Technische Universität München

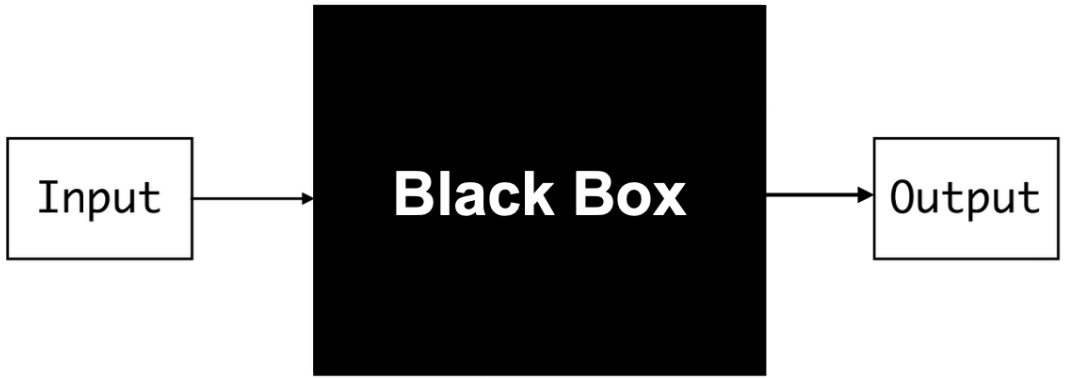
Can we trust XAI in spatial?

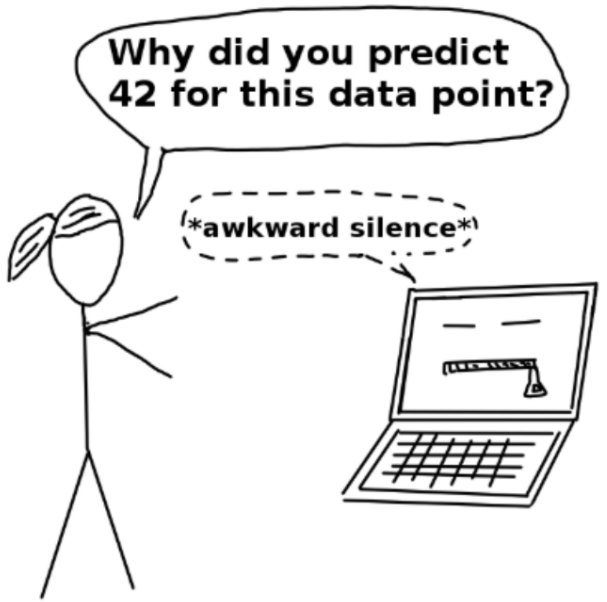
A distribution-free framework for quantifying uncertainty in explanations of spatial XAI models.

Xiayin Lou[1] Peng Luo[2,*] Song Gao[3] Ziqi Li[4] Liqiu Meng[1]

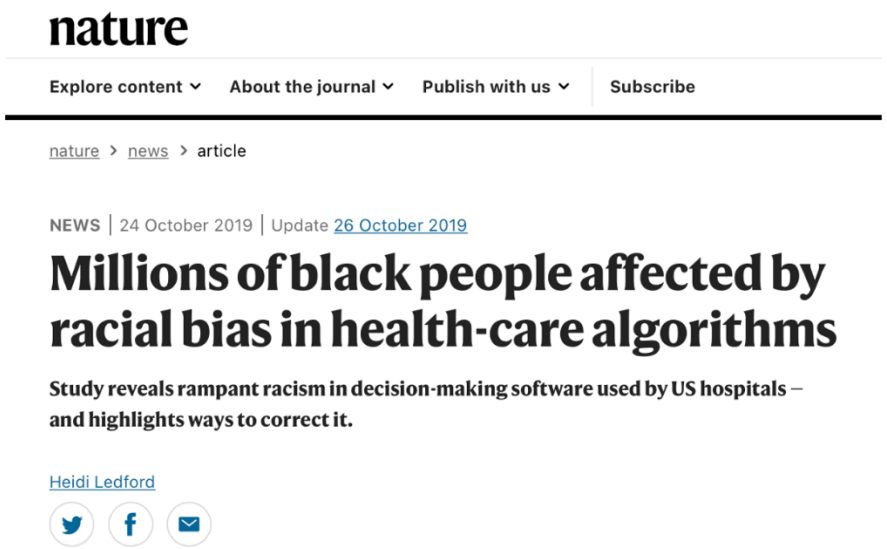
[1]Chair of Cartography and Visual Analytics, Technical University of Munich · [2]Senseable City Lab, MIT
[3]Geospatial Data Science Lab, University of Wisconsin-Madison · [4]Department of Geography, Florida State University

