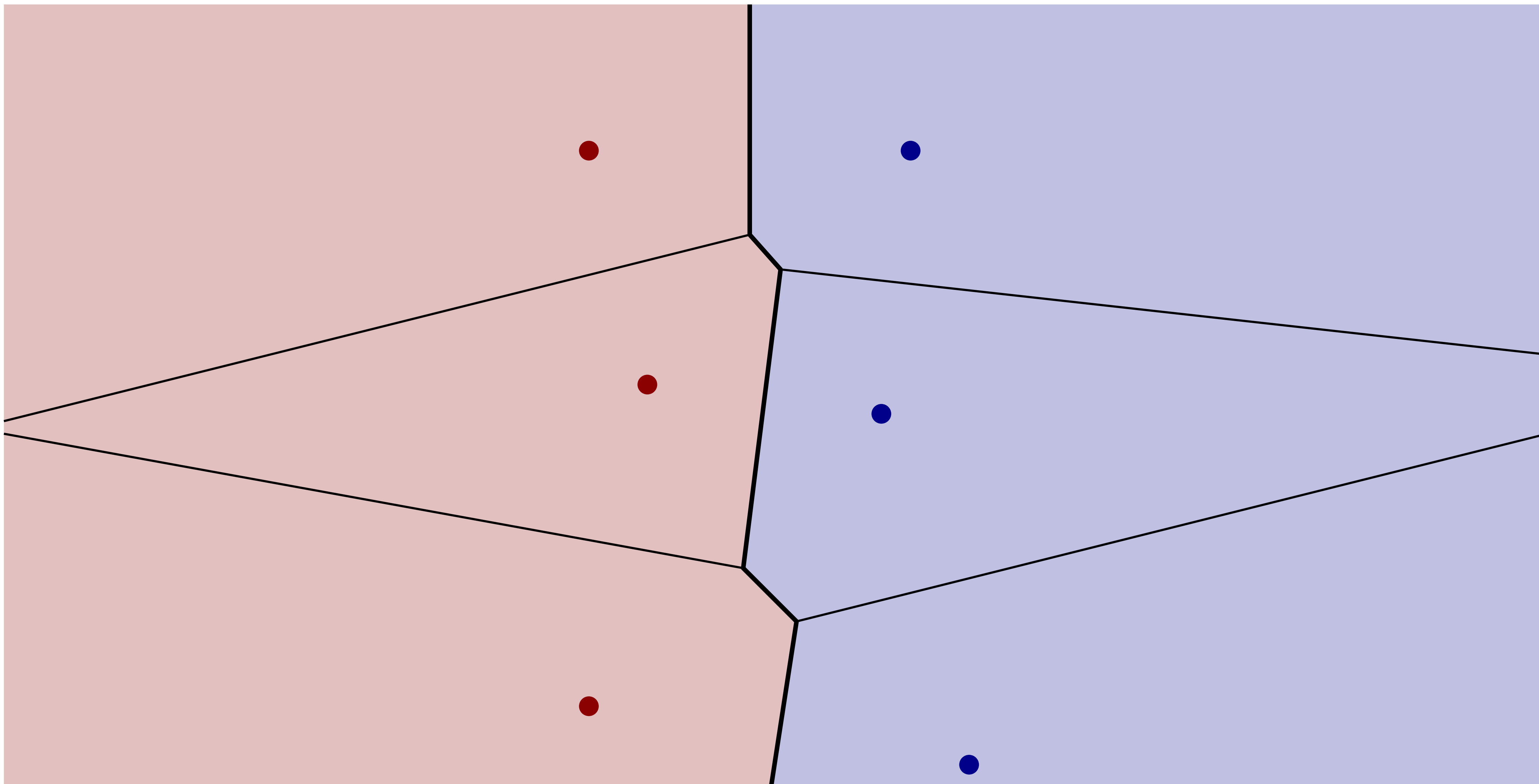
What if we reduced the size of the training data set?
"Nearest Neighbor Condensation"

We consider the *exact* case:
No point $p \in \mathbb{R}^d$ may change classification.

Reducing a Training Set



Reducing a Training Set



Reducing a Training Set

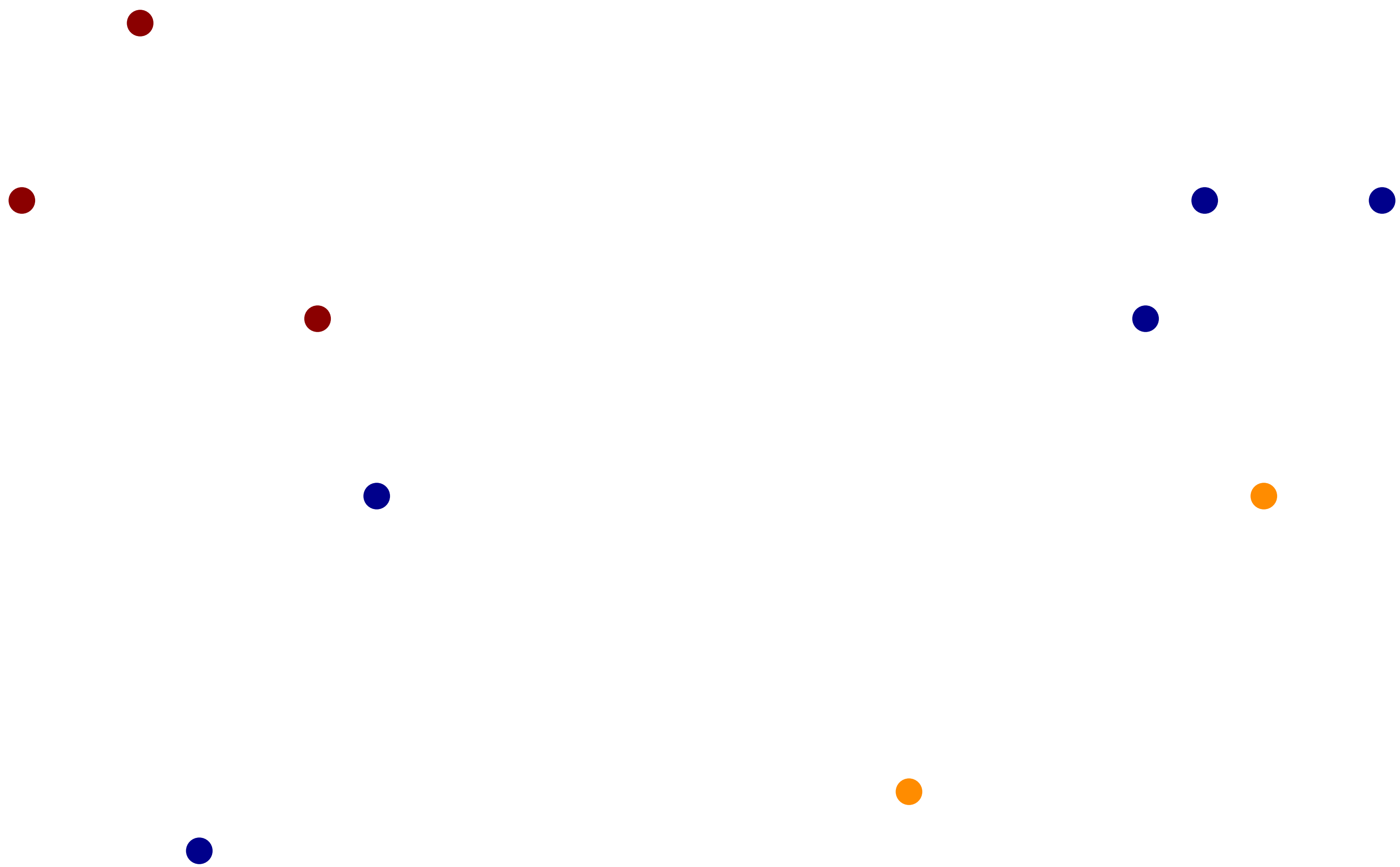
Definition: Given a labelled point set P, a point $p \in P$ is called *relevant* if the point set $P \setminus \{p\}$ induces a different nearest-neighbor classification.

Reducing a Training Set

Definition: Given a labelled point set P, a point $p \in P$ is called *relevant* if the point set $P \setminus \{p\}$ induces a different nearest-neighbor classification.

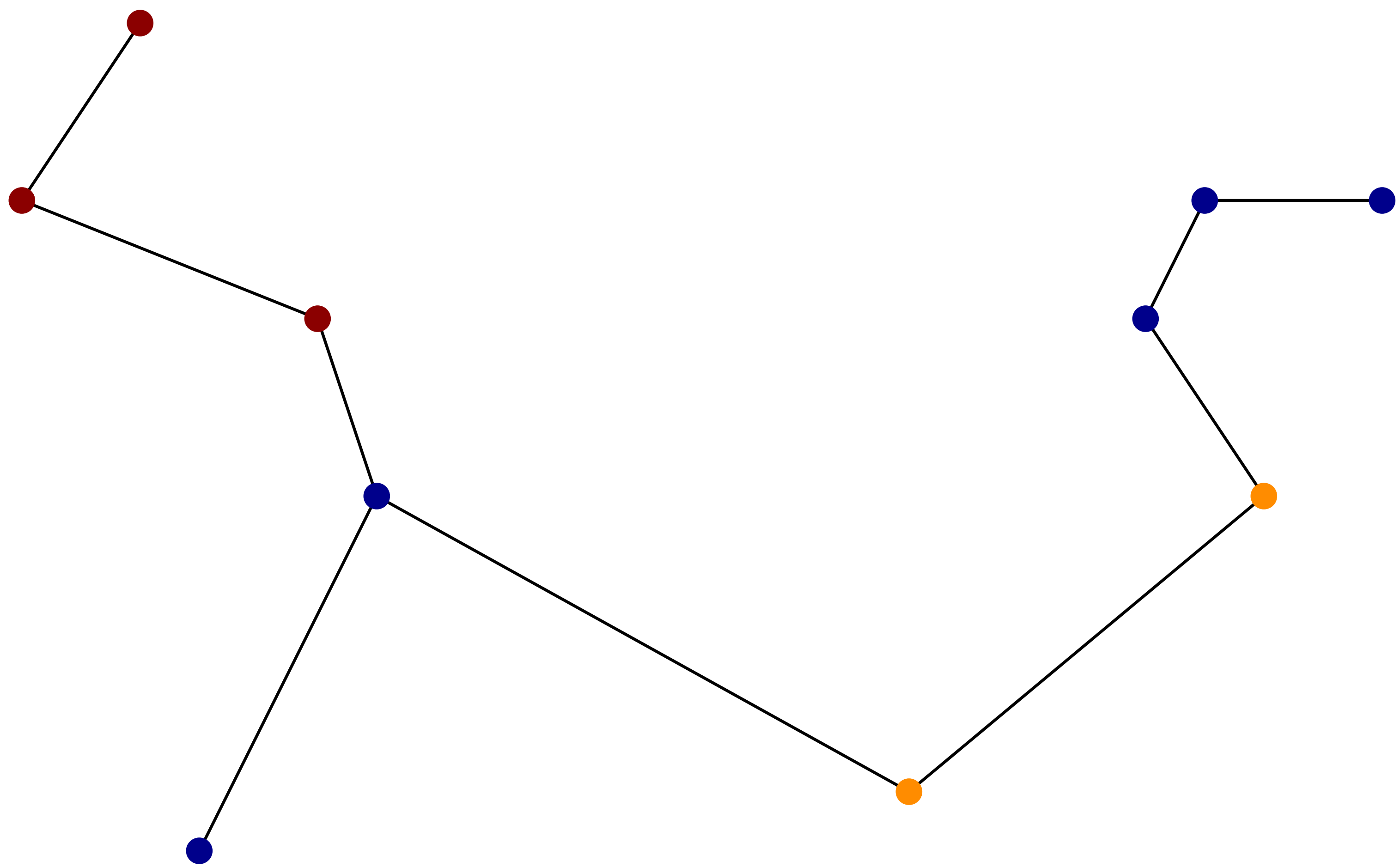
Fact: The set $rel(P) \subseteq P$ of relevant points induces the same nearest-neighbor classification as P.

