SUPER*TREE LEARNERS:

a new metalearner for local data science

LEVI JOHN WOLF levi.john.wolf@bristol.ac.uk

WHY DO WE NEED THIS?

GEOGRAPHICAL CLUSTER-REG

UNDERSTANDING QUADTREES

APPLYING QUADTREE REGRESSION

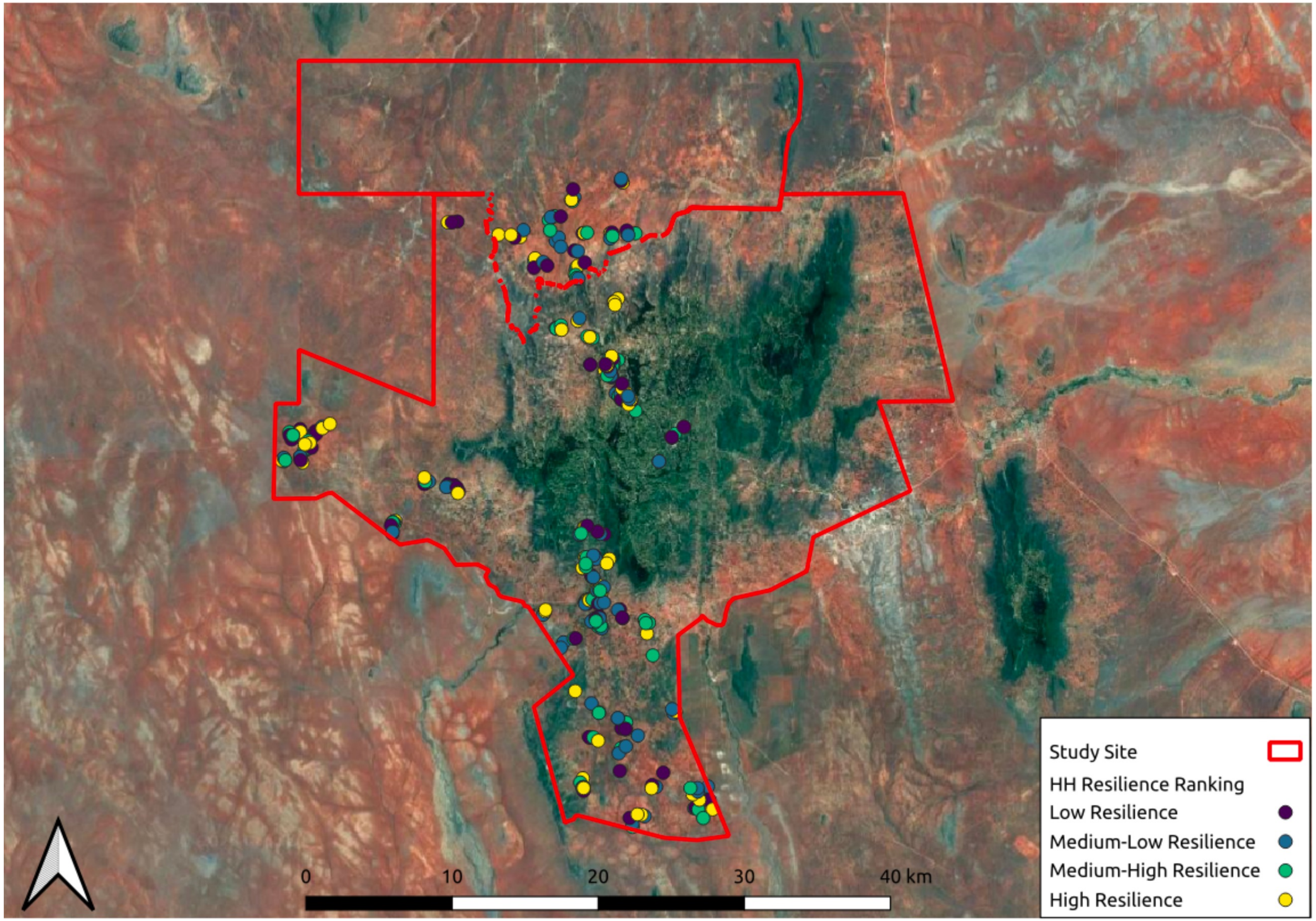
WHY DO WE NEED THIS?

place is not always so obvious

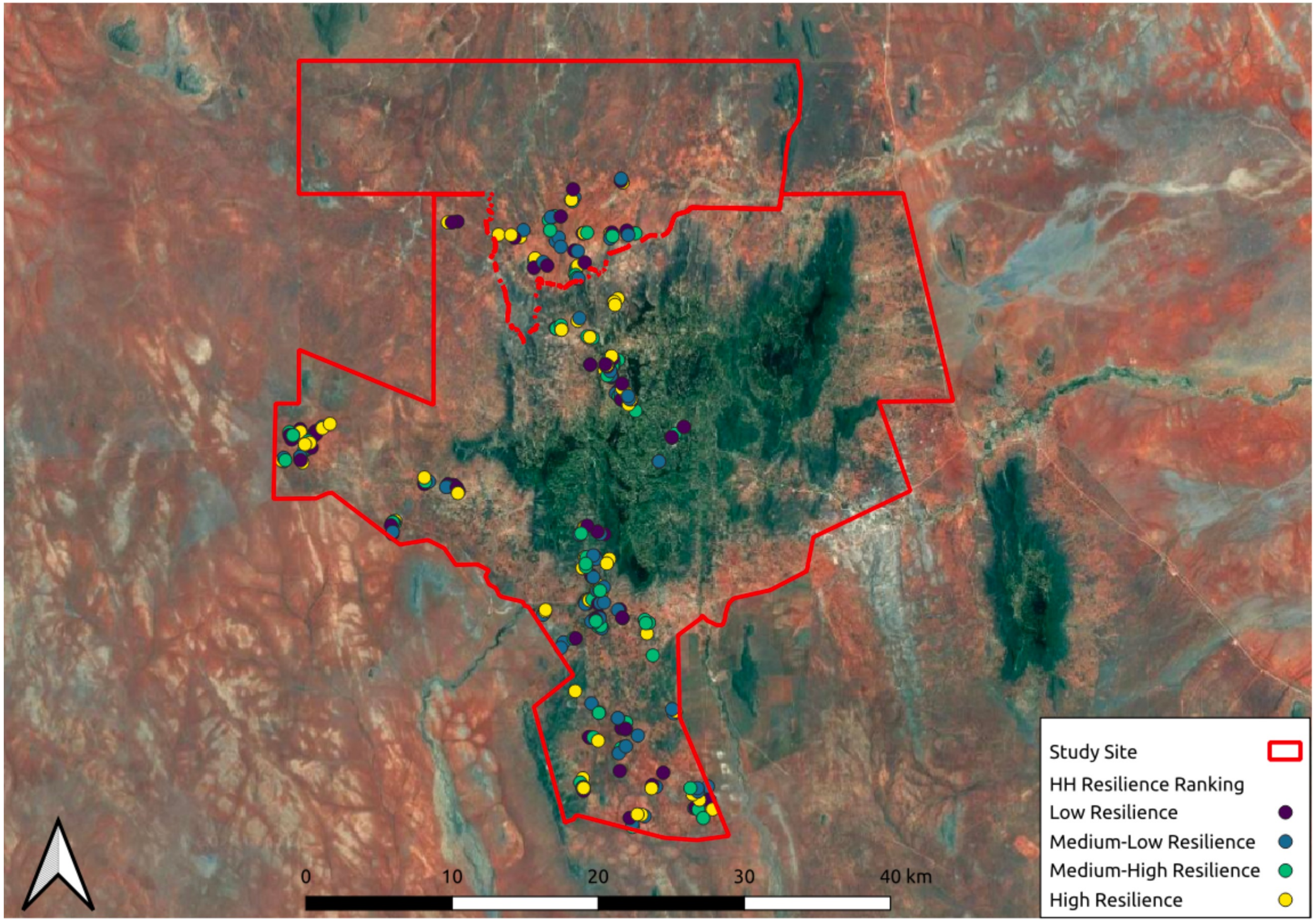
GEOGRAPHICAL CLUSTER-REG

UNDERSTANDING QUADTREES

APPLYING QUADTREE REGRESSION

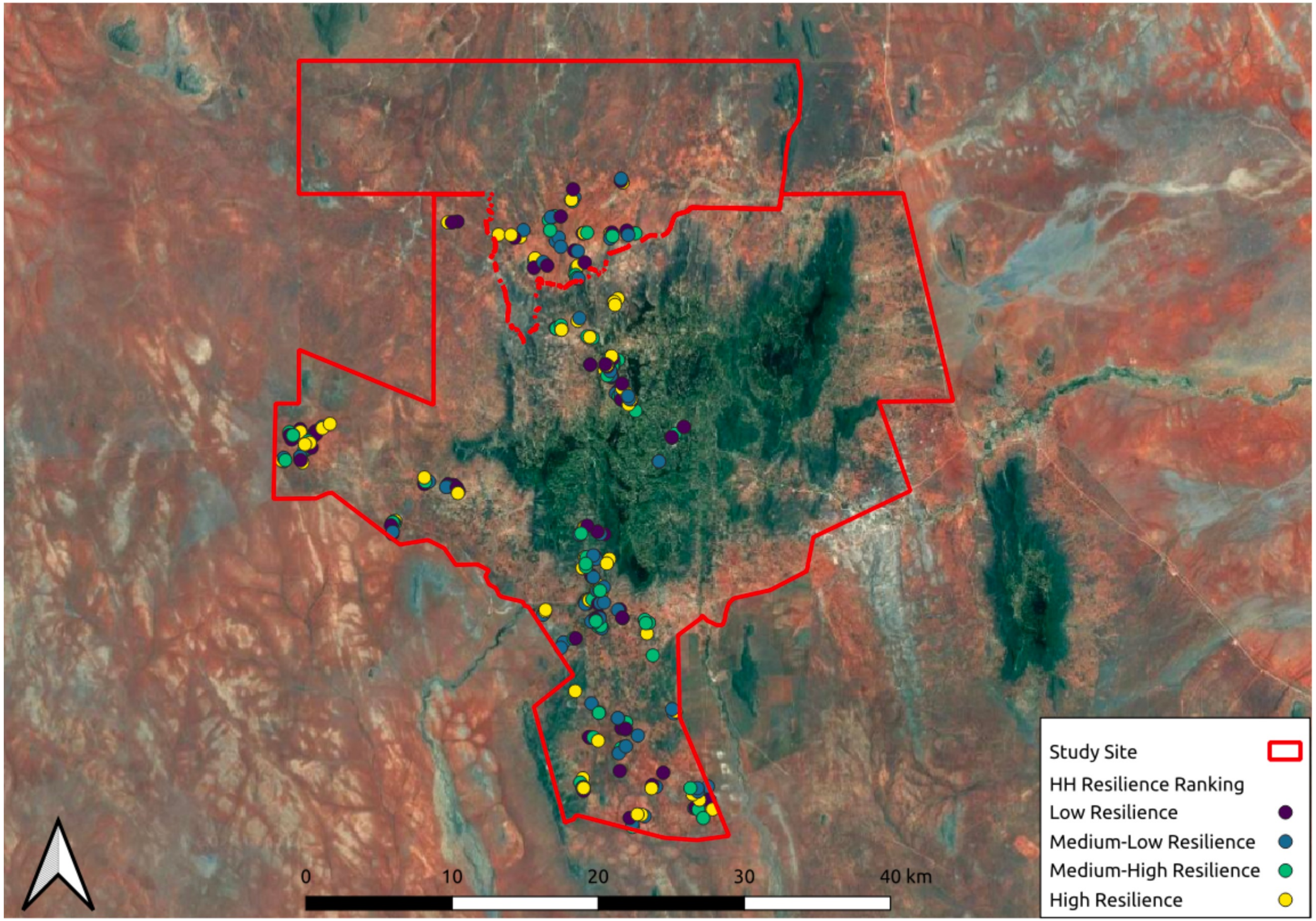


WHAT IS CLUSTERING REGRESSION



Different clusters differ:
- What is “food”?
- Do we farm or herd?
- How do we do that?
- Is water held in common?
- Is there a community safety net?

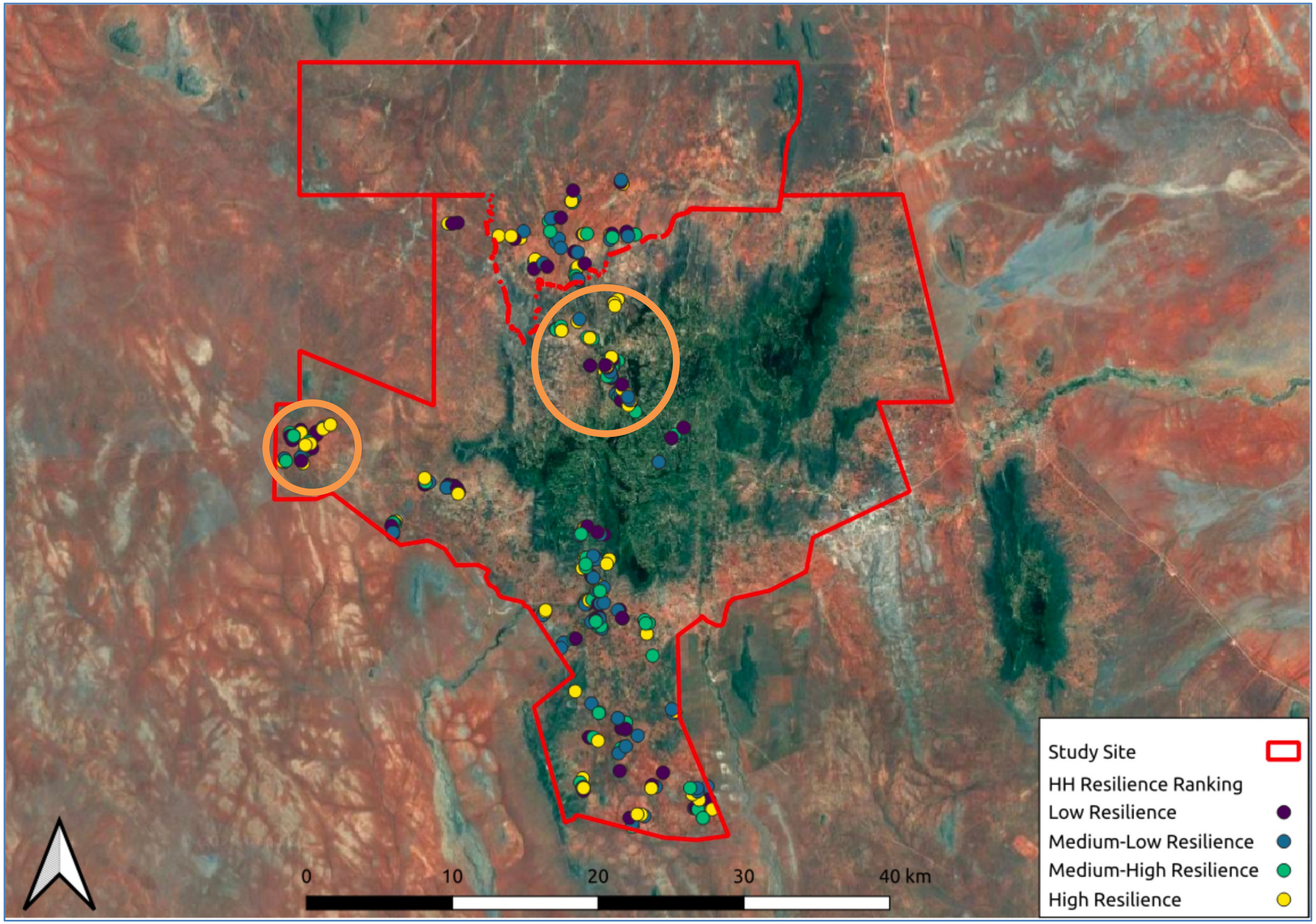
WHAT IS CLUSTERING REGRESSION



Different clusters differ:
- What is “food”?
- Do we farm or herd?
- How do we do that?
- Is water held in common?
- Is there a community safety net?

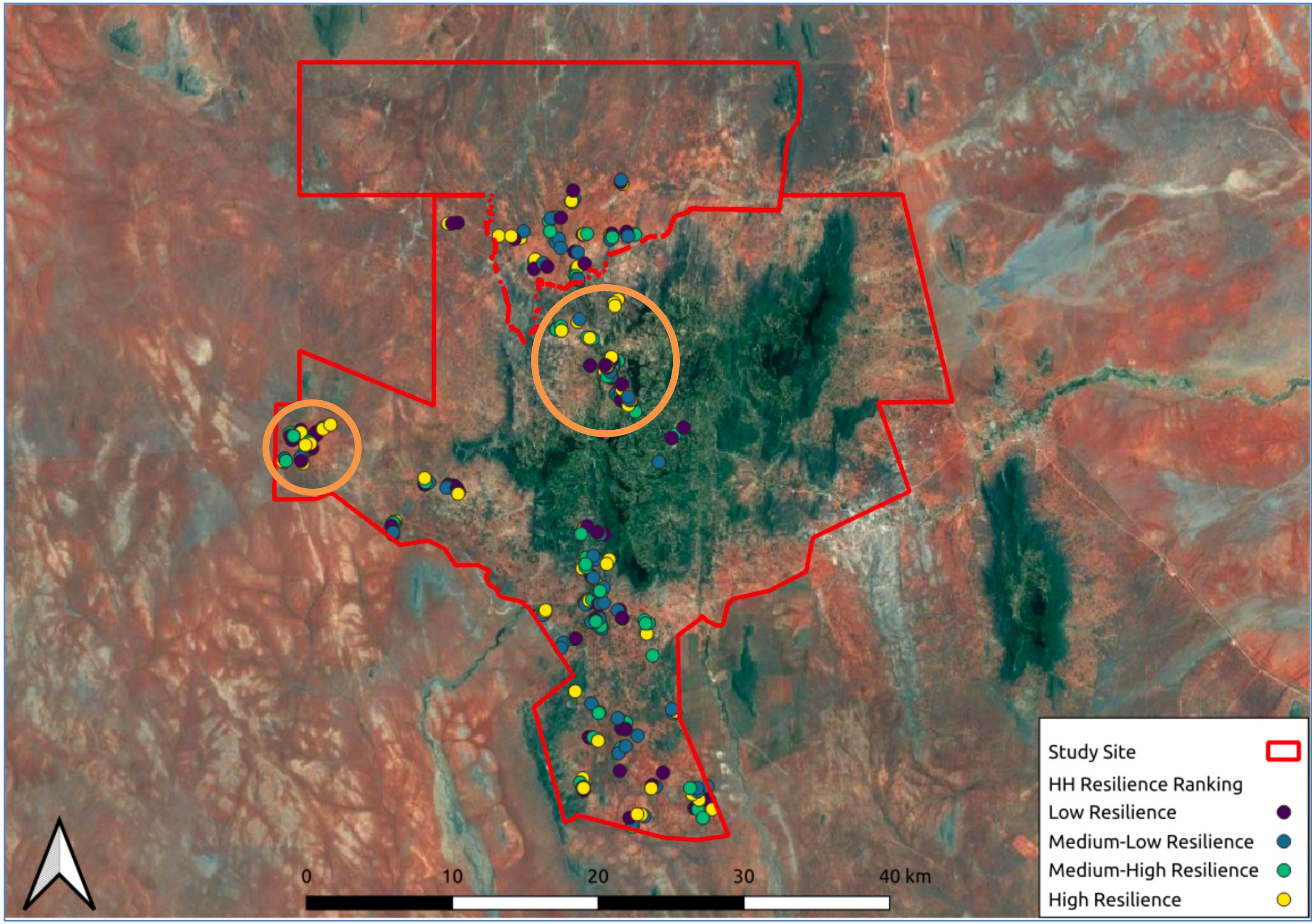
Nearby households are more similar …

WHAT IS CLUSTERING REGRESSION



Nearby households are more similar …

WHAT IS CLUSTERING REGRESSION

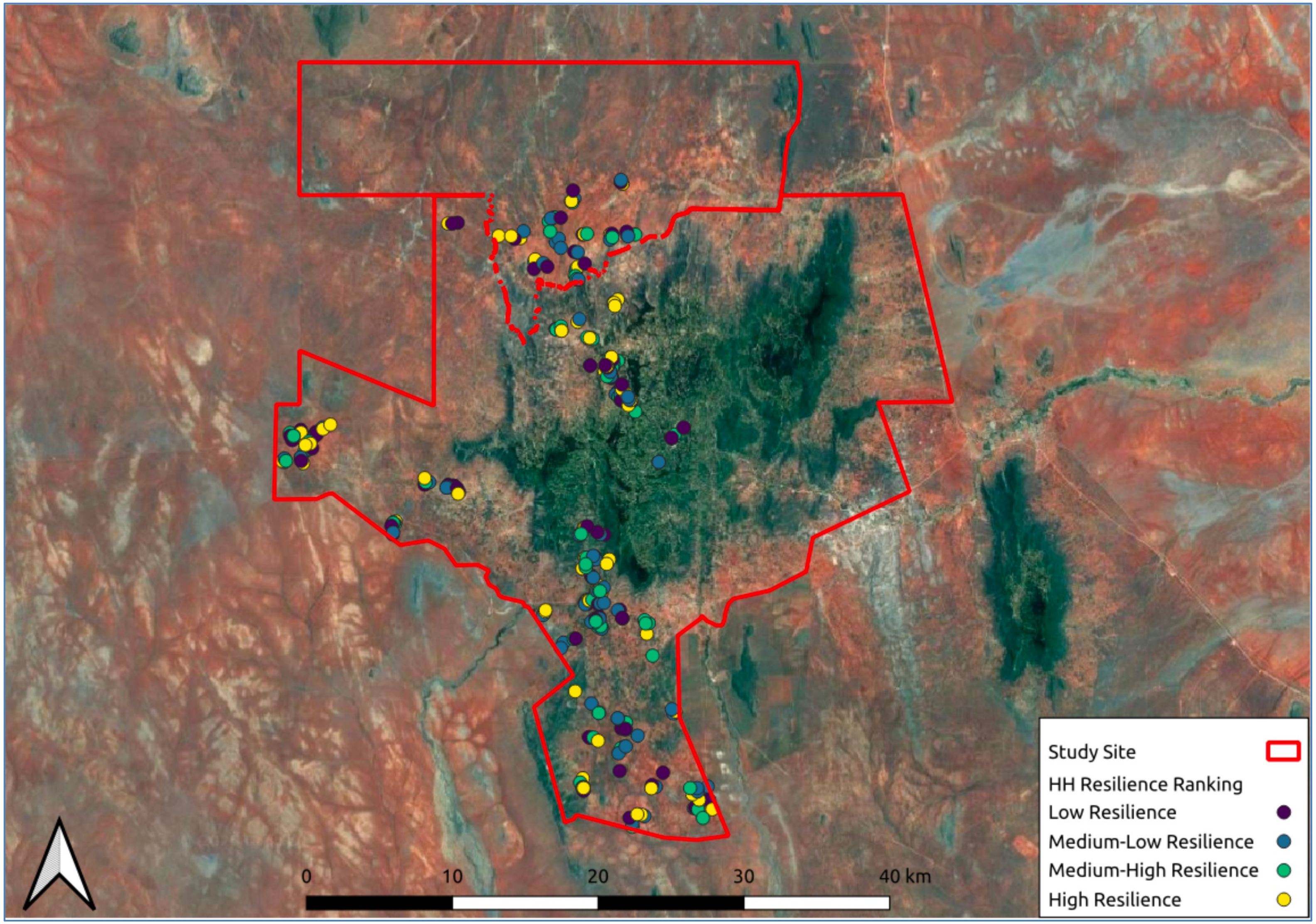


Often, separation is poor!

Nearby households are more similar …

Geographically Weighted Regression

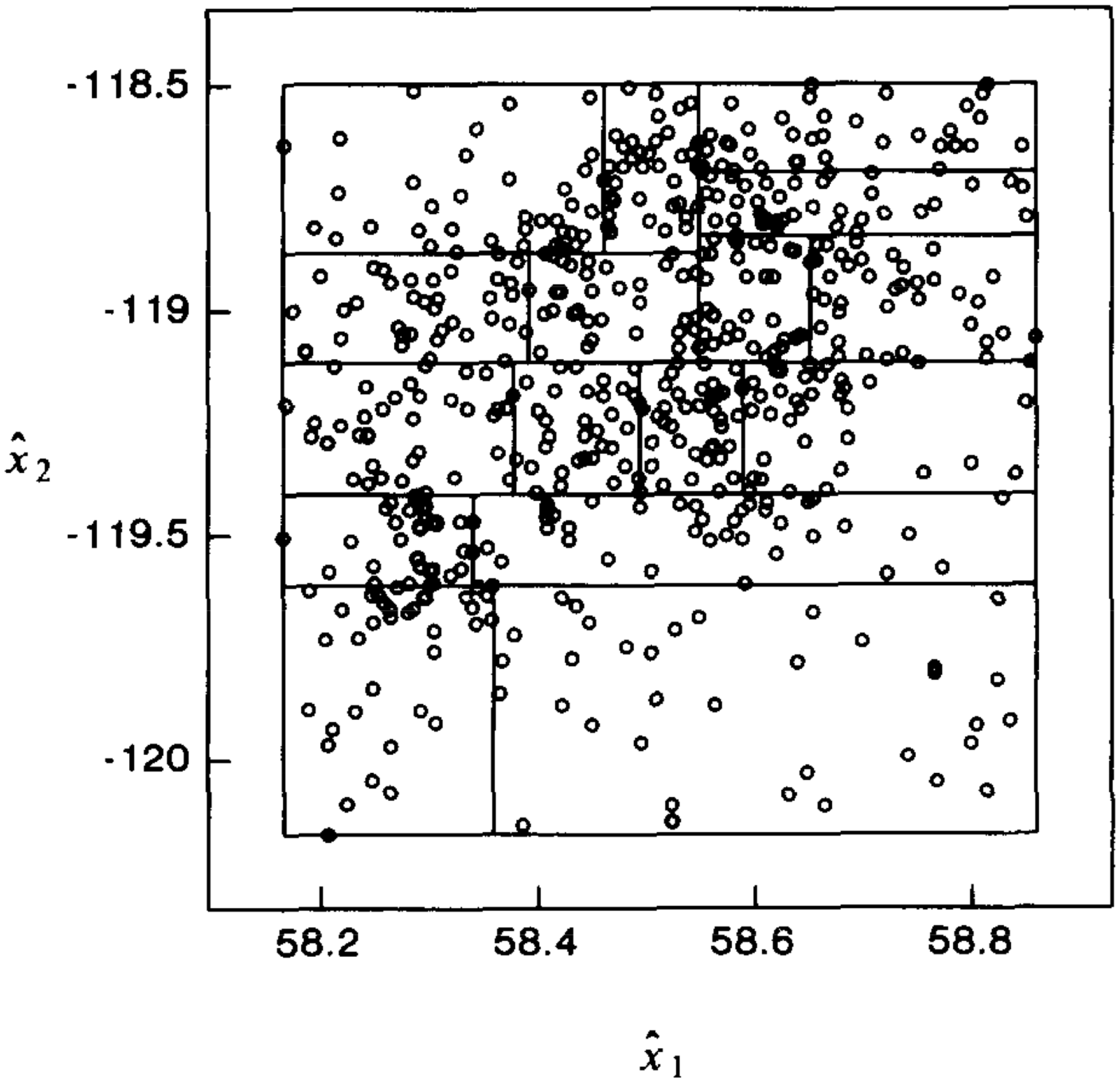
WHAT IS CLUSTERING REGRESSION



Instead, let's find who's similar in y~X and group'em!

Nearby households are more similar …

WHAT IS CLUSTERING REGRESSION



Instead, let's find who's similar in y~X and group'em!