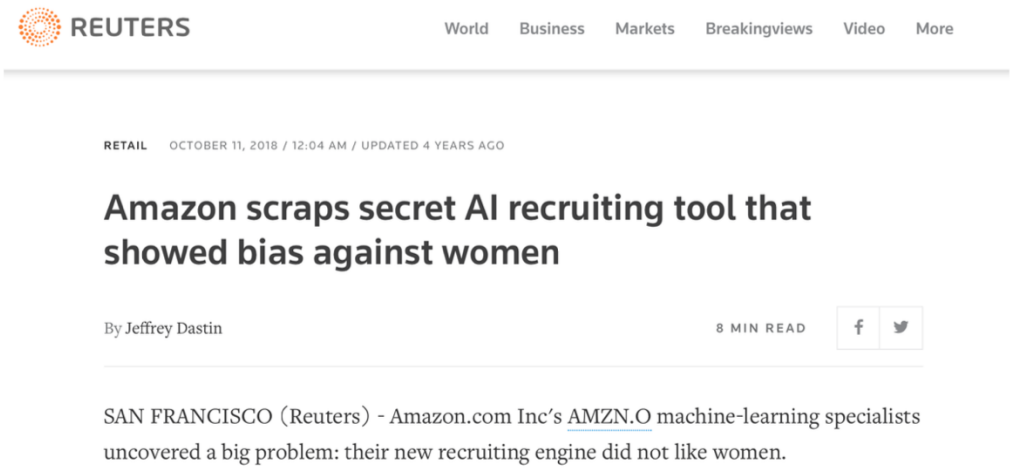
Black box AI


Input
Black Box
Output

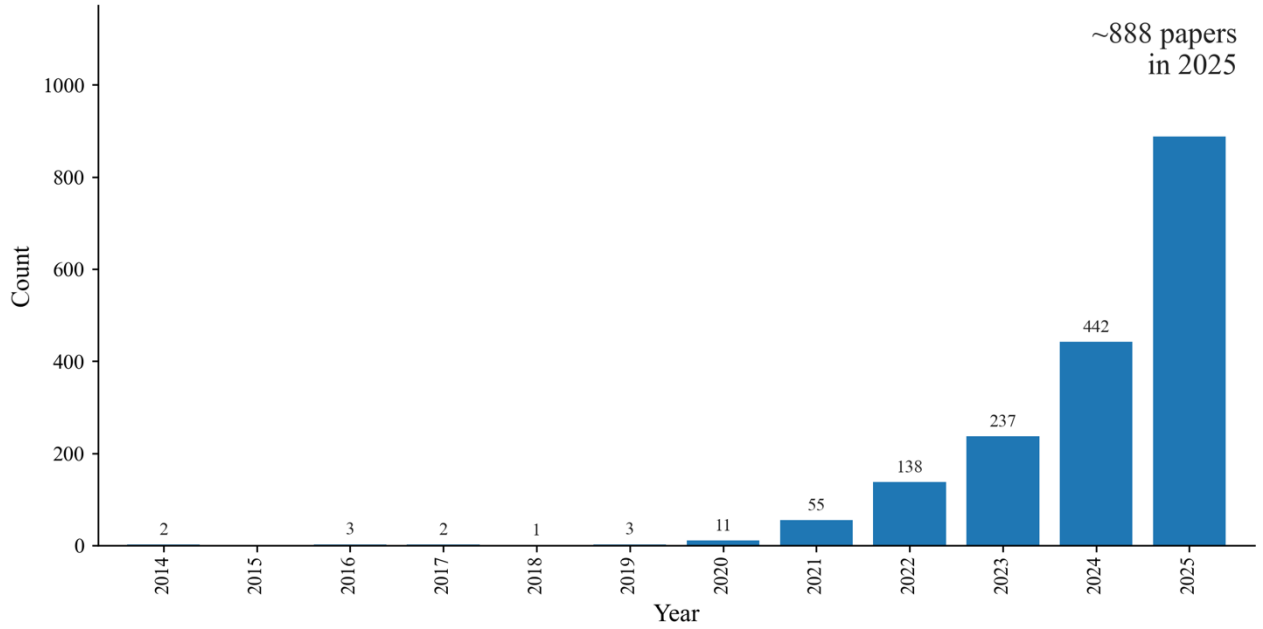

Why did you predict 42 for this data point?
awkward silence

Black box AI can cause problem

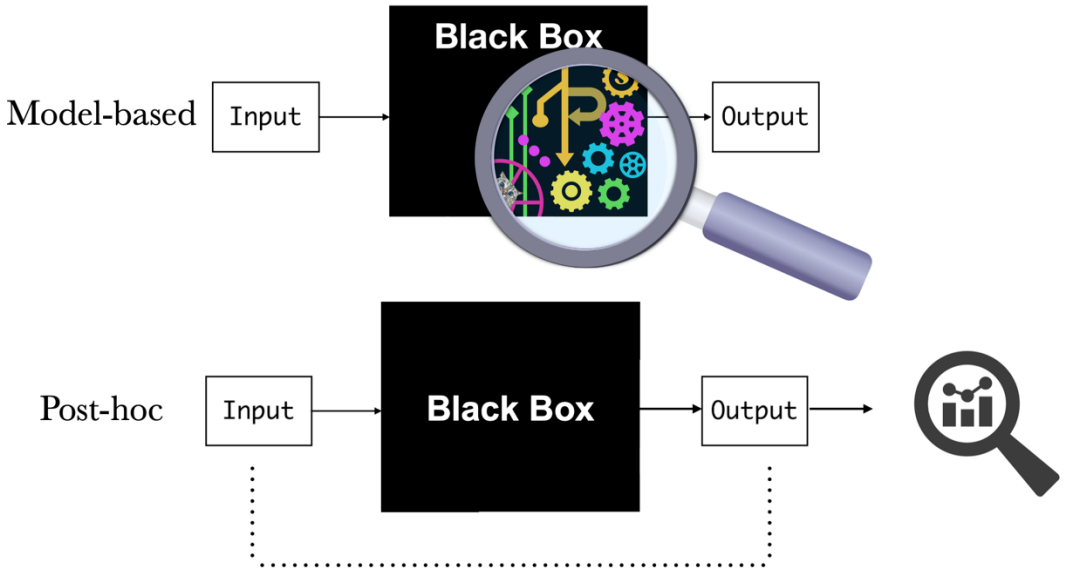




XAI is a rapid growing domain in Geography



Yearly number of papers that use XAI in several geography-related categories from Web of Science.