Eppstein's Algorithm

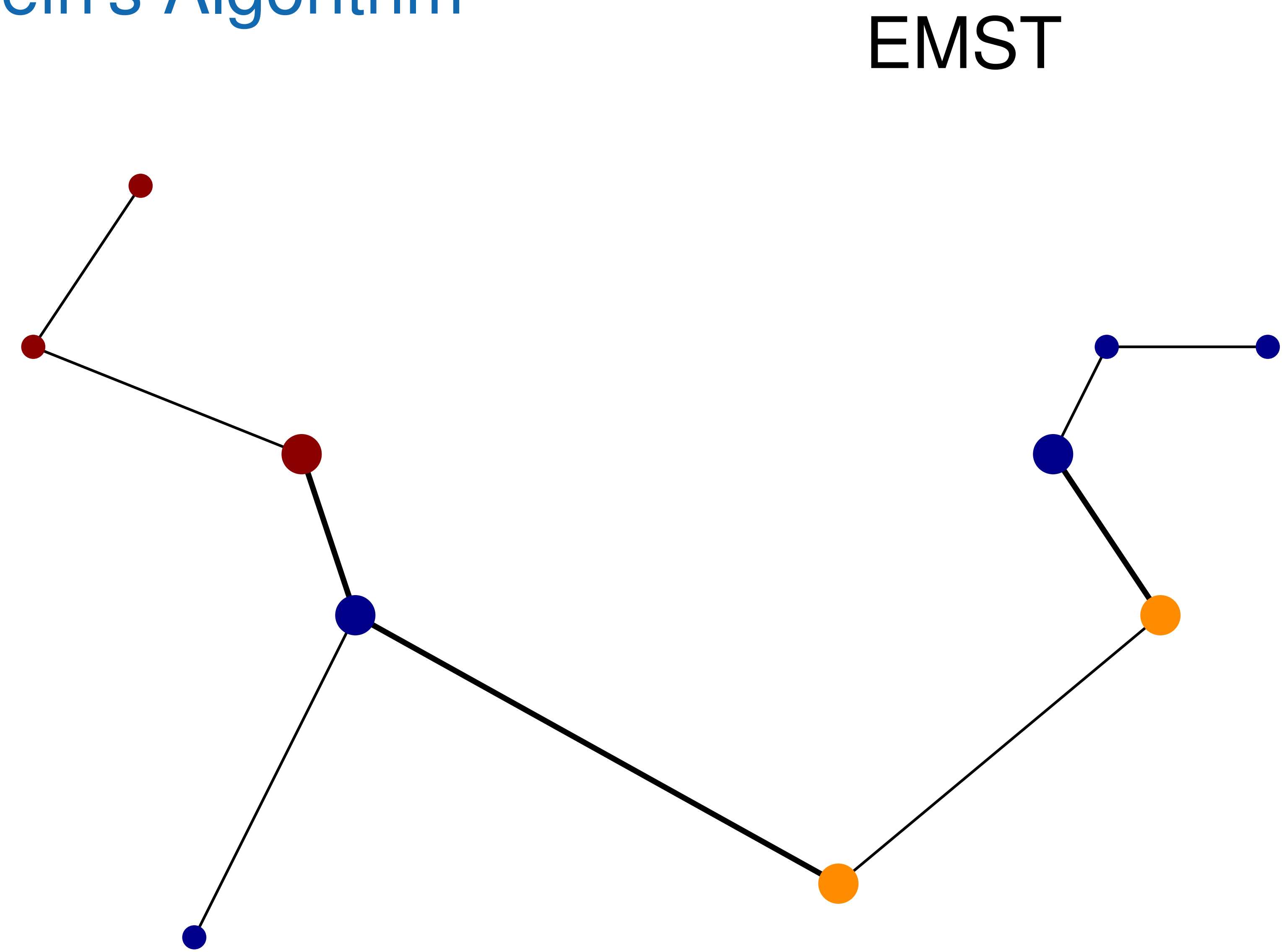


Eppstein's Algorithm

EMST

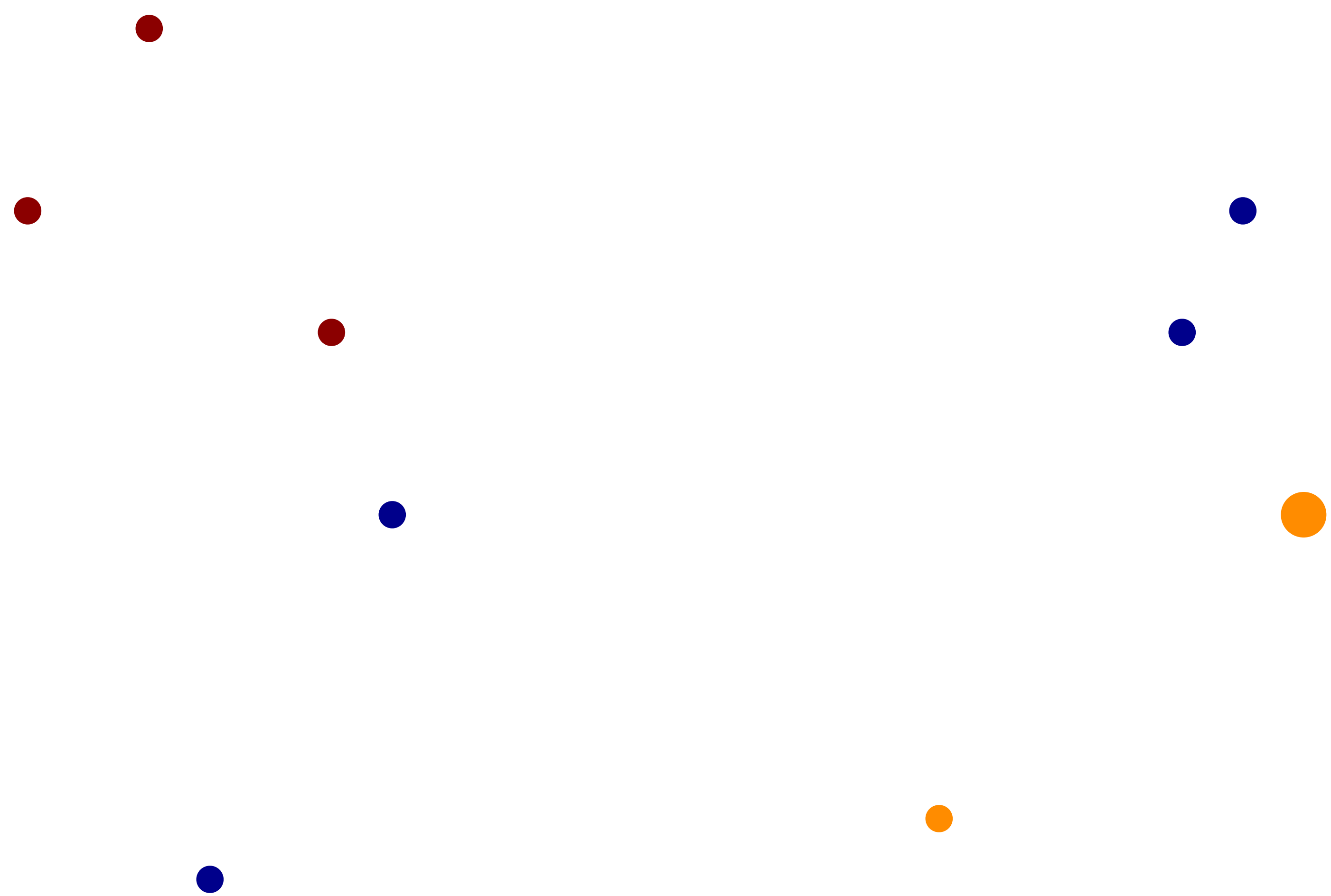


Eppstein's Algorithm

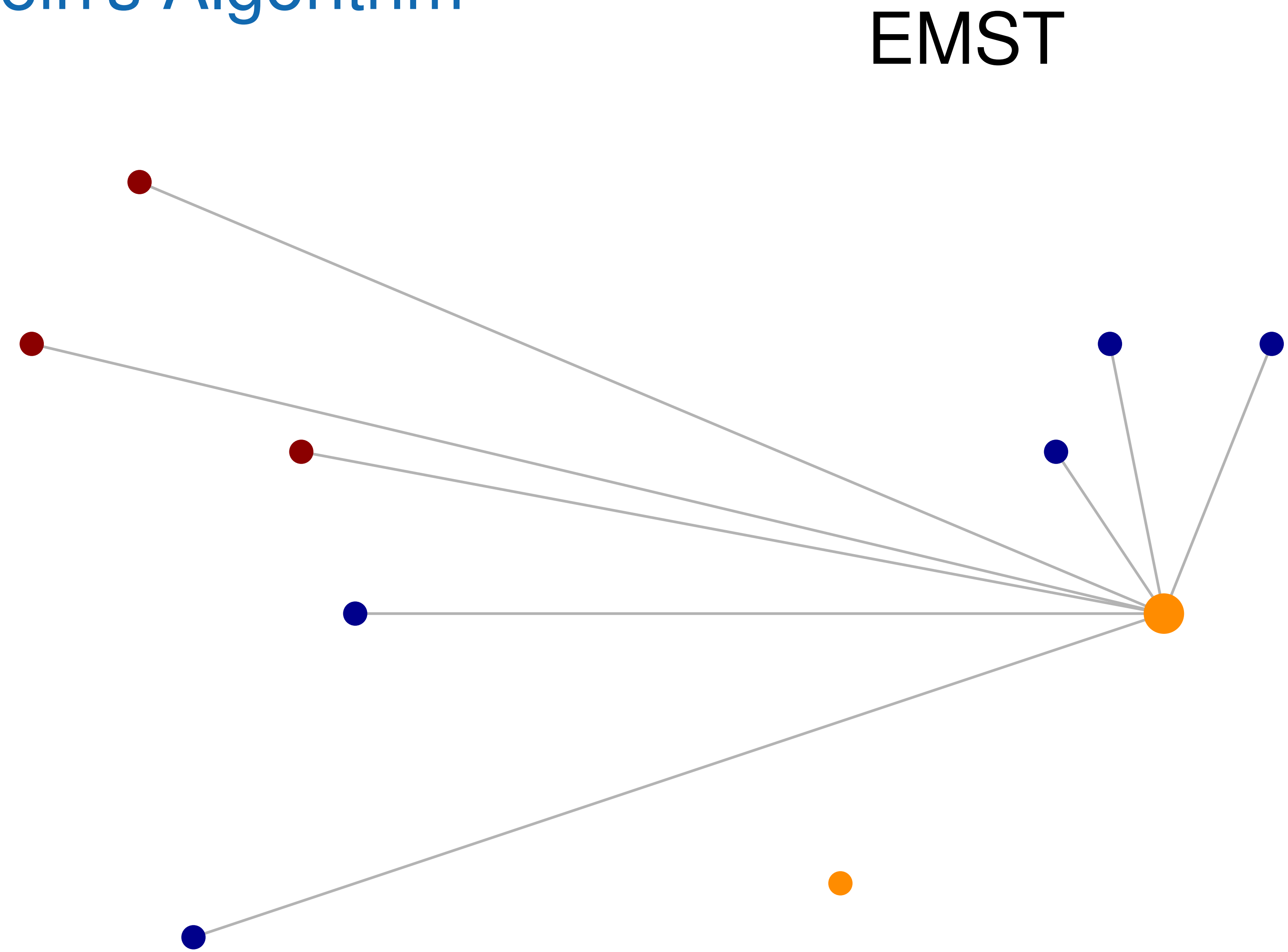


Eppstein's Algorithm

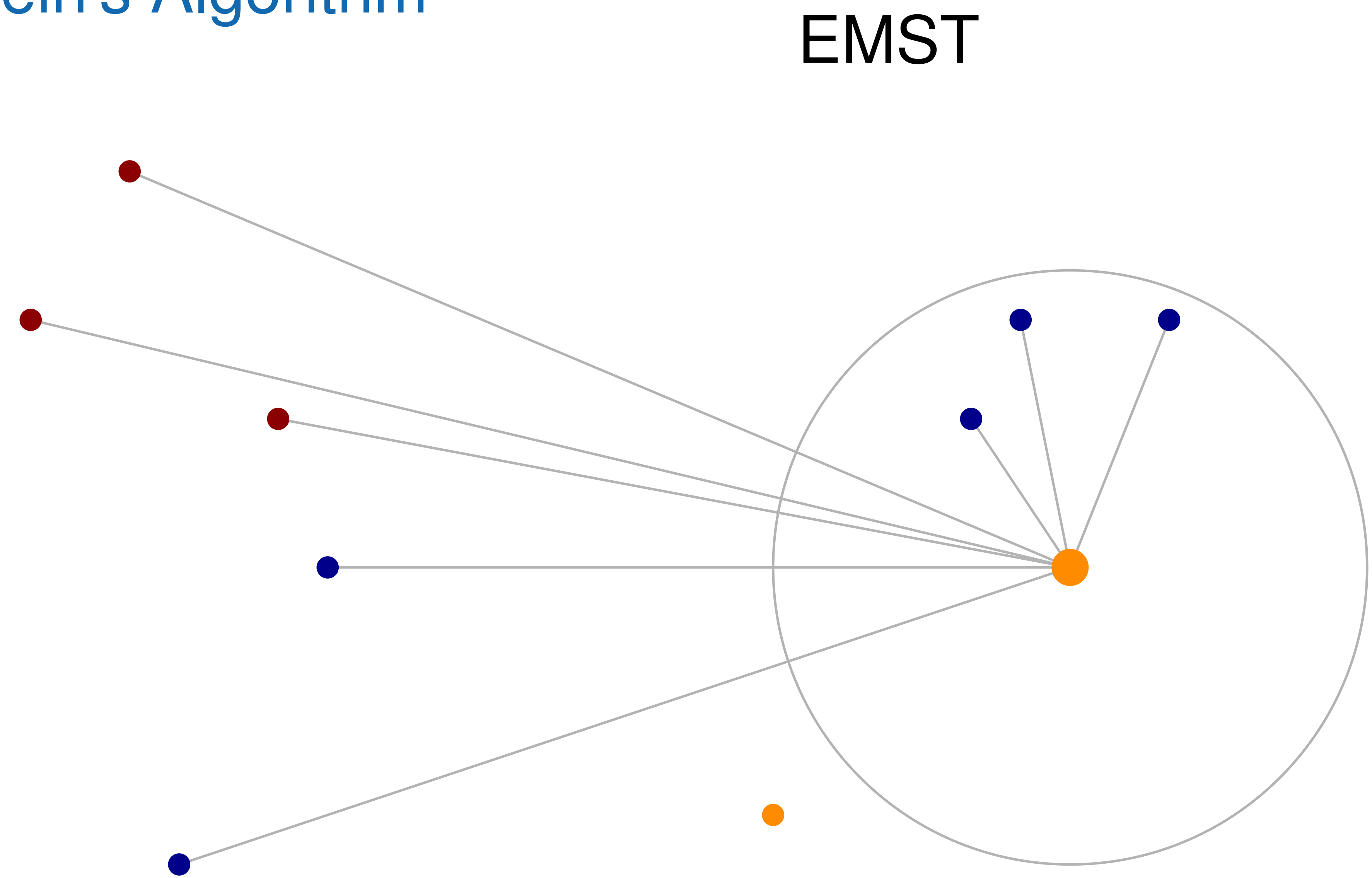
EMST



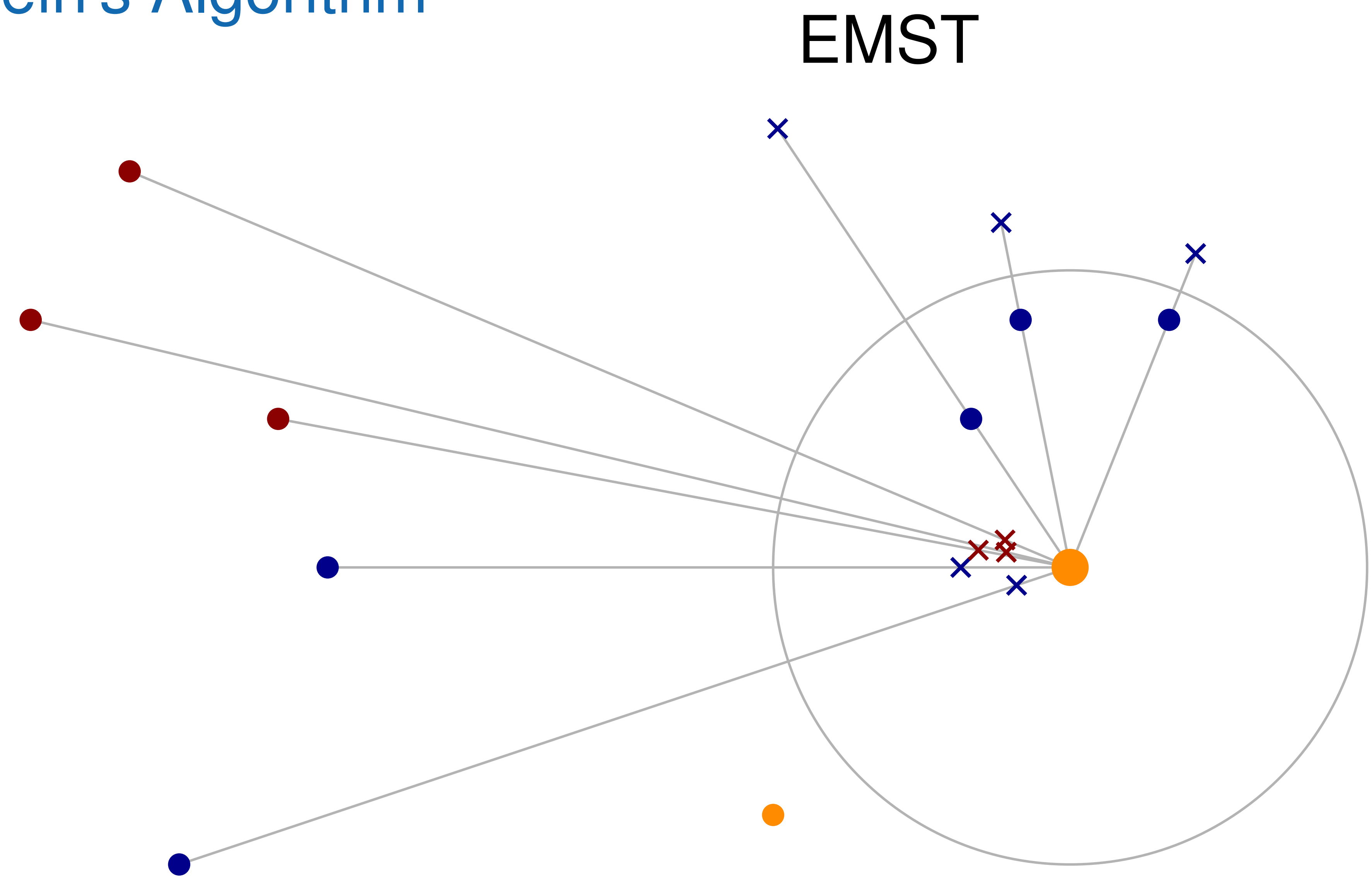
Eppstein's Algorithm



Eppstein's Algorithm



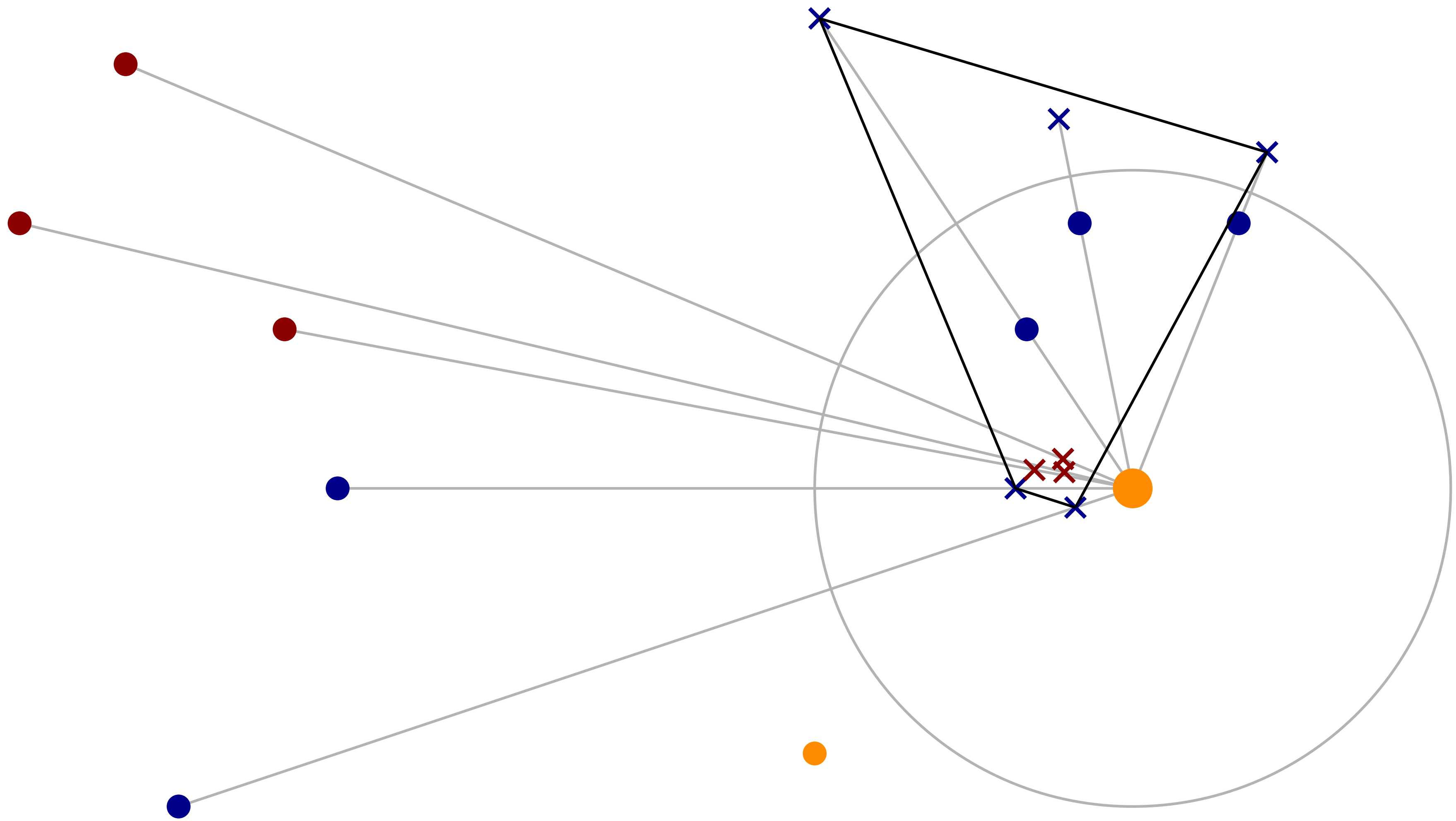
Eppstein's Algorithm



ETH zürich

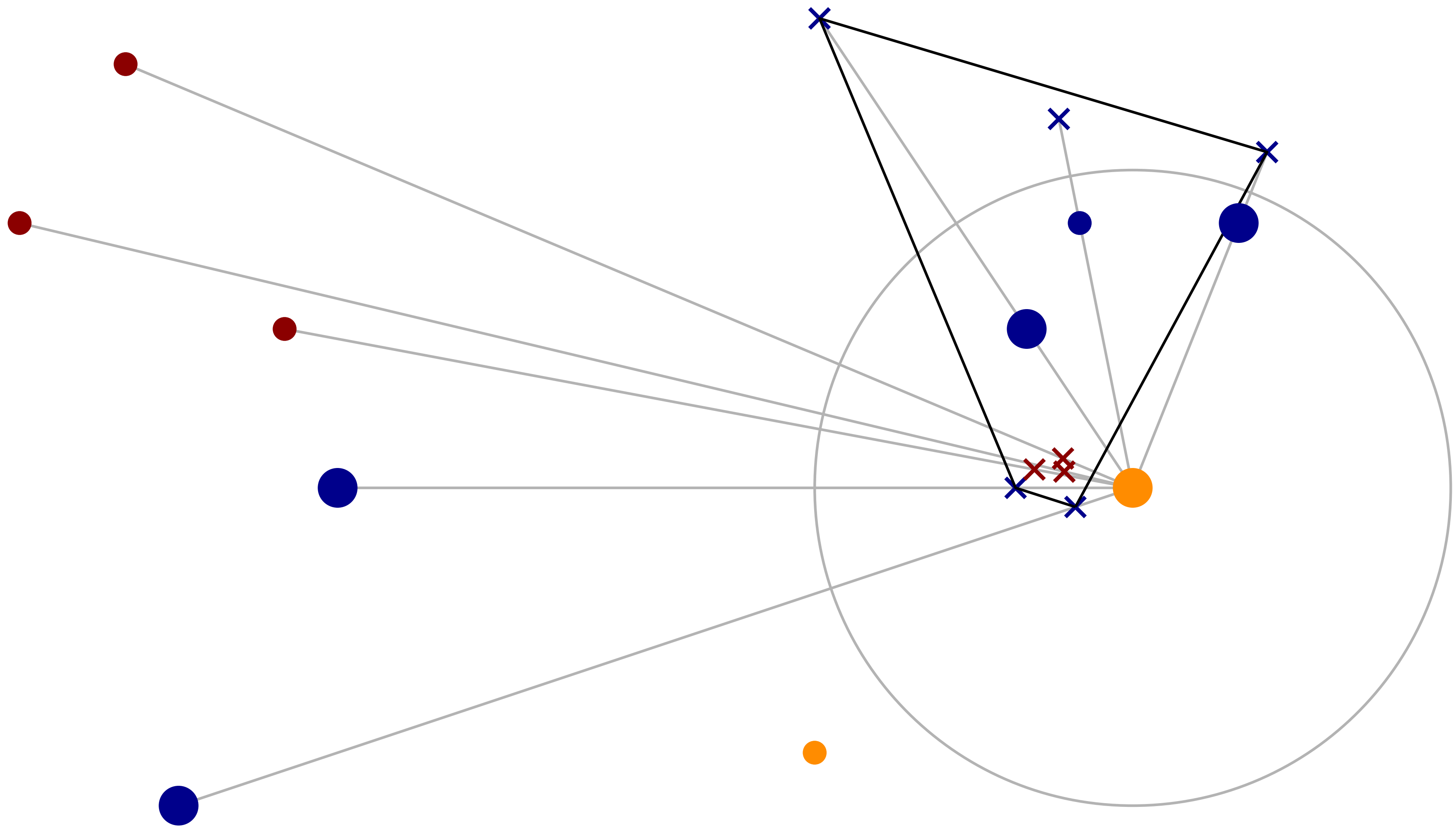
Eppstein's Algorithm

EMST + Extremal Points



Eppstein's Algorithm

EMST + Extremal Points



Eppstein's Open Question

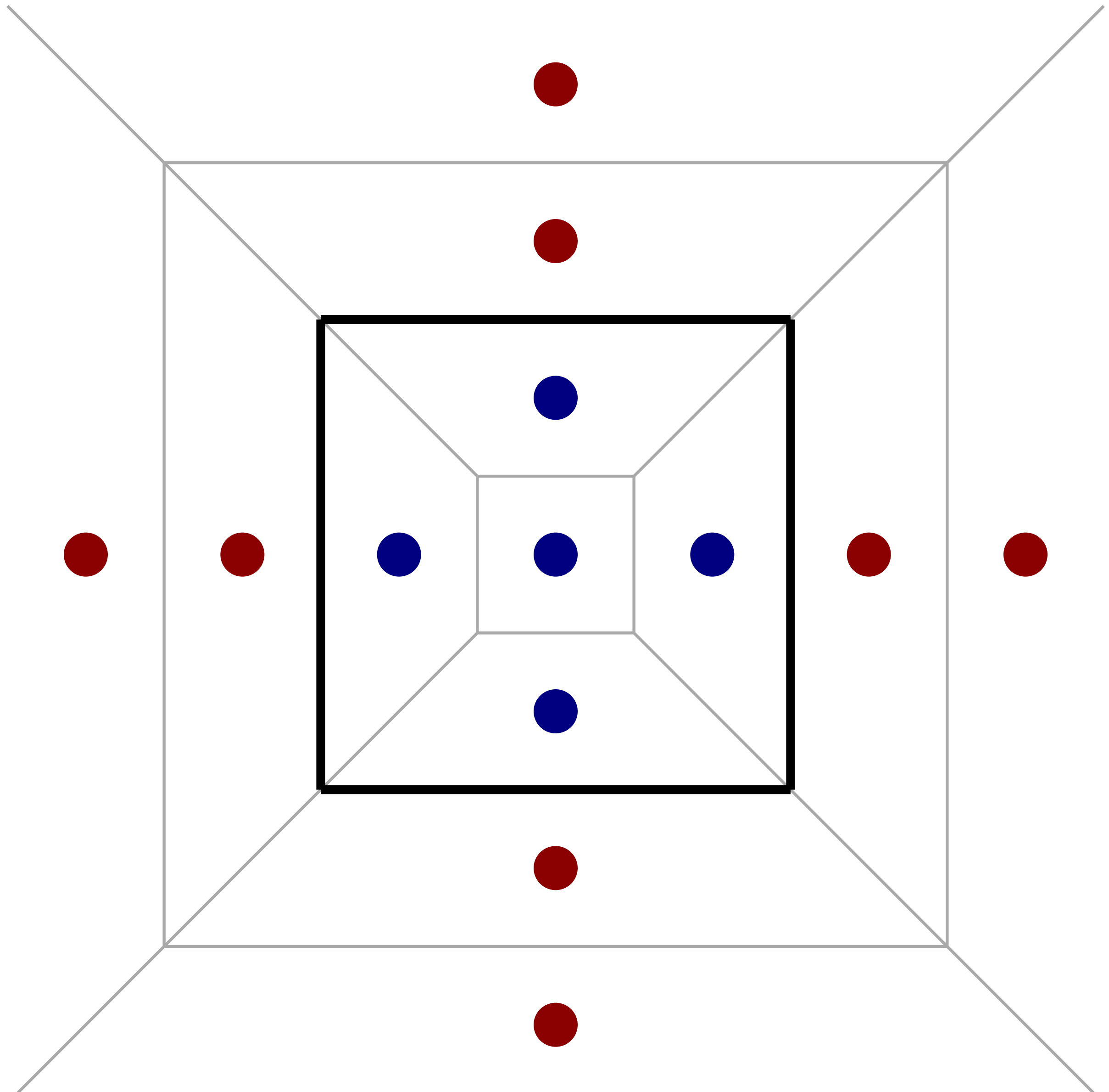
Question: Can we reduce the training set *further* than to the relevant points, without changing the resulting classification? What is the *complexity* of finding the smallest subset $Q \subseteq P$ with the same classification?