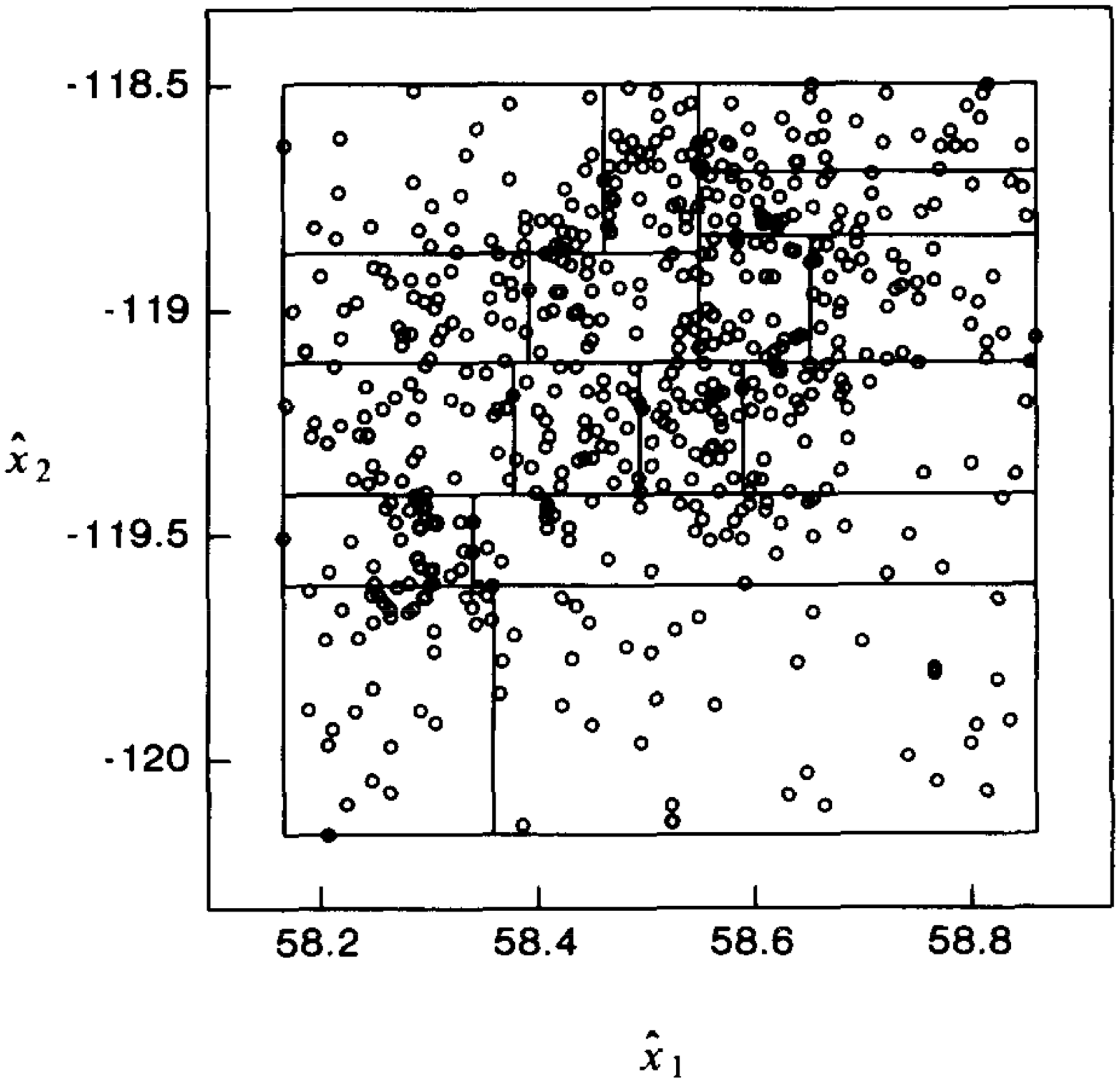
REGRESSION BY LOCAL FITTING
Methods, Properties, and Computational Algorithms

William S. CLEVELAND
AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

Susan J. DEVLIN
Bell Communications Research, Morristown, NJ 07960, USA

Eric GROSSE
AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

WHAT IS CLUSTERING REGRESSION



LOWESS Idea:
Fit a spatial tree on feature data
Estimate separately for each leaf
Smooth between nearby leaves

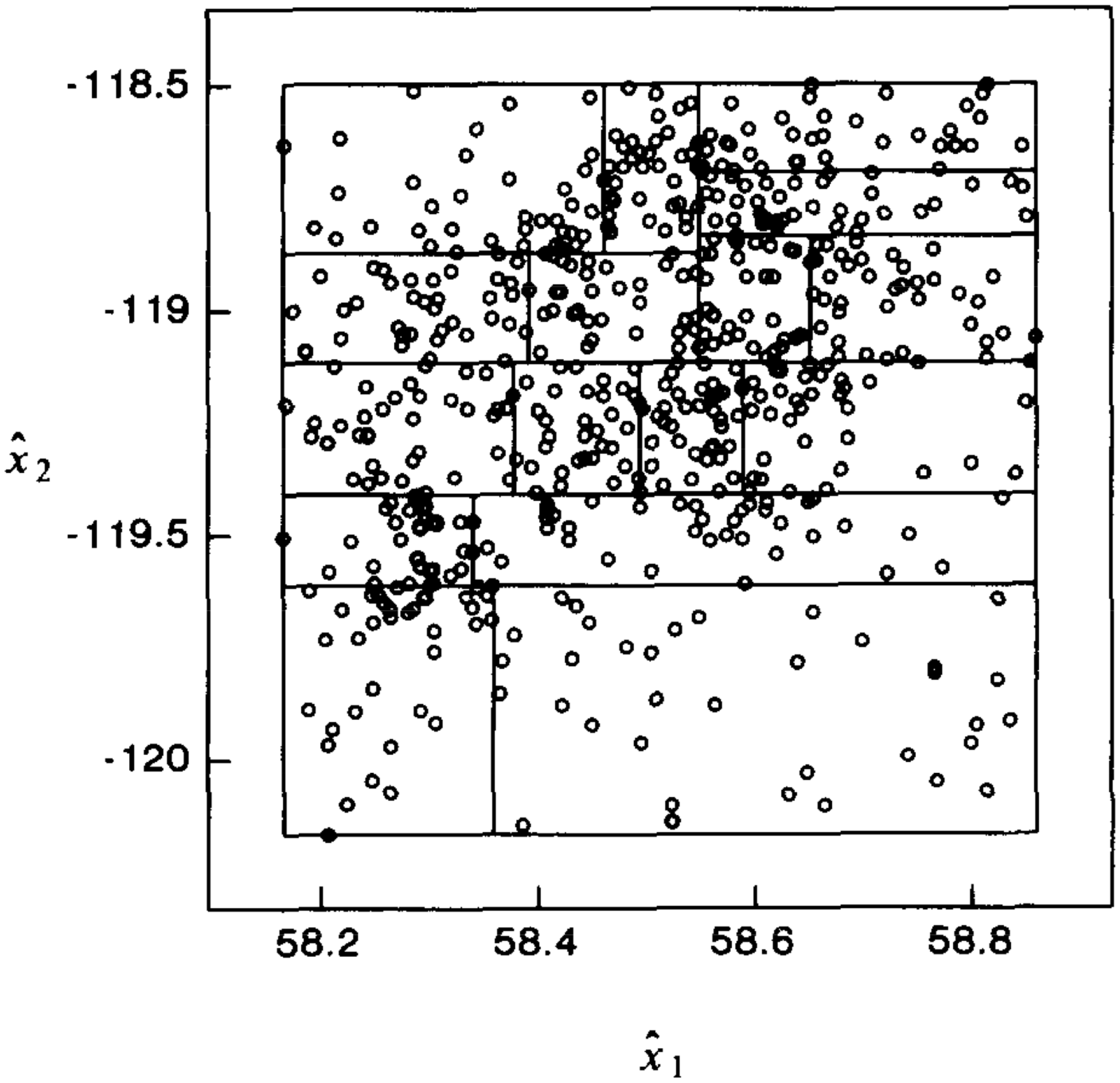
REGRESSION BY LOCAL FITTING
Methods, Properties, and Computational Algorithms

William S. CLEVELAND
AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

Susan J. DEVLIN
Bell Communications Research, Morristown, NJ 07960, USA

Eric GROSSE
AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

WHAT IS CLUSTERING REGRESSION



LOWESS Idea:

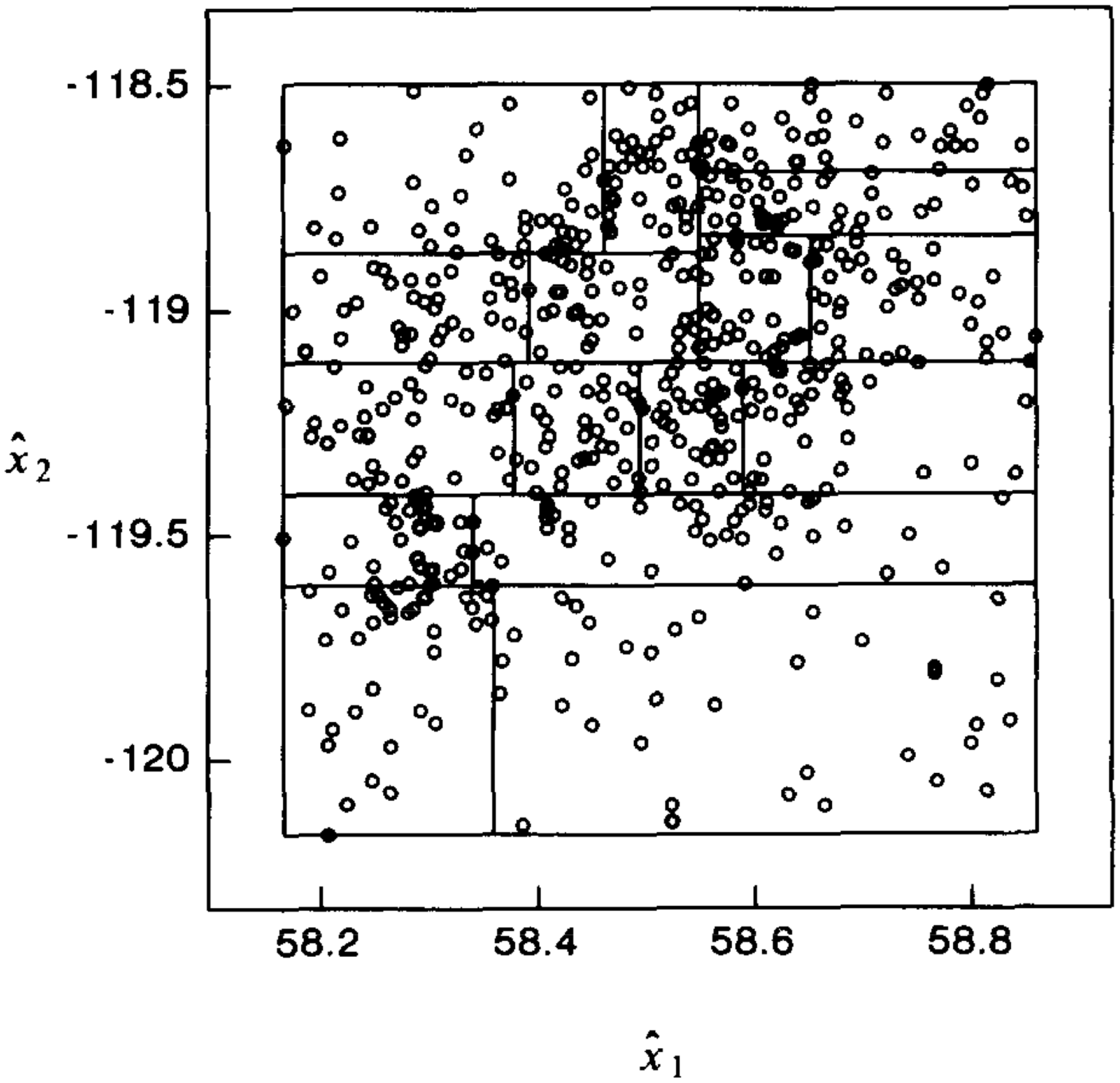
Fit a spatial tree on **feature** data

Estimate separately for each leaf

Smooth between nearby leaves

"Cluster" on X alone

WHAT IS CLUSTERING REGRESSION



LOWESS Idea:

Fit a **spatial tree** on **feature** data

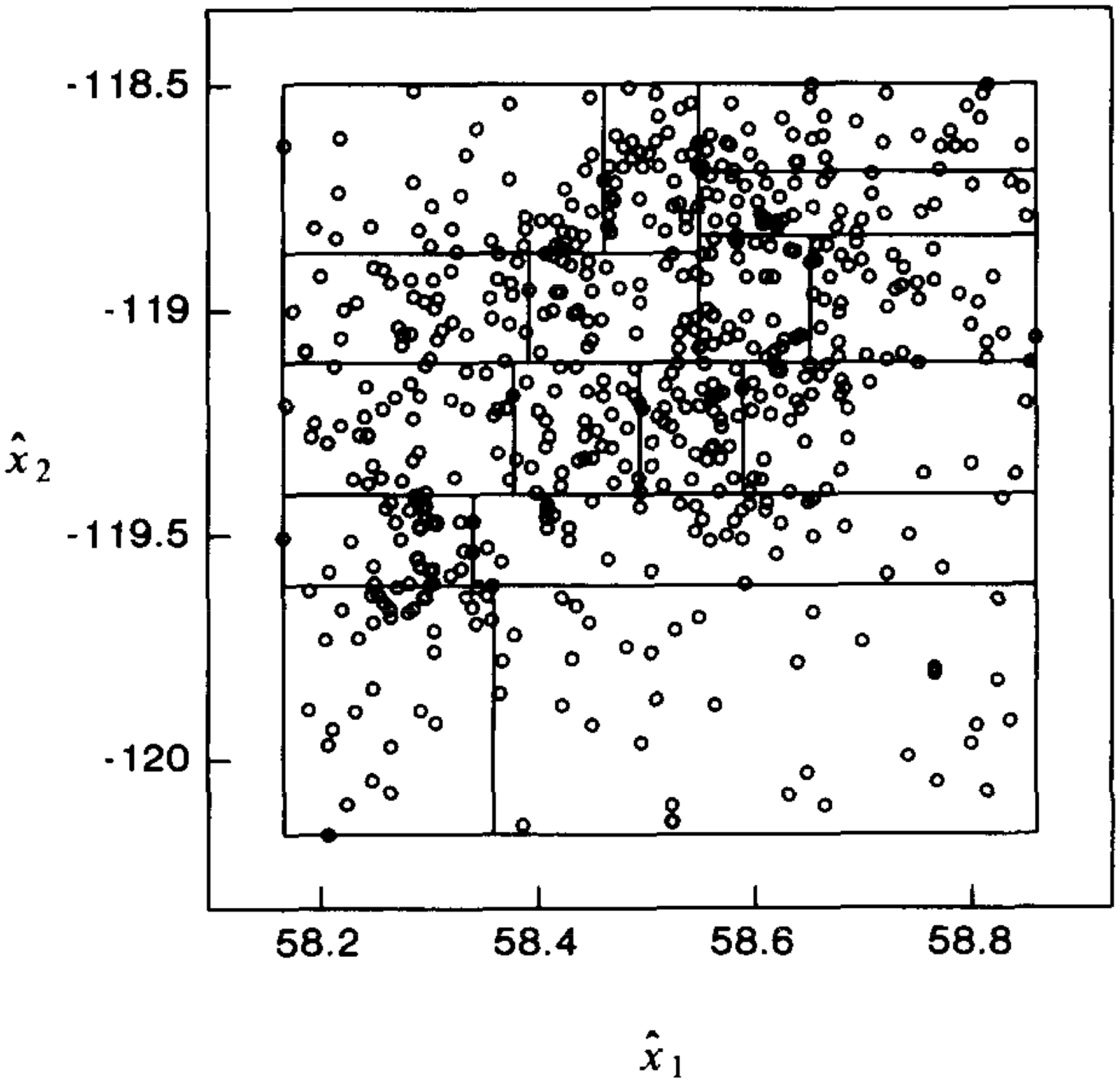
Estimate separately for each leaf

Smooth between nearby leaves

"Cluster" on X alone

Tree uses location, not regression loss

WHAT IS CLUSTERING REGRESSION



LOWESS Idea:
Fit a **spatial tree** on **feature** data
Estimate separately for each leaf
Smooth between nearby leaves

"Cluster" on X alone

Tree uses location, not regression loss

Can we do better?

WHAT IS CLUSTERING REGRESSION

CLUSTERING REGRESSION

jointly solving clustering and regression

(GEO)CLUSTERING REGRESSION

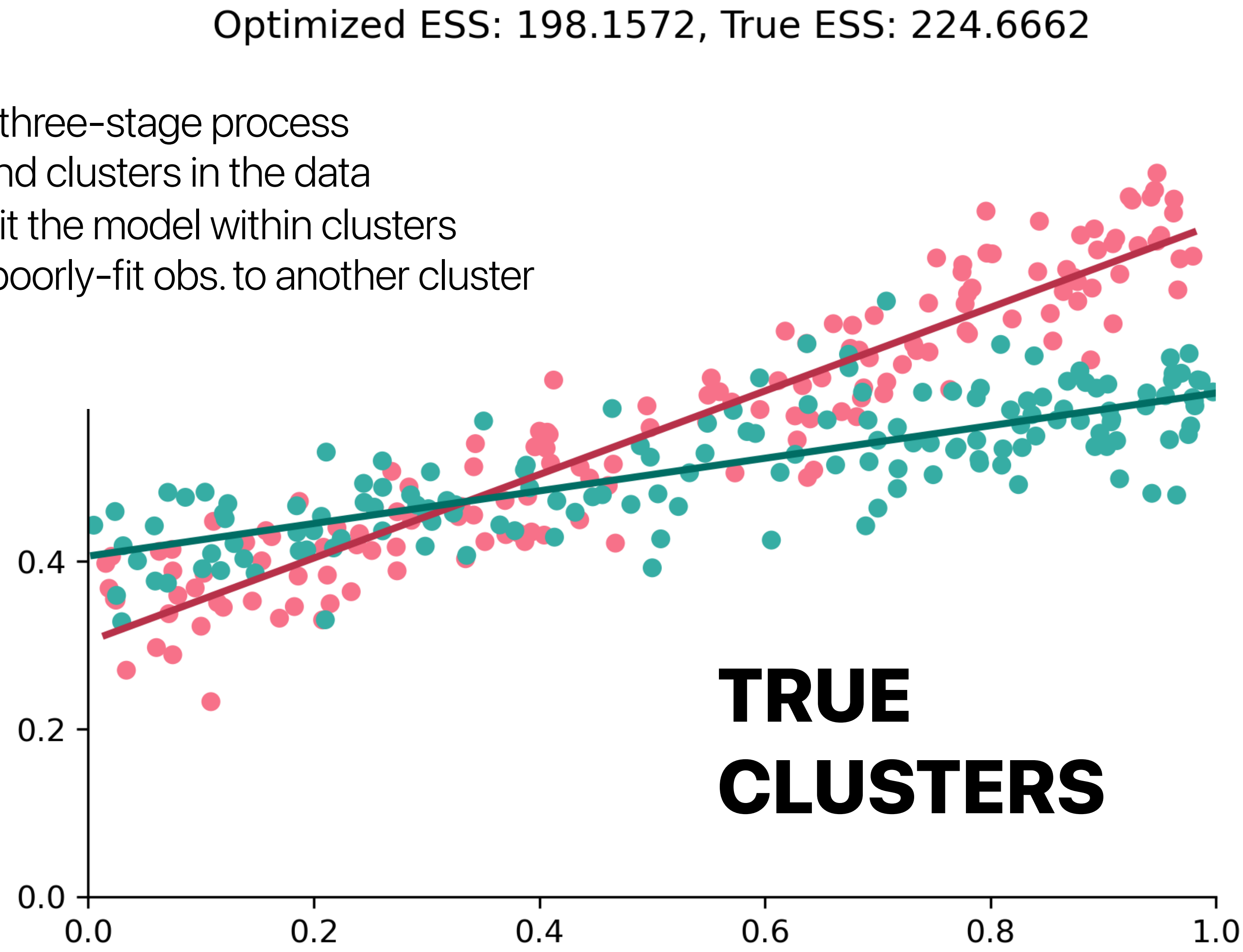
long-studied, quite recently revived

UNDERSTANDING QUADTREES

APPLYING QUADTREE REGRESSION

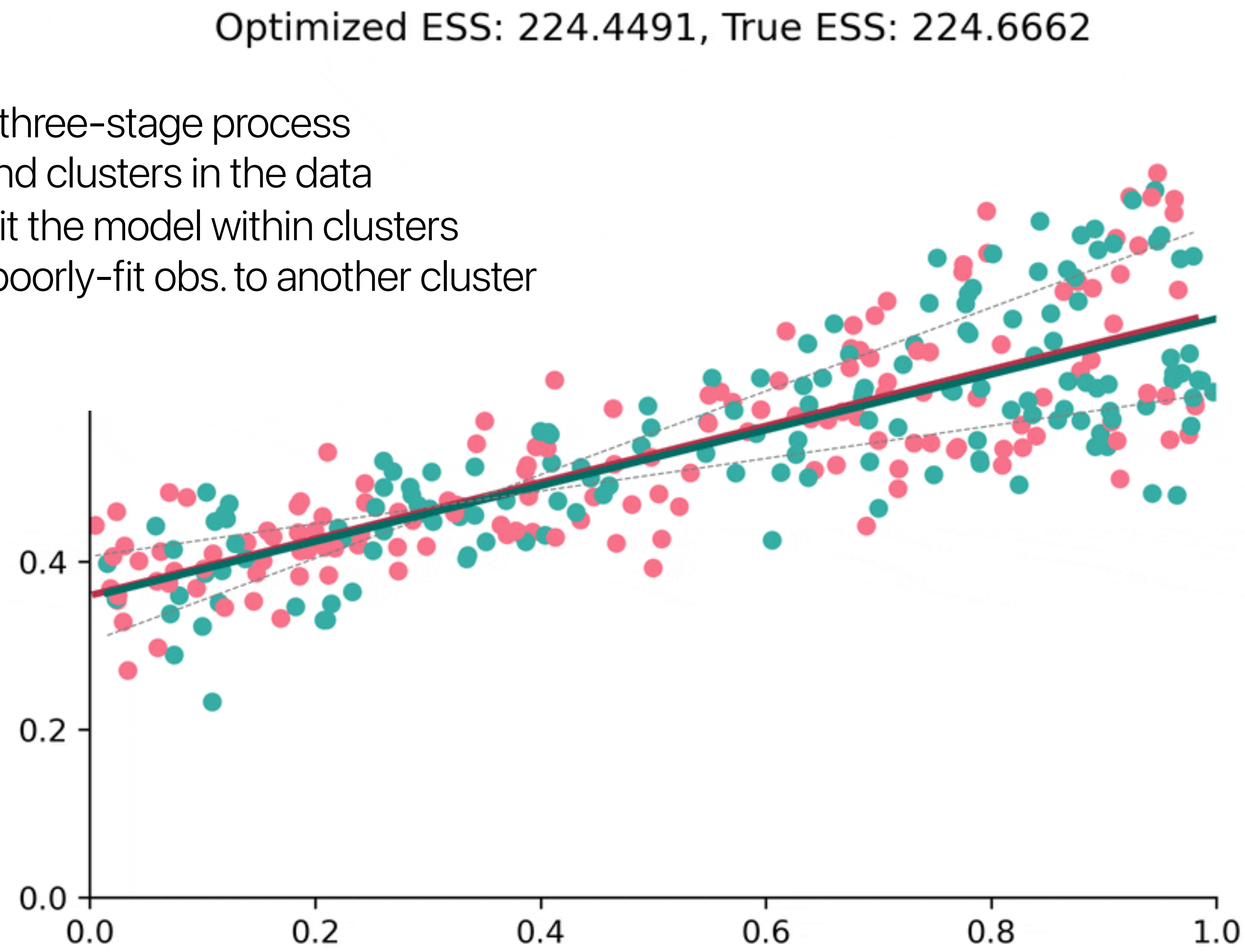
THE SUPERVISED QUADTREE

Späth (1979)'s three-stage process
- **CLUSTER:** find clusters in the data
- **ESTIMATE:** fit the model within clusters
- **REFINE:** flip poorly-fit obs. to another cluster

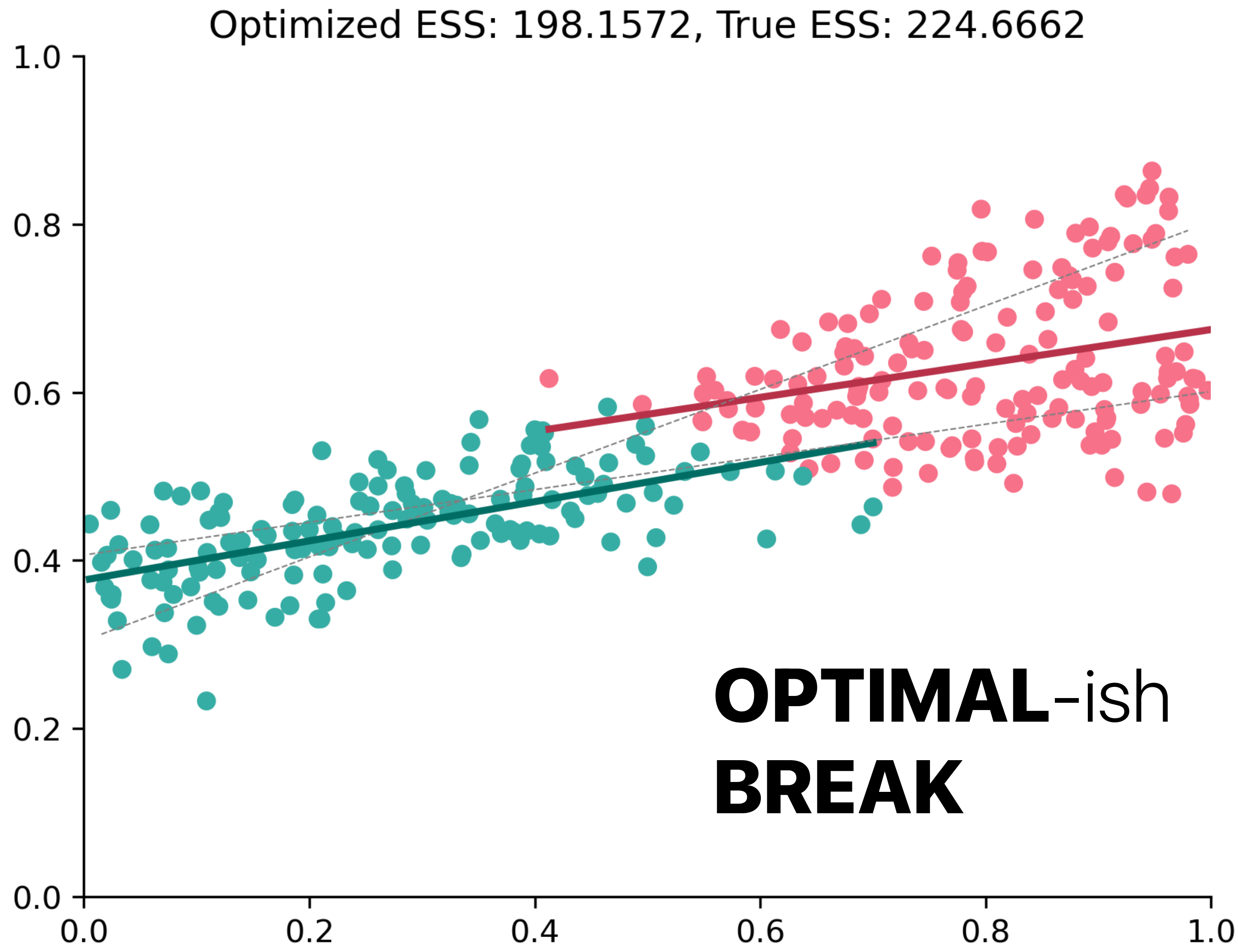


SPÄTH (1979) ITERATIVE CLUSTERING

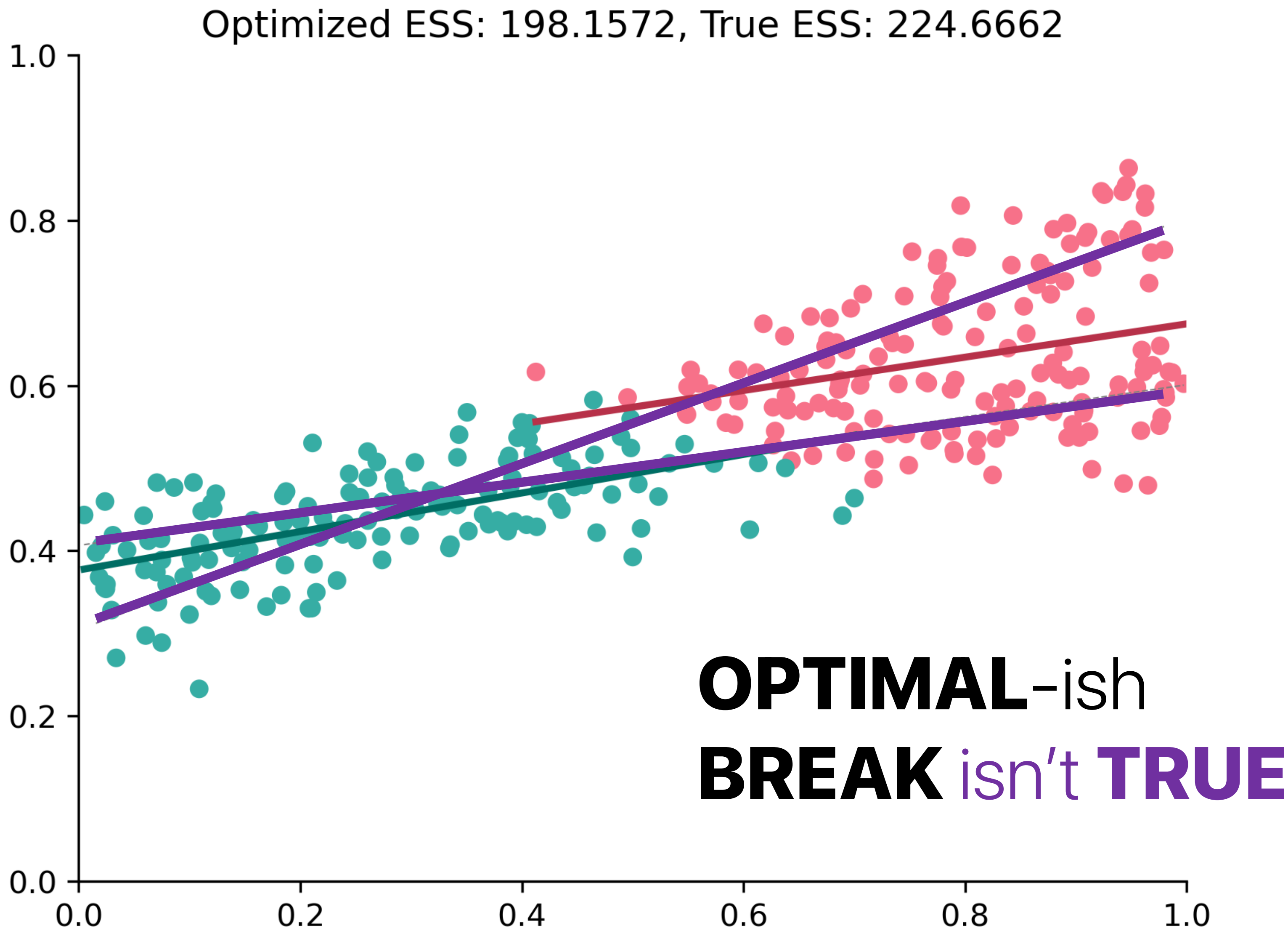
Späth (1979)'s three-stage process
- **CLUSTER:** find clusters in the data
- **ESTIMATE:** fit the model within clusters
- **REFINE:** flip poorly-fit obs. to another cluster