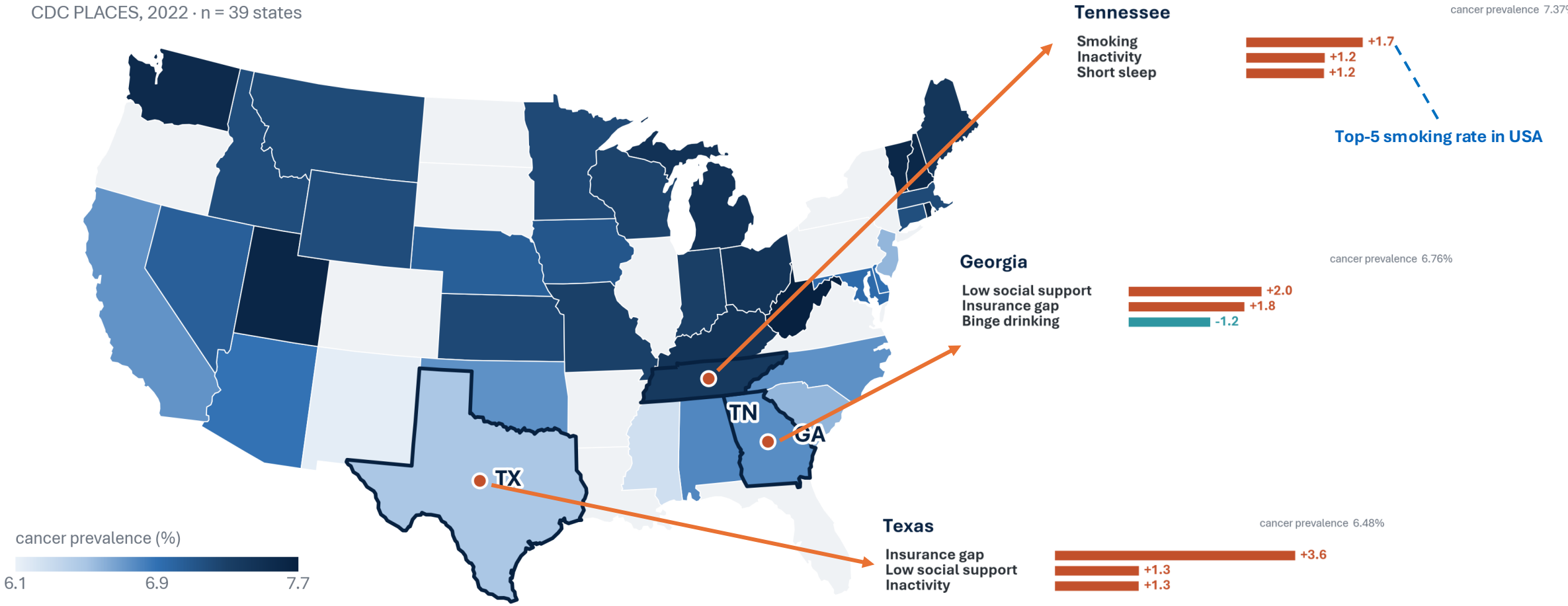
Murdoch et al., 2019

Understanding the spatial process with XAI

Cancer prevalence varies sharply across U.S. states — and a public-health officer needs the per-state lever, not a national average.