Eppstein's Open Question

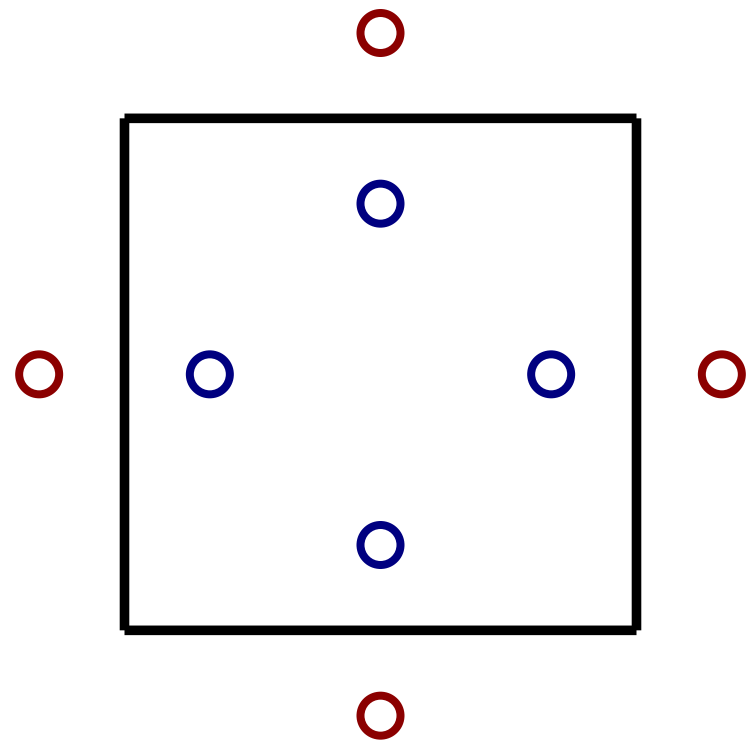
Question: Can we reduce the training set *further* than to the relevant points, without changing the resulting classification? What is the *complexity* of finding the smallest subset $Q \subseteq P$ with the same classification?

A *minimum-cardinality reduced training set*

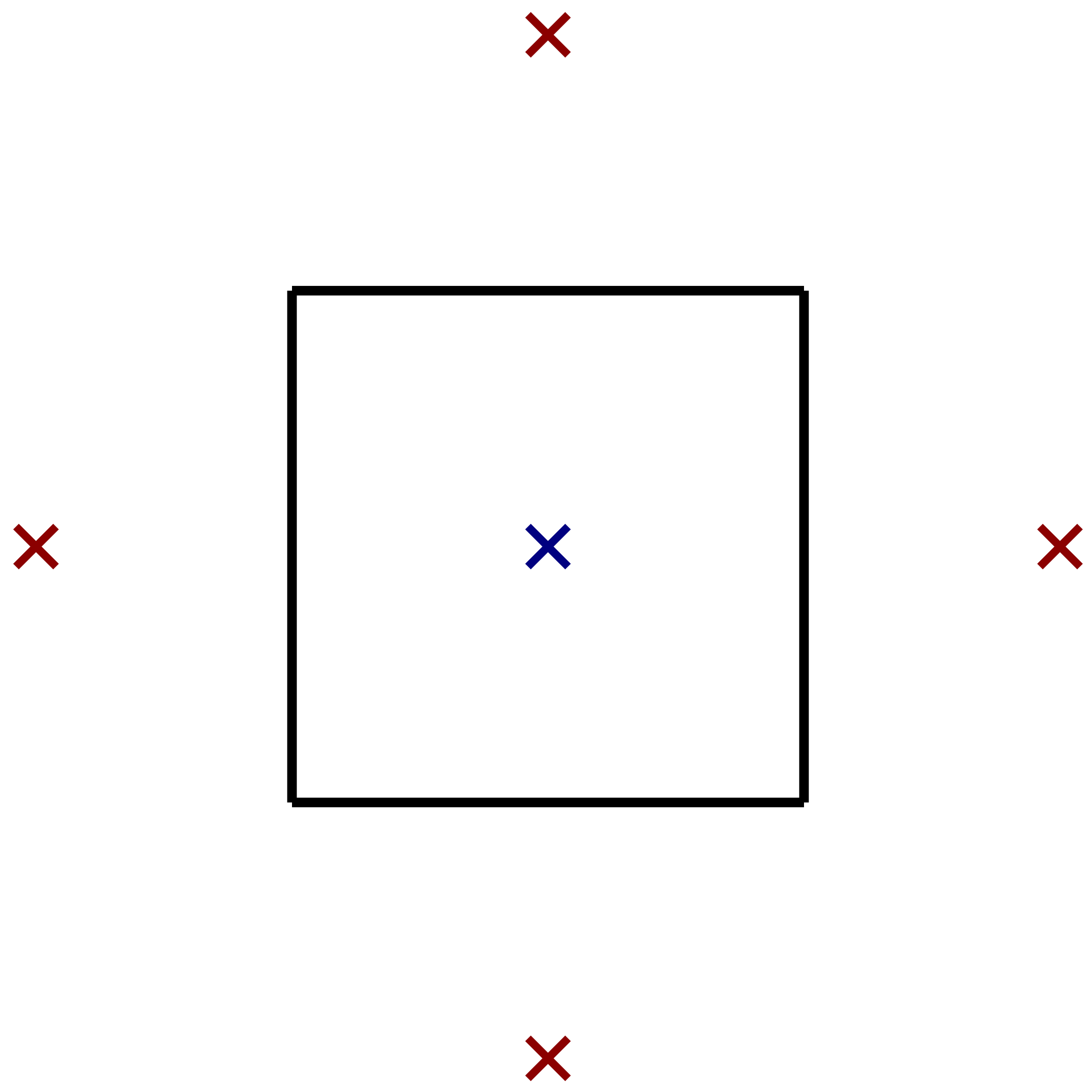
Relevant Points are not Optimal



Relevant Points are not Optimal



Relevant Points are not Optimal



Our Results

Theorem: If P is in general position, $rel(P)$ is a minimum-cardinality reduced training set.

Our Results

Theorem: If P is in general position, $rel(P)$ is a minimum-cardinality reduced training set.

Theorem: Computing a minimum-cardinality reduced training set is in P for points in $\mathbb{R}^1$.

Our Results

Theorem: If P is in general position, $rel(P)$ is a minimum-cardinality reduced training set.

Theorem: Computing a minimum-cardinality reduced training set is in P for points in $\mathbb{R}^1$.

Theorem: Computing a minimum-cardinality reduced training set is NP-complete for points in $\mathbb{R}^d$ for $d \geq 2$, even if there are only two colors.

Points in General Position

Points in General Position

no 3 collinear, no 4 cocircular

Points in General Position

no 3 collinear, no 4 cocircular

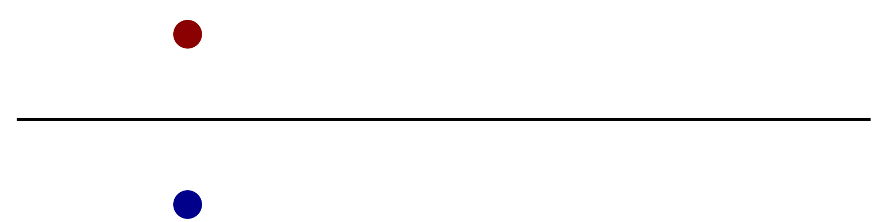
Observation: Every Voronoi wall between differently classified cells must lie in the bisecting hyperplane of some $a, b \in Q \subseteq P$.

Points in General Position

no 3 collinear, no 4 cocircular

Observation: Every Voronoi wall between differently classified cells must lie in the bisecting hyperplane of some $a, b \in Q \subseteq P$.

Claim: Every hyperplane is the bisecting hyperplane of at most one pair of points $a, b \in P$.

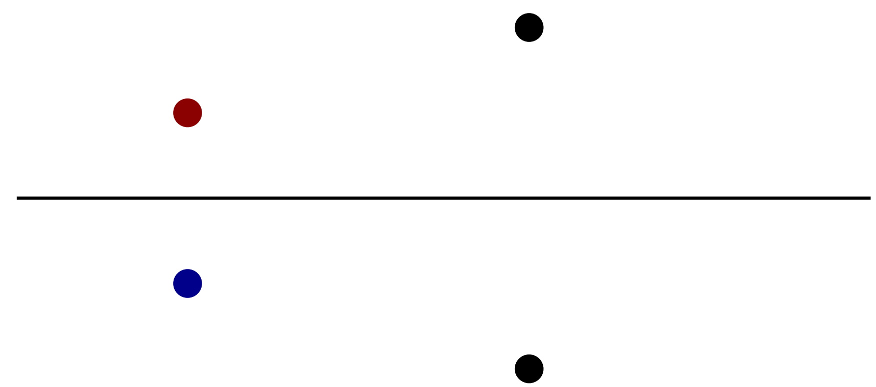


Points in General Position

no 3 collinear, no 4 cocircular

Observation: Every Voronoi wall between differently classified cells must lie in the bisecting hyperplane of some $a, b \in Q \subseteq P$.

Claim: Every hyperplane is the bisecting hyperplane of at most one pair of points $a, b \in P$.

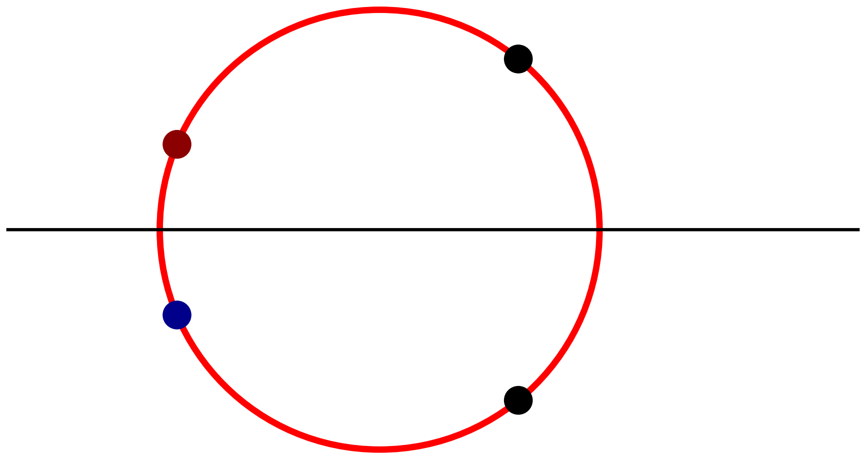


Points in General Position

no 3 collinear, no 4 cocircular

Observation: Every Voronoi wall between differently classified cells must lie in the bisecting hyperplane of some $a, b \in Q \subseteq P$.

Claim: Every hyperplane is the bisecting hyperplane of at most one pair of points $a, b \in P$.

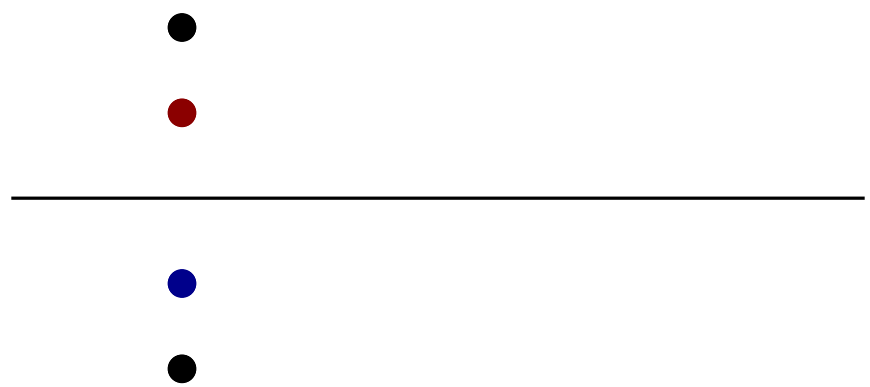


Points in General Position

no 3 collinear, no 4 cocircular

Observation: Every Voronoi wall between differently classified cells must lie in the bisecting hyperplane of some $a, b \in Q \subseteq P$.

Claim: Every hyperplane is the bisecting hyperplane of at most one pair of points $a, b \in P$.

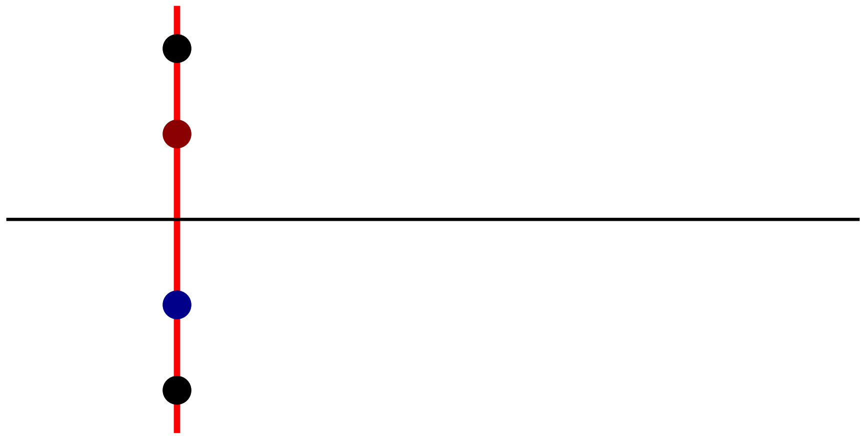


Points in General Position

no 3 collinear, no 4 cocircular

Observation: Every Voronoi wall between differently classified cells must lie in the bisecting hyperplane of some $a, b \in Q \subseteq P$.

Claim: Every hyperplane is the bisecting hyperplane of at most one pair of points $a, b \in P$.



Points in General Position

no 3 collinear, no 4 cocircular

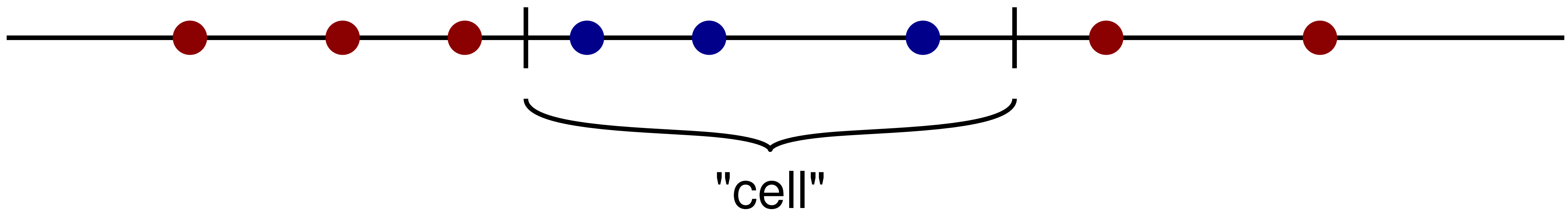
Observation: Every Voronoi wall between differently classified cells must lie in the bisecting hyperplane of some $a, b \in Q \subseteq P$.

Claim: Every hyperplane is the bisecting hyperplane of at most one pair of points $a, b \in P$.

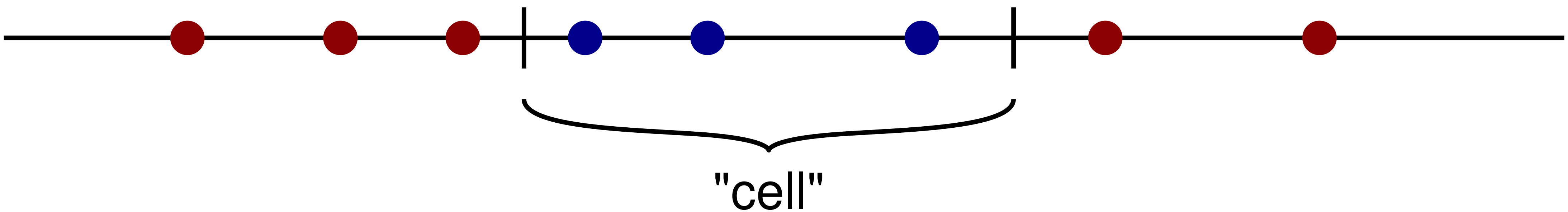
Fact: A point p is relevant if and only if it shares a $(d-1)$-dimensional Voronoi wall with a point of different classification.

Finding a minimum-cardinality reduced training set in $\mathbb{R}^1$

Finding a minimum-cardinality reduced training set in $\mathbb{R}^1$



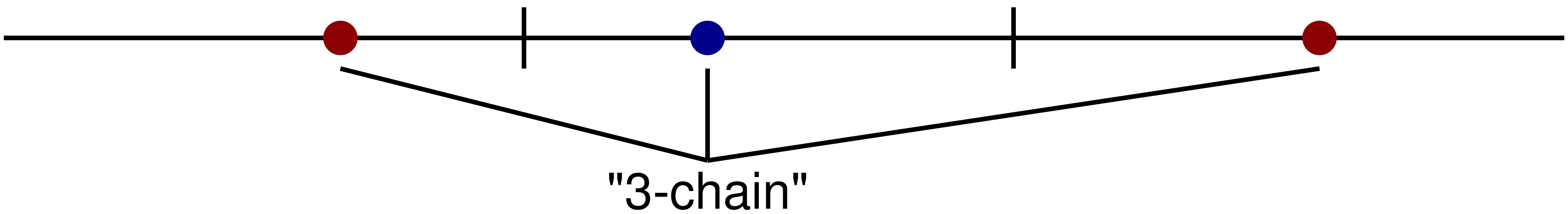
Finding a minimum-cardinality reduced training set in $\mathbb{R}^1$



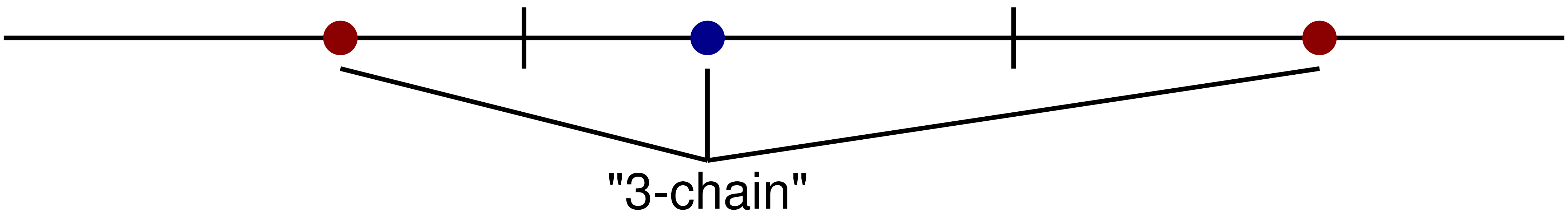
Observation: A minimum-cardinality reduced training set has either 1 or 2 points per cell.

Finding a minimum-cardinality reduced training set in $\mathbb{R}^1$

Finding a minimum-cardinality reduced training set in $\mathbb{R}^1$



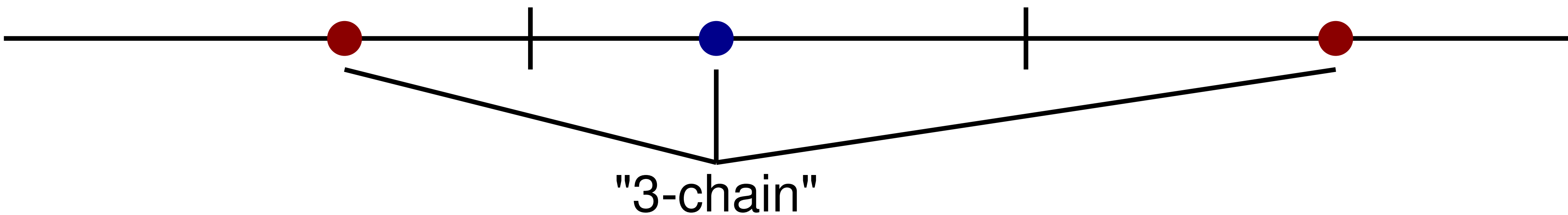
Finding a minimum-cardinality reduced training set in $\mathbb{R}^1$



Reduction to *maximum weight independent set on interval graphs*:

Theorem [Hsiao, Tang, Chang, '92]:
MaxW-IS on interval graphs is in P.