SPÄTH (1979) ITERATIVE CLUSTERING



SPÄTH (1979) ITERATIVE CLUSTERING

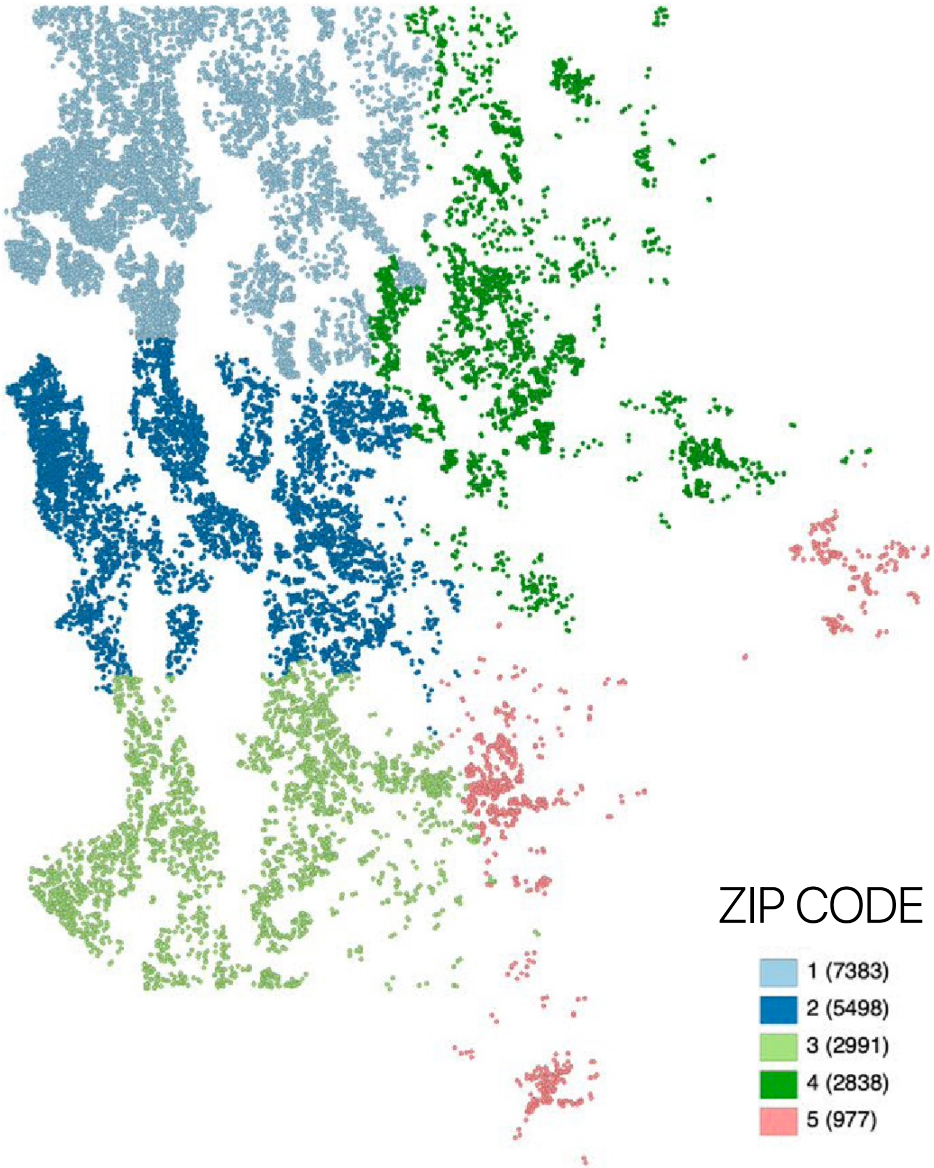


SPÄTH (1979) ITERATIVE CLUSTERING

Späth (1979)'s three-stage process
- **CLUSTER:** find clusters in the data
- **ESTIMATE:** fit the model using clusters as a feature
- **REFINE:** flip poorly-fit observations to a different cluster

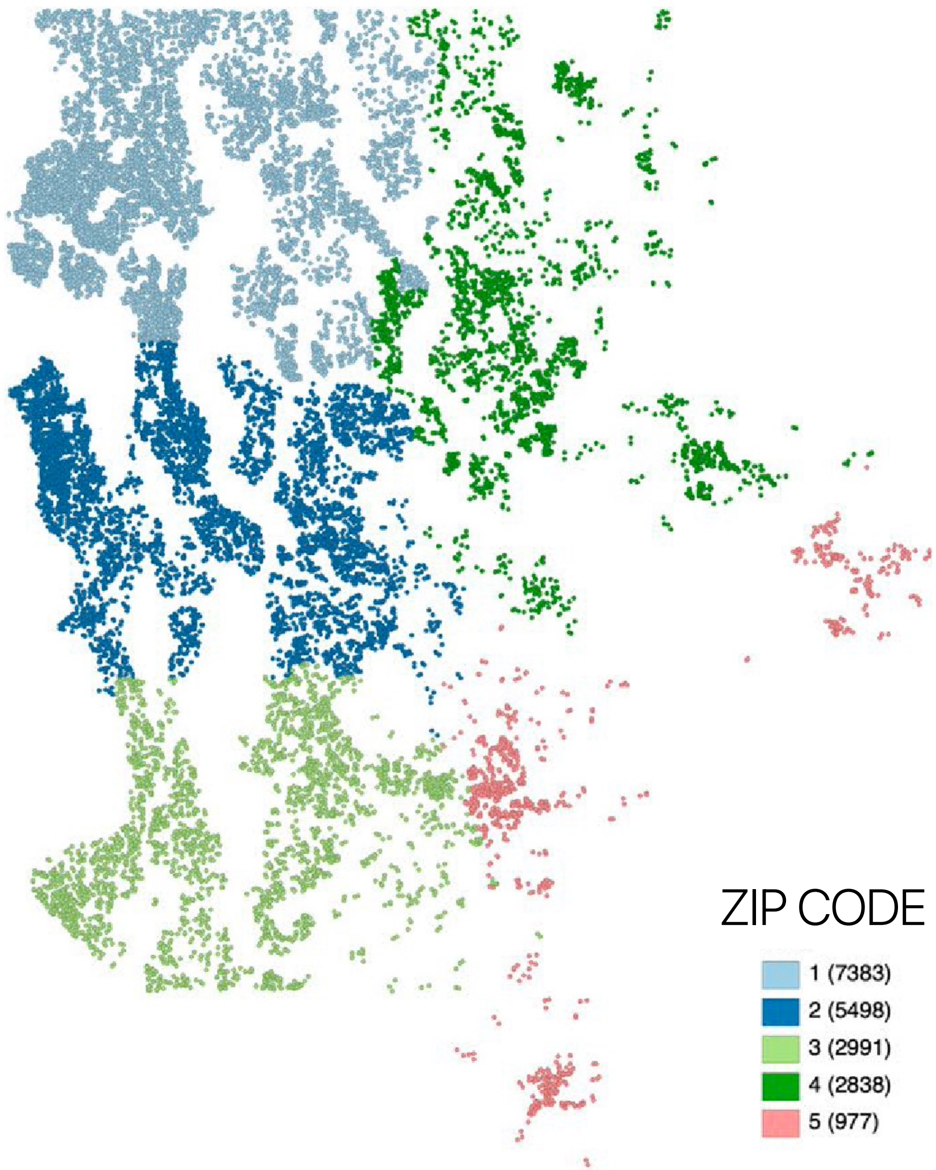
Heuristic (i.e. not optimal), slow+computationally intensive, even in the seriously improved spatial variant provided by Sugasawa and Murakami (2021)
No uncertainty for checking assignment, hence no guarantee of recovering the true classes

BEFORE THE SUPER*TREE LEARNER



Spatial SUR Models

WHAT IS CLUSTERING REGRESSION



Spatial SUR Models

Stack each group's y~X model together

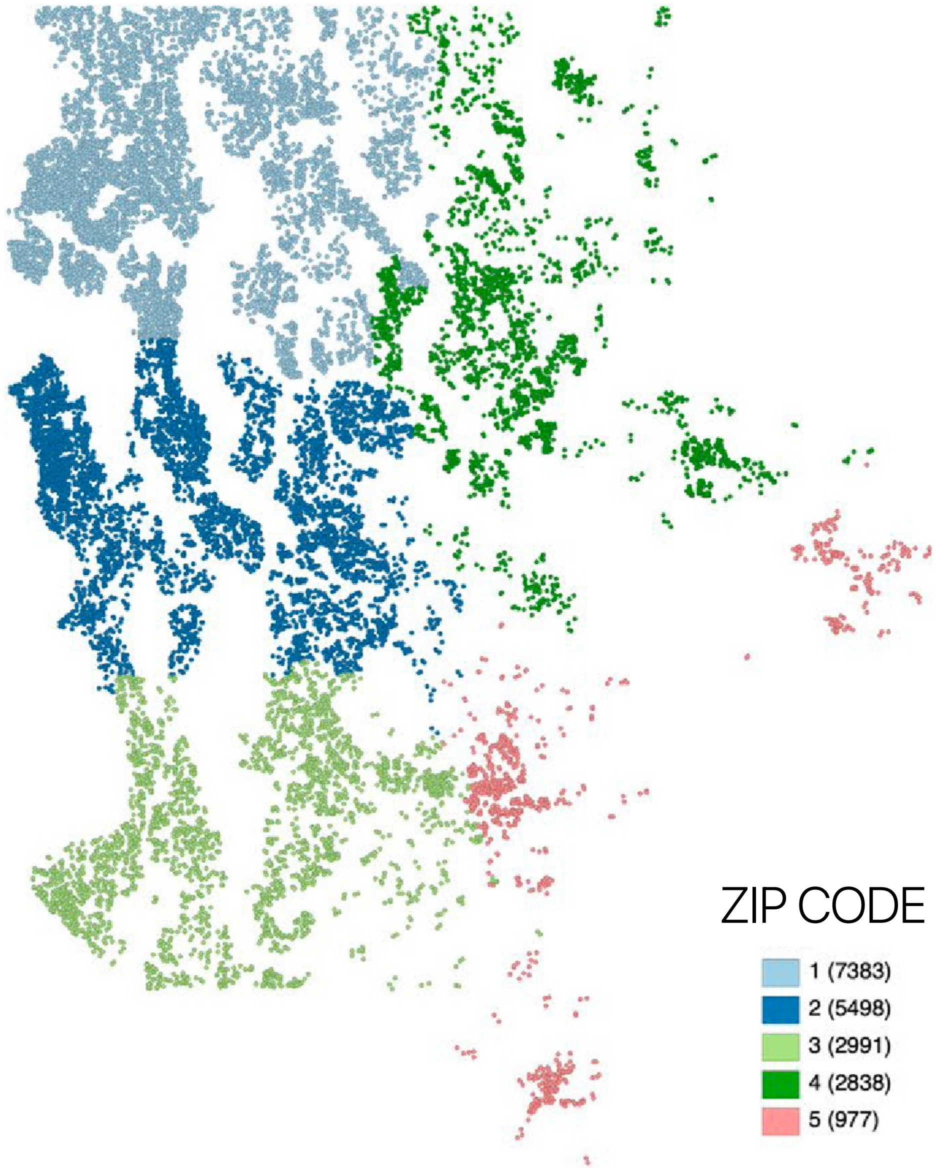
JOURNAL OF REGIONAL SCIENCE, VOL. 30, NO. 2, 1990, pp. 185–207

SPATIAL DEPENDENCE AND SPATIAL STRUCTURAL INSTABILITY IN APPLIED REGRESSION ANALYSIS*

Luc Anselin†

Test to check if separate models are needed!

WHAT IS CLUSTERING REGRESSION



Spatial SUR Models

Stack each group's y~X model together

JOURNAL OF REGIONAL SCIENCE, VOL. 30, NO. 2, 1990, pp. 185–207

SPATIAL DEPENDENCE AND SPATIAL STRUCTURAL INSTABILITY IN APPLIED REGRESSION ANALYSIS*

Luc Anselin†

Test to check if separate models are needed!

Can we integrate *that* test during fitting?

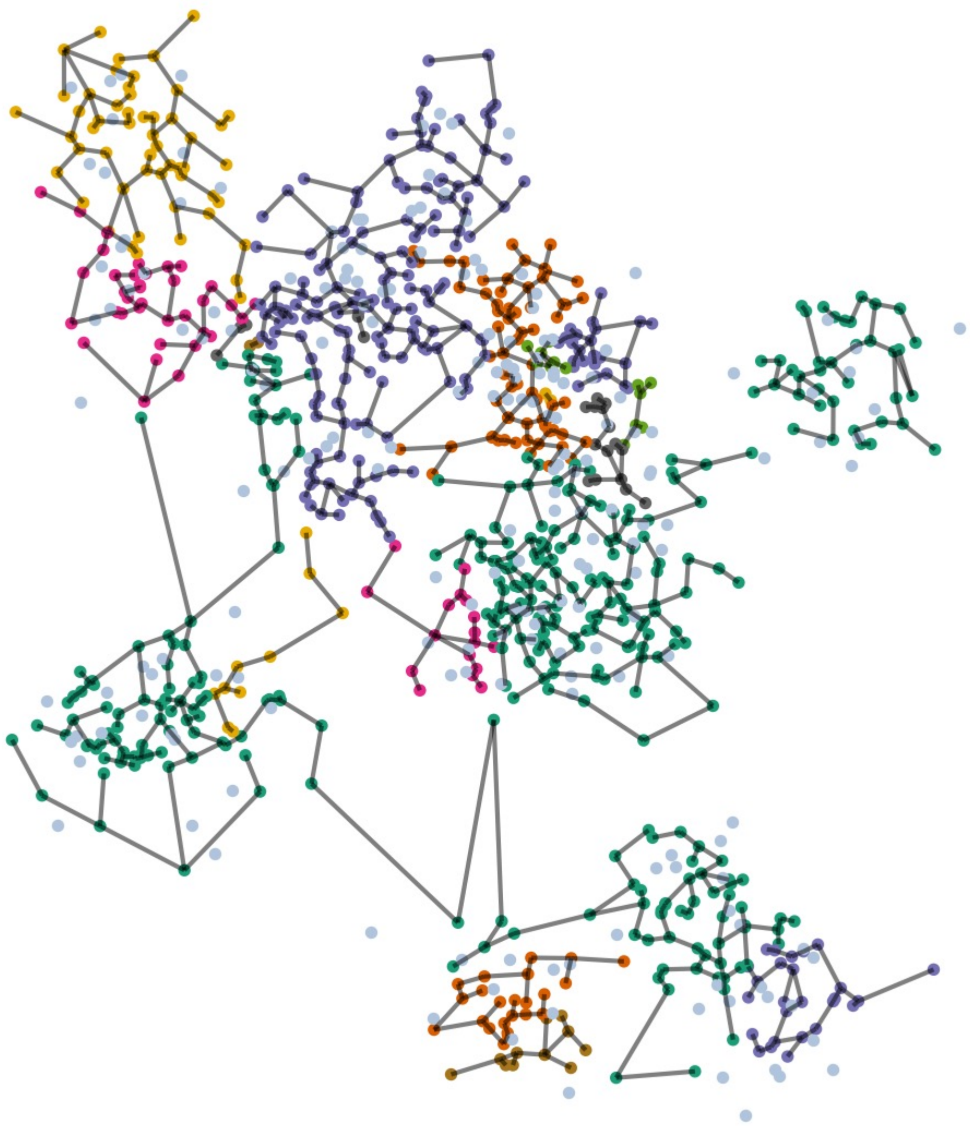
WHAT IS CLUSTERING REGRESSION

Späth (1979)'s three-stage process
- **CLUSTER:** find clusters in the data
- **ESTIMATE:** fit the model using clusters as a feature
- **REFINE:** flip poorly-fit observations to a different cluster

Heuristic (i.e. not optimal), slow+computationally intensive, likelihood-dependent, even in the seriously improved spatial variant provided by Sugasawa and Murakami (2021)
No guarantee of recovering the original classes

Anselin & Amaral (2023)
- **COMBINE:** Combine spatial and feature similarity
- **SPAN:** compute a MST over the combined affinity matrix
- **PRUNE:** remove MST links to create clusters according to Chow test on $(SSR - (SSR_a + SSR_b))$

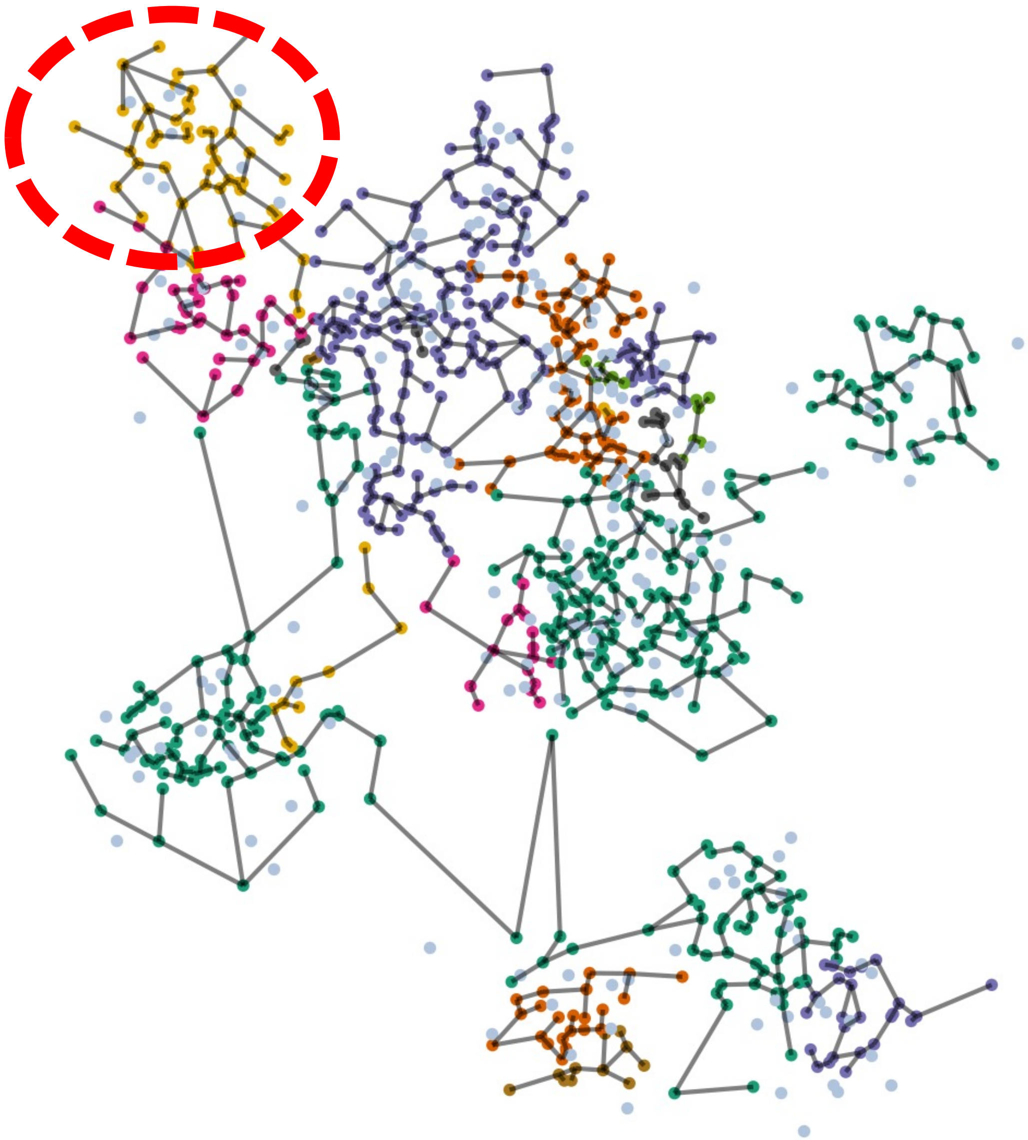
BEFORE THE SUPER*TREE



SKATER REGRESSION

each cluster gets
one regression

Clusters are subtrees
with good regressions.



SKATER REGRESSION

Späth (1979)'s three-stage process
- **CLUSTER:** find clusters in the data
- **ESTIMATE:** fit the model using clusters as a feature
- **REFINE:** flip poorly-fit observations to a different cluster

Heuristic (i.e. not optimal), slow+computationally intensive, likelihood-dependent, even in the seriously improved spatial variant provided by Sugasawa and Murakami (2021)
No guarantee of recovering the original classes

Anselin & Amaral (2023)
- **COMBINE:** Combine spatial and feature similarity
- **SPAN:** compute a MST over the combined affinity matrix
- **PRUNE:** remove MST links to create clusters according to Chow test on $(SSR - (SSR_a + SSR_b))$

BEFORE THE SUPERVISED QUADTREE

Späth (1979)'s three-stage process
- **CLUSTER:** find clusters in the data
- **ESTIMATE:** fit the model using clusters as a feature
- **REFINE:** flip poorly-fit observations to a different cluster

Heuristic (i.e. not optimal), slow+computationally intensive, likelihood-dependent, even in the seriously improved spatial variant provided by Sugasawa and Murakami (2021)
No guarantee of recovering the original classes

SKATER Additive Learner (STAKERAL): A Data-driven Method for Delineating Covariate-specific Endogenous Spatial Regimes

Sui Zhang[*1], Wei Zheng[†1] and Cecilia Wong[‡1]

[1]Spatial Policy and Analysis Lab, Department of Planning, Property and Environmental Management, the University of Manchester

GISRUK 2025

BEFORE THE SUPERVISED QUADTREE

Späth (1979)'s three-stage process
- **CLUSTER:** find clusters in the data
- **ESTIMATE:** fit the model using clusters as a feature
- **REFINE:** flip poorly-fit observations to a different cluster

Heuristic (i.e. not optimal), slow+computationally intensive, likelihood-dependent, even in the seriously improved spatial variant provided by Sugasawa and Murakami (2021)
No guarantee of recovering the original classes

SKATER Additive Learner (STAKERAL): A Data-driven Method for Delineating Covariate-specific Endogenous Spatial Regimes

Sui Zhang*[1], Wei Zheng†[1] and Cecilia Wong‡[1]

[1]Spatial Policy and Analysis Lab, Department of Planning, Property and Environmental Management, the University of Manchester