a **Age-adjusted cancer prevalence (%) by state**

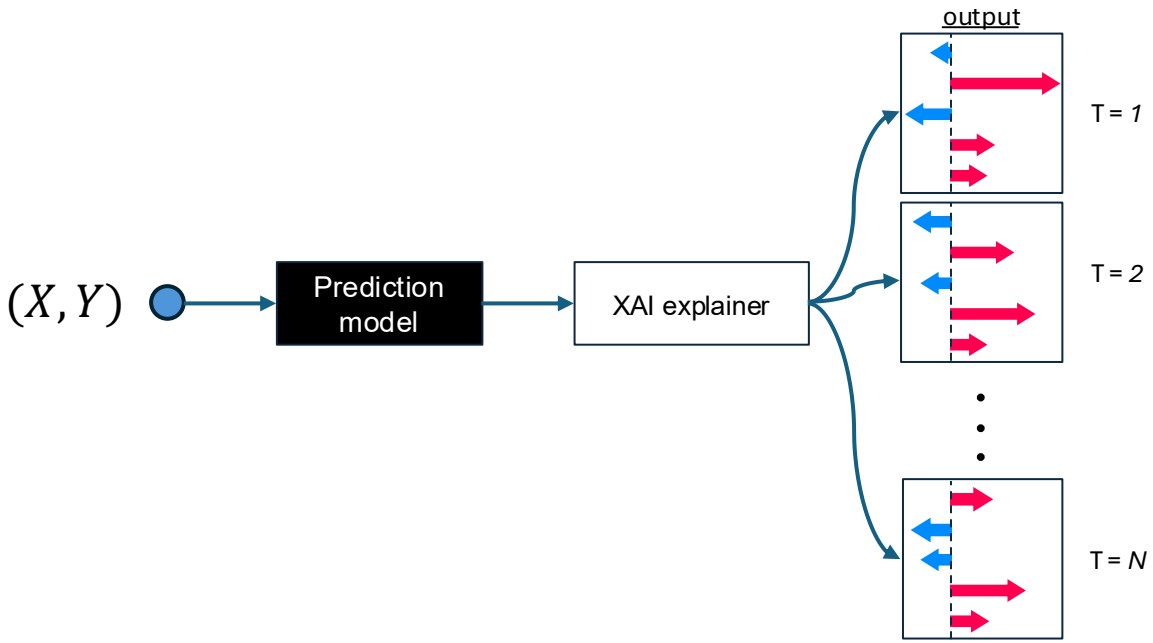
CDC PLACES, 2022 · n = 39 states



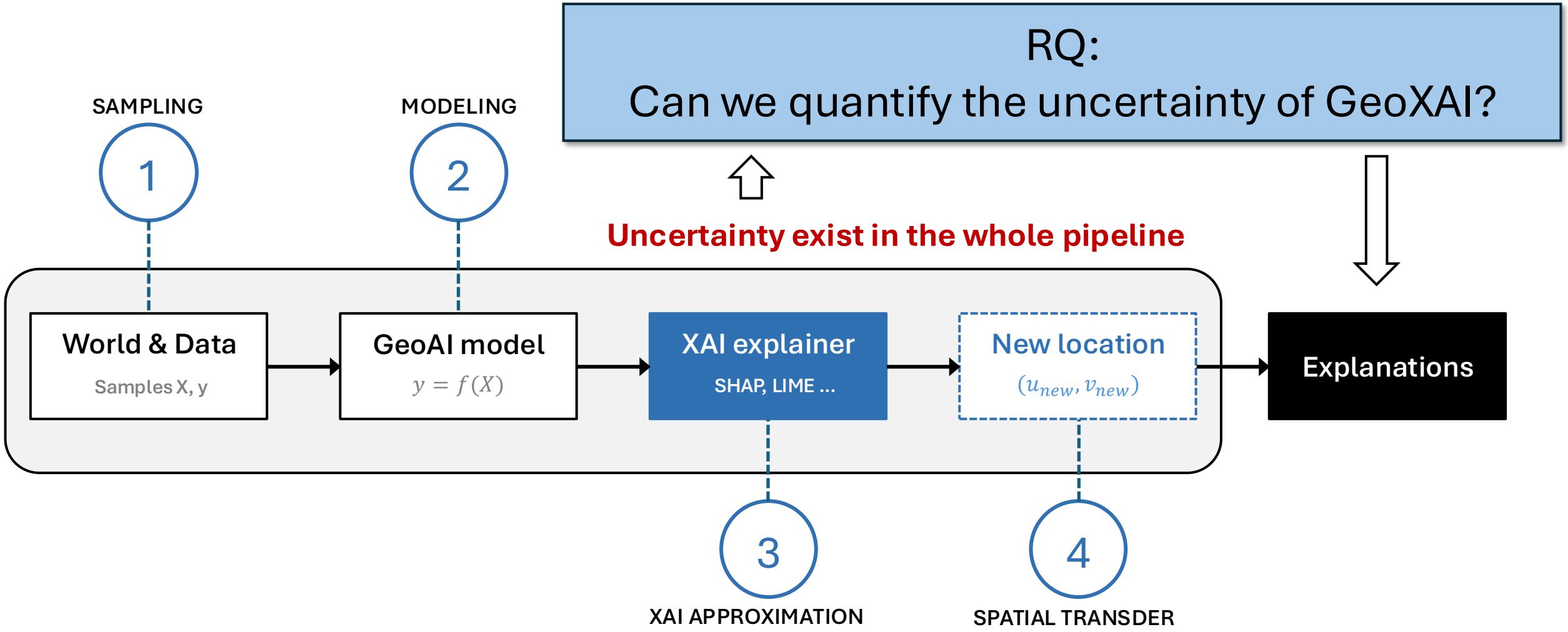
XAI is unstable in spatial tasks

A public-health officer in West Virginia runs SHAP three times — and gets three different answers.

> **The brief.** West Virginia has the highest cancer prevalence in the U.S. Which risk factor should the state act on first? The team trains an ML model on state-level data, runs KernelSHAP, and ships the per-state lever map. To be safe, they re-run it twice more — same data, same code, same seed unset.



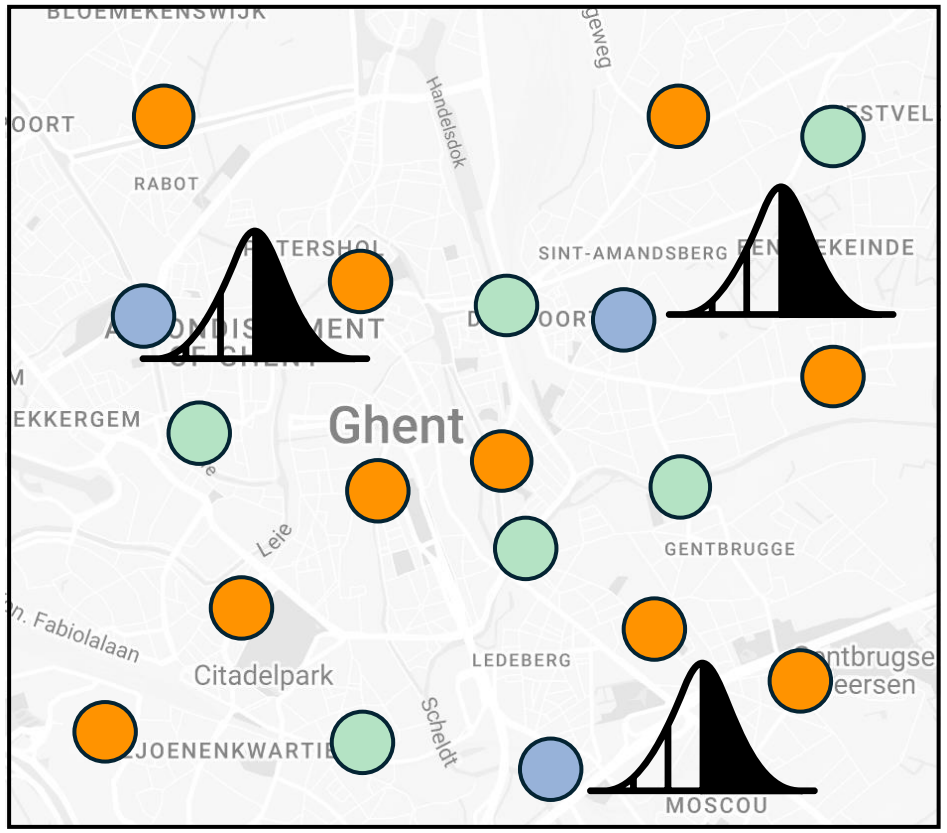
Uncertainty enters the pipeline at four points.