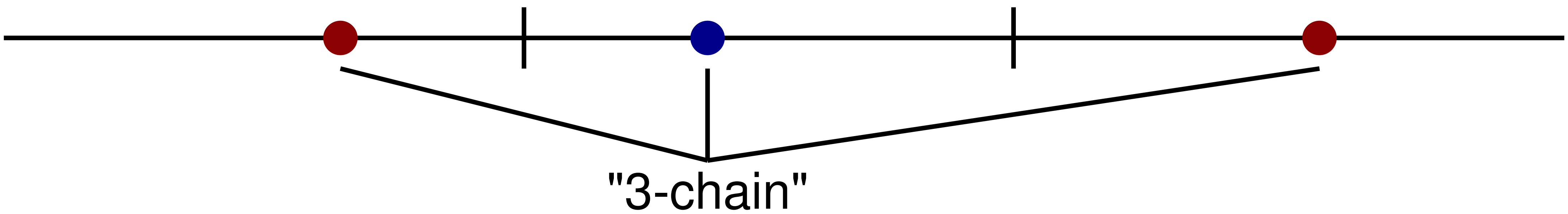
Finding a minimum-cardinality reduced training set in $\mathbb{R}^1$



Reduction to *maximum weight independent set on interval graphs*:

- Find all chains (including the non-maximal ones)

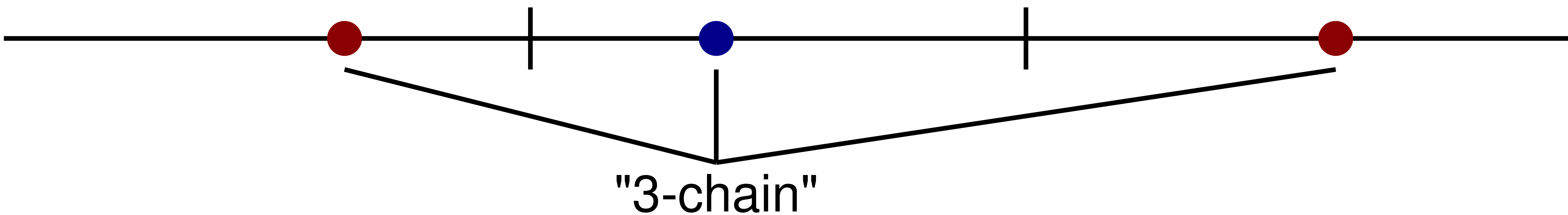
Finding a minimum-cardinality reduced training set in $\mathbb{R}^1$



Reduction to *maximum weight independent set on interval graphs*:

- Find all chains (including the non-maximal ones)
- Associate each chain with its convex hull interval

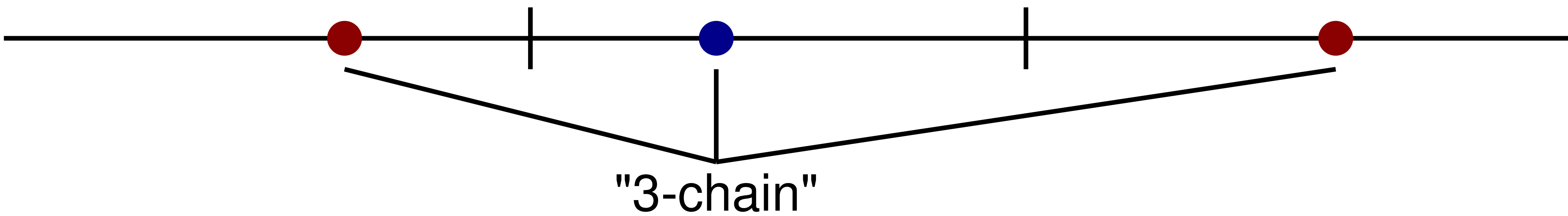
Finding a minimum-cardinality reduced training set in $\mathbb{R}^1$



Reduction to *maximum weight independent set on interval graphs*:

- Find all chains (including the non-maximal ones)
- Associate each chain with its convex hull interval
- Give every k-chain weight $max(k-2, \epsilon)$

Finding a minimum-cardinality reduced training set in $\mathbb{R}^1$



Reduction to *maximum weight independent set on interval graphs*:

- Find all chains (including the non-maximal ones)
- Associate each chain with its convex hull interval
- Give every k-chain weight $max(k-2, \epsilon)$

Observation: MaxW-IS $\Leftrightarrow$ minimum-cardinality reduced training set

NP-Hardness for $d \geq 2$

Definition: An instance of *V-cycle Max2SAT* is given by

- a 2-CNF formula $\phi = C_1 \wedge \ldots \wedge C_b$ over the variables $x_1, \ldots, x_a$, such that $G_{cyc}(\phi)$, the bipartite clause-variable graph of ϕ with an additional Hamiltonian cycle $(x_1, \ldots, x_a, x_1)$, is planar

NP-Hardness for $d \geq 2$

Definition: An instance of *V-cycle Max2SAT* is given by

- a 2-CNF formula $\phi = C_1 \wedge \ldots \wedge C_b$ over the variables $x_1, \ldots, x_a$, such that $G_{cyc}(\phi)$, the bipartite clause-variable graph of ϕ with an additional Hamiltonian cycle $(x_1, \ldots, x_a, x_1)$, is planar

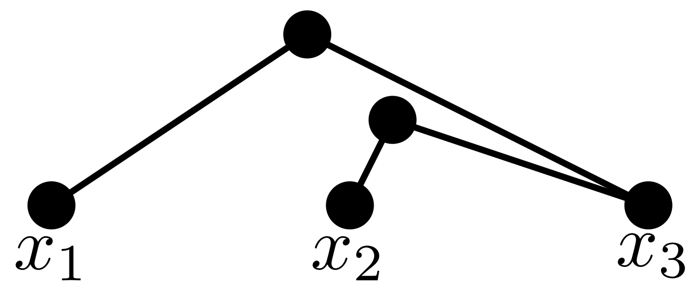
$$\phi = (x_1 \vee \neg x_3) \wedge (\neg x_2 \vee x_3)$$

NP-Hardness for $d \geq 2$

Definition: An instance of *V-cycle Max2SAT* is given by

- a 2-CNF formula $\phi = C_1 \wedge \ldots \wedge C_b$ over the variables $x_1, \ldots, x_a$, such that $G_{cyc}(\phi)$, the bipartite clause-variable graph of ϕ with an additional Hamiltonian cycle $(x_1, \ldots, x_a, x_1)$, is planar

$$\phi = (x_1 \vee \neg x_3) \wedge (\neg x_2 \vee x_3)$$

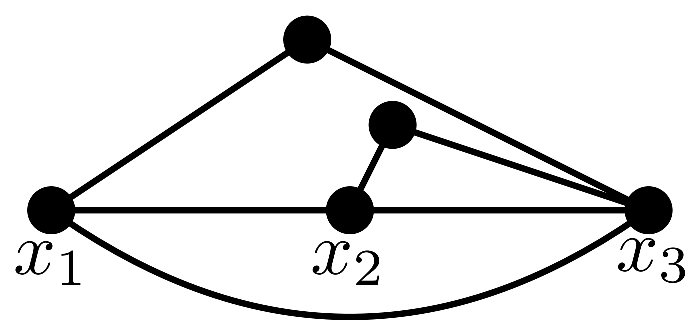


NP-Hardness for $d \geq 2$

Definition: An instance of *V-cycle Max2SAT* is given by

- a 2-CNF formula $\phi = C_1 \wedge \ldots \wedge C_b$ over the variables $x_1, \ldots, x_a$, such that $G_{cyc}(\phi)$, the bipartite clause-variable graph of ϕ with an additional Hamiltonian cycle $(x_1, \ldots, x_a, x_1)$, is planar

$$\phi = (x_1 \vee \neg x_3) \wedge (\neg x_2 \vee x_3)$$



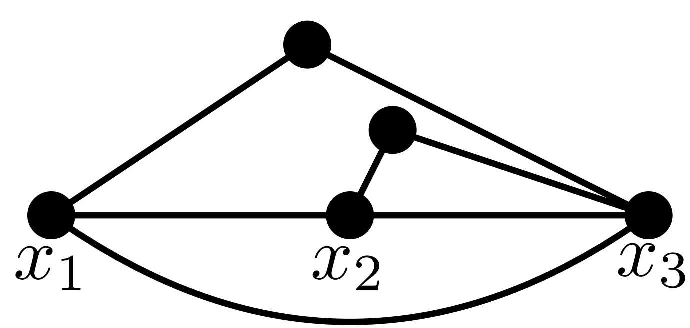
NP-Hardness for $d \geq 2$

Definition: An instance of *V-cycle Max2SAT* is given by

- a 2-CNF formula $\phi = C_1 \wedge \ldots \wedge C_b$ over the variables $x_1, \ldots, x_a$, such that $G_{cyc}(\phi)$, the bipartite clause-variable graph of ϕ with an additional Hamiltonian cycle $(x_1, \ldots, x_a, x_1)$, is planar
- an integer k

The task is to decide whether there exists a variable assignment fulfilling at least k clauses of ϕ.

$$\phi = (x_1 \vee \neg x_3) \wedge (\neg x_2 \vee x_3)$$



NP-Hardness for $d \geq 2$

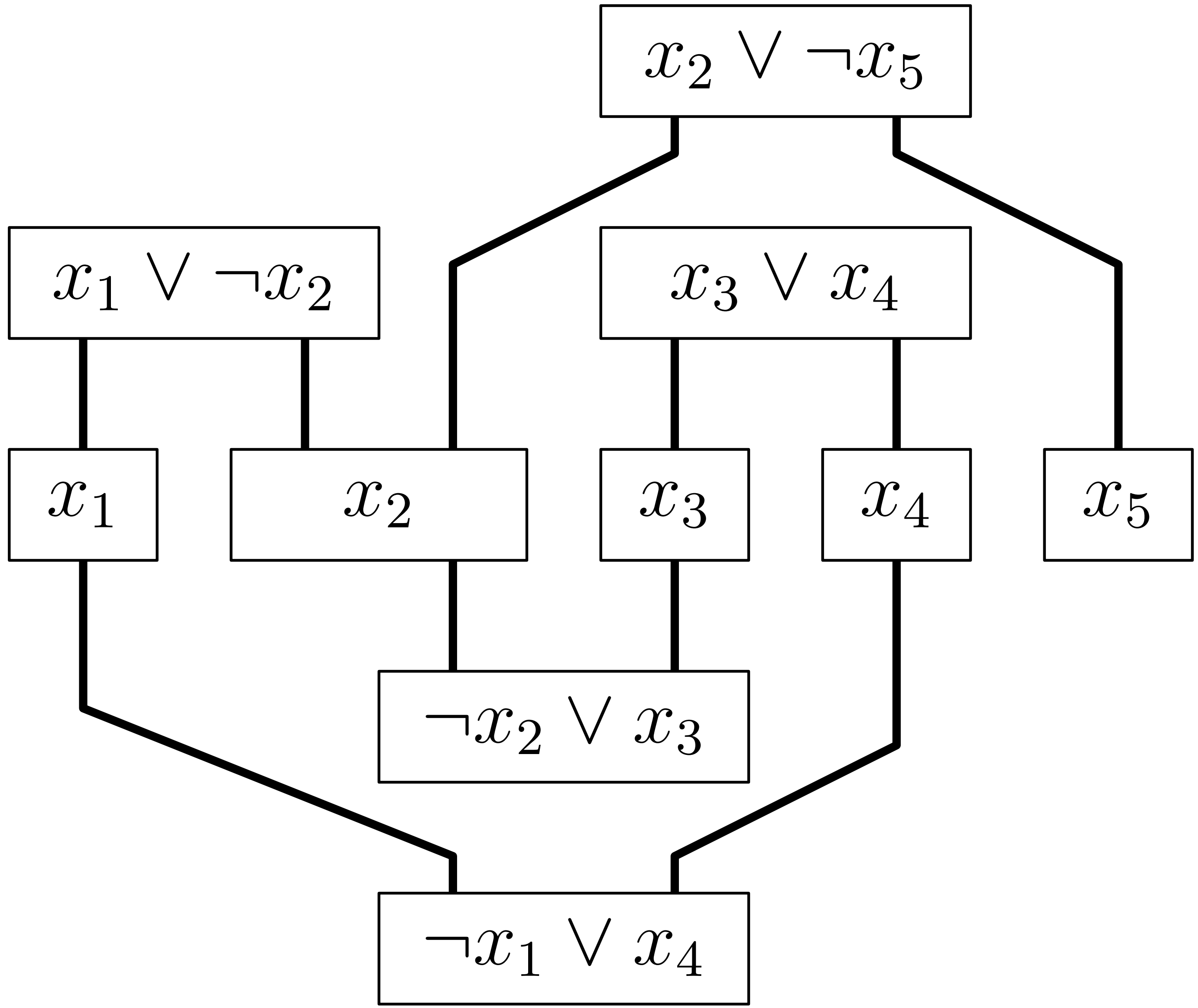
Definition: An instance of *V-cycle Max2SAT* is given by

- a 2-CNF formula $\phi = C_1 \wedge \ldots \wedge C_b$ over the variables $x_1, \ldots, x_a$, such that $G_{cyc}(\phi)$, the bipartite clause-variable graph of ϕ with an additional Hamiltonian cycle $(x_1, \ldots, x_a, x_1)$, is planar
- an integer k

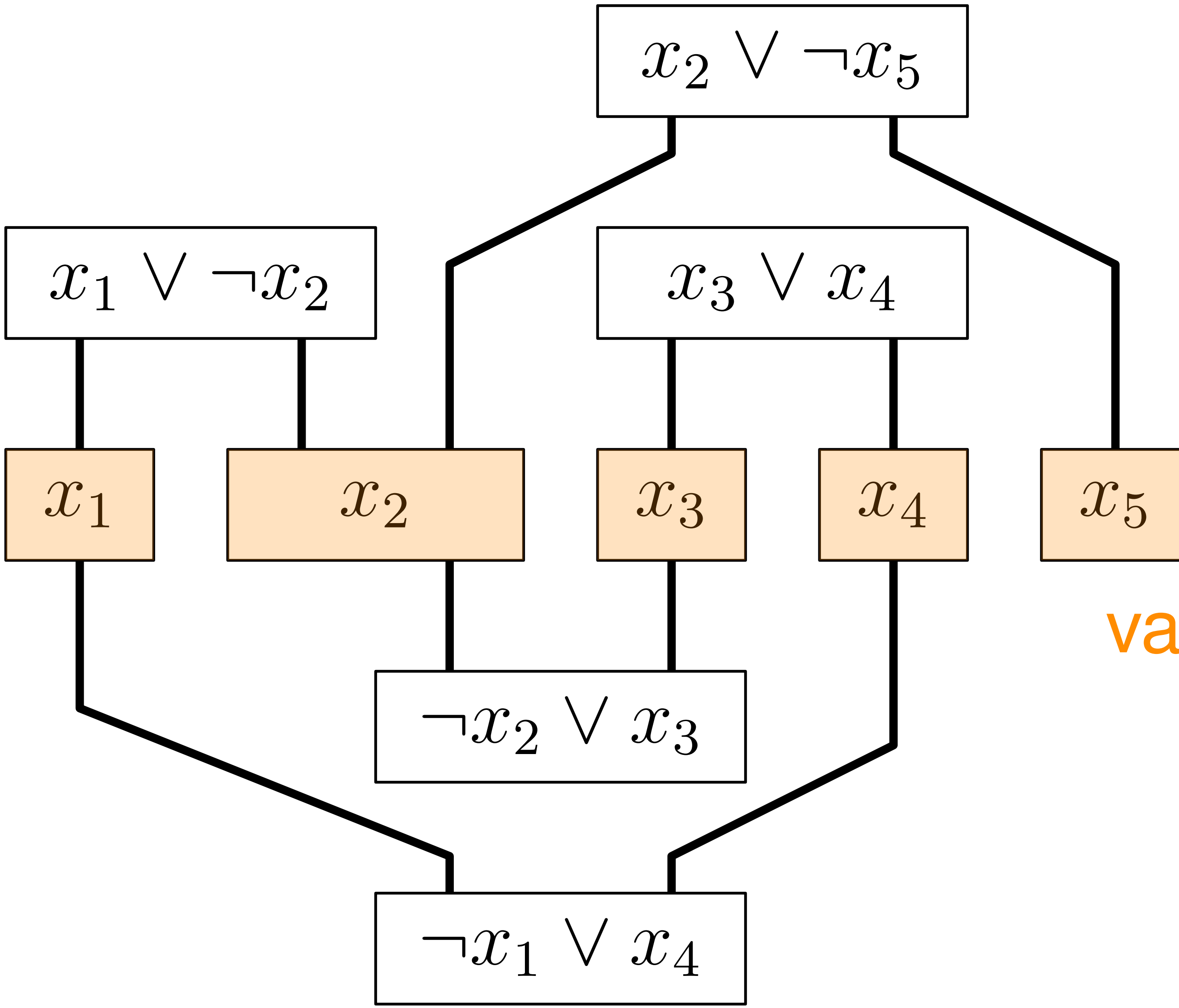
The task is to decide whether there exists a variable assignment fulfilling at least k clauses of ϕ.

Theorem [Buchin et al., 2020]: V-cycle Max2SAT is NP-hard.