GISRUK 2025

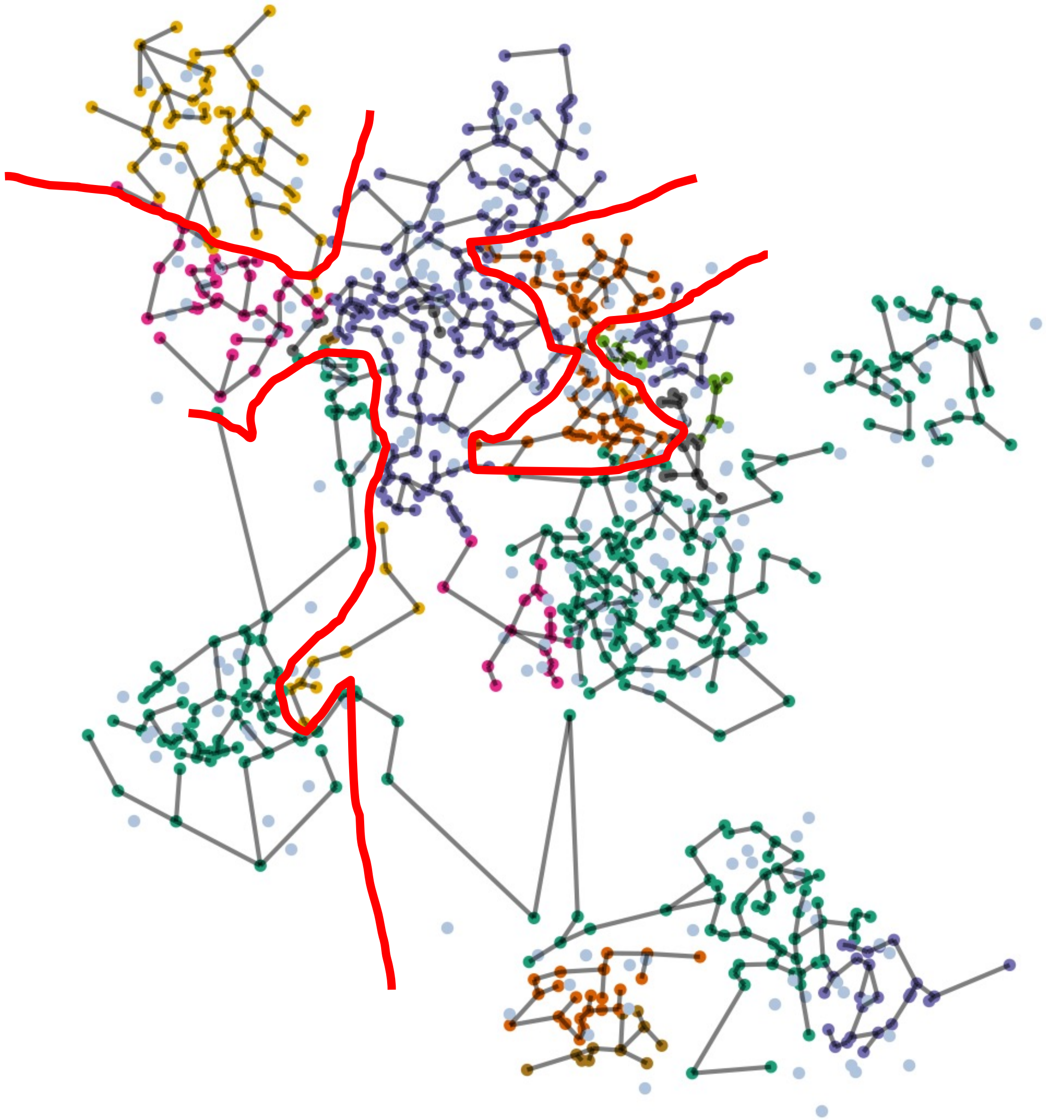
BEFORE THE SUPERVISED QUADTREE

each cluster gets
one ~~regression~~
dummy term

stack these into a
dummy matrix, $\mathbf{D}$

Regression becomes
$y \sim \mathbf{D} + \mathbf{DX}$

clusters are regions
where interaction term
improves model fit

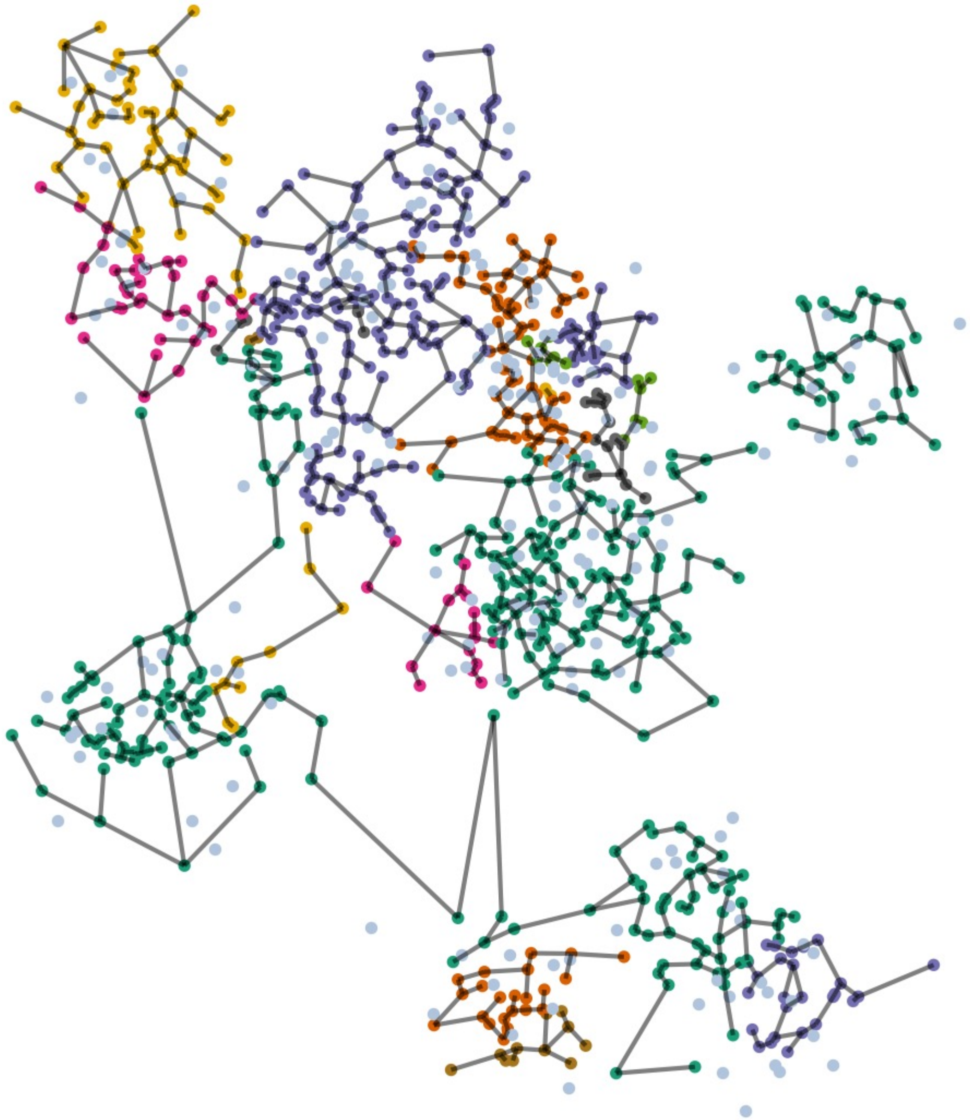


SKATER REGRESSION

Strengths:

Judicious use of degrees of freedom

Uses LRT to find good clusters

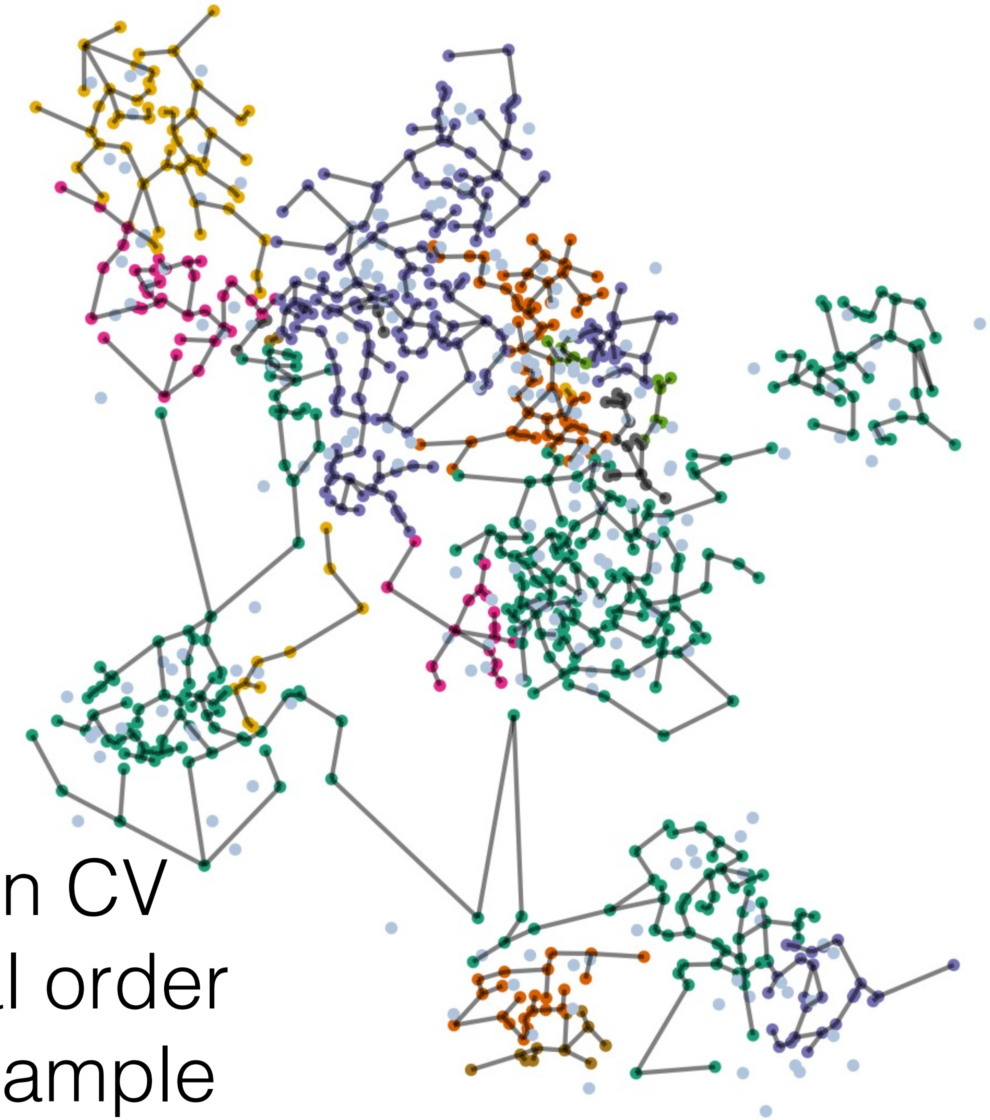


SKATER REGRESSION IN PRINCIPLE

Strengths:

Judicious use of degrees of freedom

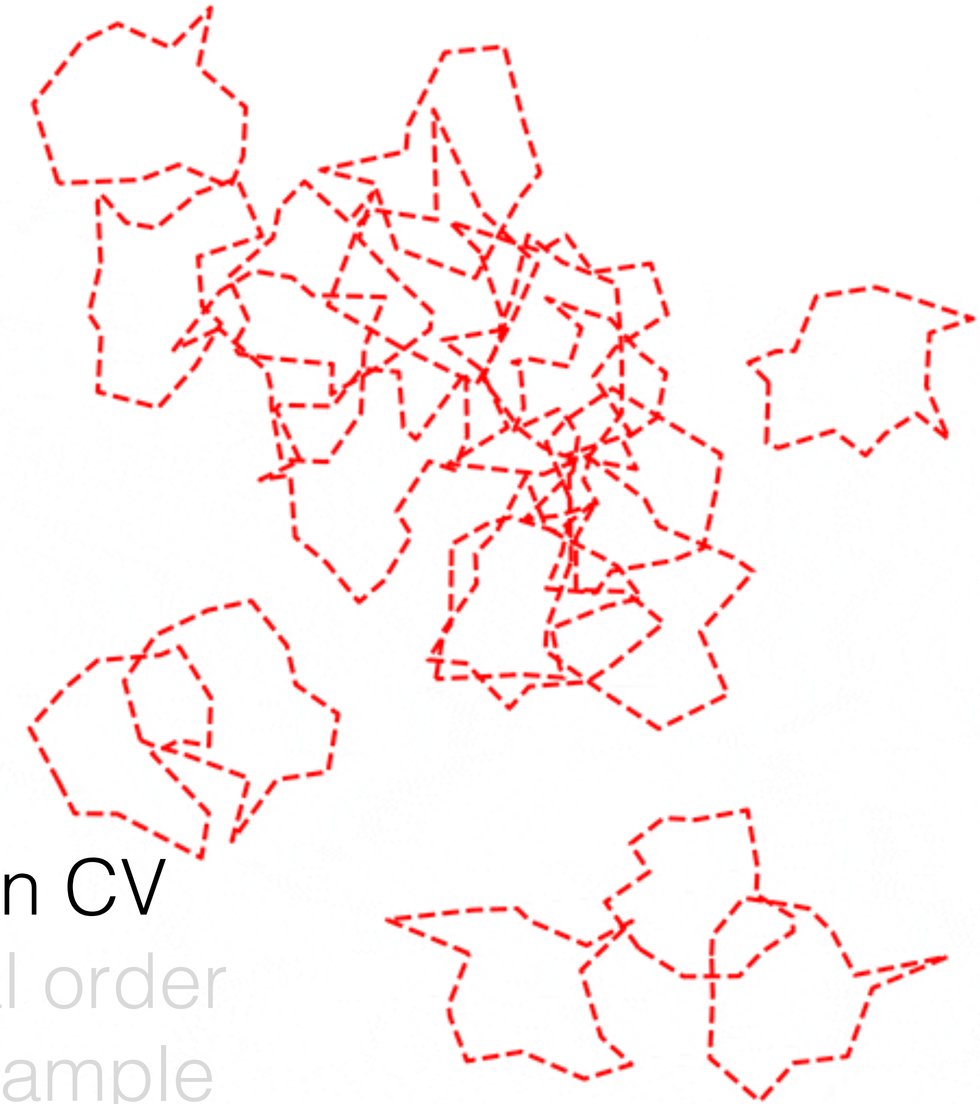
Uses LRT to find good clusters



Weaknesses:

- MST is finnicky in CV
- MST has no total order
- can't do out of sample

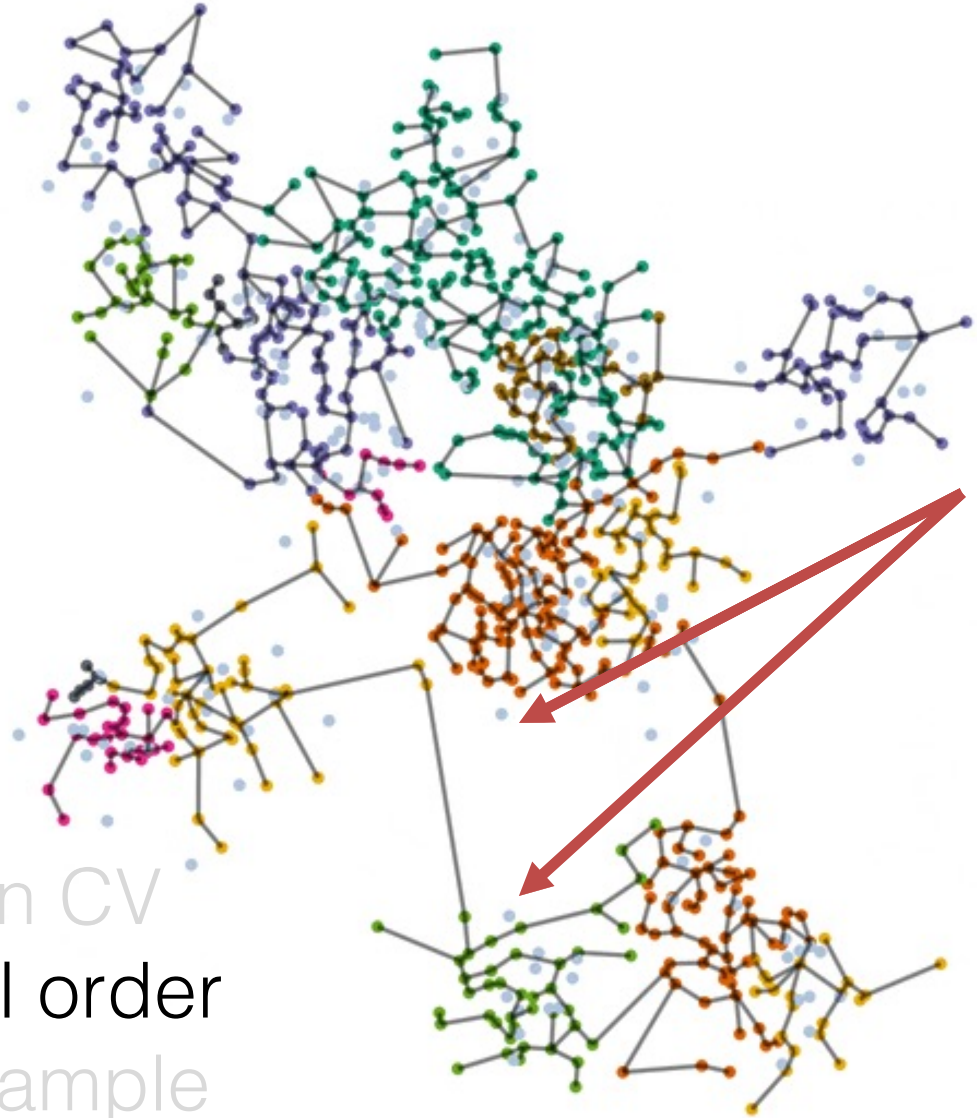
SKATER REGRESSION IN PRINCIPLE



Weaknesses:

- MST is finnicky in CV
- MST has no total order
- can't do out of sample

MSTs ARE VERY SENSITIVE OBJECTS

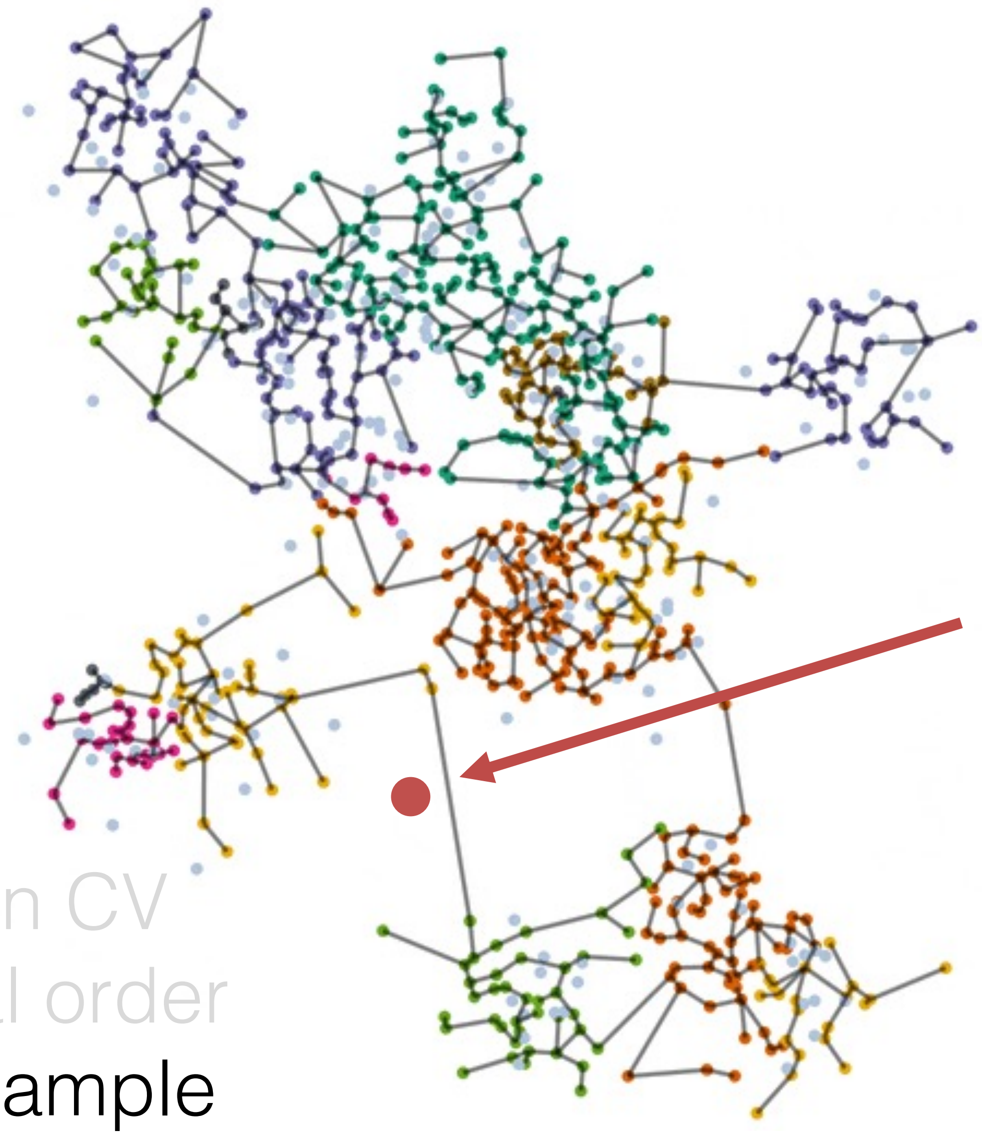


Very distant in some splits, neighbors in others!

Weaknesses:

- MST is finnicky in CV
- MST has no total order
- can't do out of sample

TREE ADJACENCY IS NOT SIMILARITY



How should this connect?

Nearest leaf?
Nearest node?
Re-tree everybody?

Weaknesses:

- MST is finnicky in CV
- MST has no total order
- can’t do out of sample

NEW POINTS BREAK MST STRUCTURE

CLUSTERING REGRESSION

jointly solving clustering and regression

GEOGRAPHICAL CLUSTER-REG

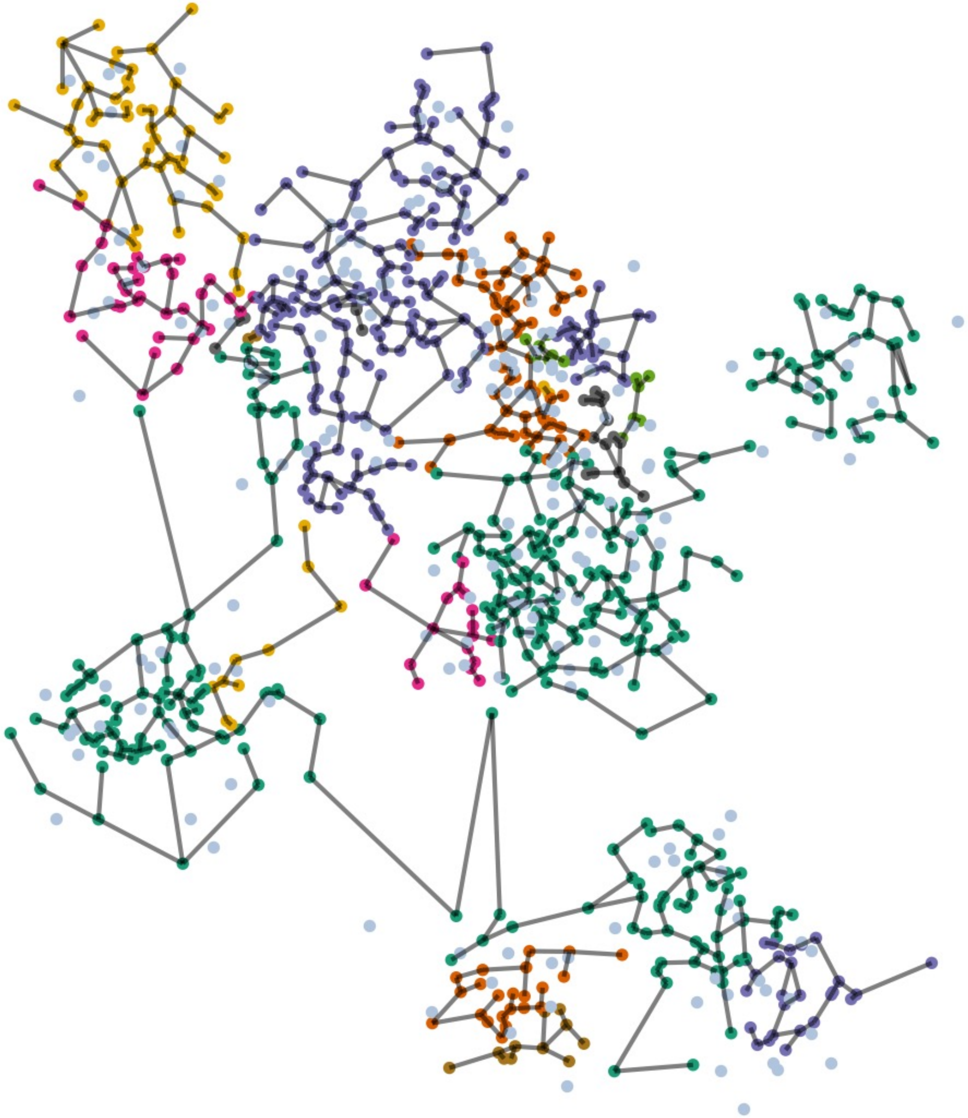
jointly solving clustering and regression

UNDERSTANDING SUPER*TREES

spatial splits for spatial fits

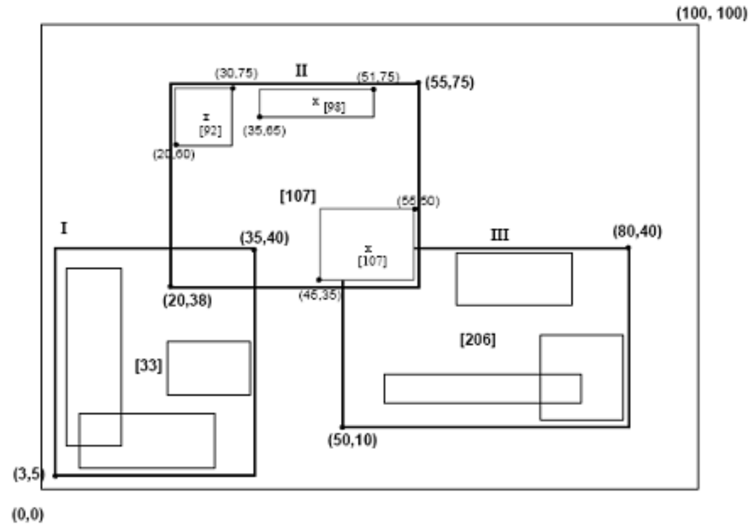
APPLYING QUADTREE REGRESSION

Build a Minimum Spanning Tree
Split tree when it improves predictions
Aggregate subtrees to make predictions

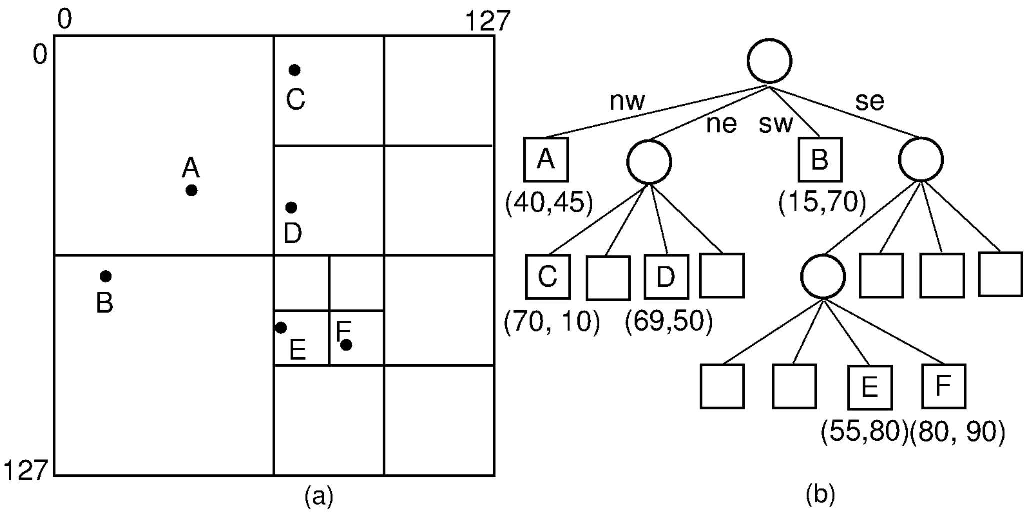


BEFORE THE SUPER*TREE

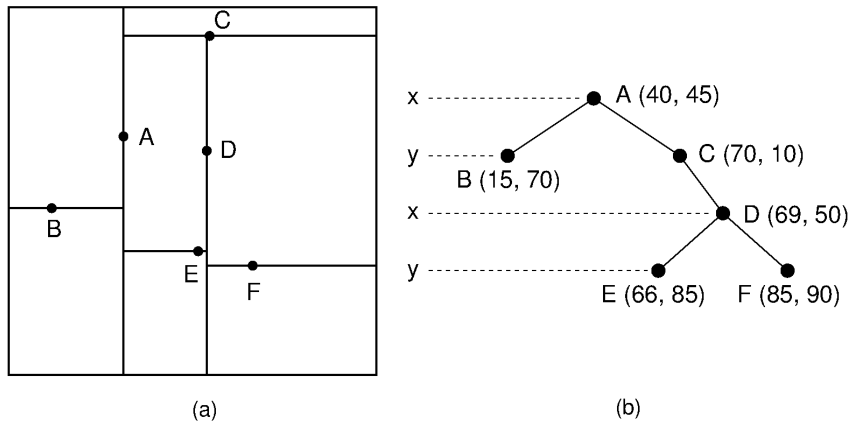
Build a ~~Minimum Spanning~~ Tree
Split tree when it improves predictions
Aggregate subtrees to make predictions