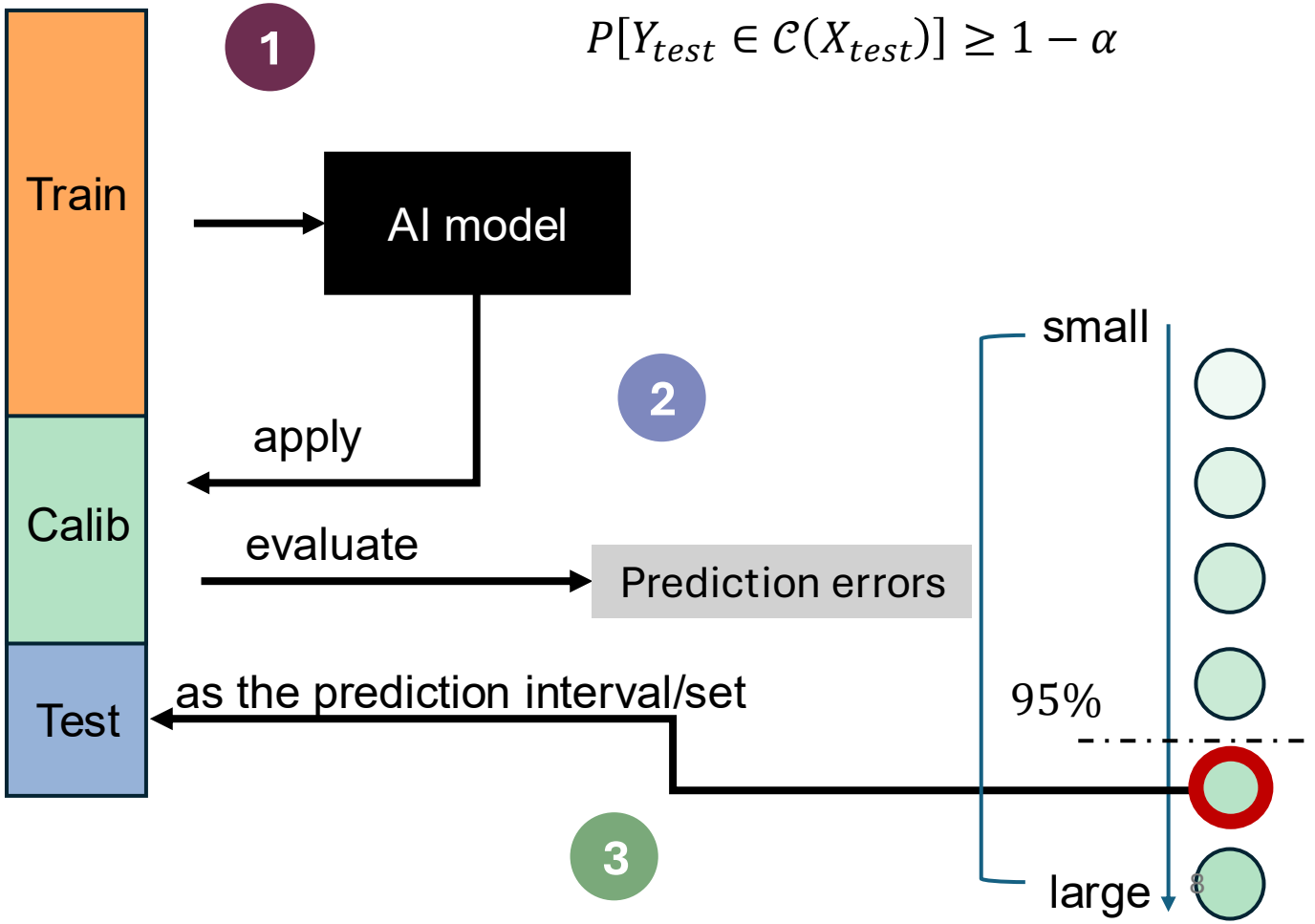


Conformal prediction

Exchangeability Assumption

Assume that housing prices everywhere come from the same underlying distribution





$$P[Y_{test} \in \mathcal{C}(X_{test})] \geq 1 - \alpha$$

Challenges of CP in GeoXAI

OBSTACLE 01

No ground truth for explanations.

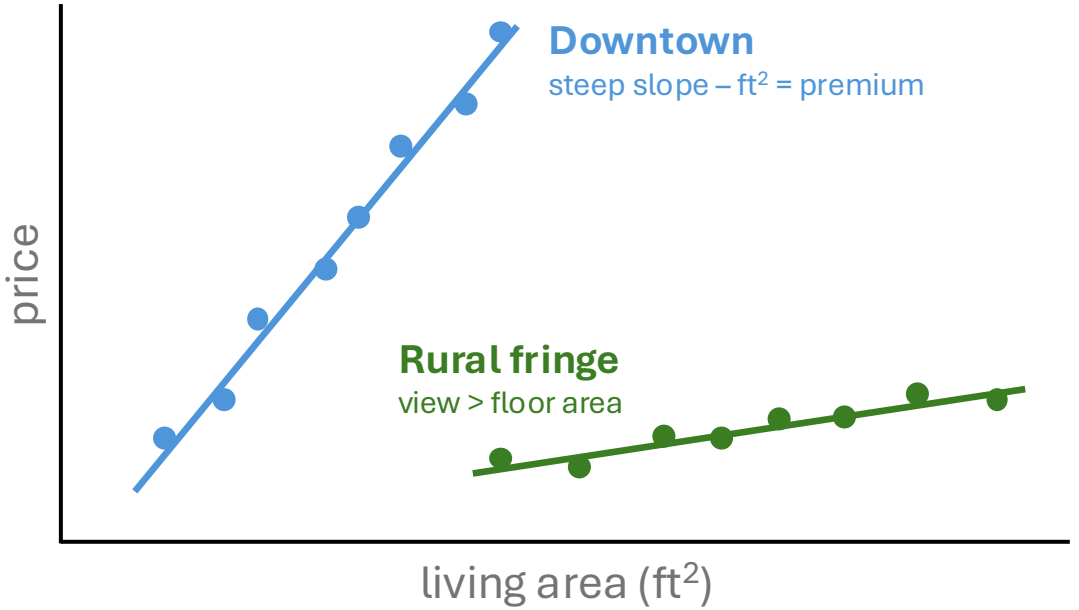
CP needs a nonconformity score that compares predicted to observed. For an explanation s, no observable target exists.

PREDICTION	$\alpha = \lvert y - \hat{y} \rvert$	**works !** 😀
EXPLANATION	$\alpha = \lvert s - ? \rvert$	no truth 😫

OBSTACLE 02

Space breaks exchangeability.

The same feature drives differently in different places – you can't shuffle calibration points without changing the model.