The Reduction

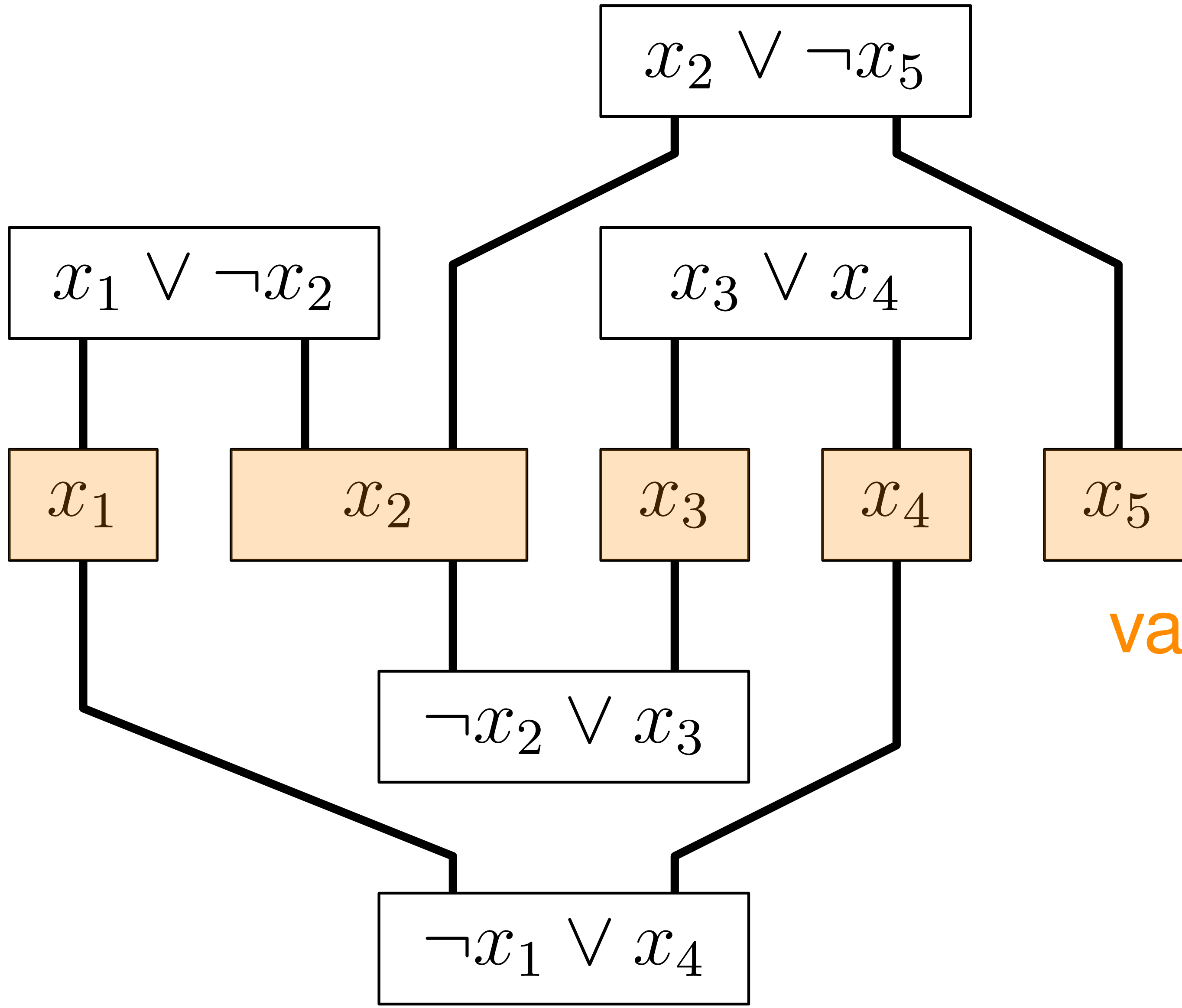


The Reduction



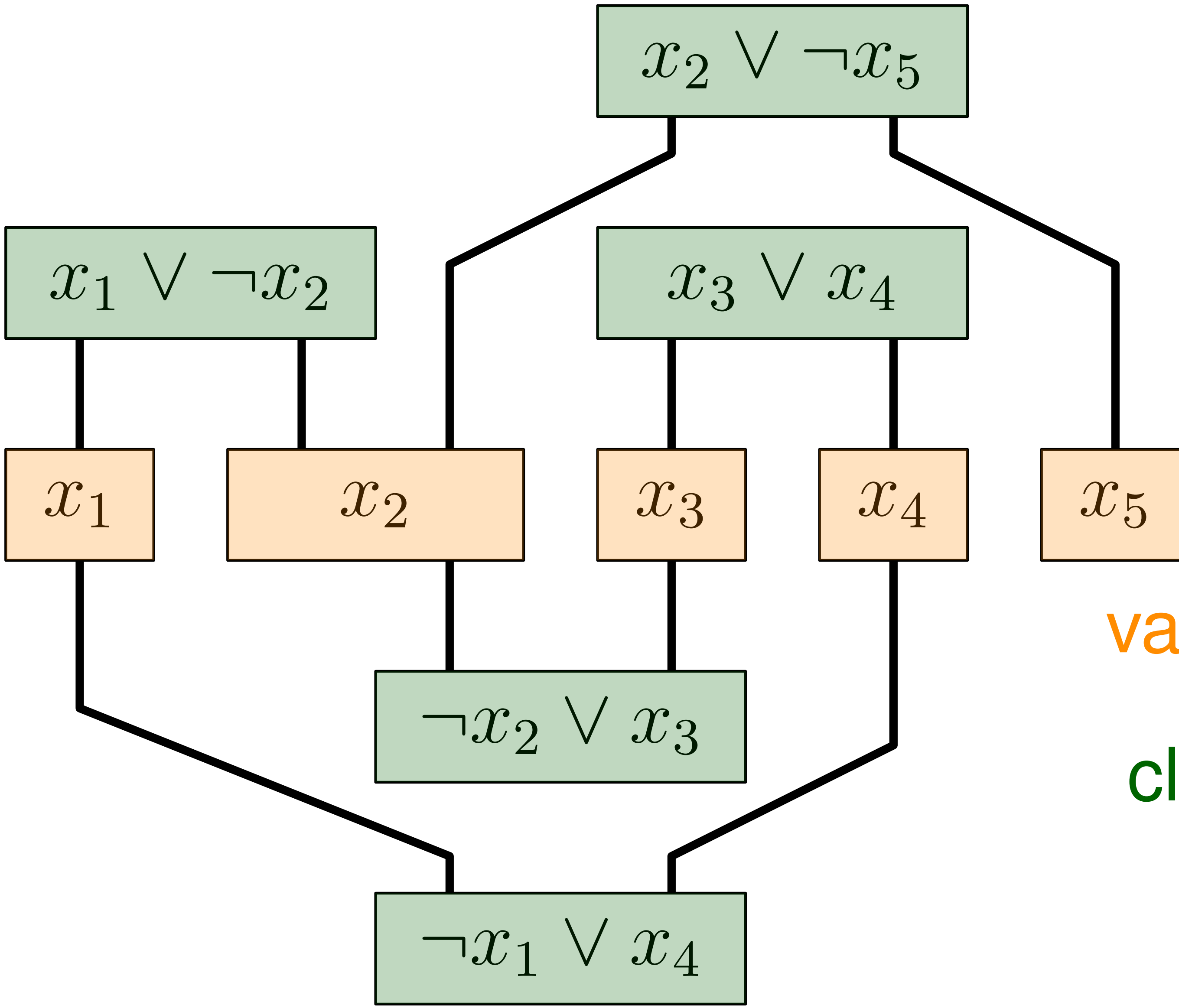
variable gadgets

The Reduction



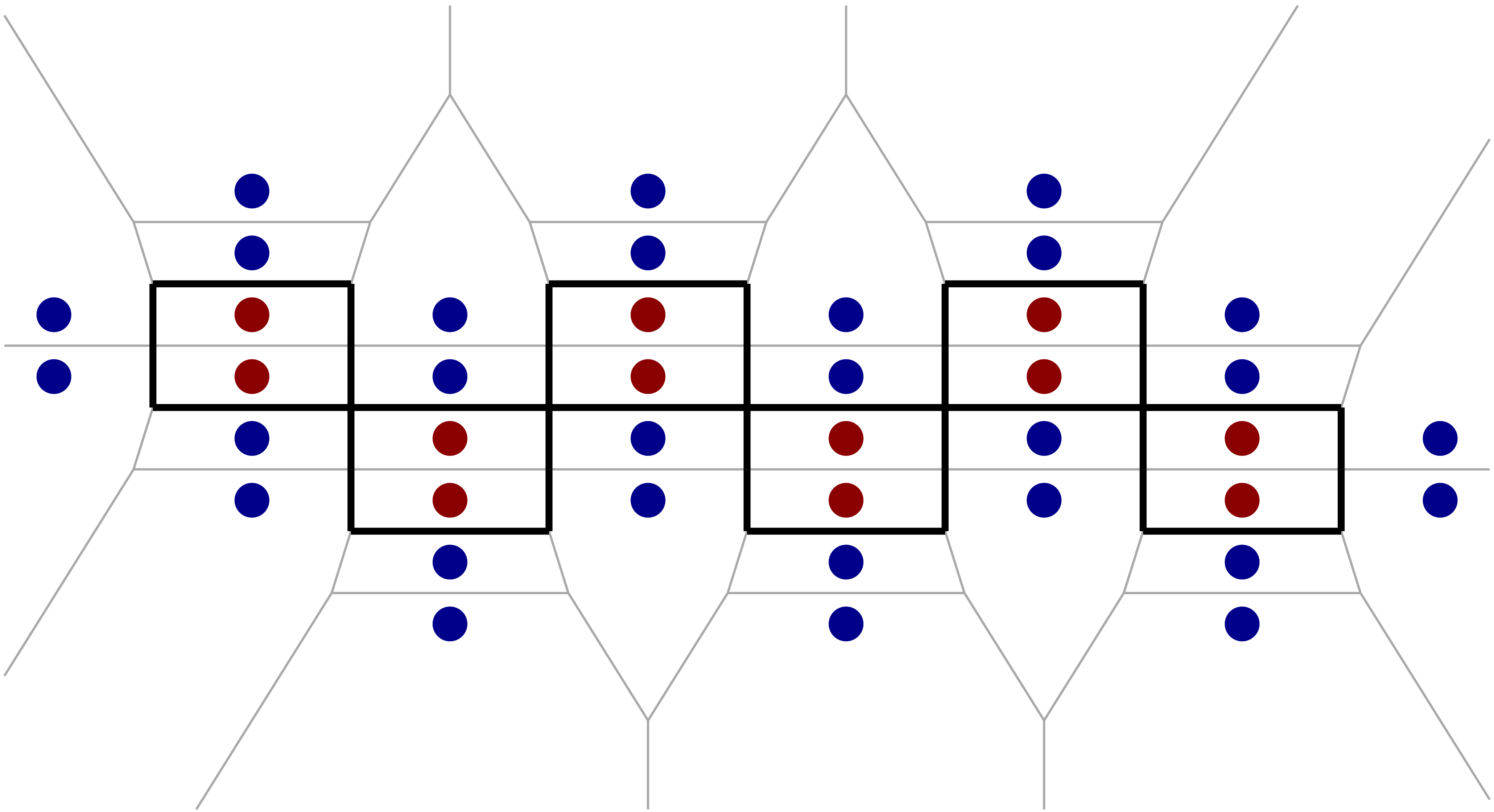
variable gadgets
channels

The Reduction

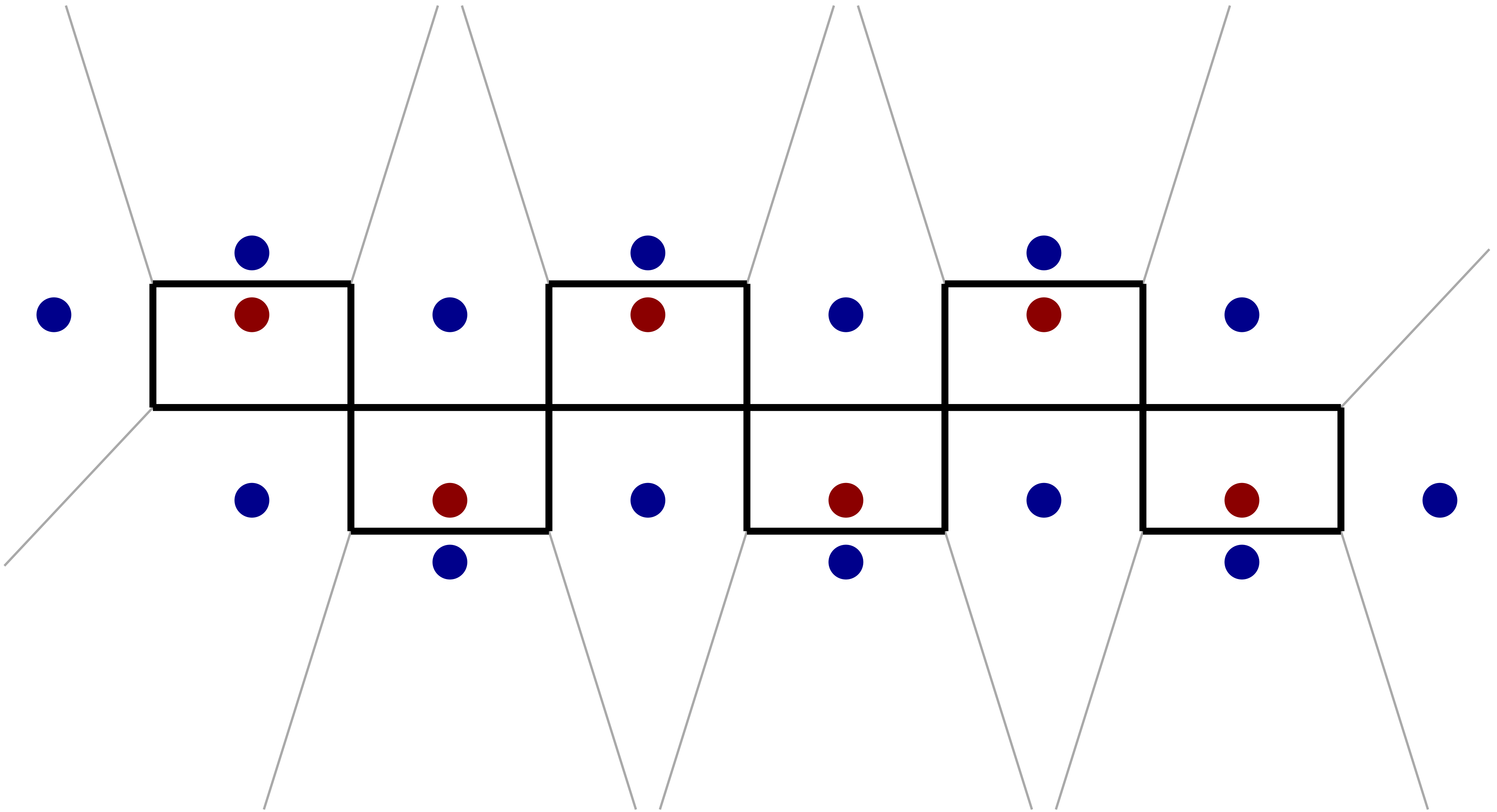


variable gadgets
channels
clause gadgets

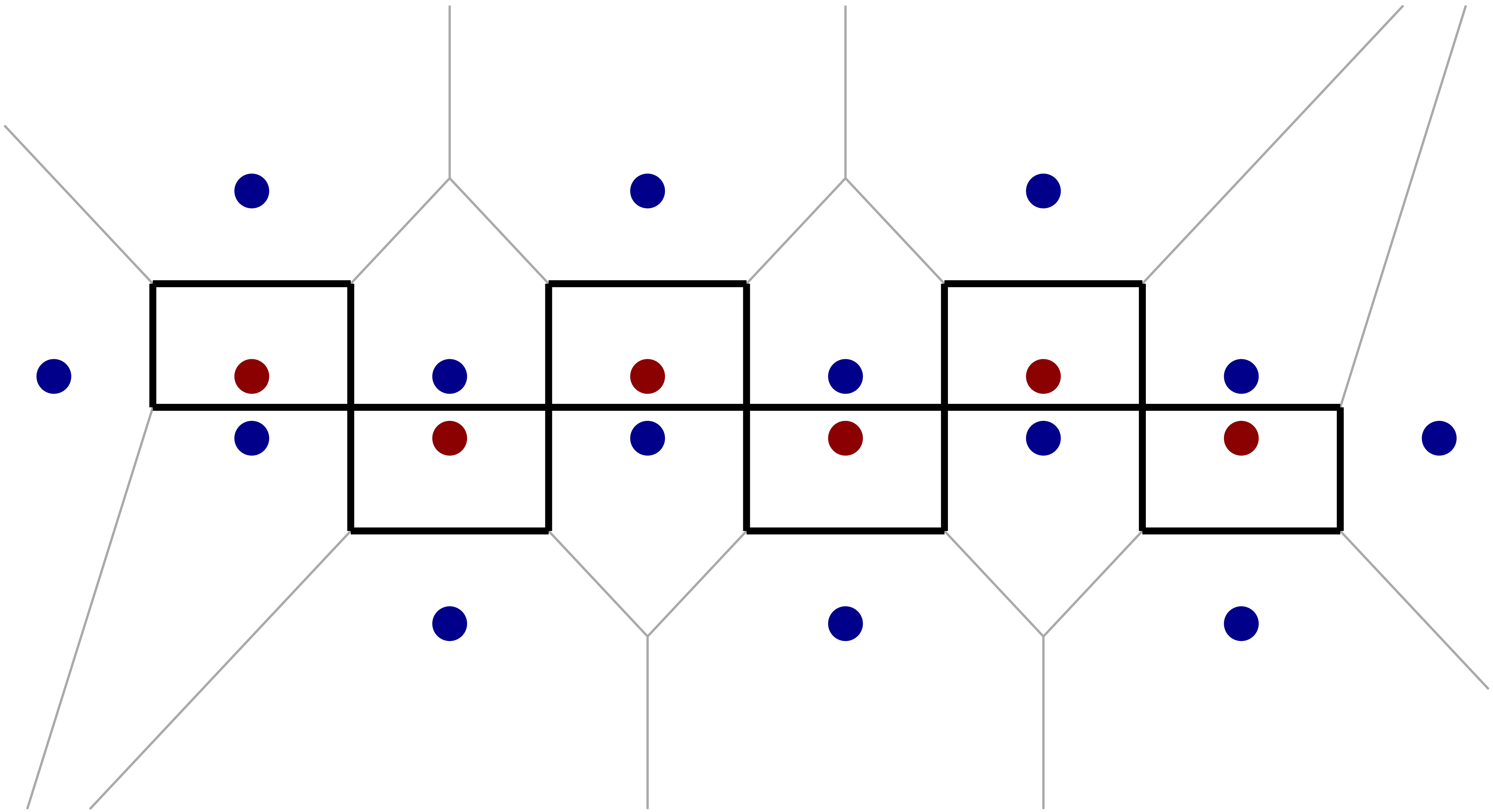
Variable Gadgets



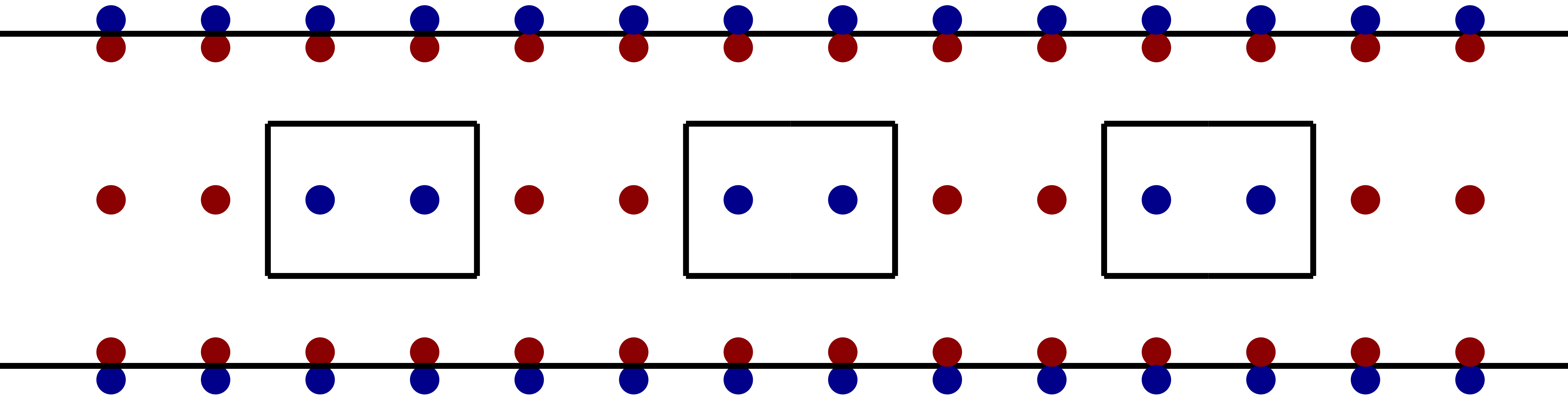
Variable Gadgets



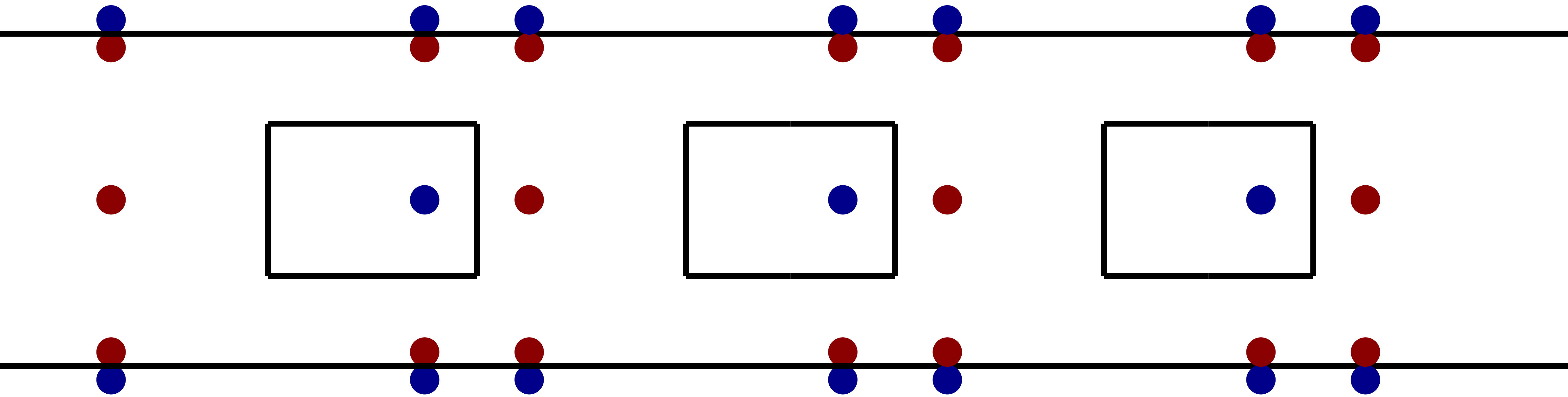
Variable Gadgets



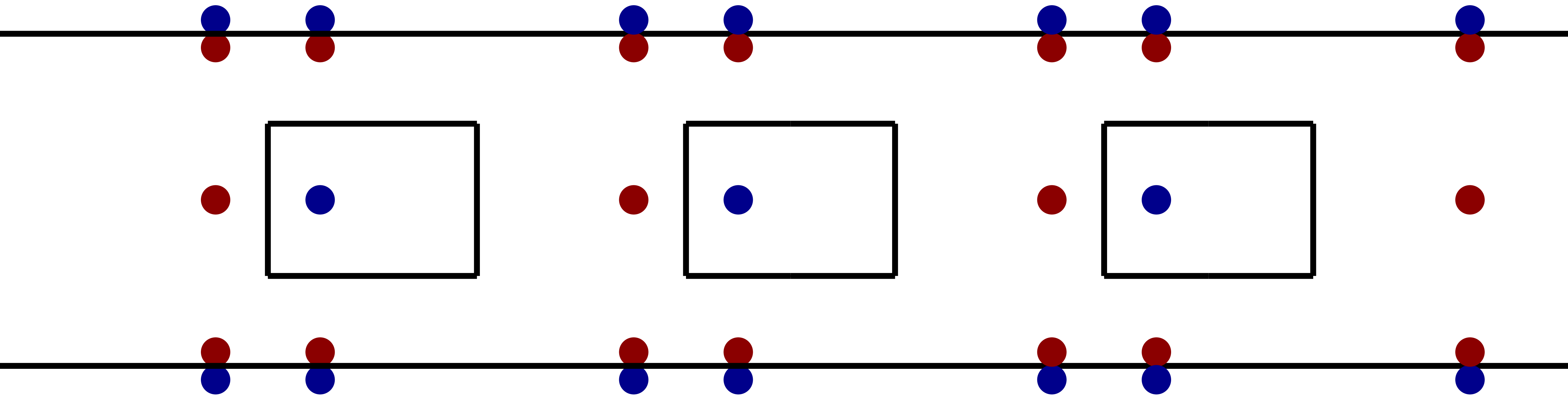
Channels



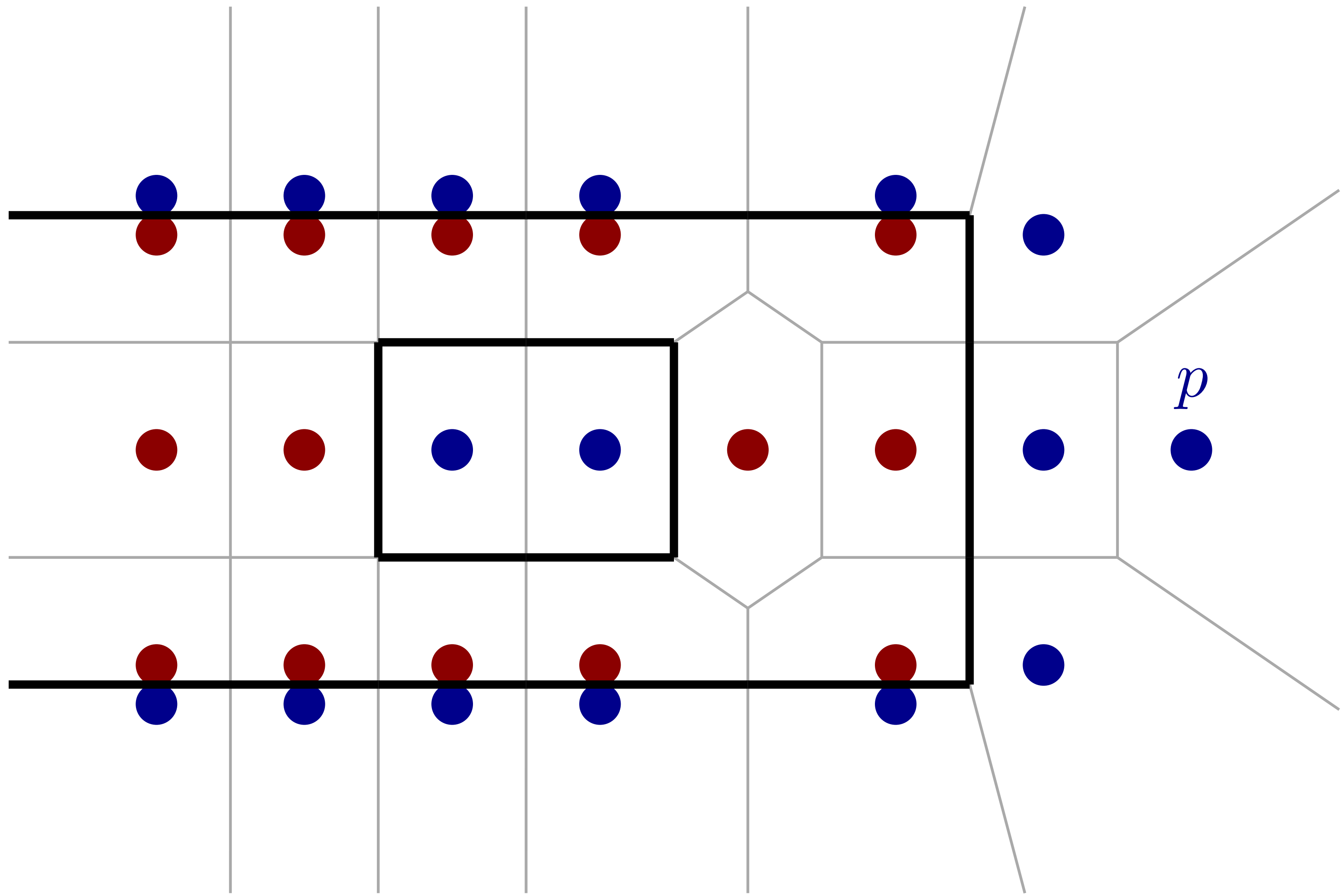
Channels



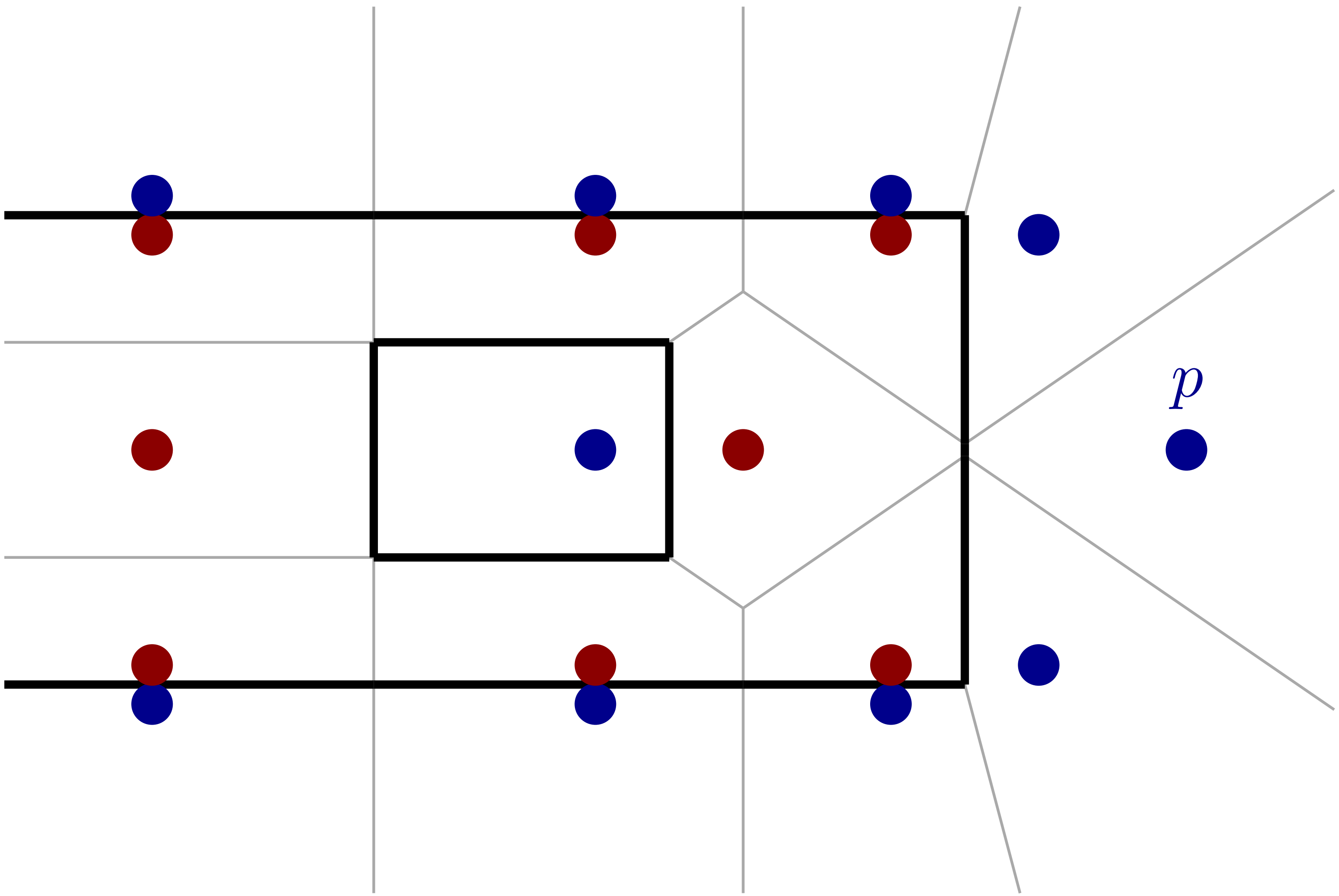
Channels



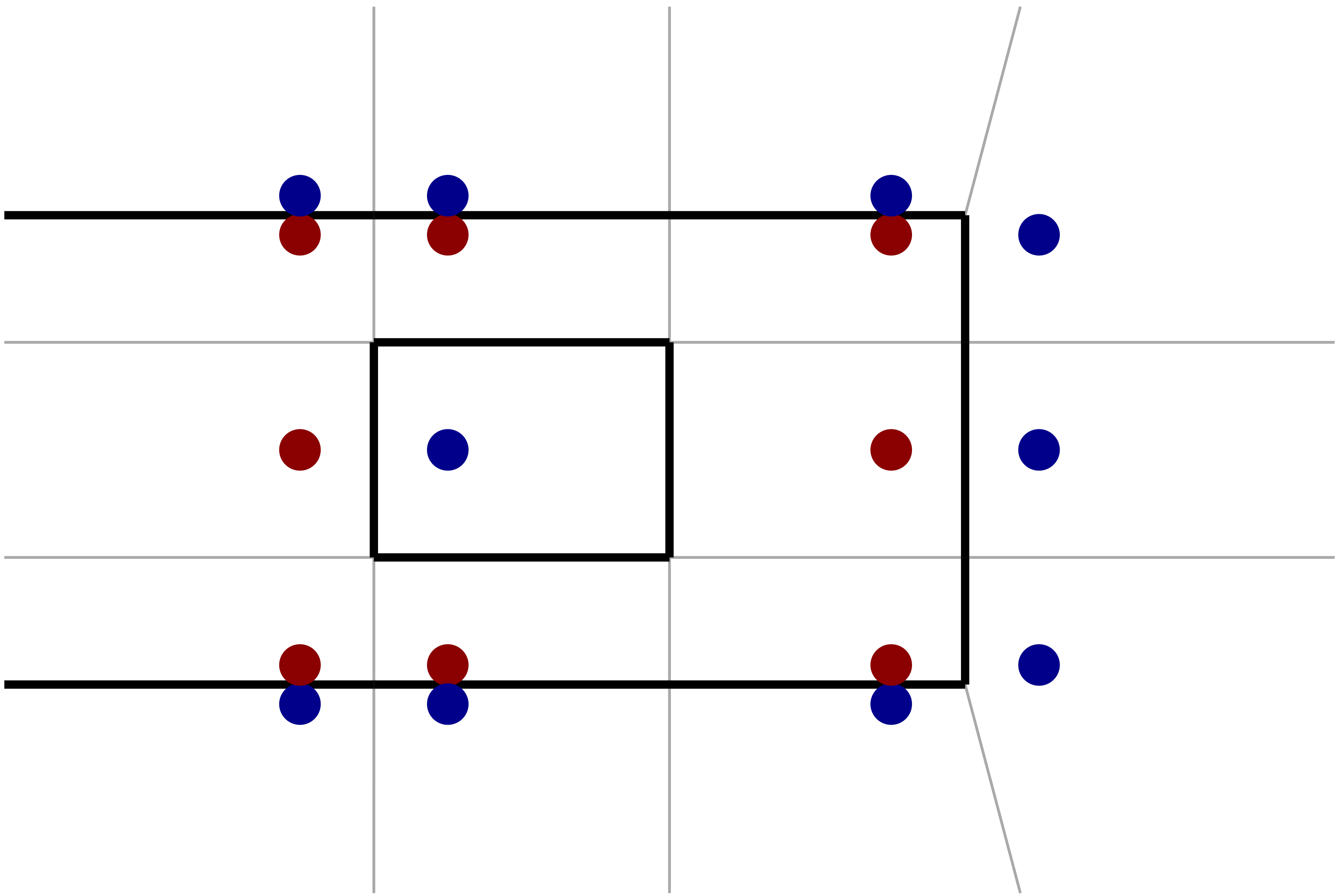