Hilbert R-Tree



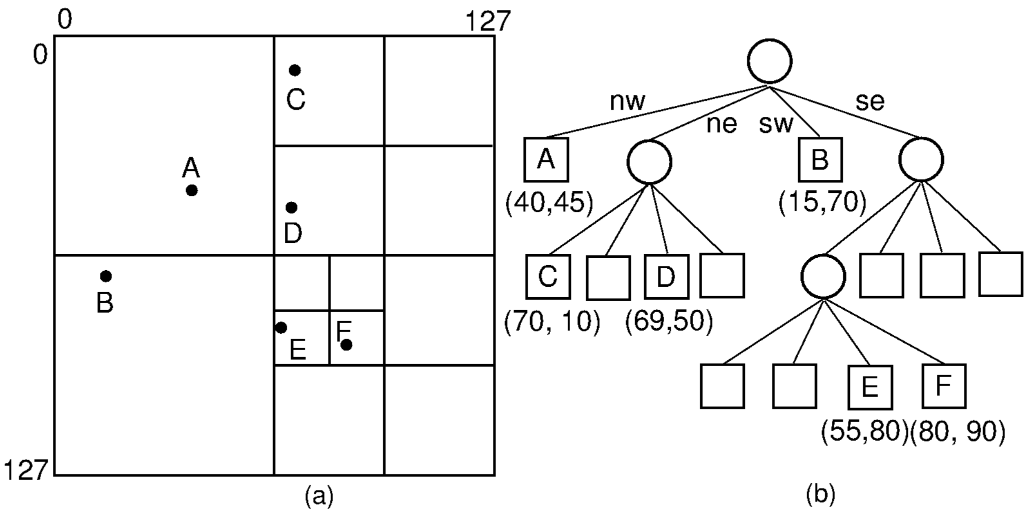
QuadTree



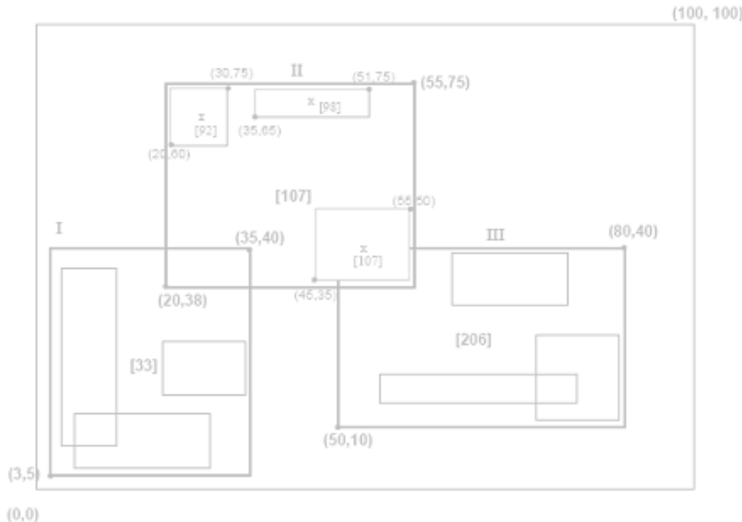
kd-Tree

THE SUPER*TREE

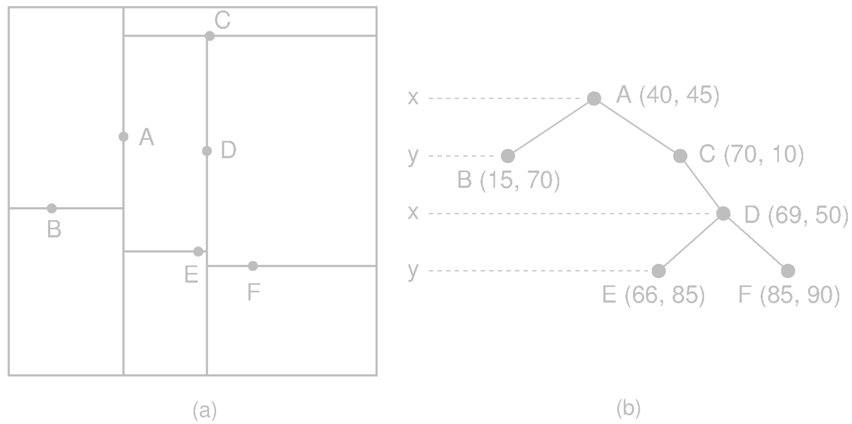
Build a **Quad**Tree
Split tree when it improves predictions
Aggregate subtrees to make predictions



QuadTree



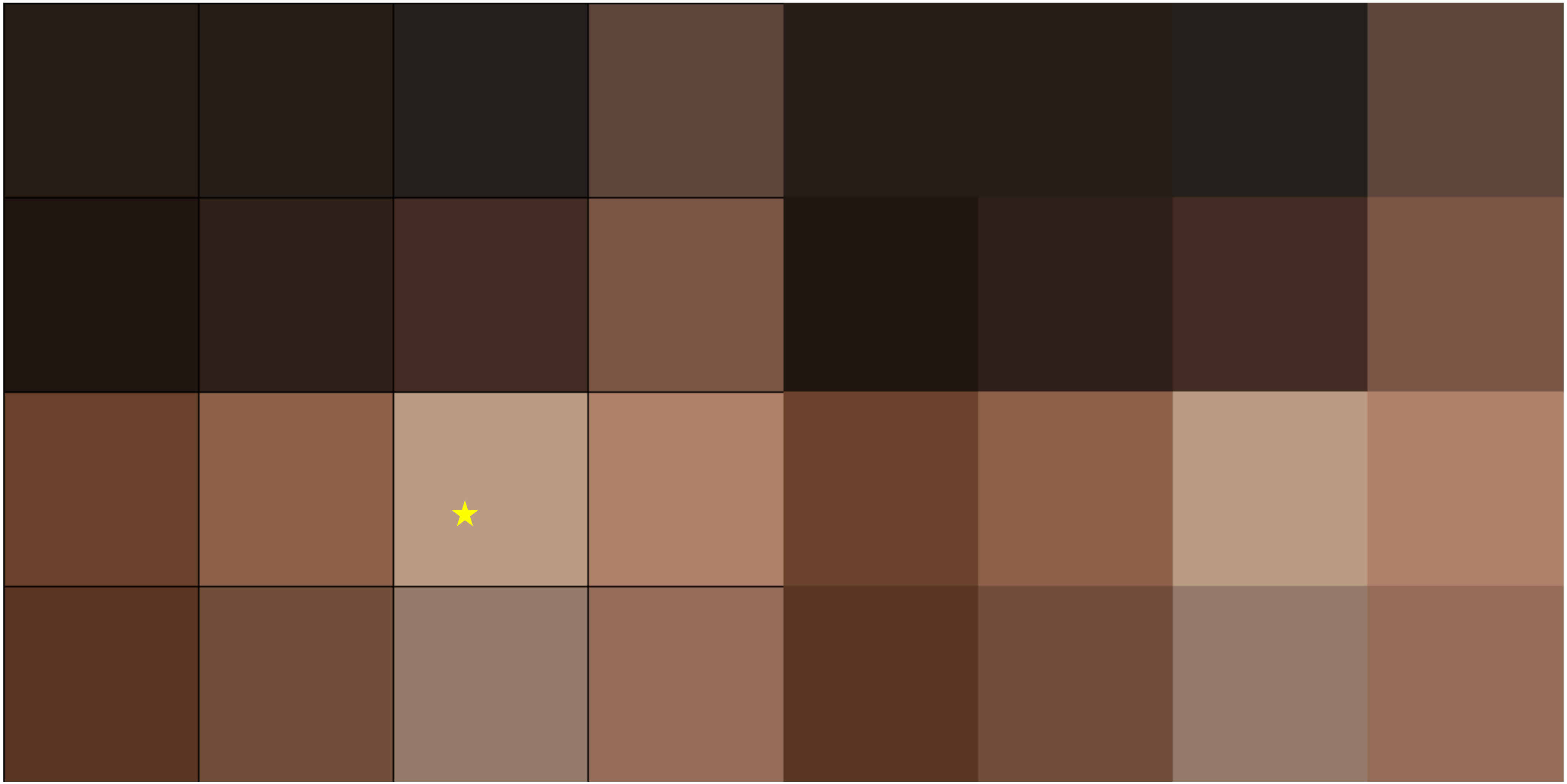
Hilbert R-Tree



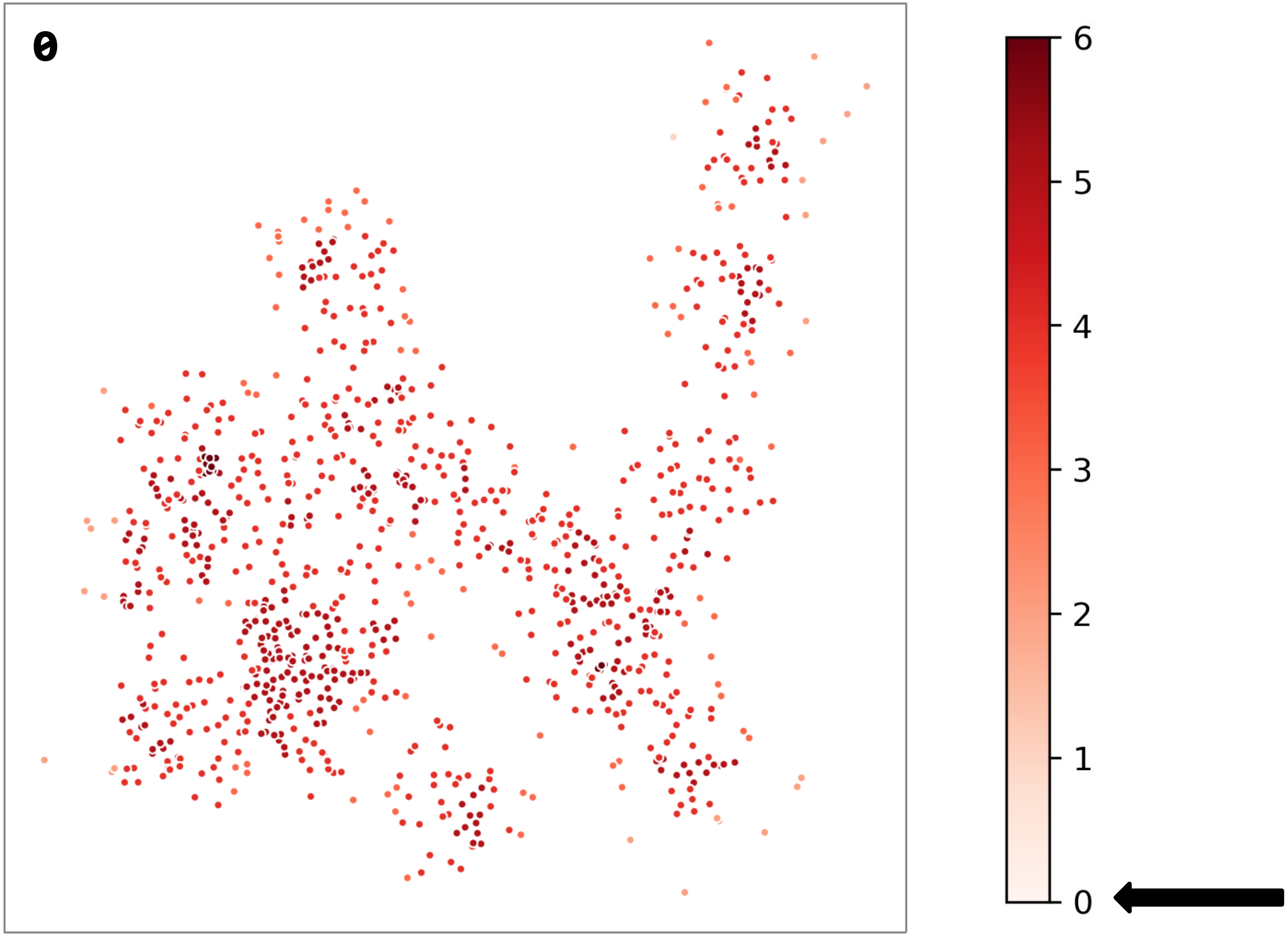
kd-Tree

THE SUPER QUADTREE (SUPERQT)

QUADTREES IN IMAGE COMPRESSION

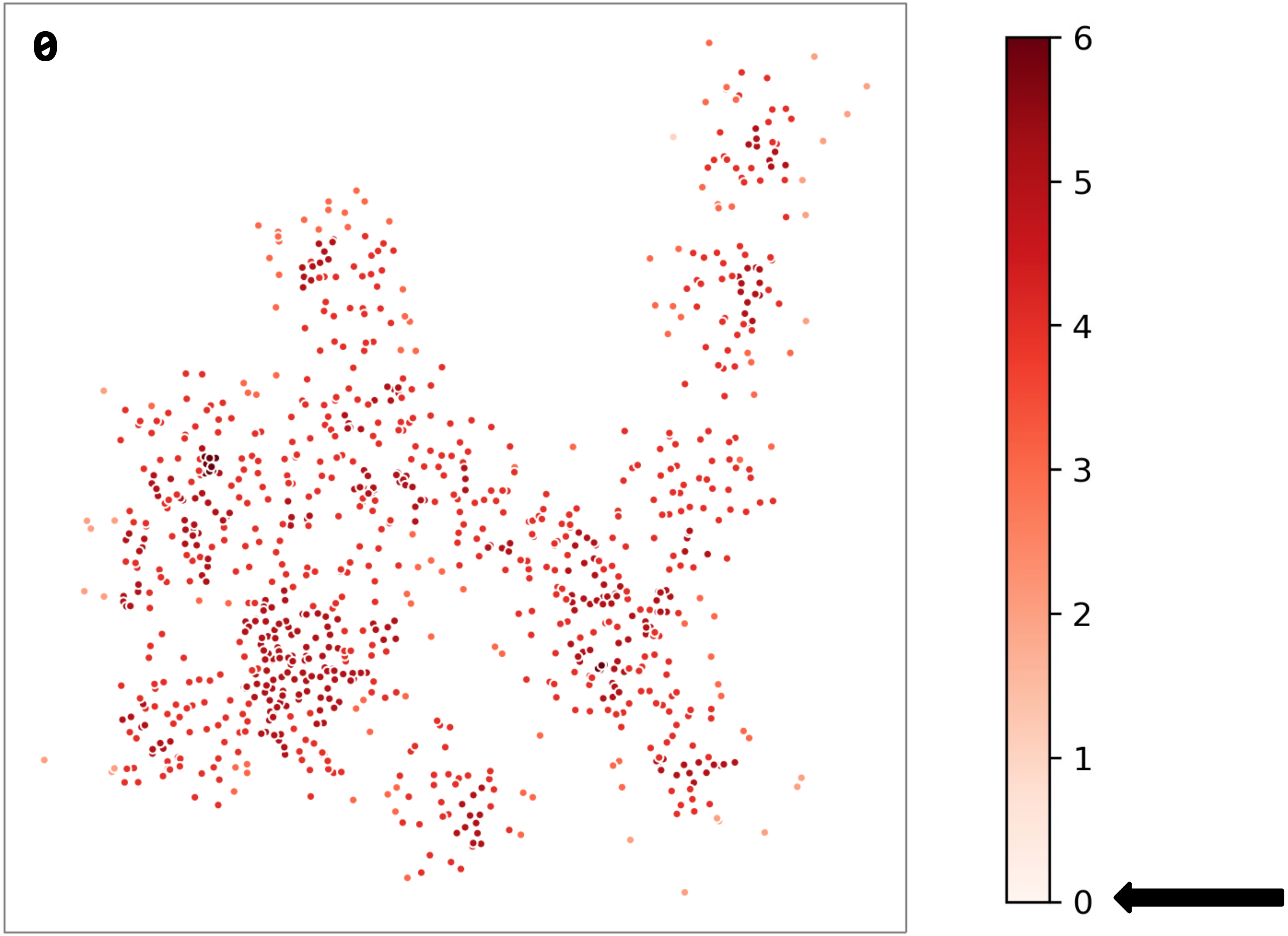


SPEND DETAIL WHERE ITS NEEDED MOST

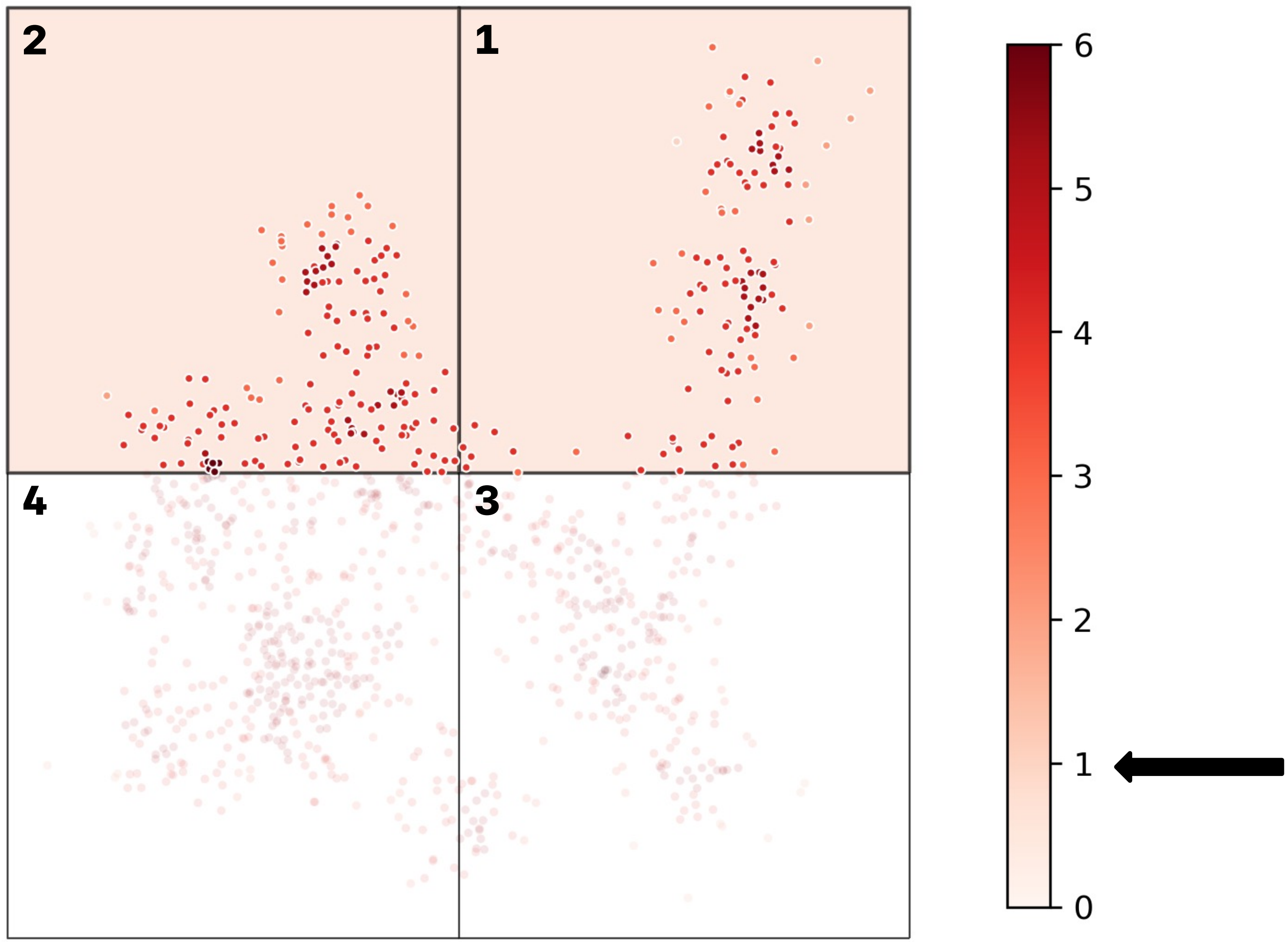


WALK DOWN THE DEPTH LADDER

c	x_1
1	x_{11}
1	x_{12}
1	x_{13}
1	x_{14}
1	x_{15}
...	...
1	x_{1n}

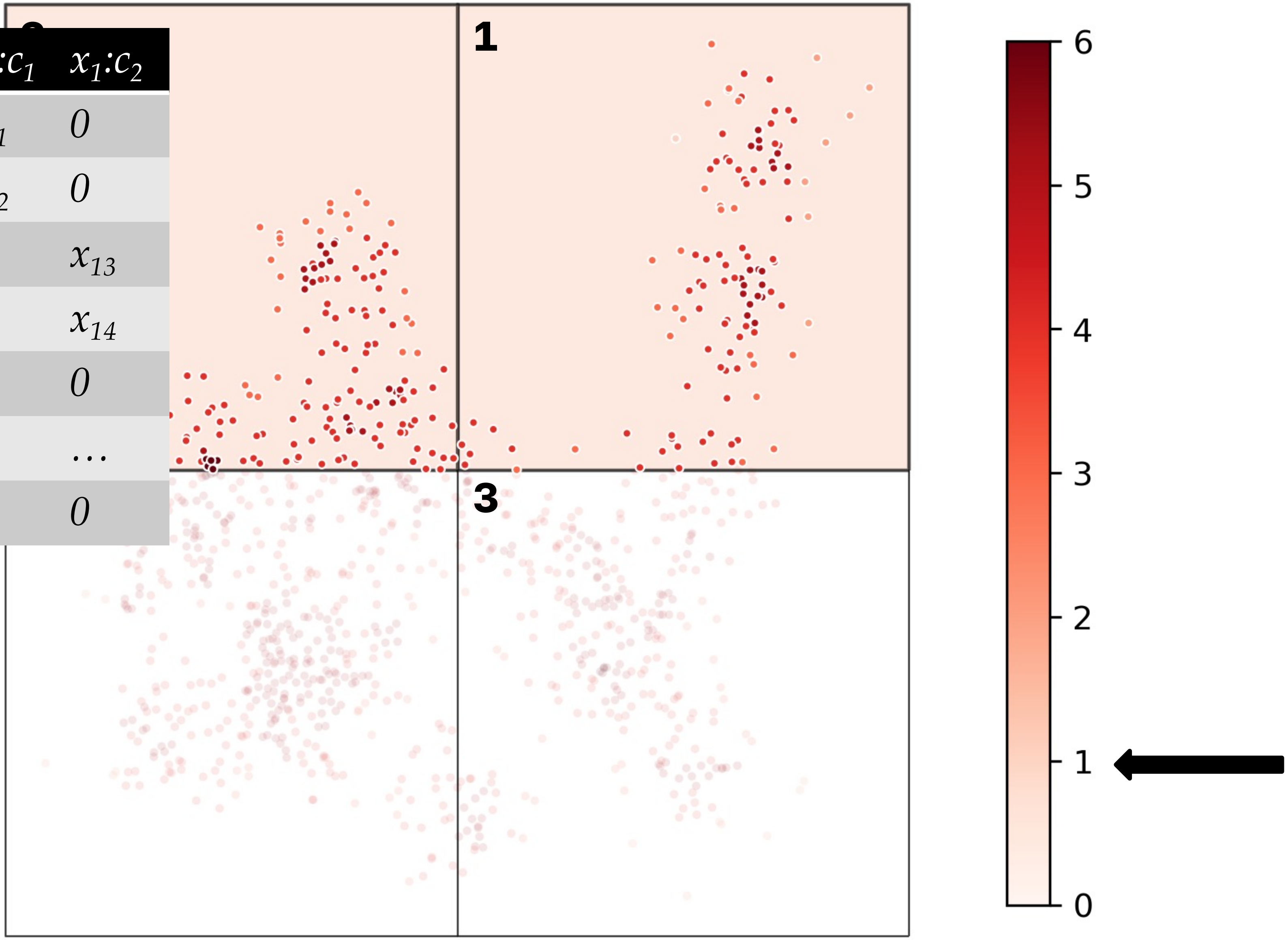


WALK DOWN THE DEPTH LADDER



WALK DOWN THE DEPTH LADDER

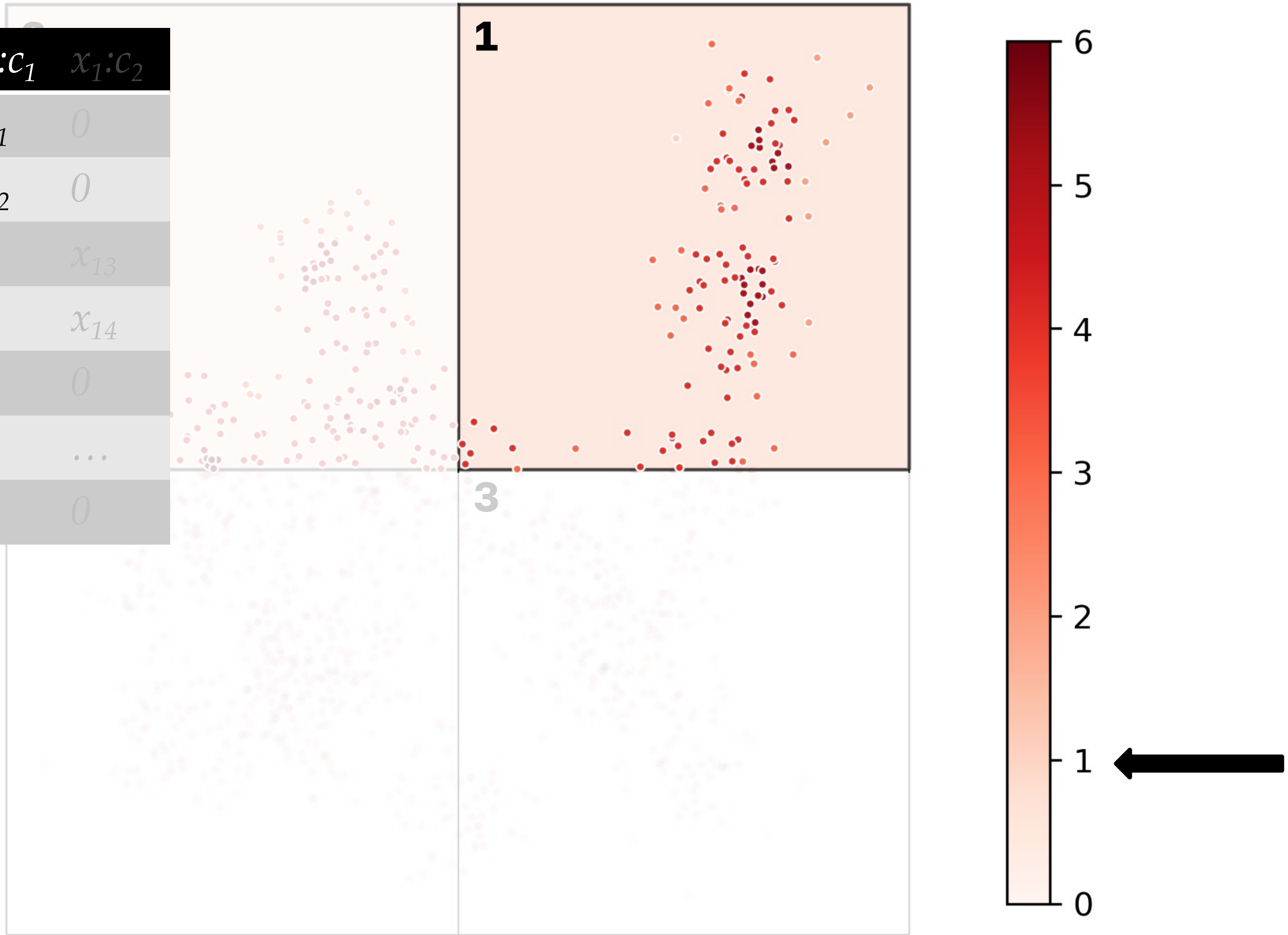
c	c_1	c_2	$x_1{:}c$	$x_1{:}c_1$	$x_1{:}c_2$
0	1	0	0	x_{11}	0
0	1	0	0	x_{12}	0
0	0	1	0	0	x_{13}
0	0	1	0	0	x_{14}
1	0	0	x_{15}	0	0
…	…	…	…	…	…
1	0	0	x_{1n}	0	0