Solution #1: Train an XAI predictor to approximate the explainer

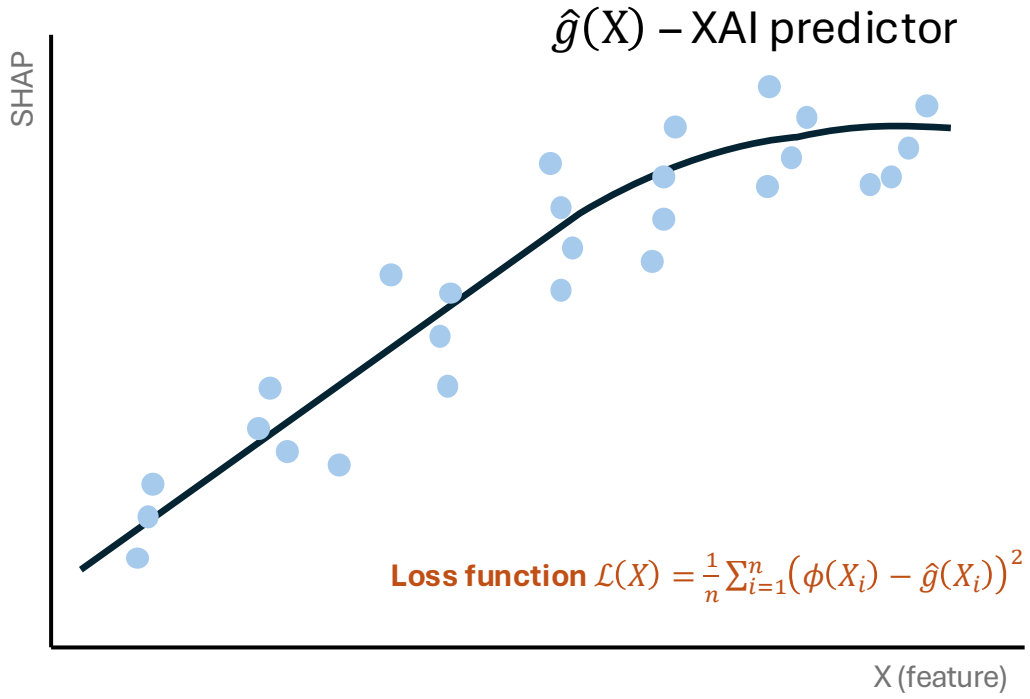
The XAI explainer Φ is stochastic: every run draws a different SHAP estimate.

We treat $\Phi: X \rightarrow \mathbb{R}^k$ as a multi-target regression and fit a model $\hat{g}$ on its outputs.

WHY IT WORKS

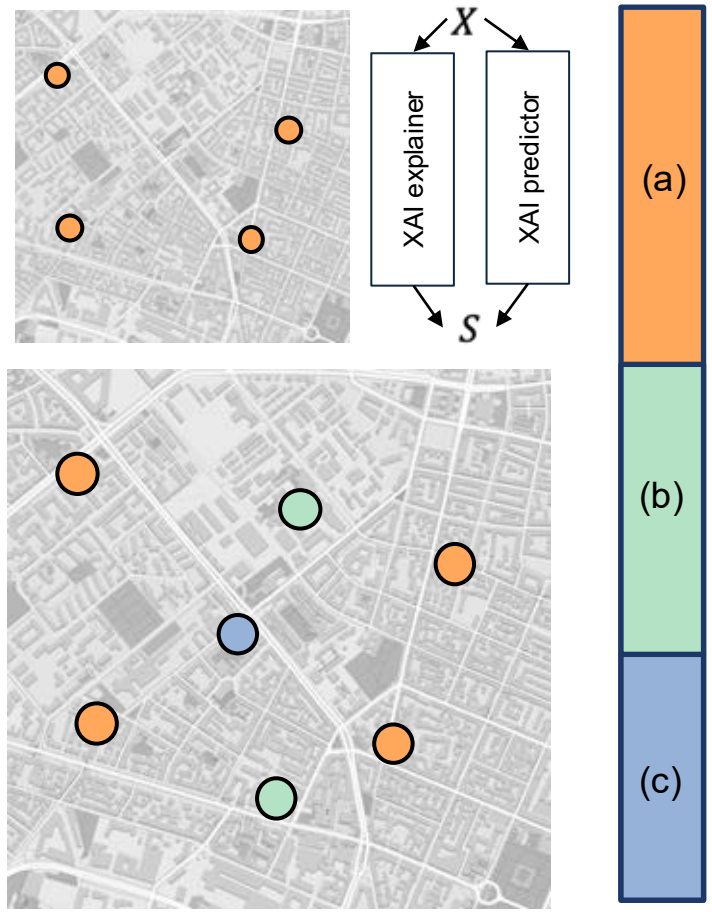
Minimizing mean squared loss leads to a conditional expectation.

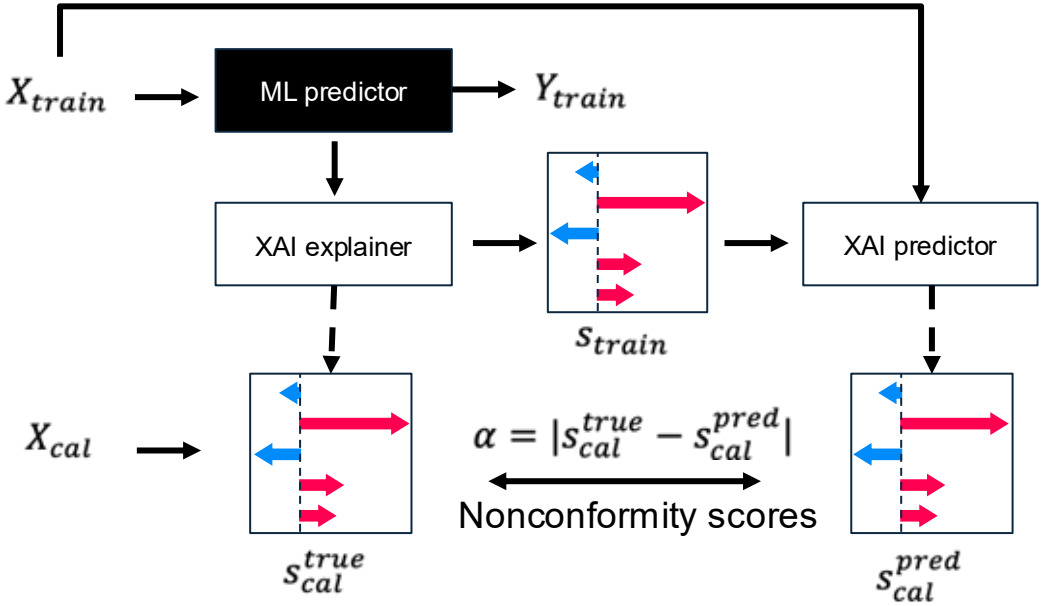
$$E[Y|X] = \arg\min_f E\left[\left(Y - f(X)\right)^2\right]$$