Reading the Value Off a Channel



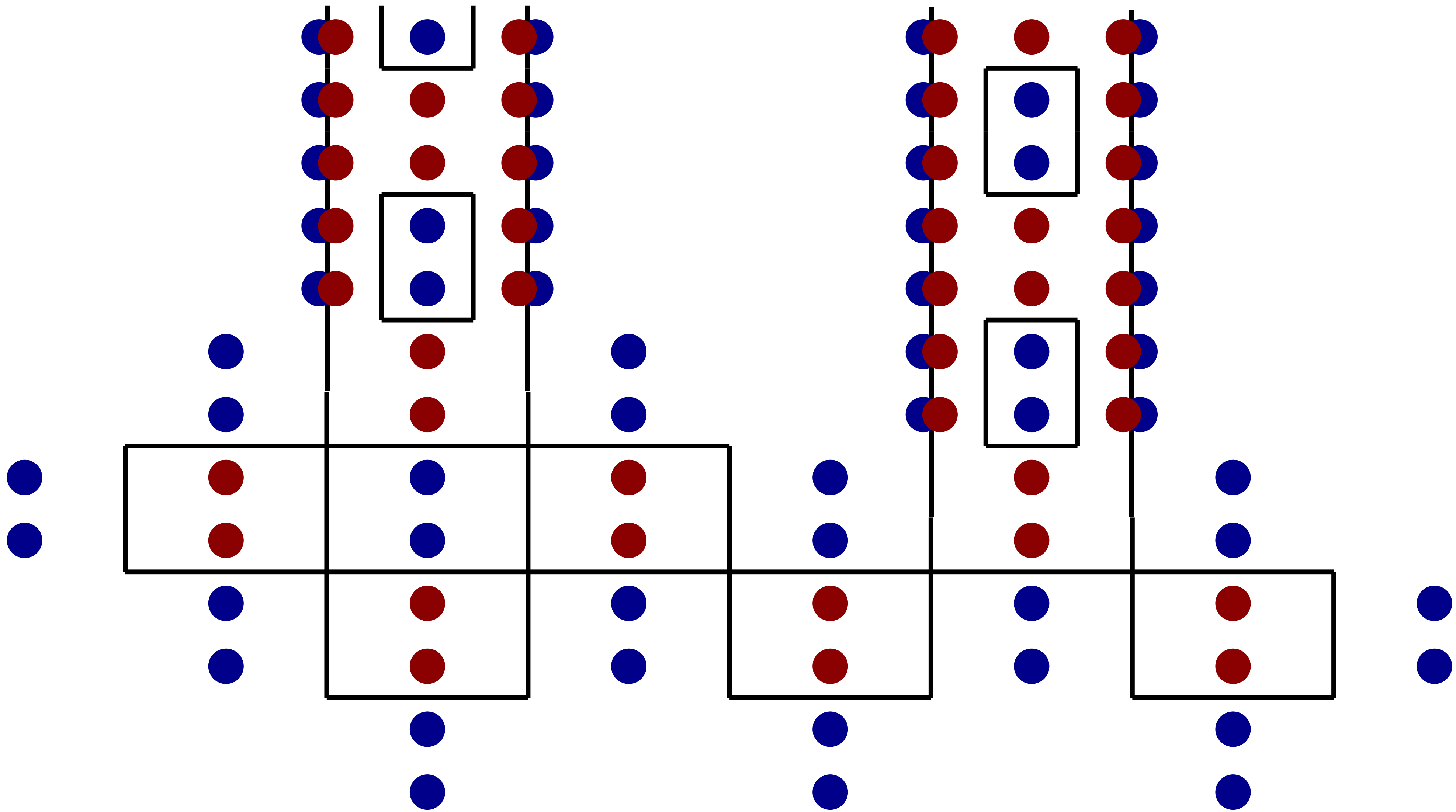
Reading the Value Off a Channel



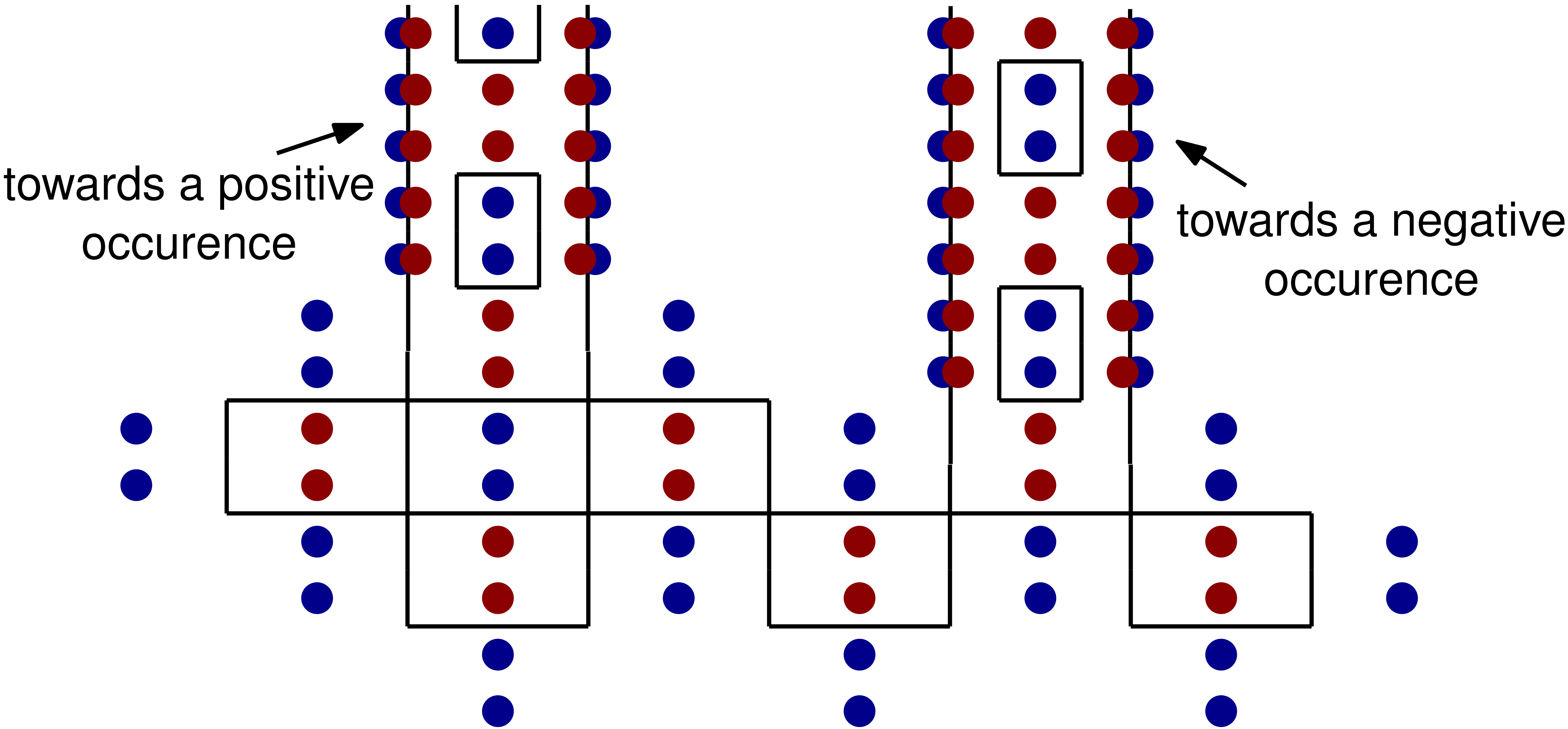
Reading the Value Off a Channel



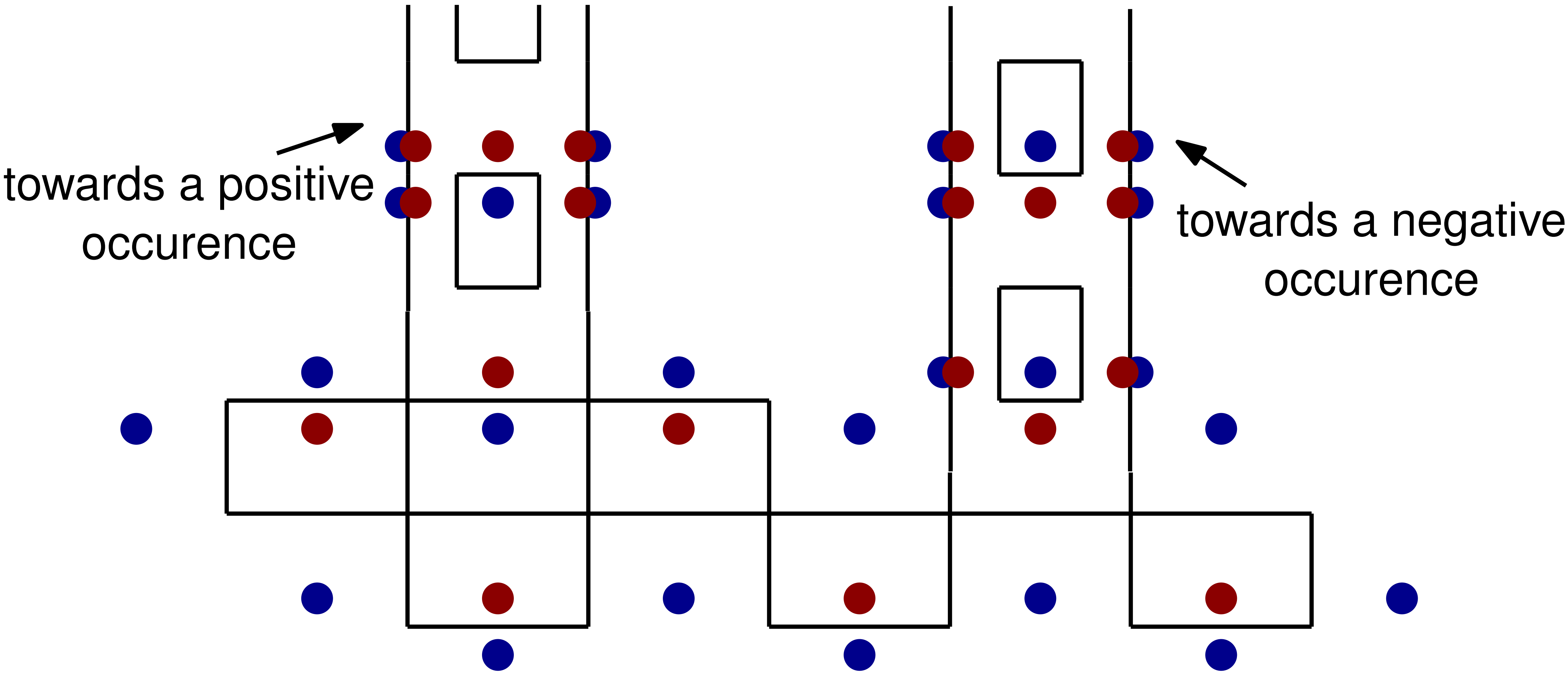
Getting a Value Onto a Channel



Getting a Value Onto a Channel

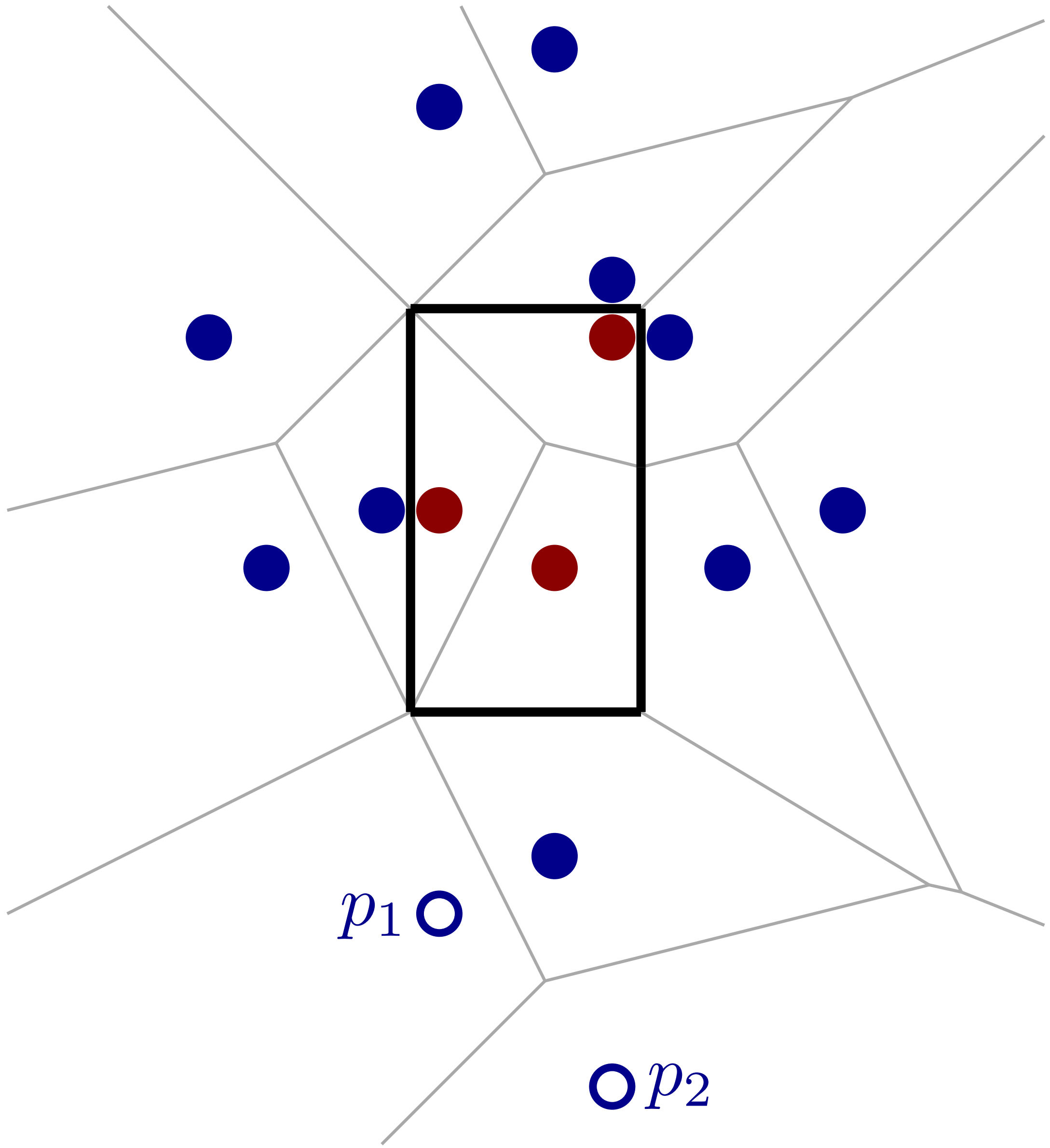


Getting a Value Onto a Channel



Clause Gadget

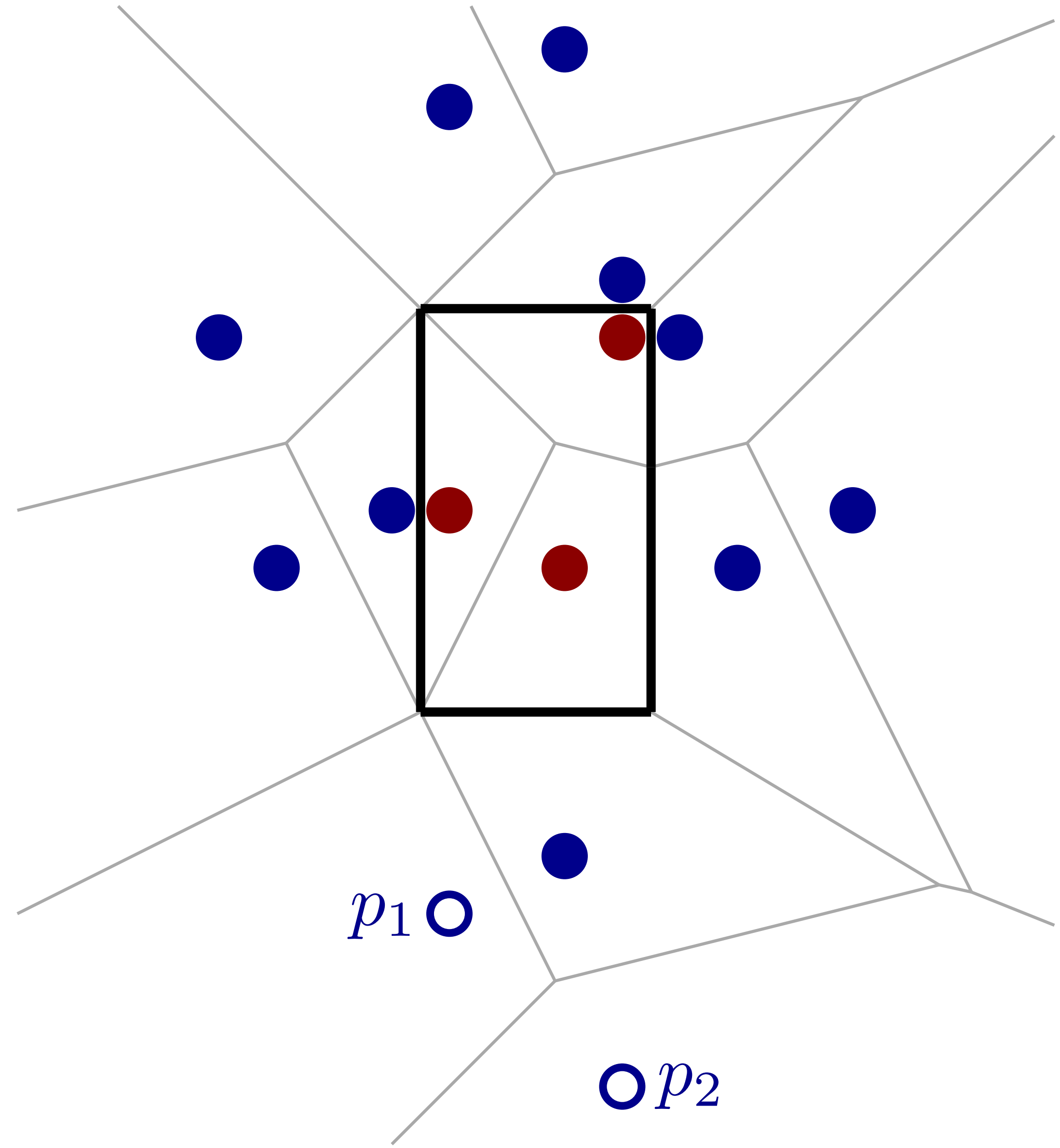
p_i is present “for free” if the i-th literal is fulfilled



Clause Gadget

p_i is present "for free" if the i-th literal is fulfilled

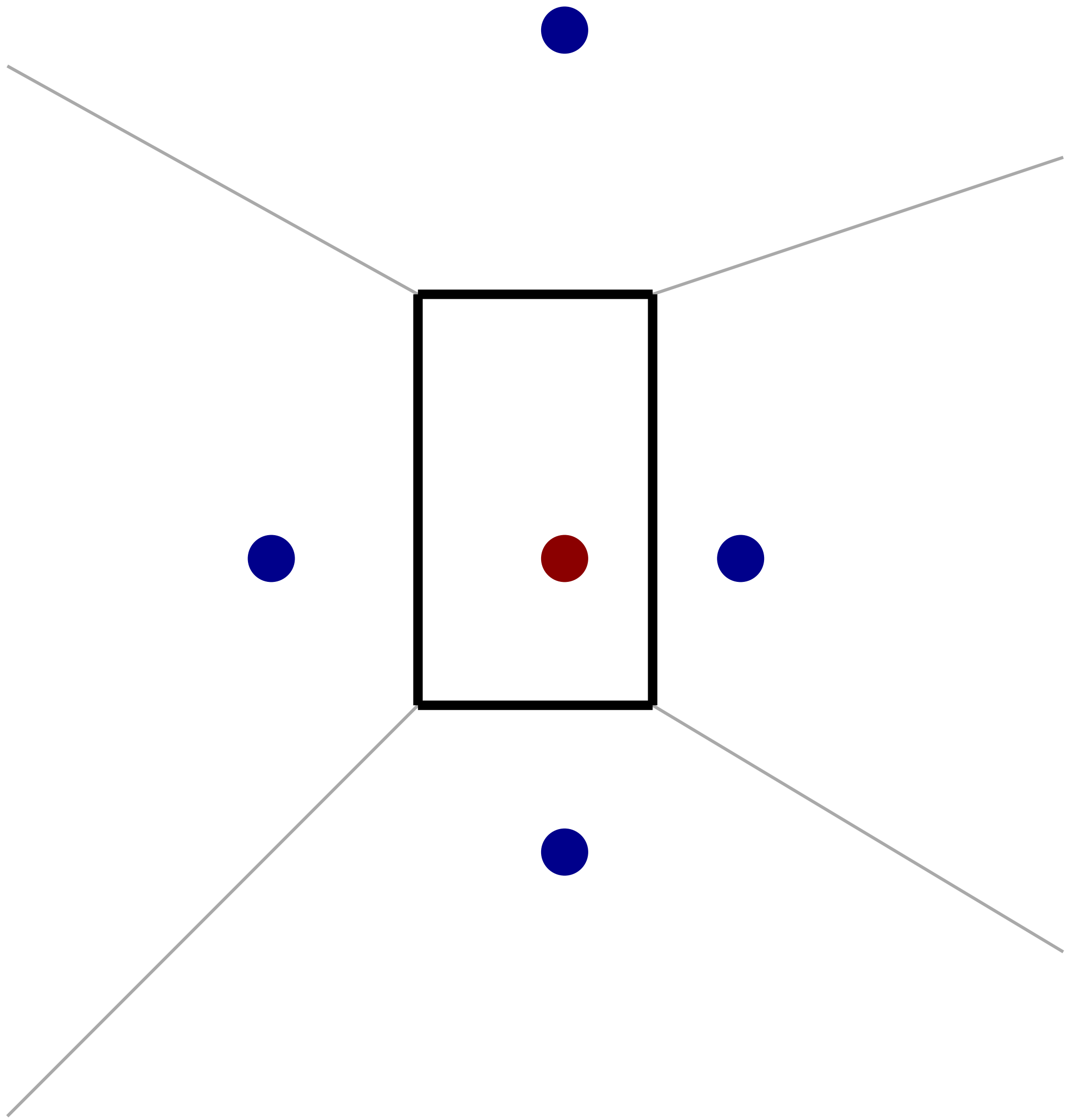
Idea: If p_1 or p_2 (or both) is present, we only need 4 points for the clause gadget. Otherwise, we need 5.