WALK DOWN THE DEPTH LADDER

c	c_1	c_2
0	1	0
0	1	0
0	0	1
0	0	1
1	0	0
…	…	…
1	0	0

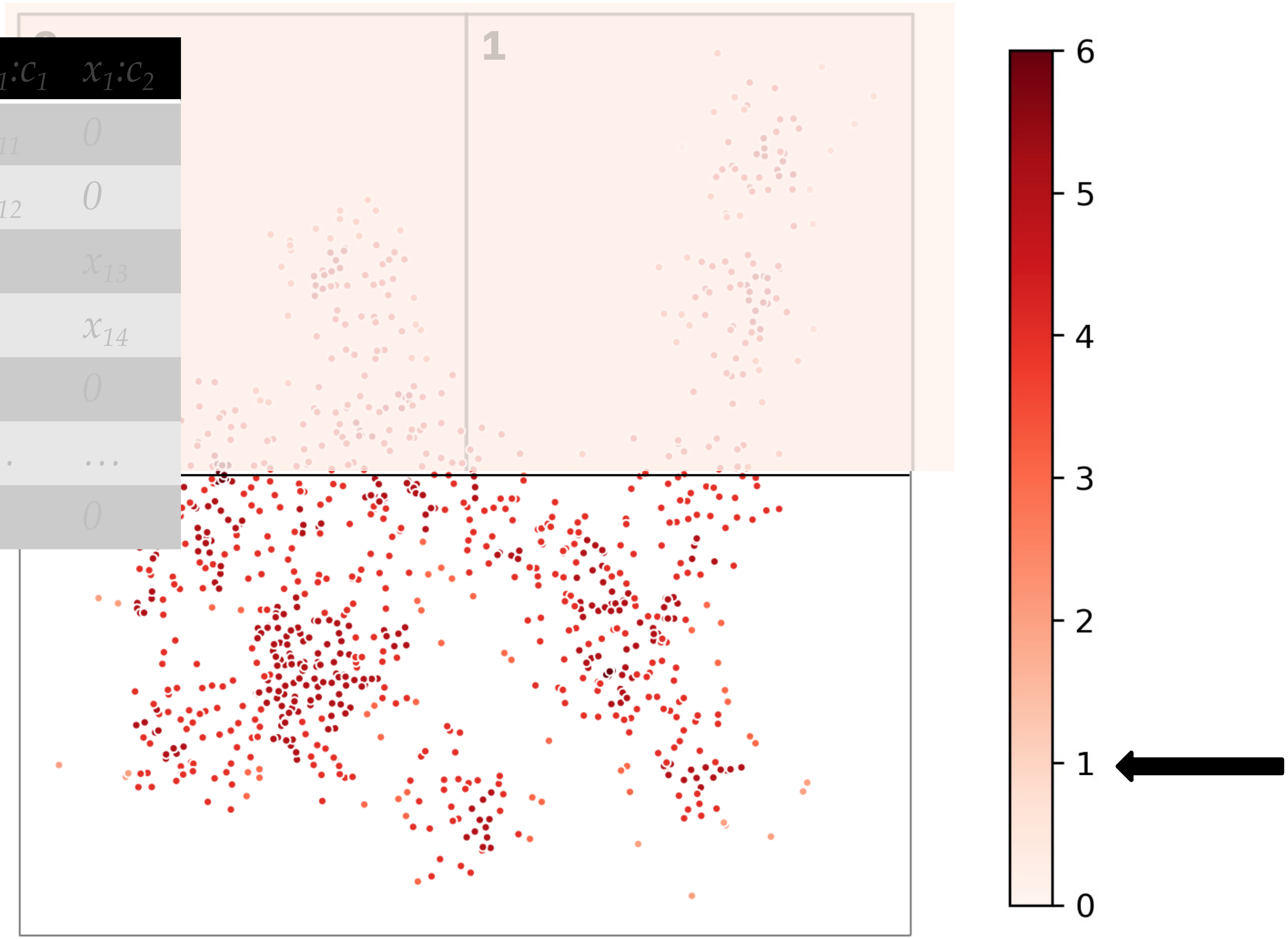
$x_1{:}c$	$x_1{:}c_1$	$x_1{:}c_2$
0	x_{11}	0
0	x_{12}	0
0	0	x_{13}
0	0	x_{14}
x_{15}	0	0
…	…	…
x_{1n}	0	0



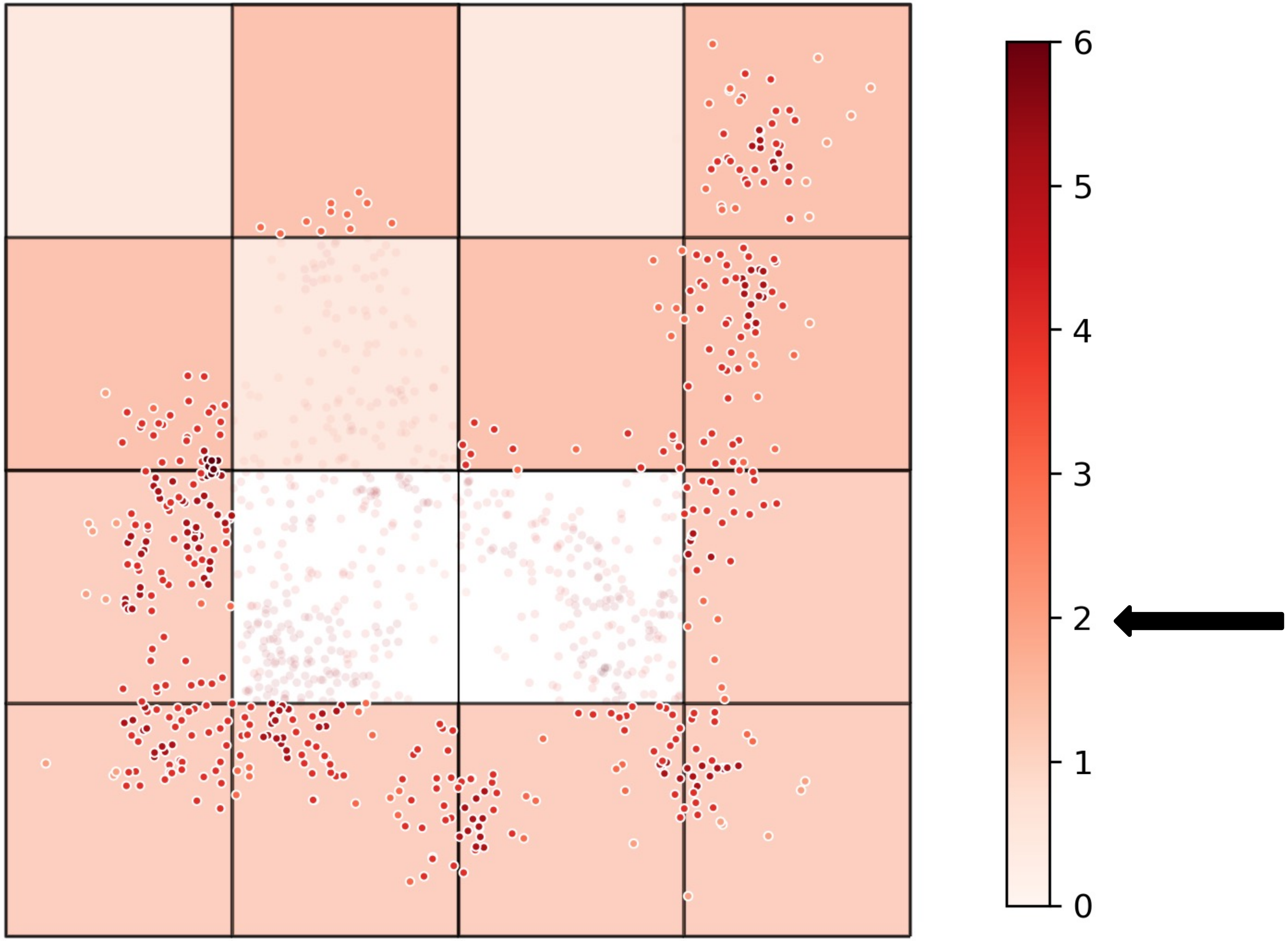
WALK DOWN THE DEPTH LADDER

c	c_1	c_2
0	1	0
0	1	0
0	0	1
0	0	1
1	0	0
…	…	…
1	0	0

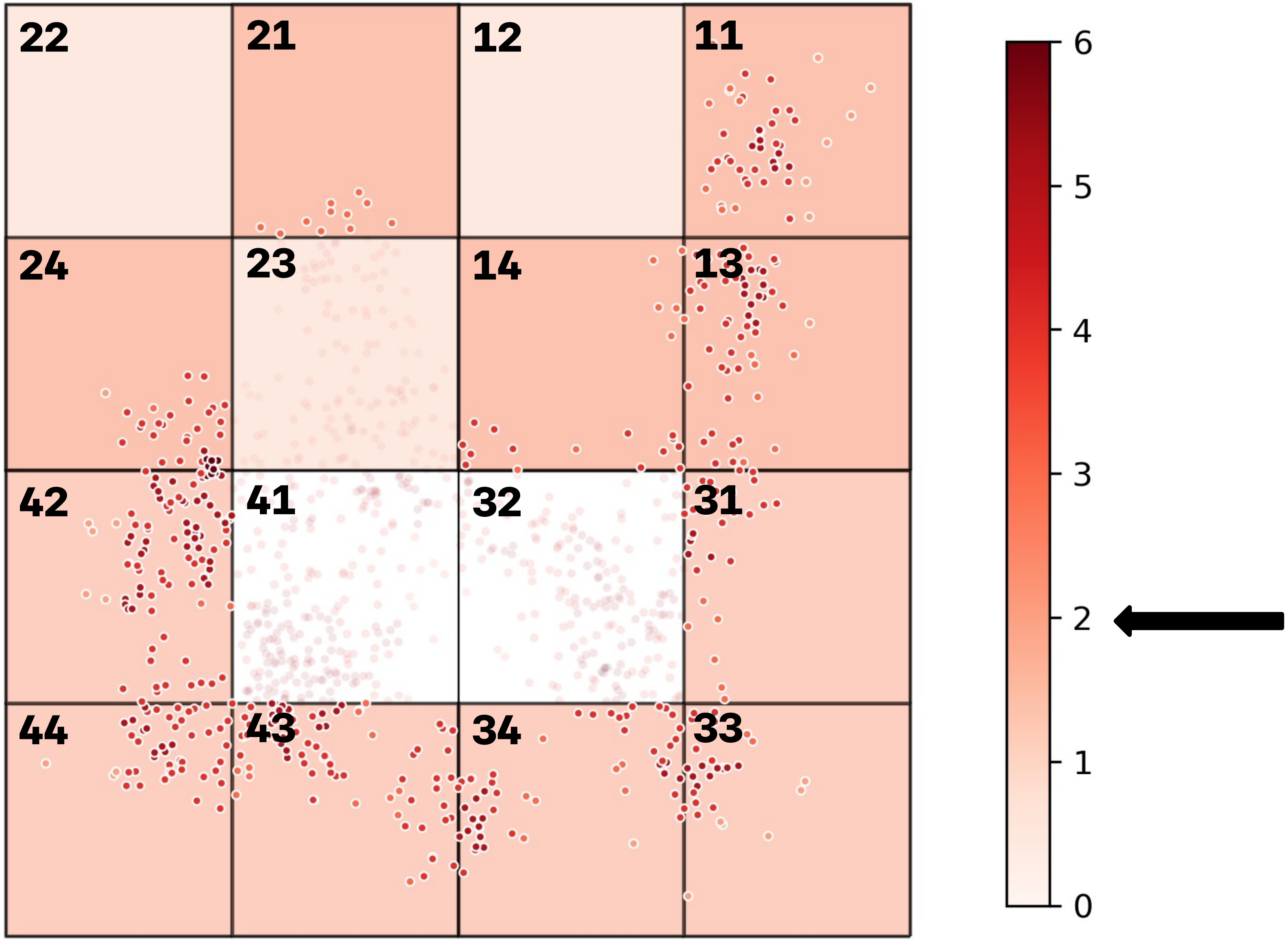
$x_1{:}c$	$x_1{:}c_1$	$x_1{:}c_2$
0	x_{11}	0
0	x_{12}	0
0	0	x_{13}
0	0	x_{14}
x_{15}	0	0
…	…	…
x_{1n}	0	0