Solution #1: Train an XAI predictor to approximate the explainer

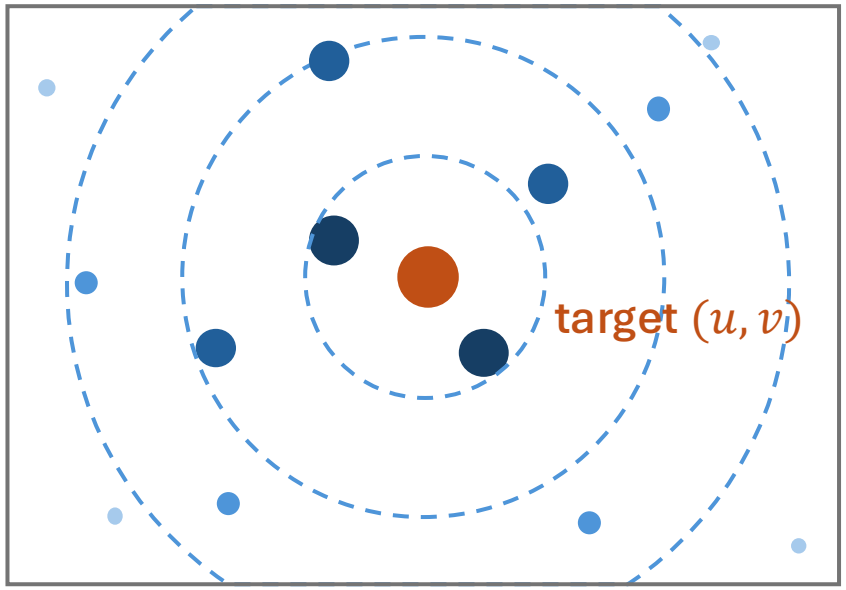
(a) Train two XAI models

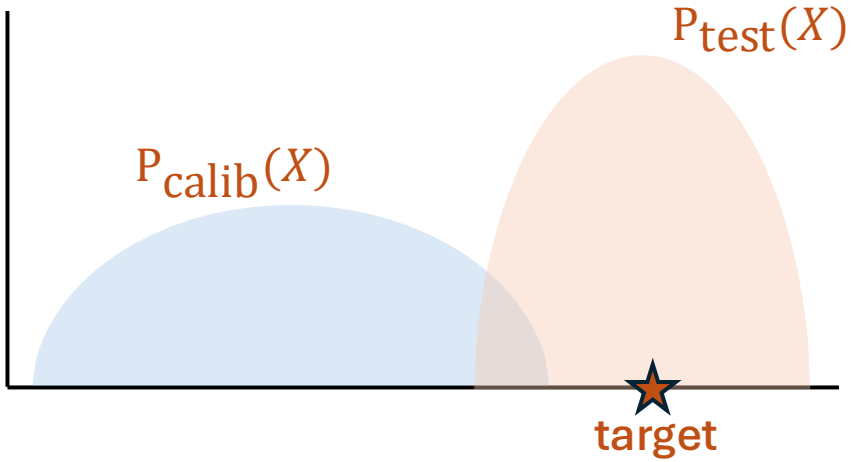




Solution #2: Geo-weighted quantile

Closer points count more. Tobler, doing the work.





GAUSSIAN KERNEL

$$w_i = \exp\left(-\frac{d_i^2}{b^2}\right)$$

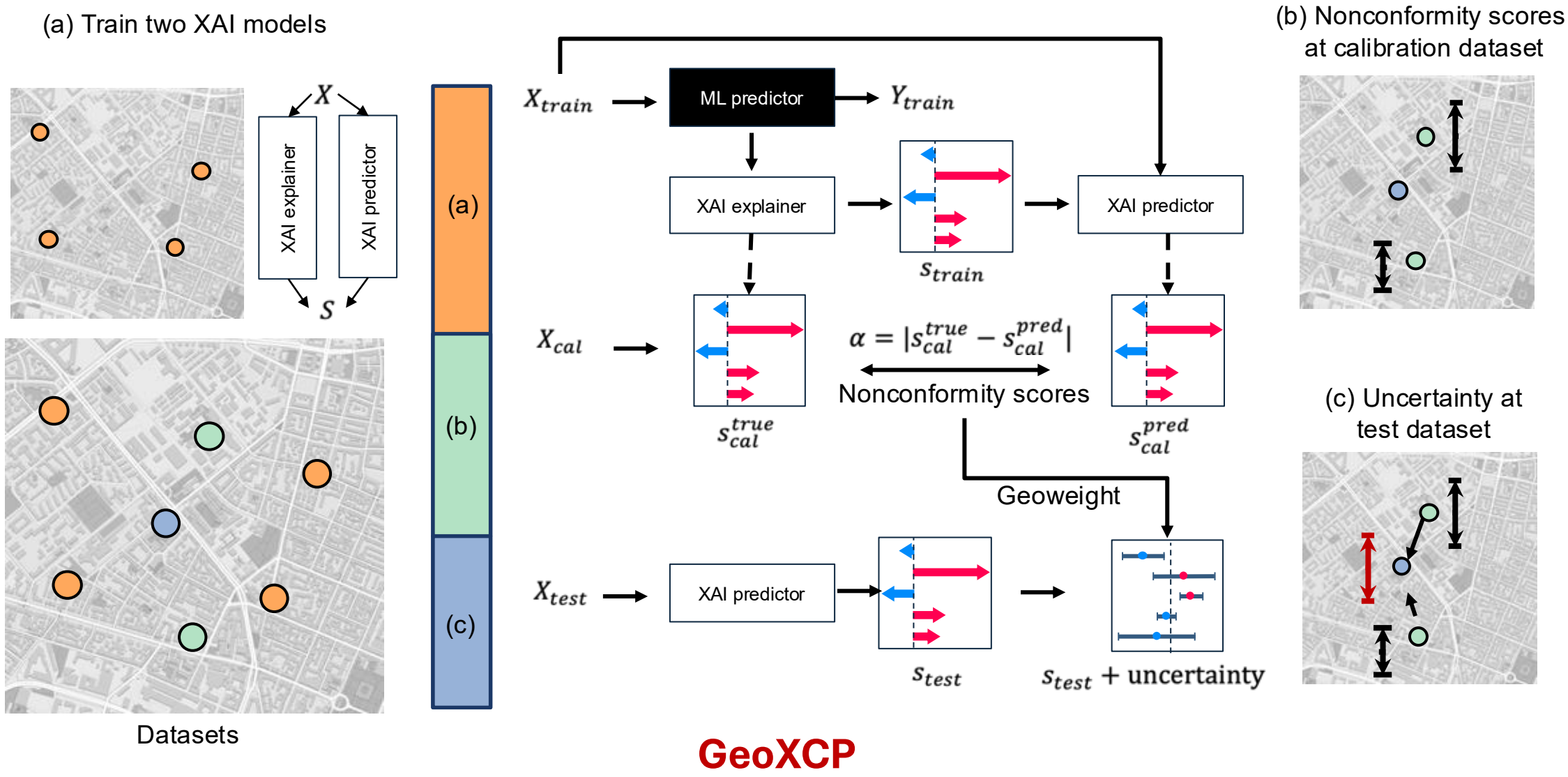
For target location (u, v), weight every calibration point i by its geographic distance.