Clause Gadget

p_i is present “for free” if the i-th literal is fulfilled

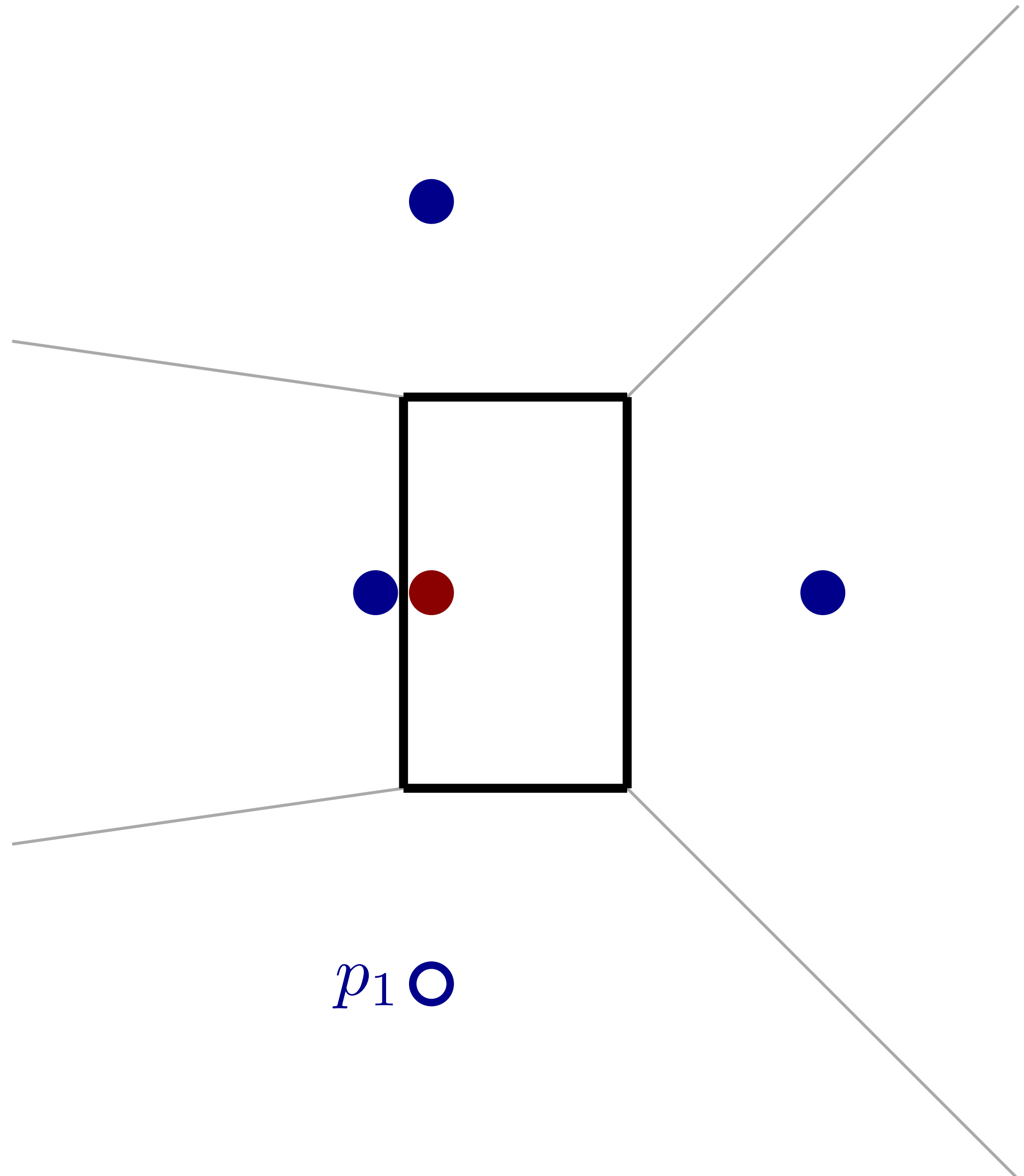
Idea: If p_1 or p_2 (or both) is present, we only need 4 points for the clause gadget. Otherwise, we need 5.



Clause Gadget

p_i is present "for free" if the i-th literal is fulfilled

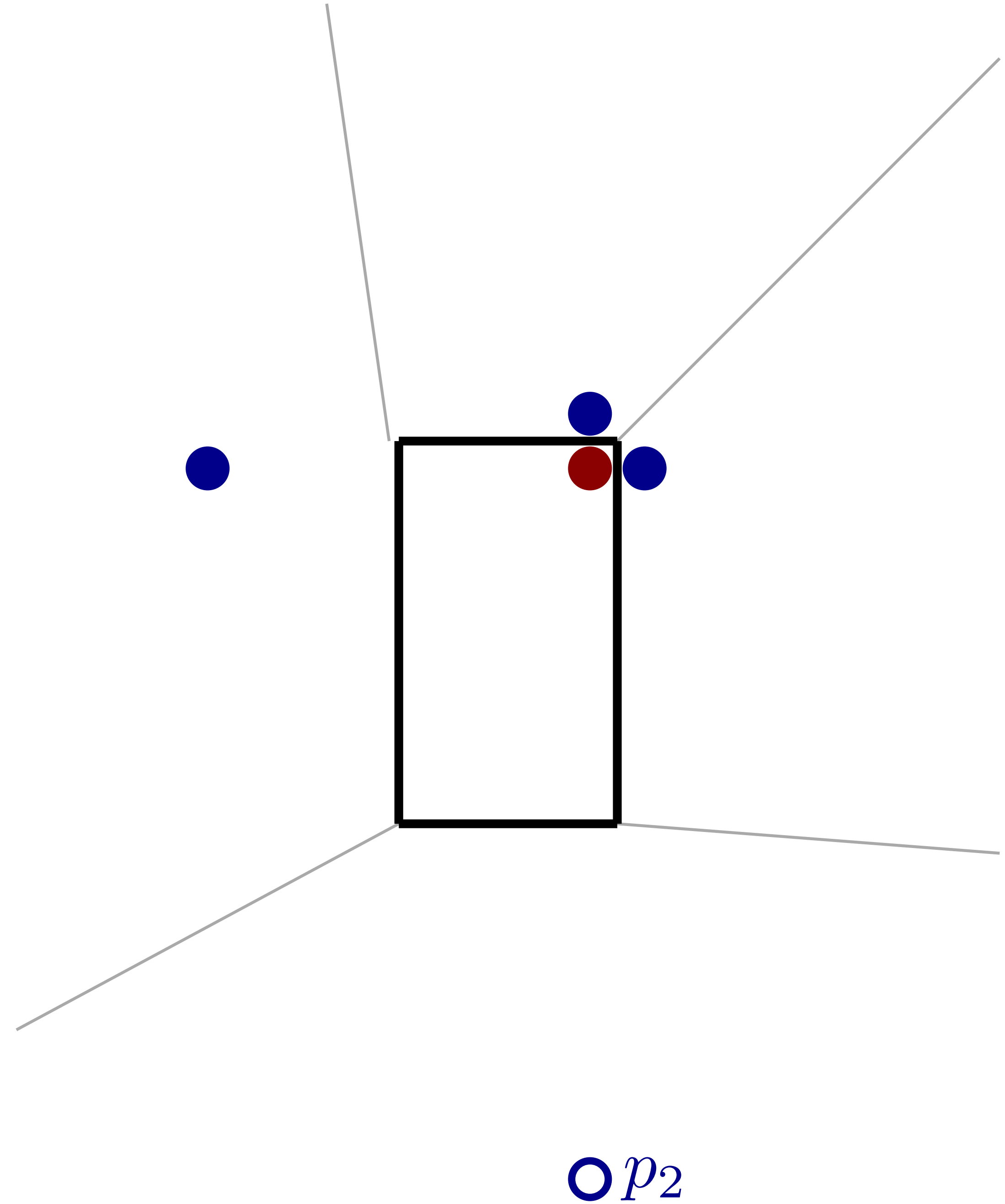
Idea: If p_1 or p_2 (or both) is present, we only need 4 points for the clause gadget. Otherwise, we need 5.



Clause Gadget

p_i is present "for free" if the i-th literal is fulfilled

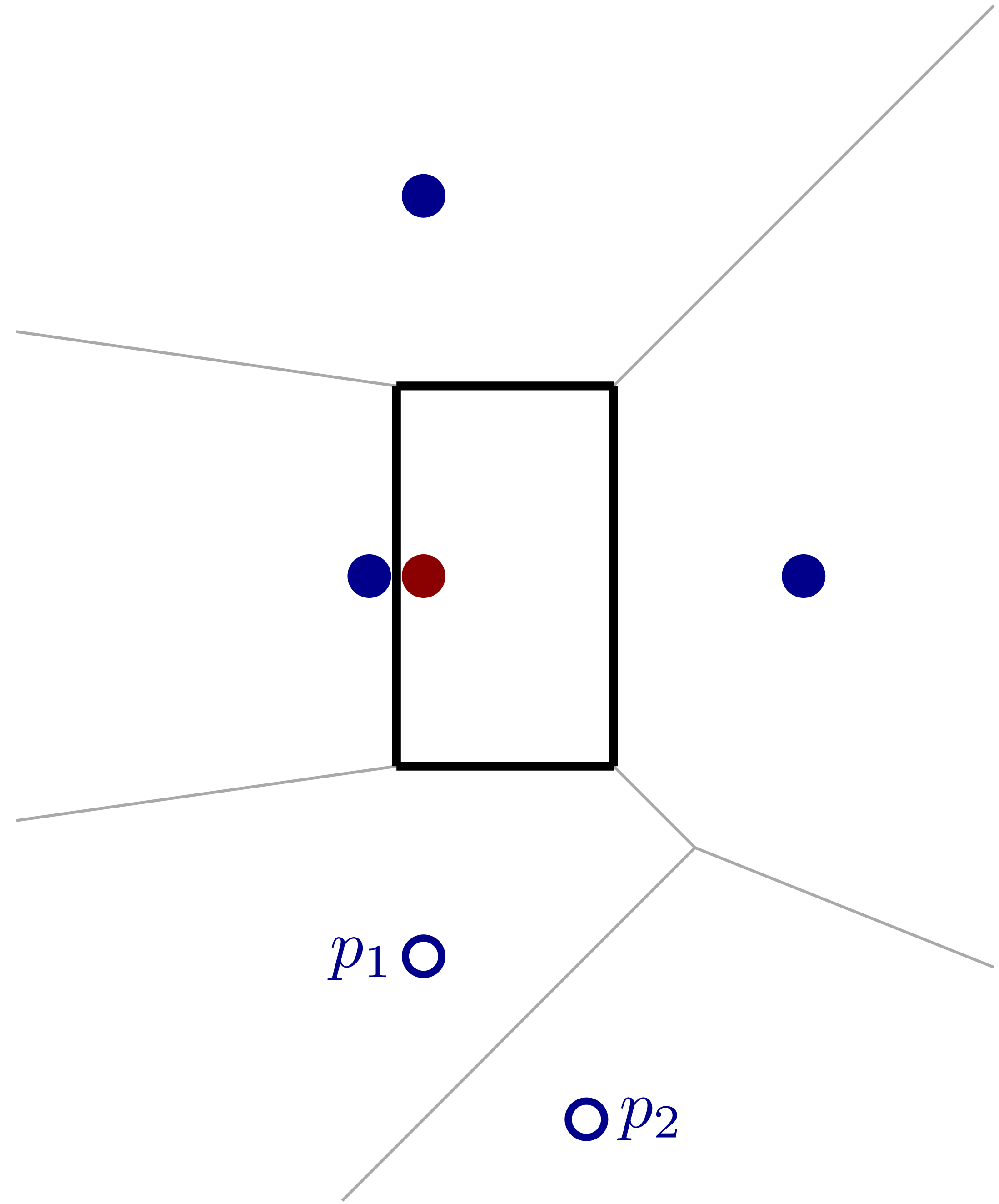
Idea: If p_1 or p_2 (or both) is present, we only need 4 points for the clause gadget. Otherwise, we need 5.