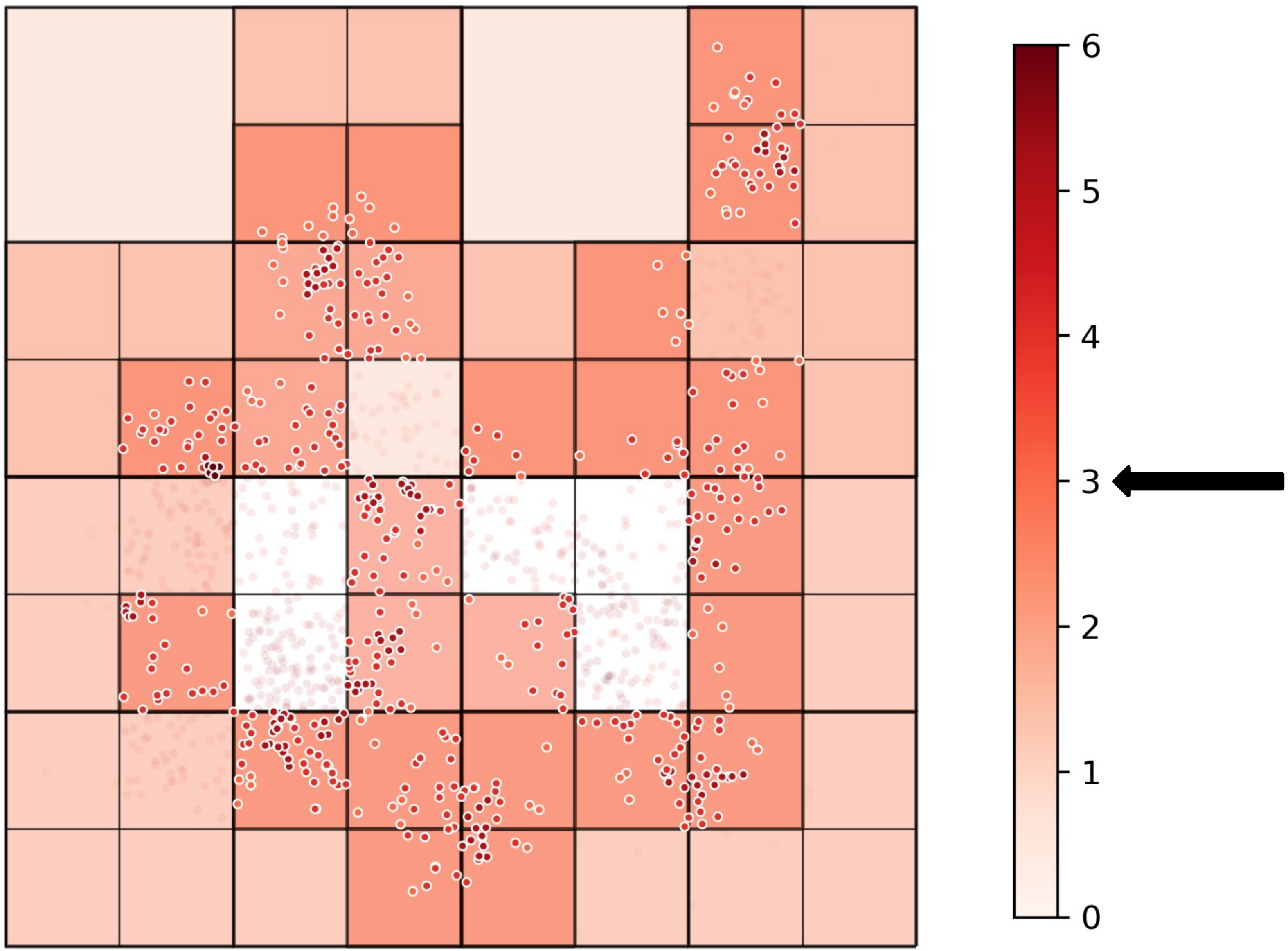
WALK DOWN THE DEPTH LADDER



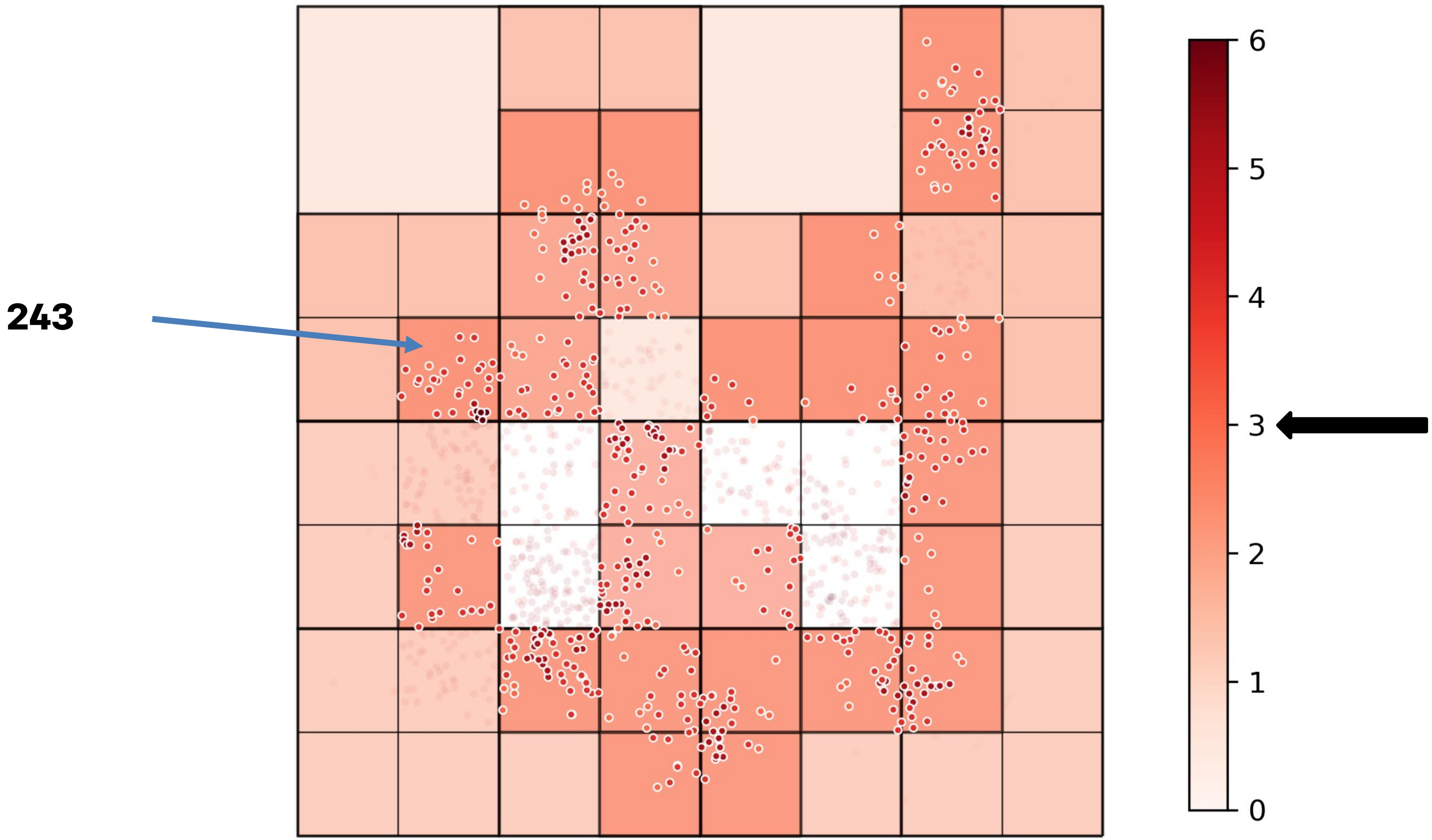
WALK DOWN THE DEPTH LADDER



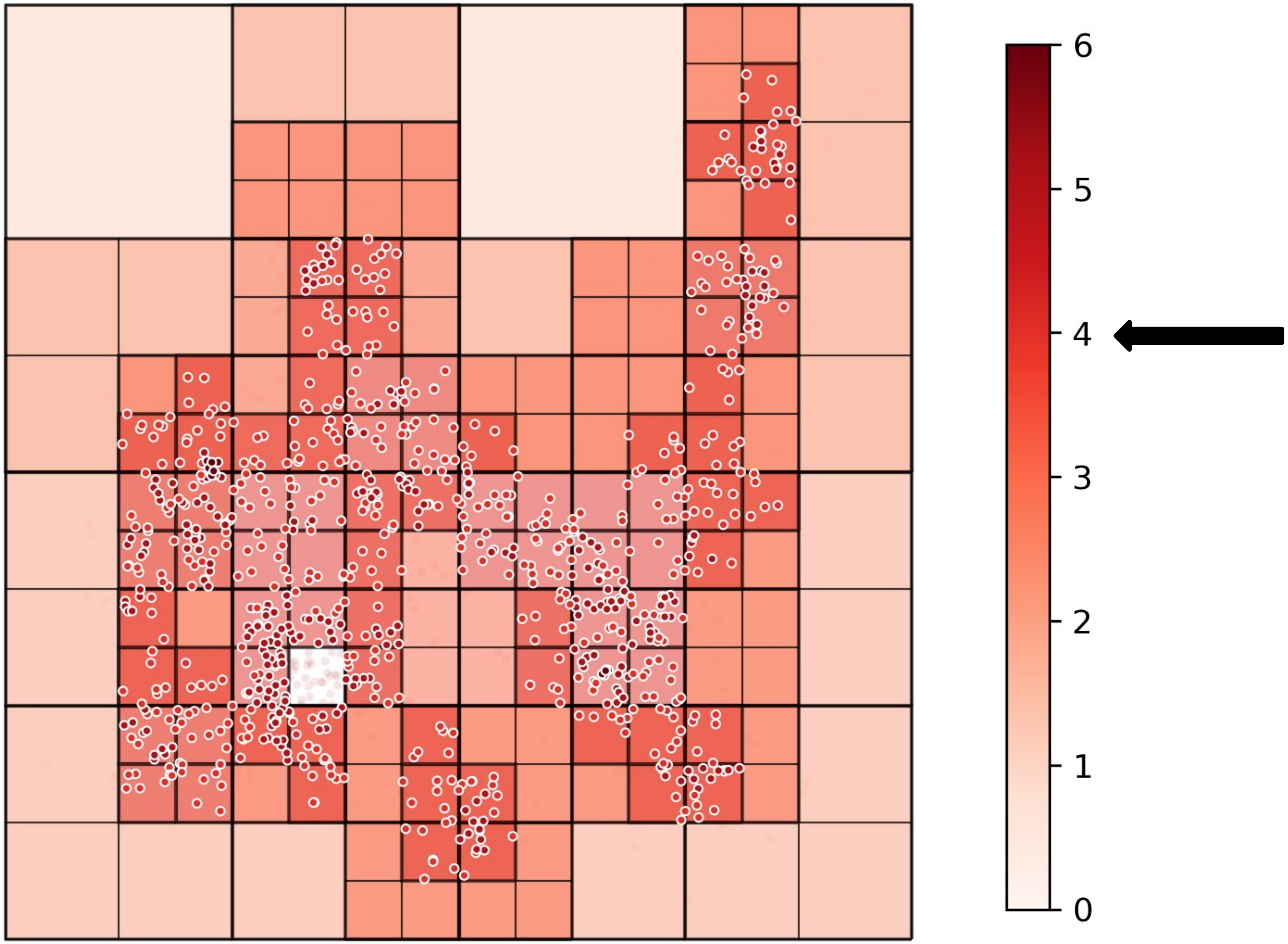
WALK DOWN THE DEPTH LADDER



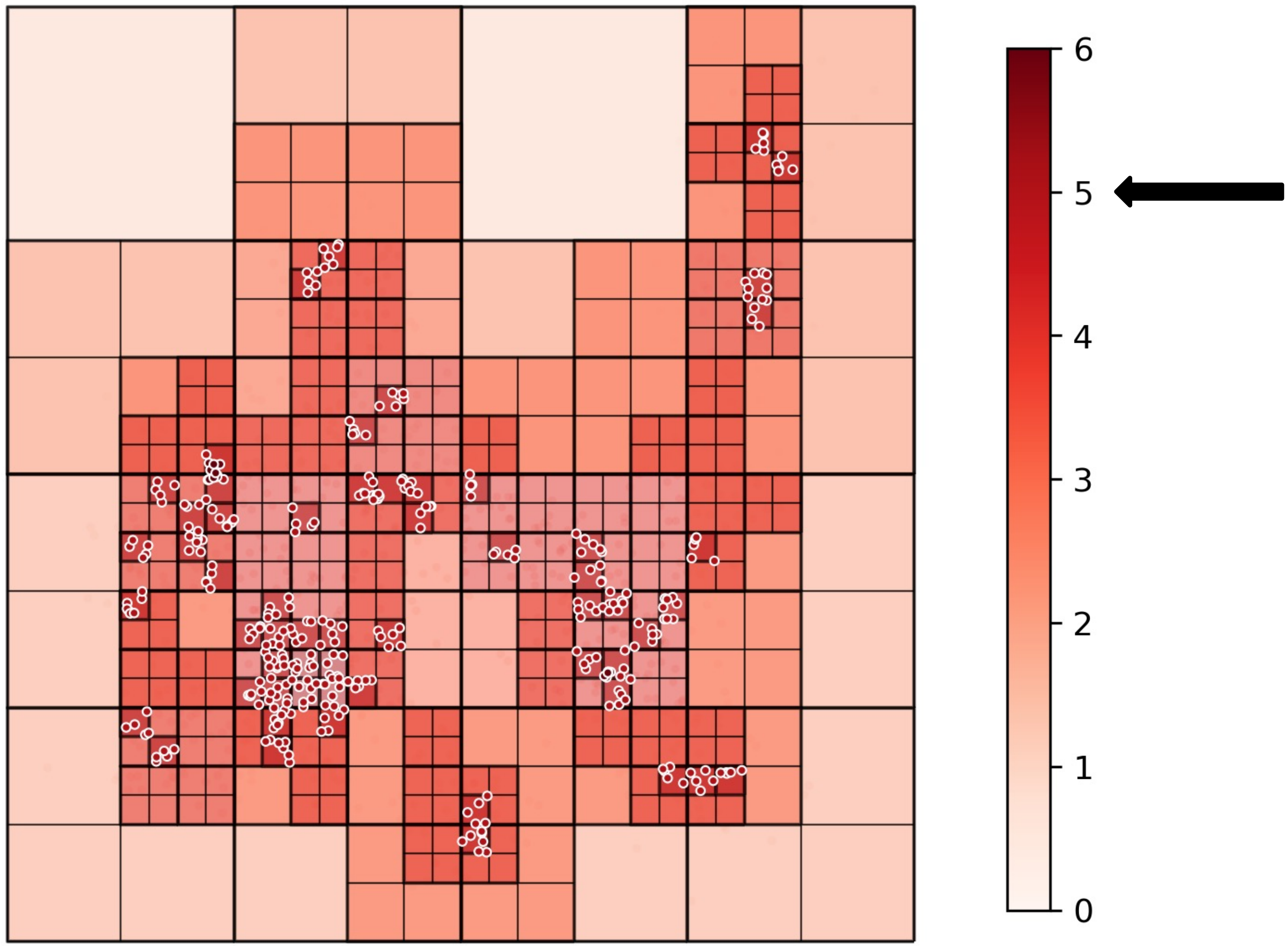
WALK DOWN THE DEPTH LADDER



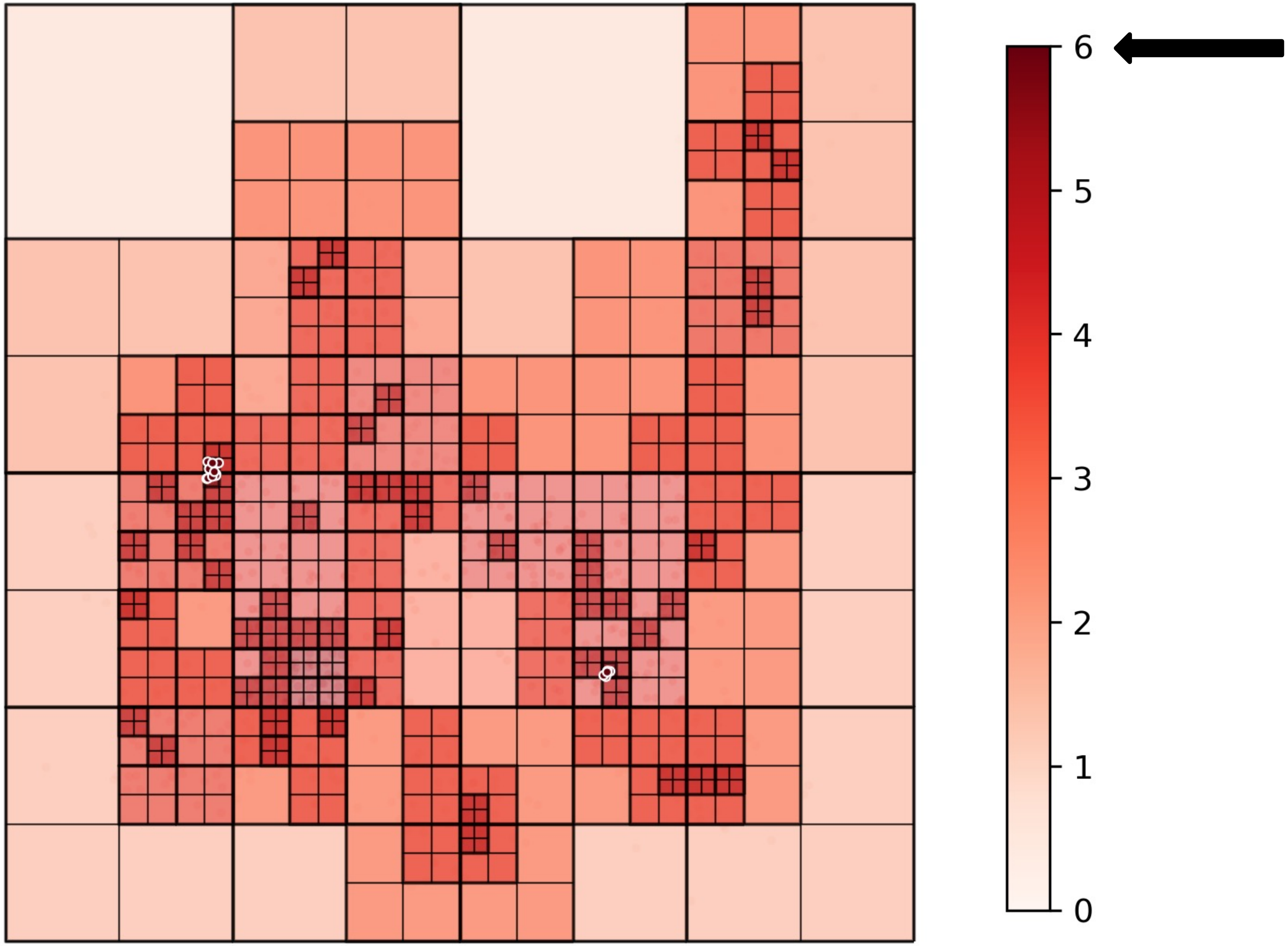
WALK DOWN THE DEPTH LADDER



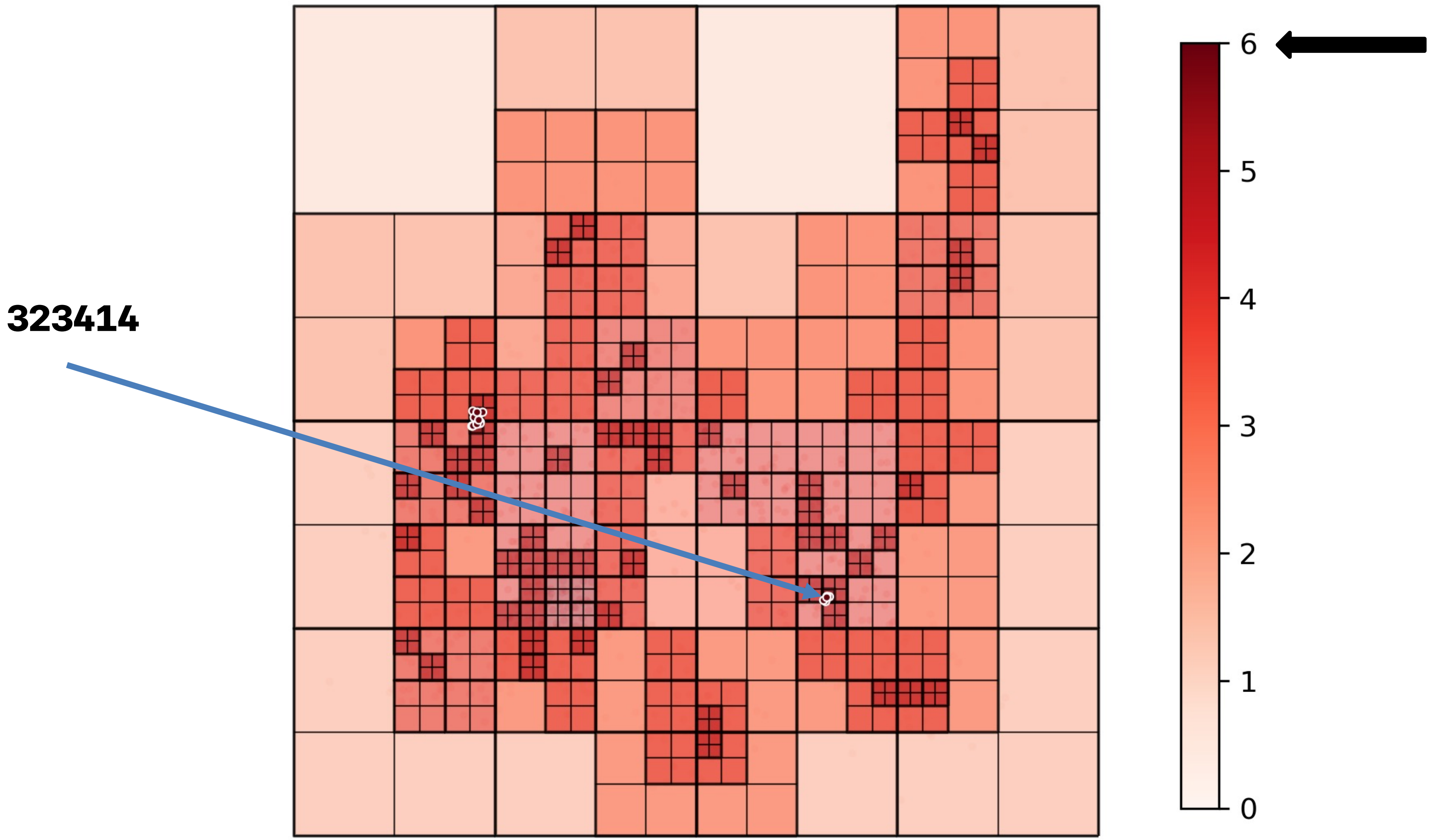
WALK DOWN THE DEPTH LADDER



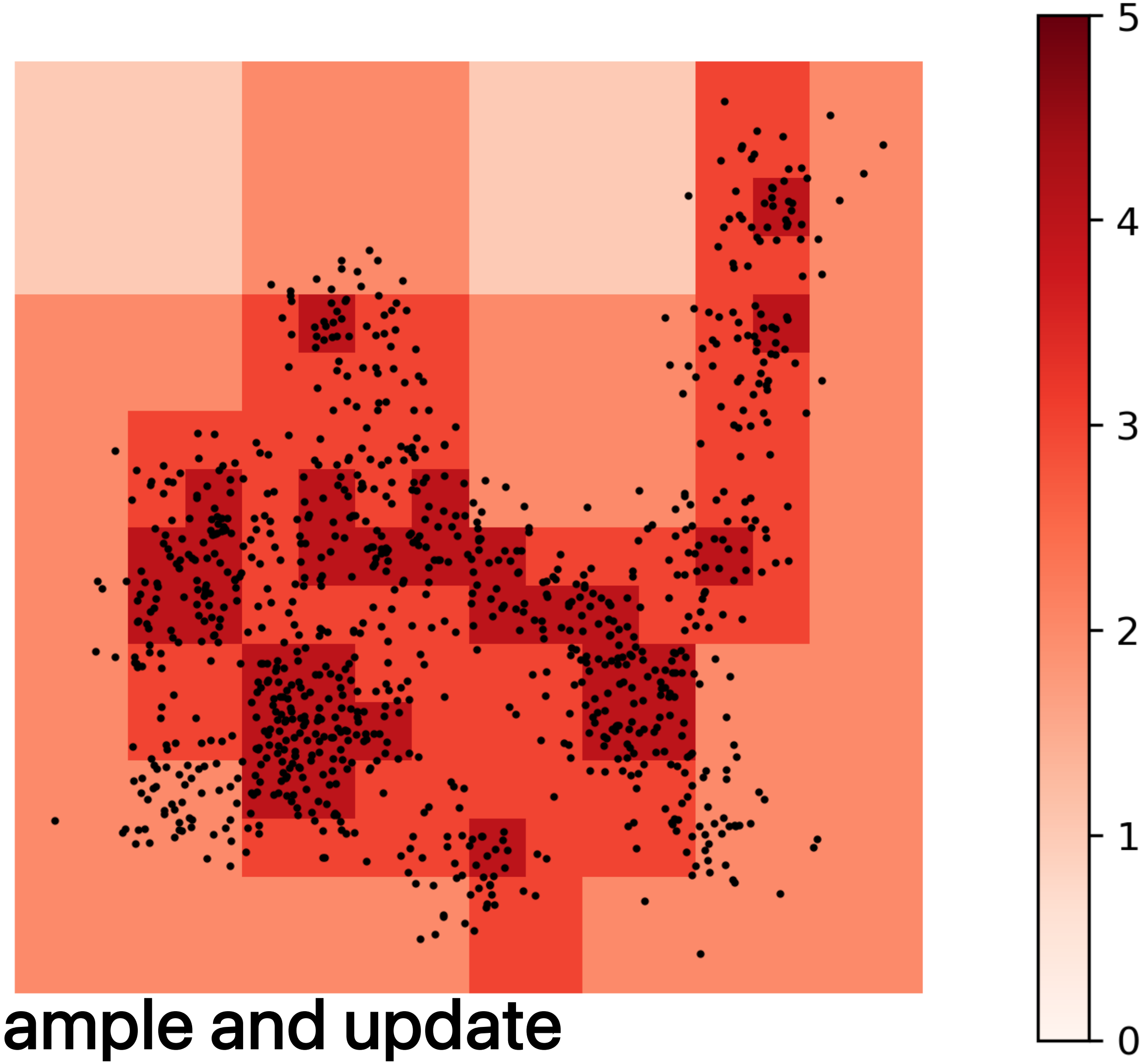
WALK DOWN THE DEPTH LADDER



WALK DOWN THE DEPTH LADDER



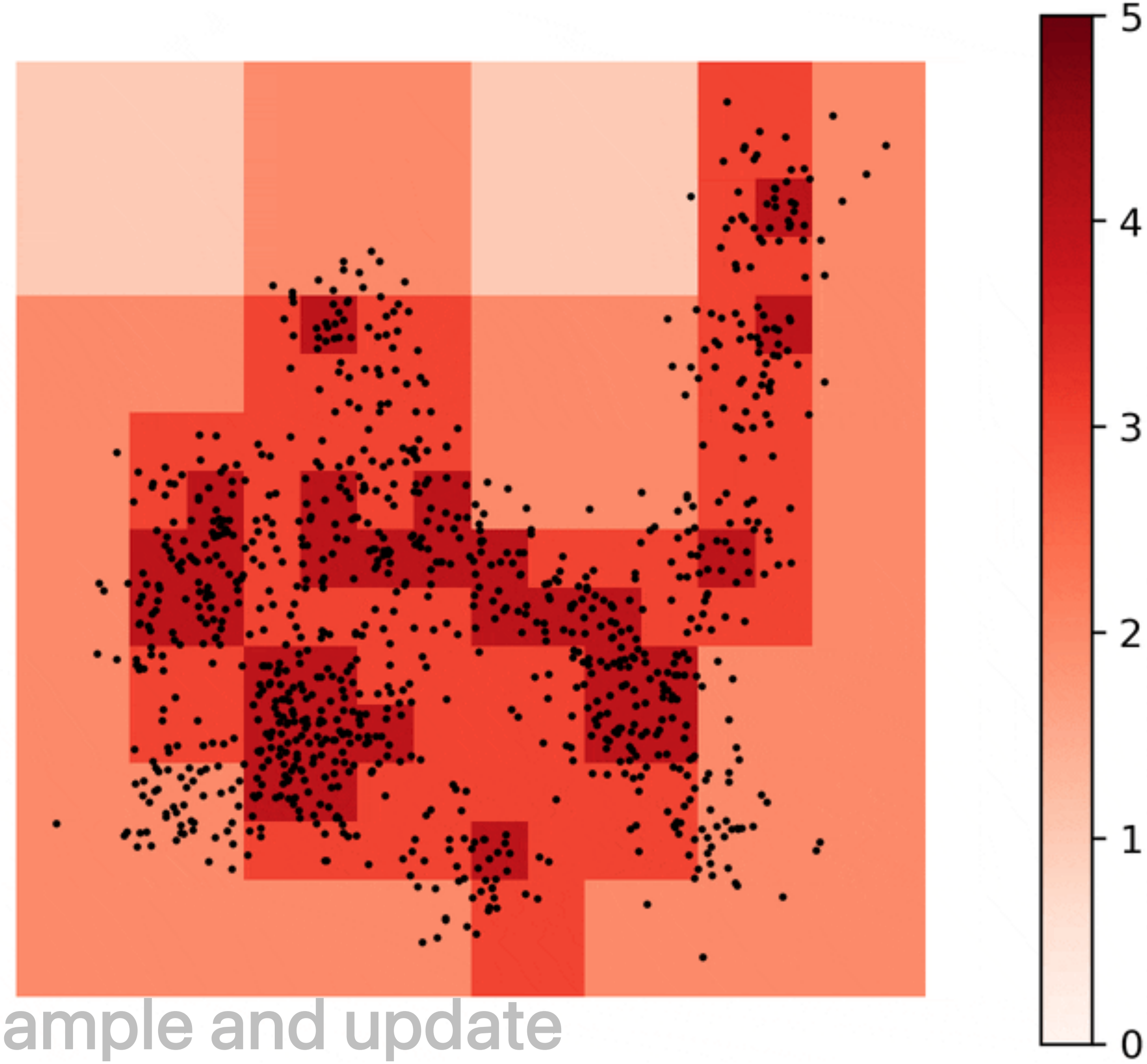
WALK DOWN THE DEPTH LADDER



SuperQT

- stable in CV
- totally ordered
- simple out of sample and update

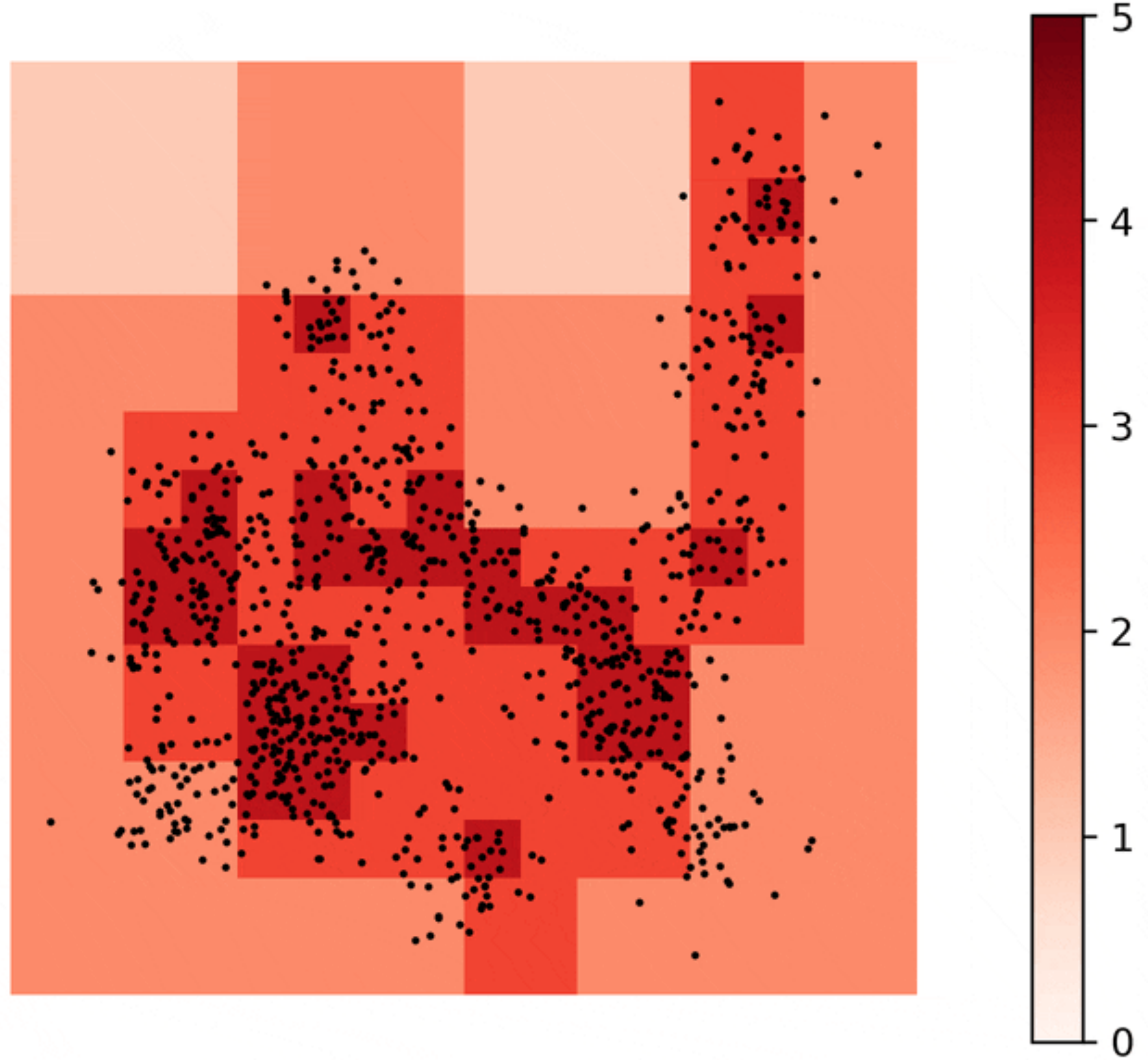
QUADTREES ARE FAIRLY STABLE IN CV



SuperQT
- stable in CV
- totally ordered
- simple out of sample and update

QUADTREES ARE FAIRLY STABLE IN CV

TOTALLY ORDERED
An observation will nest exactly into one cell at any given depth; depth is a function of both fit and density, and big jumps in depth are rare.



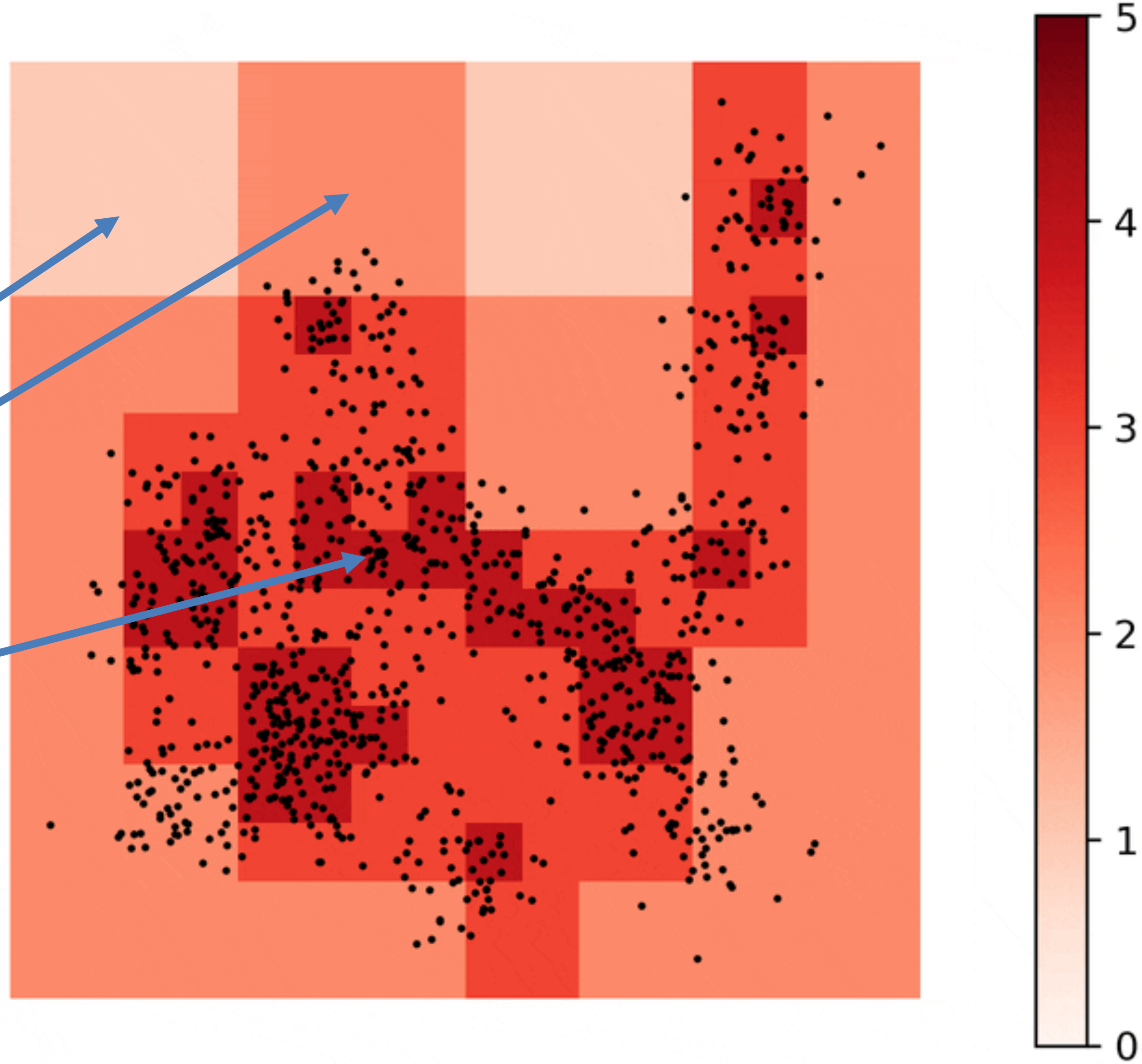
QUADTREES ARE FAIRLY STABLE IN CV

OUT-OF-BAG
For interpolating predictions *within* frame, we always have a model

Always **2**

2 or **21**

4112
or
411

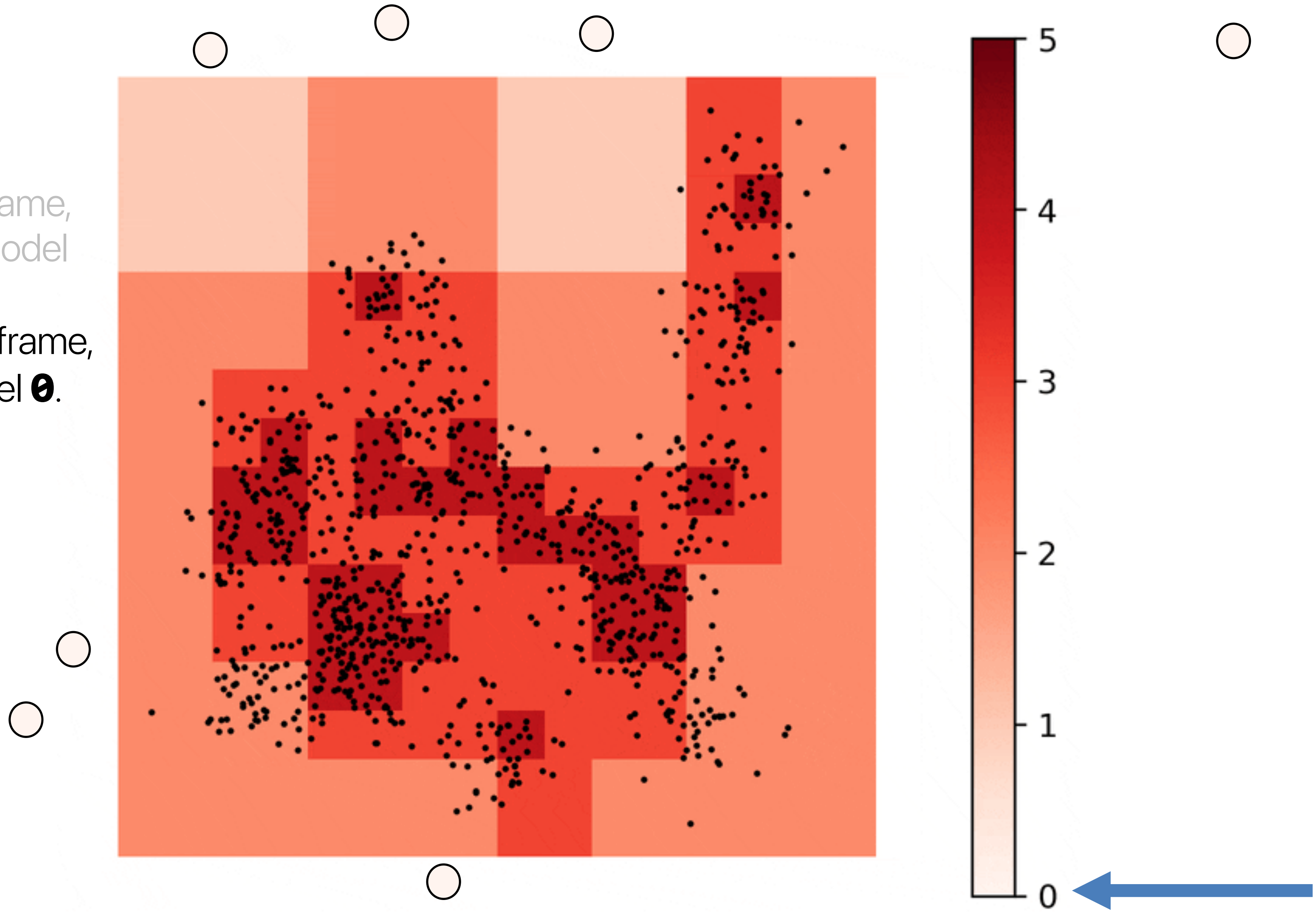


QUADTREES ARE FAIRLY STABLE IN CV

OUT-OF-BAG

For interpolating predictions *within* frame, we always have a model

For *extrapolating* predictions outside frame, we always use model **0**.