Clause Gadget

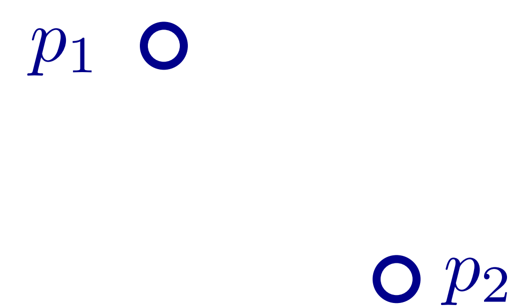
p_i is present "for free" if the i-th literal is fulfilled

Idea: If p_1 or p_2 (or both) is present, we only need 4 points for the clause gadget. Otherwise, we need 5.



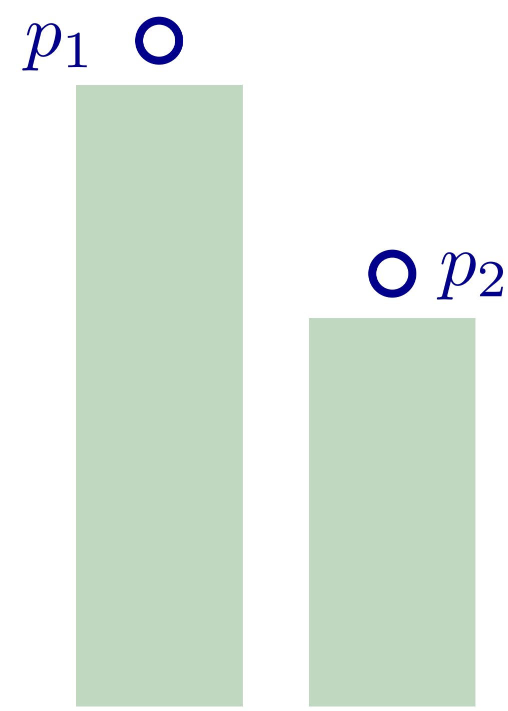
The Ugly Truth

p_1 and p_2 need to be endpoints of channels!



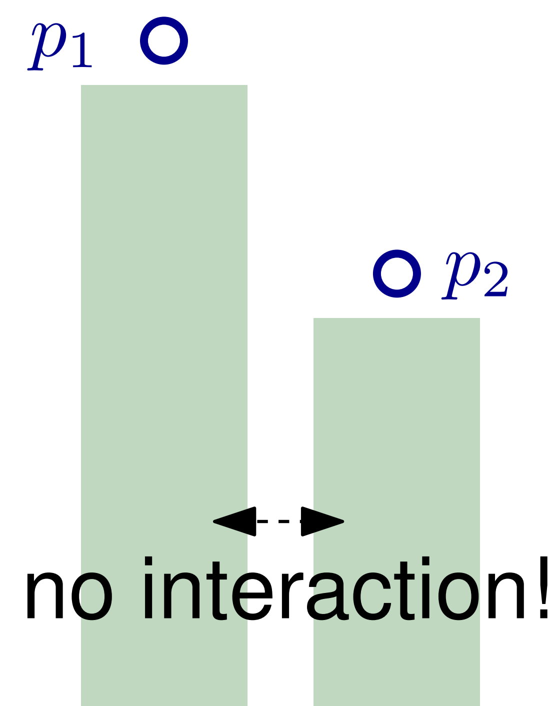
The Ugly Truth

p_1 and p_2 need to be endpoints of channels!



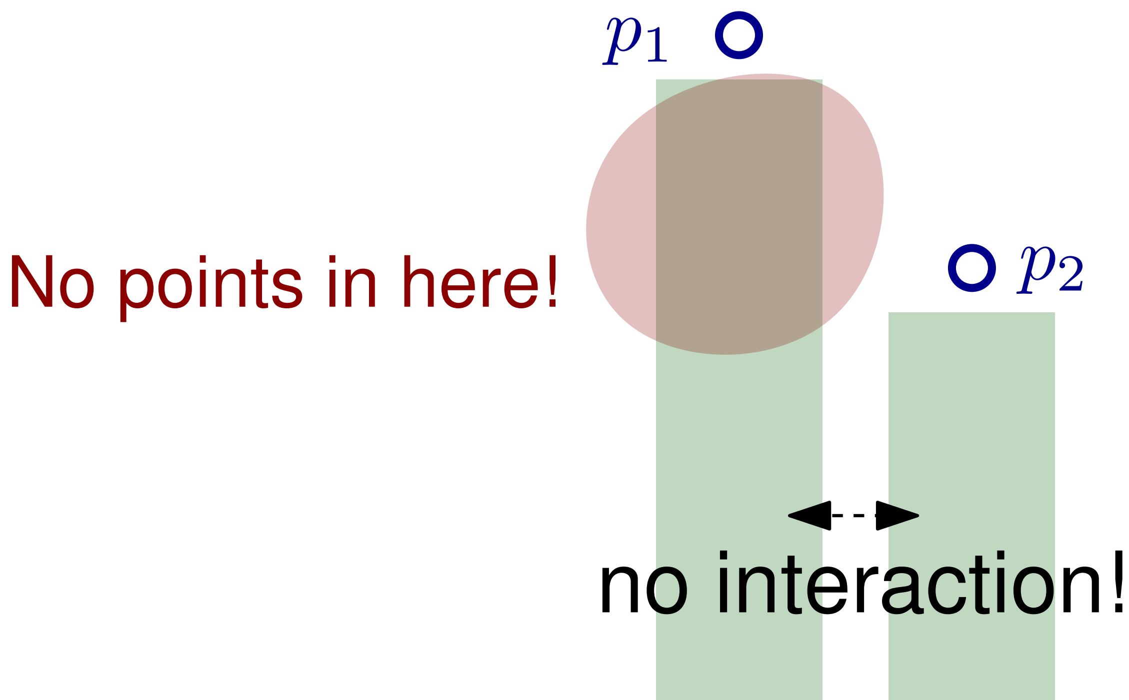
The Ugly Truth

p_1 and p_2 need to be endpoints of channels!



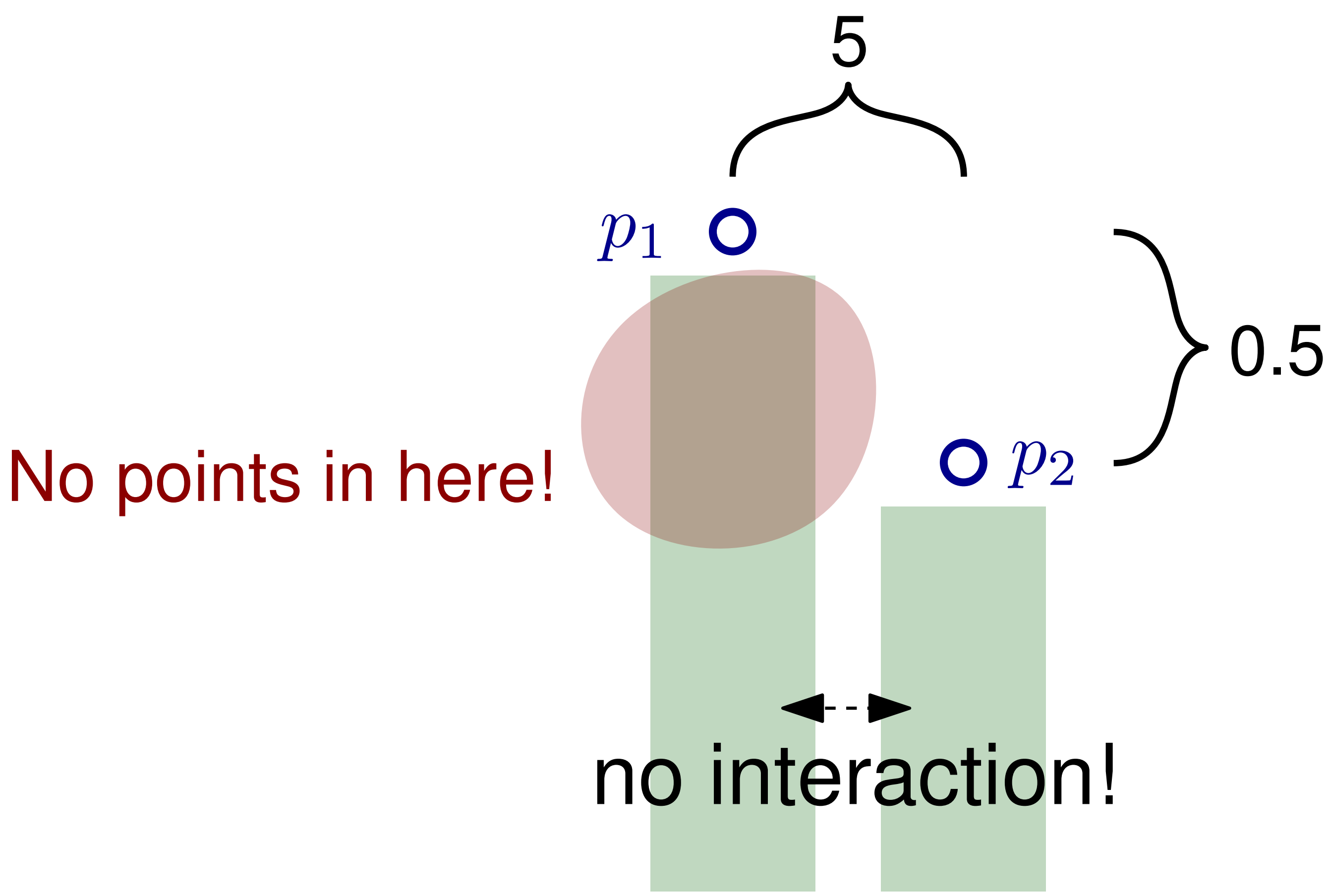
The Ugly Truth

p_1 and p_2 need to be endpoints of channels!

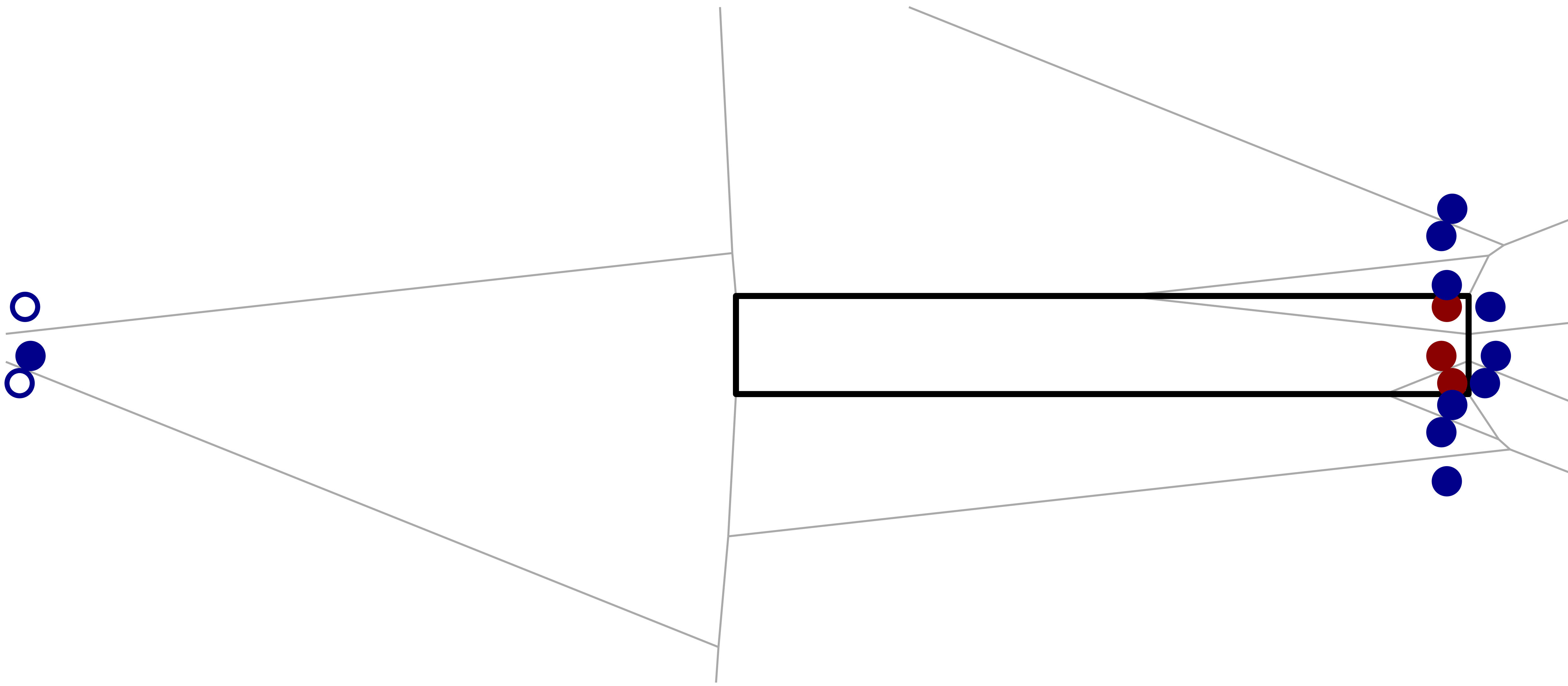


The Ugly Truth

p_1 and p_2 need to be endpoints of channels!

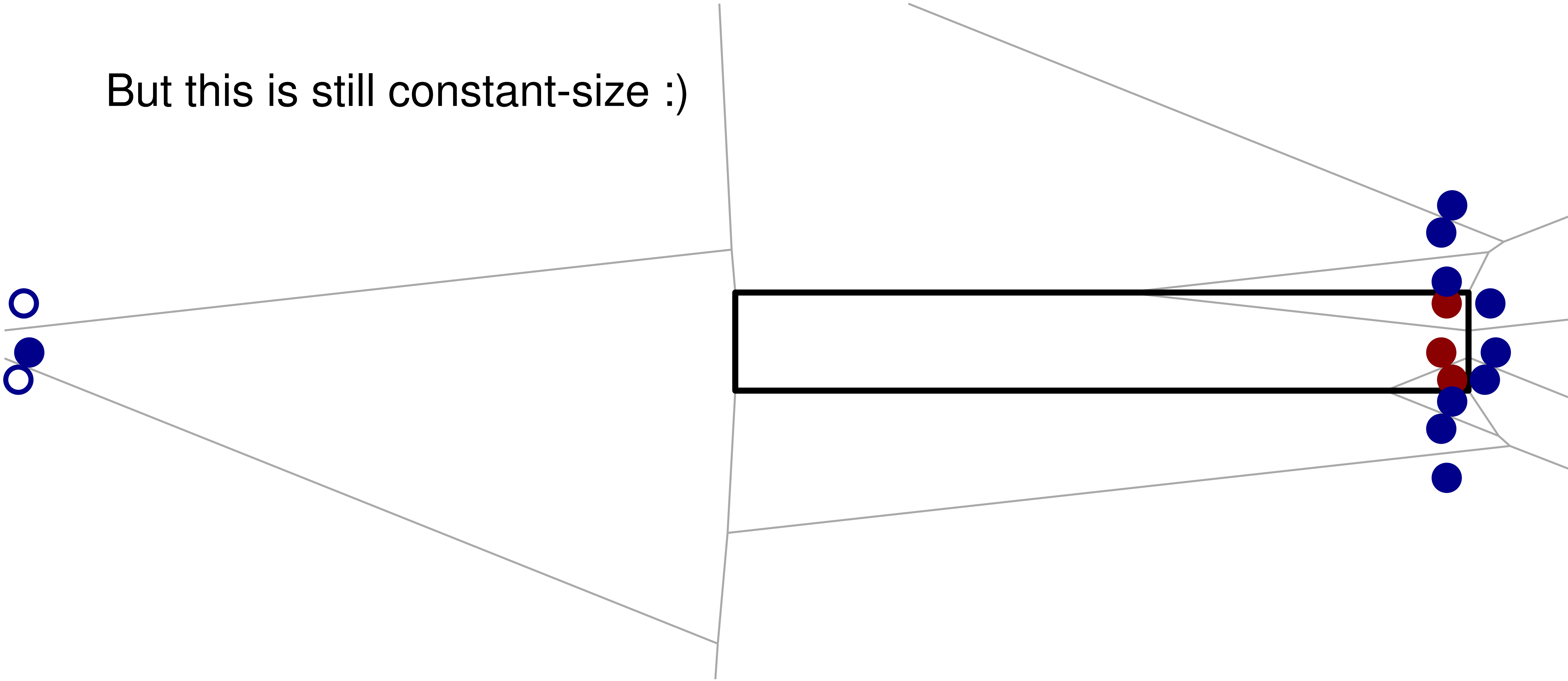


The Ugly Truth



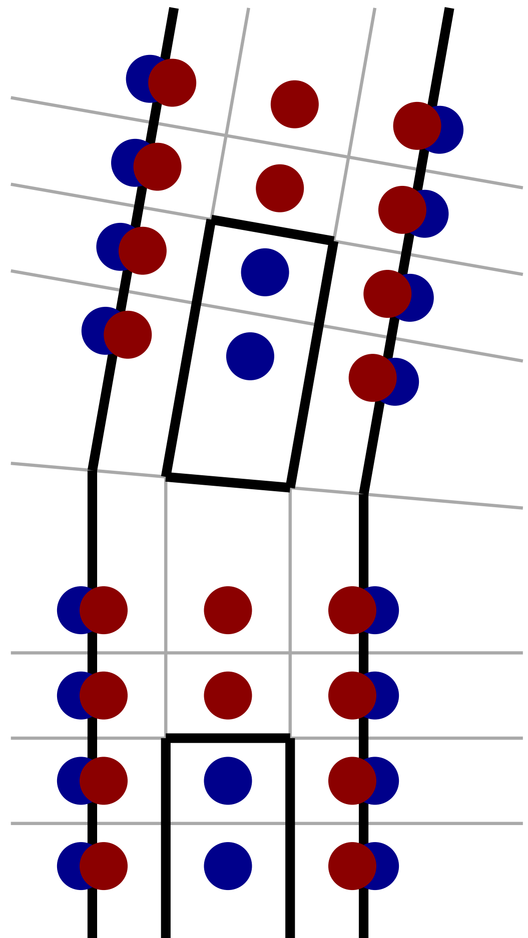
The Ugly Truth

But this is still constant-size :)

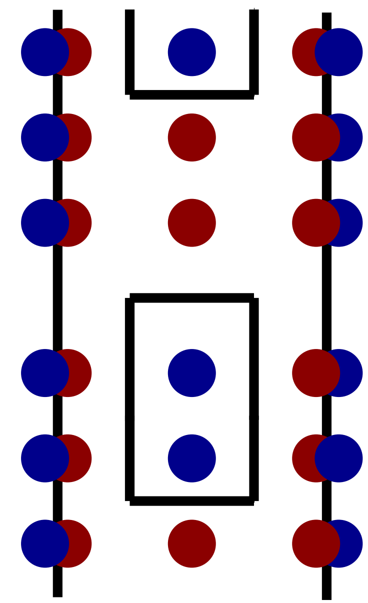


Some Missing Ingredients

Bends



Stretching



The Final Argument

Claim: There exists an integer $N(\phi)$, such that there exists an assignment fulfilling at least k clauses of ϕ if and only if there exists a reduced training set $Q \subseteq P(\phi)$ with $|Q| \leq N(\phi) - k$.

The Final Argument

Claim: There exists an integer $N(\phi)$, such that there exists an assignment fulfilling at least k clauses of ϕ if and only if there exists a reduced training set $Q \subseteq P(\phi)$ with $|Q| \leq N(\phi) - k$.

assignment $\Rightarrow$ reduced training set: clear

The Final Argument

Claim: There exists an integer $N(\phi)$, such that there exists an assignment fulfilling at least k clauses of ϕ if and only if there exists a reduced training set $Q \subseteq P(\phi)$ with $|Q| \leq N(\phi) - k$.

assignment $\Rightarrow$ reduced training set: clear
reduced training set $\Rightarrow$ assignment:

The Final Argument

Claim: There exists an integer $N(\phi)$, such that there exists an assignment fulfilling at least k clauses of ϕ if and only if there exists a reduced training set $Q \subseteq P(\phi)$ with $|Q| \leq N(\phi) - k$.

assignment $\Rightarrow$ reduced training set: clear
reduced training set $\Rightarrow$ assignment:

"Cheating is not beneficial"

Conclusion

Finding the minimum-cardinality reduced training set is NP-complete for $d \geq 2$ and any number of colors $k \geq 2$.

Open Question: For many “lossy” notions of nearest-neighbor condensation even *approximating* the minimum-cardinality subset fulfilling the required guarantees is NP-hard.
What about our exact setting?