QUADTREES ARE FAIRLY STABLE IN CV

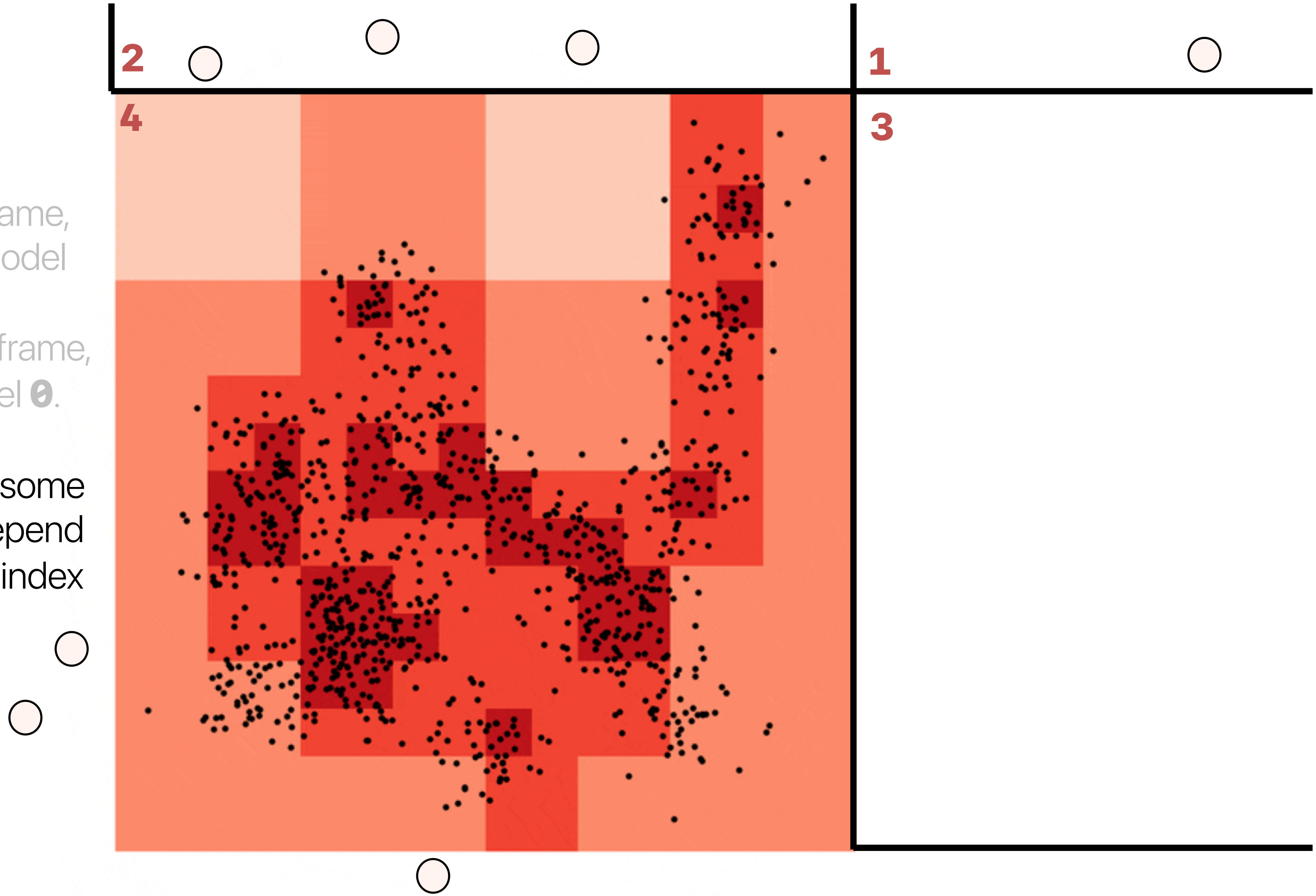
OUT-OF-BAG

For interpolating predictions *within* frame, we always have a model

For *extrapolating* predictions outside frame, we always use model **0**.

UPDATING

Treat 0 as a child of some larger frame and prepend all cells with its new index



QUADTREES ARE FAIRLY STABLE IN CV

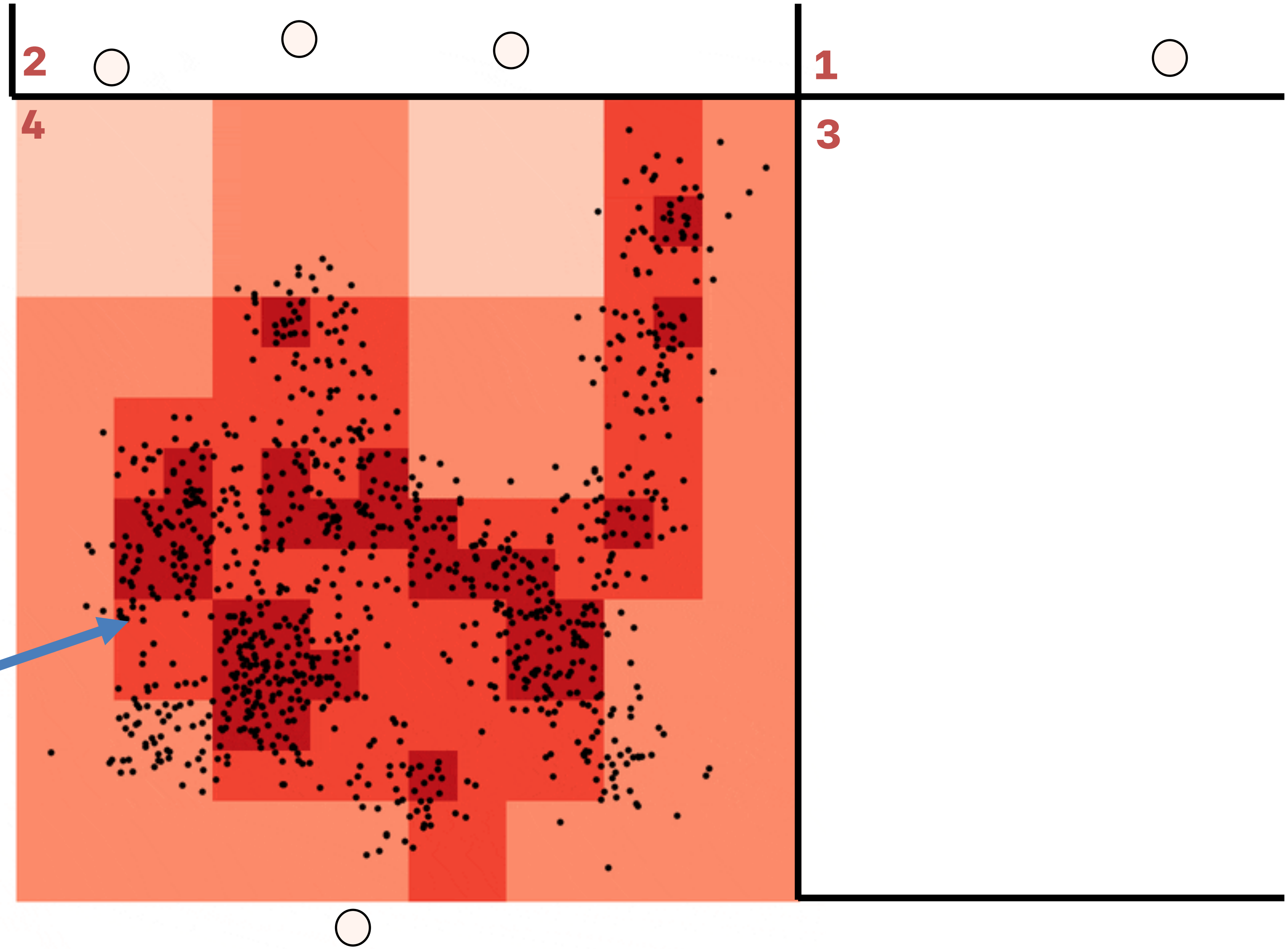
OUT-OF-BAG

For interpolating predictions *within* frame, we always have a model

For *extrapolating* predictions outside frame, we always use model **0**.

UPDATING

Treat 0 as a child of some larger frame and prepend all cells with its new index



424

updates to

4424

QUADTREES ARE FAIRLY STABLE IN CV

0. INITIALIZE: add root node to queue with label 0 and global model predicting $y \sim X$

1. SPLIT: pop the first node from our queue and split. For each child in that split:

1-1. fit a submodel predicting only child's outcomes using child's data.

1-2. update predictions for the whole map.

1-3. if child model is significantly more skillful than parent **within child (out of sample)**, label samples with child label, return child to queue, & go to 1; otherwise assign all samples to parent's label & drop child.

2. PRUNE: fit a model with all current feature:label interactions (one hot).

2-1. if all terms are helpful, go to 3, otherwise find the depth (d^*) of the least helpful feature:label term

2-2. for each unhelpful term at d^*, merge the feature:child_label term into feature:parent_label, and return to 2.

3. FINALIZE: Fit a final model on final retained feature:label interaction terms

THE SUPERVISED QUADTREE

0. INITIALIZE: add root node to queue with label 0 and global model predicting $y \sim X$
1. SPLIT: pop the first node from our queue and split. For each child in that split:
1-1. fit a submodel predicting only child's outcomes using child's data.
1-2. update predictions for the whole map.
1-3. if child model is significantly more skillful than parent **within child (out of sample)**, label samples with child label, return child to queue, & go to 1; otherwise assign all samples to parent's label & drop child.
2. PRUNE: fit a model with all current feature:label interactions (one hot).
2-1. if all terms are helpful, go to 3, otherwise find the depth (d^*) of the least helpful feature:label term
2-2. for each unhelpful term at d^*, merge the feature:child_label term into feature:parent_label, and return to 2.
3. FINALIZE: Fit a final model on final retained feature:label interaction terms

Note that this is *model agnostic*: can use any learner, classification or regression.
- GLMs can split using LRT (χ^2) prune using Wald (t) tests, placing a natural limit on n_{child}
- Always can split using score improvement ($\Delta MSE > \varepsilon$) prune using permutation feature importance ($X_p * I(leaf)$ in best q% of unpruned features)

THE SUPERVISED QUADTREE

0. INITIALIZE: add root node to queue with label 0 and global model predicting $y \sim X$

1. SPLIT: pop the first node from our queue and split. For each child in that split:

1-1. fit a submodel predicting only child's outcomes using child's data.

1-2. update predictions for the whole map.

1-3. if child model is significantly more skillful than parent **within child (out of sample)**, label samples with child label, return child to queue, & go to 1; otherwise assign all samples to parent's label & drop child.

2. PRUNE: fit a model with all current feature:label interactions (one hot).

2-1. if all terms are helpful, go to 3, otherwise find the depth (d^*) of the least helpful feature:label term

2-2. for each unhelpful term at d^*, merge the feature:child_label term into feature:parent_label, and return to 2.

3. FINALIZE: Fit a final model on final retained feature:label interaction terms

Note that this is ***tree agnostic***: can use any spatial tree index

- Pruning is most useful with hierarchical trees, where parent is still "local"
- Overlap (e.g. in RTree) introduces intersection terms (feature:label1:label2).
- Balance is optional, but helpful for submodel efficiency

THE SUPER*TREE

CLUSTERING REGRESSION

jointly solving clustering and regression

GEOGRAPHICAL CLUSTER-REG

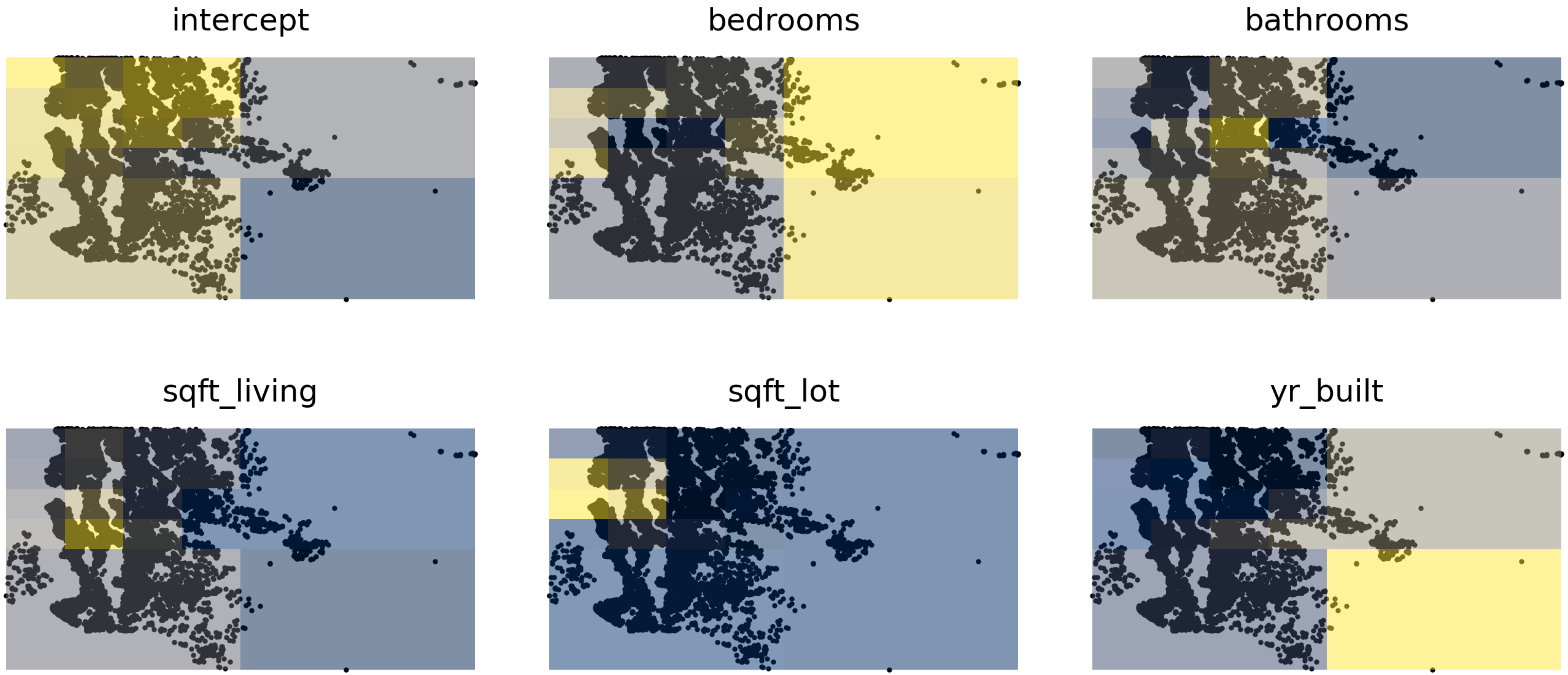
jointly solving clustering and regression

UNDERSTANDING QUADTREES

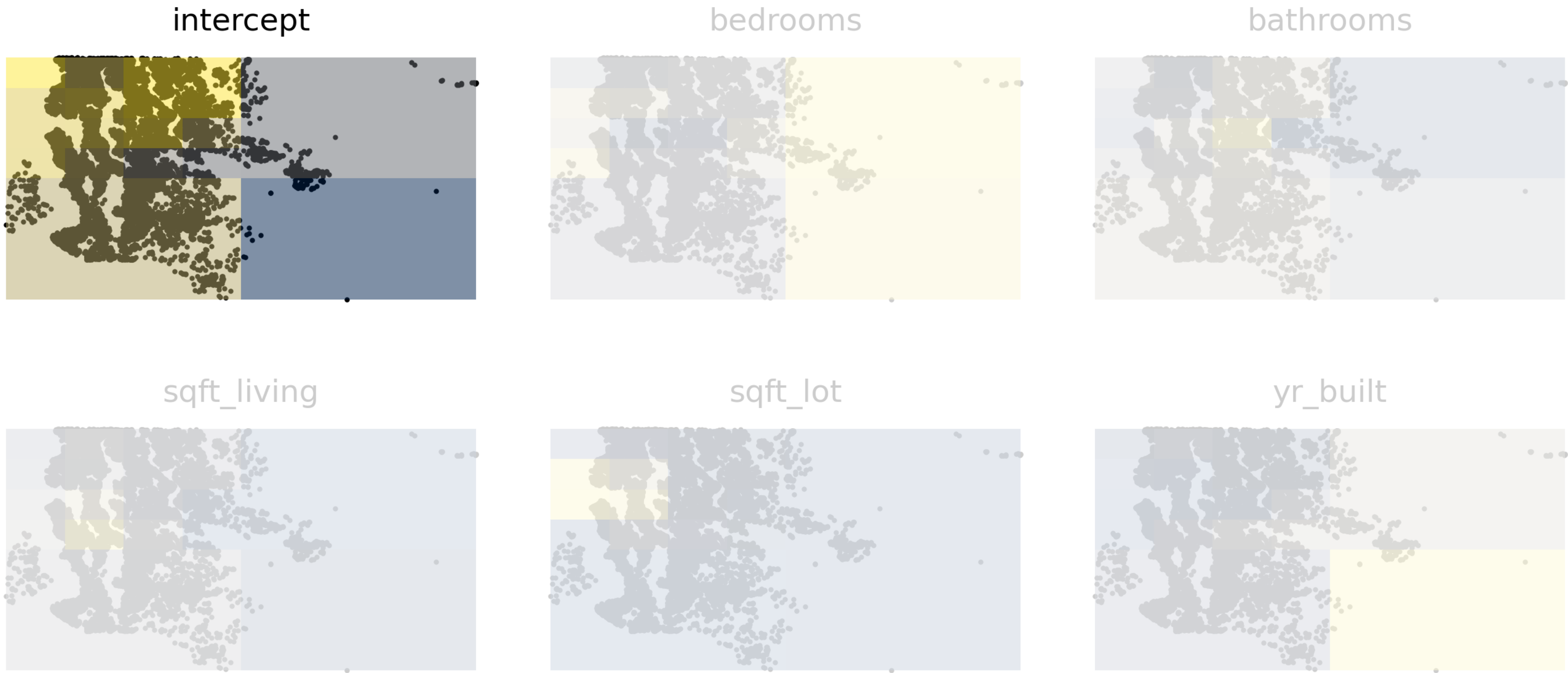
spatial splits for spatial fits

APPLYING QUADTREE REGRESSION

model comparison & basic metrics

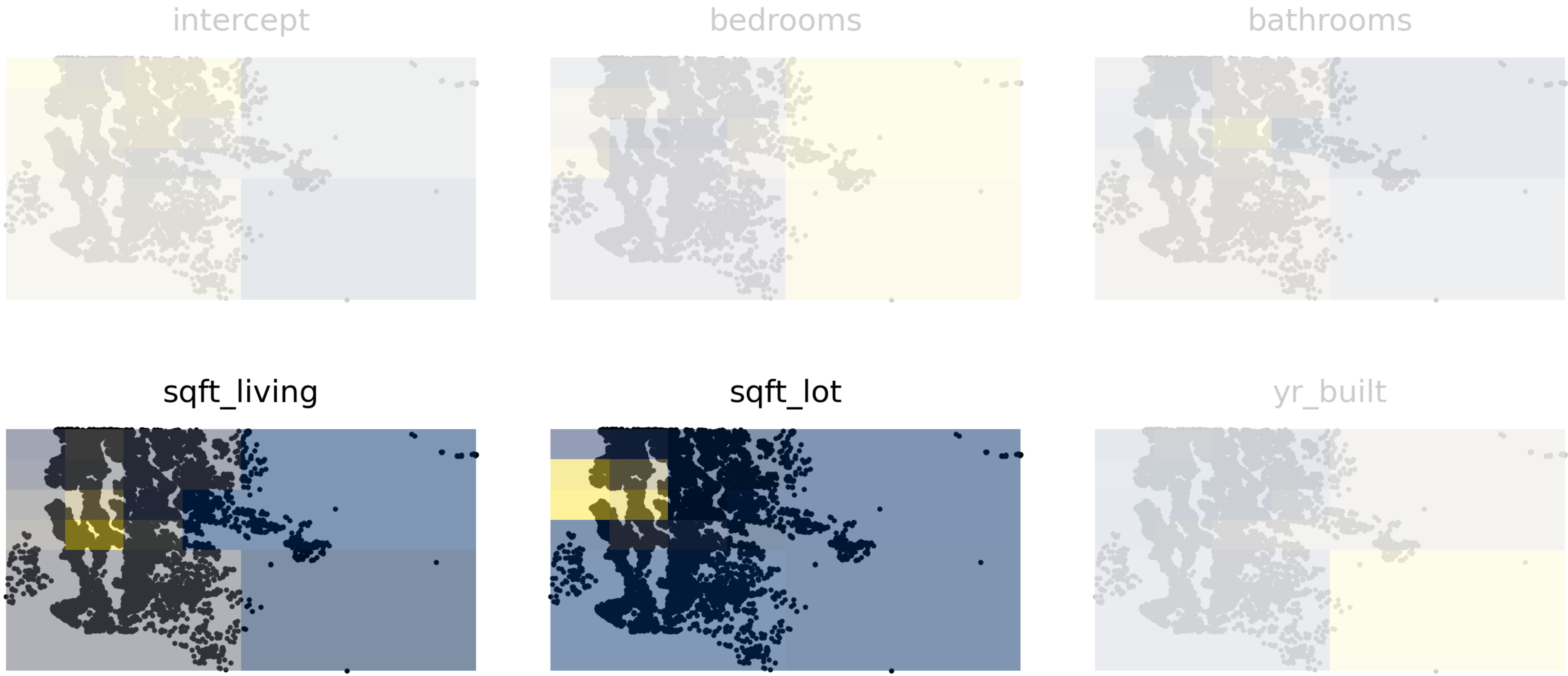


COEFFICIENT SURFACES LOOK LIKE THIS



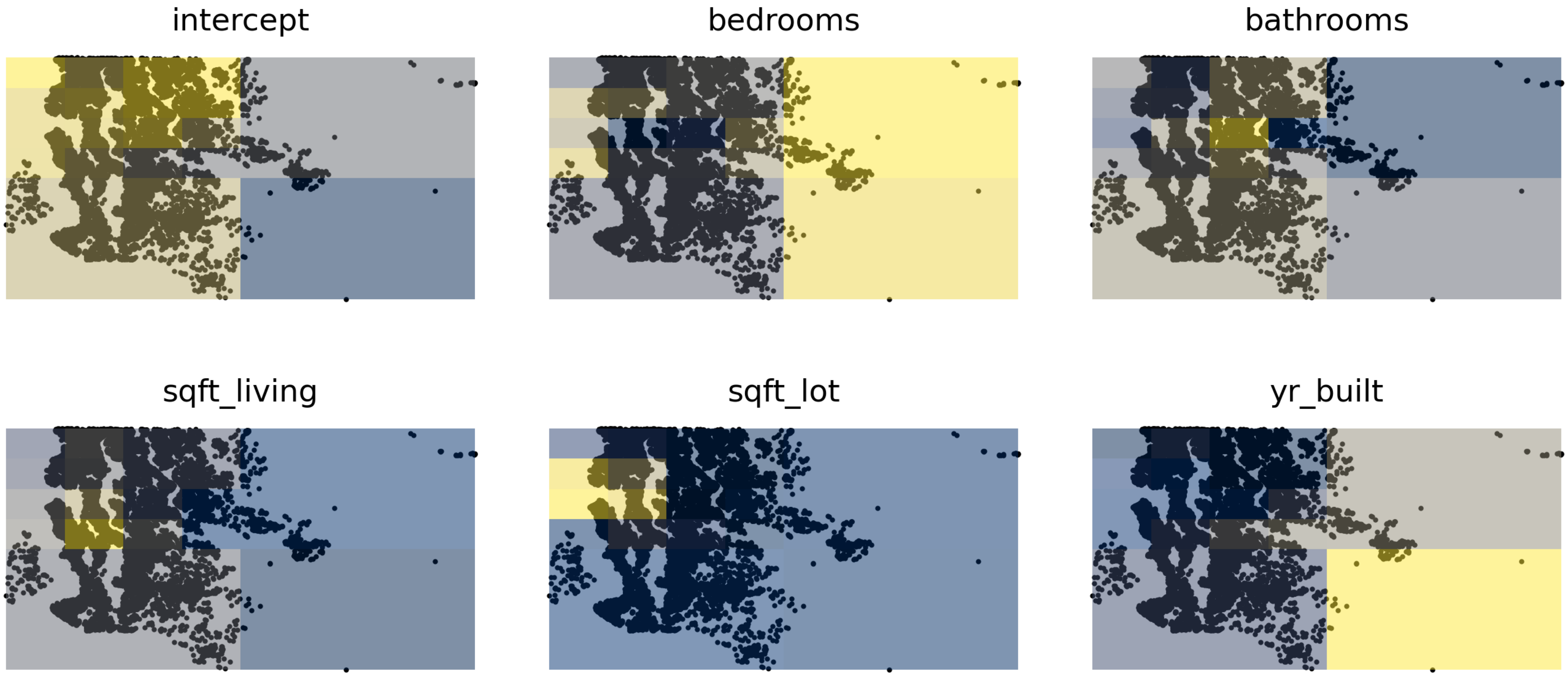
Prices in general are expensive towards the coast, also in the techy areas in the northwest.

COEFFICIENT SURFACES LOOK LIKE THIS



Space is at a premium, with living space slightly more expensive south-center.

COEFFICIENT SURFACES LOOK LIKE THIS



COEFFICIENT SURFACES LOOK LIKE THIS

PLACE	METHOD	MSE	R^2	TIME
Baltimore	OLS	.16	.49	Sub-seconds
n=211	Skater 15 Regimes	.02	.93	Seconds
	MGWR	.01	.95	Seconds
	SuperQT (no prune)	.04 (.02)	.89 (.94)	Seconds
King's Co.	OLS	.13	.53	Seconds
n=31k	Skater 15 Regimes	.05	.811	Hours
	MGWR	DNF	DNF	DNF
	SuperQT (no prune)	.06 (.06)	.78 (.78)	Minutes

COMPARISONS

PLACE	METHOD	MSE	R^2	TIME
Baltimore	OLS	.16	.49	Sub-seconds
n=211	Skater 15 Regimes	.02	.93	Seconds
	MGWR	.01	.95	Seconds
	SuperQT (no prune)	.04 (.02)	.89 (.94)	Seconds
King's Co.	OLS	.13	.53	Seconds
n=31k	Skater 15 Regimes	.05	.811	Hours
	MGWR	DNF	DNF	DNF
	SuperQT (no prune)	.06 (.06)	.78 (.78)	Minutes

Small n: usually need eps+perm (lrt+wald dof issues)

Big n: both lrt+wald and eps+perm are useful

COMPARISONS

PLACE	METHOD	MSE	R^2	TIME
Baltimore	OLS	.16	.49	Sub-seconds
n=211	Skater 15 Regimes	.02	.93	Seconds
	MGWR	.01	.95	Seconds
	SuperQT (no prune)	.04 (.02)	.89 (.94)	Seconds
King's Co.	OLS	.13	.53	Seconds
n=31k	Skater 15 Regimes	.05	.811	Hours
	MGWR	DNF	DNF	DNF
	SuperQT (no prune)	.06 (.06)	.78 (.78)	Minutes

COMPETITIVE EVEN AFTER PRUNING

PLACE	METHOD	MSE	R^2	TIME
Baltimore	OLS	.16	.49	Seconds
n=211	Skater 15 Regimes	.02	.93	Seconds
	SuperQT (no prune)	.04 (.02)	.89 (.94)	Seconds
King's Co.	OLS	.13	.53	Seconds
n=31k	Skater 15 Regimes	.05	.811	Hours
	MGWR	DNF	DNF	DNF
	SuperQT (no prune)	.06 (.06)	.78 (.78)	Minutes

Pruning is especially powerful in big data:

~90 feature x leaf interactions in Skater & SuperQT

~60 post pruning w/ very small change to fit (for both)

COMPETITIVE EVEN AFTER PRUNING

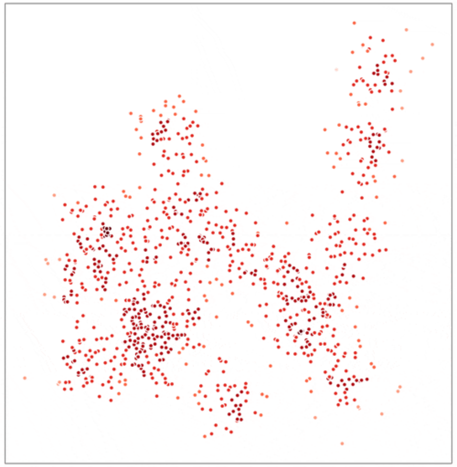
MIT-Licensed

High performance w/ `scipy.sparse`

Available in PySAL by December 2025

SUPER*TREE LEARNERS:

a new metalearner for local data science



LEVI JOHN WOLF levi.john.wolf@bristol.ac.uk