Take the weighted $\lceil (n+1)(1-\varepsilon) \rceil / n$ quantile of $\{\alpha_i\} \rightarrow \hat{q}(u, v)$.

GUARANTEE

Under *local exchangeability*, coverage is preserved location-by-location.

Conformal Prediction for GeoXAI



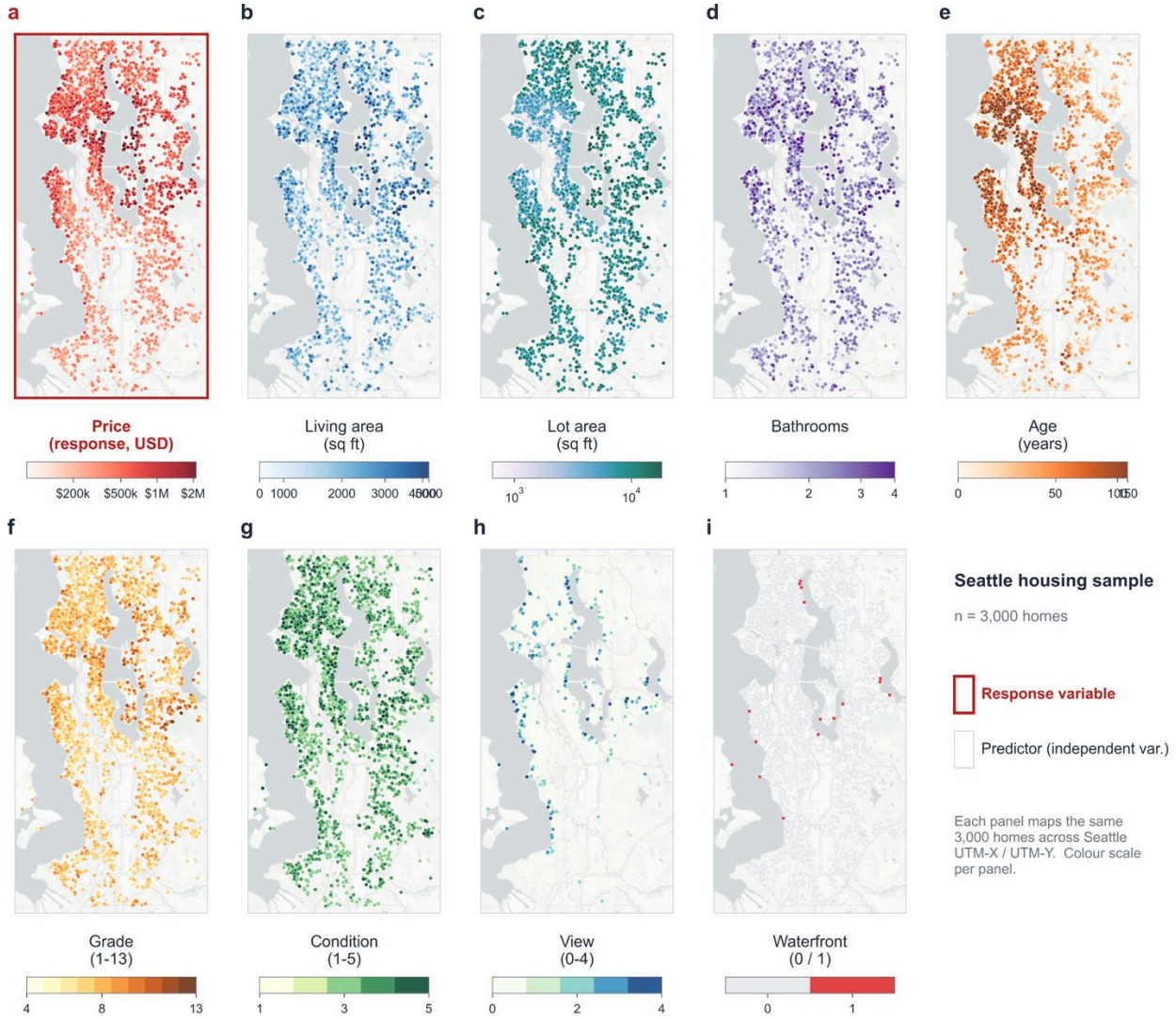
GeoXCP

Lou, X., Luo, P., Li, Z., Gao, S. and Meng, L., 2025. GeoXCP: uncertainty quantification of spatial explanations in explainable AI. *International Journal of Geographical Information Science*, pp.1-31.

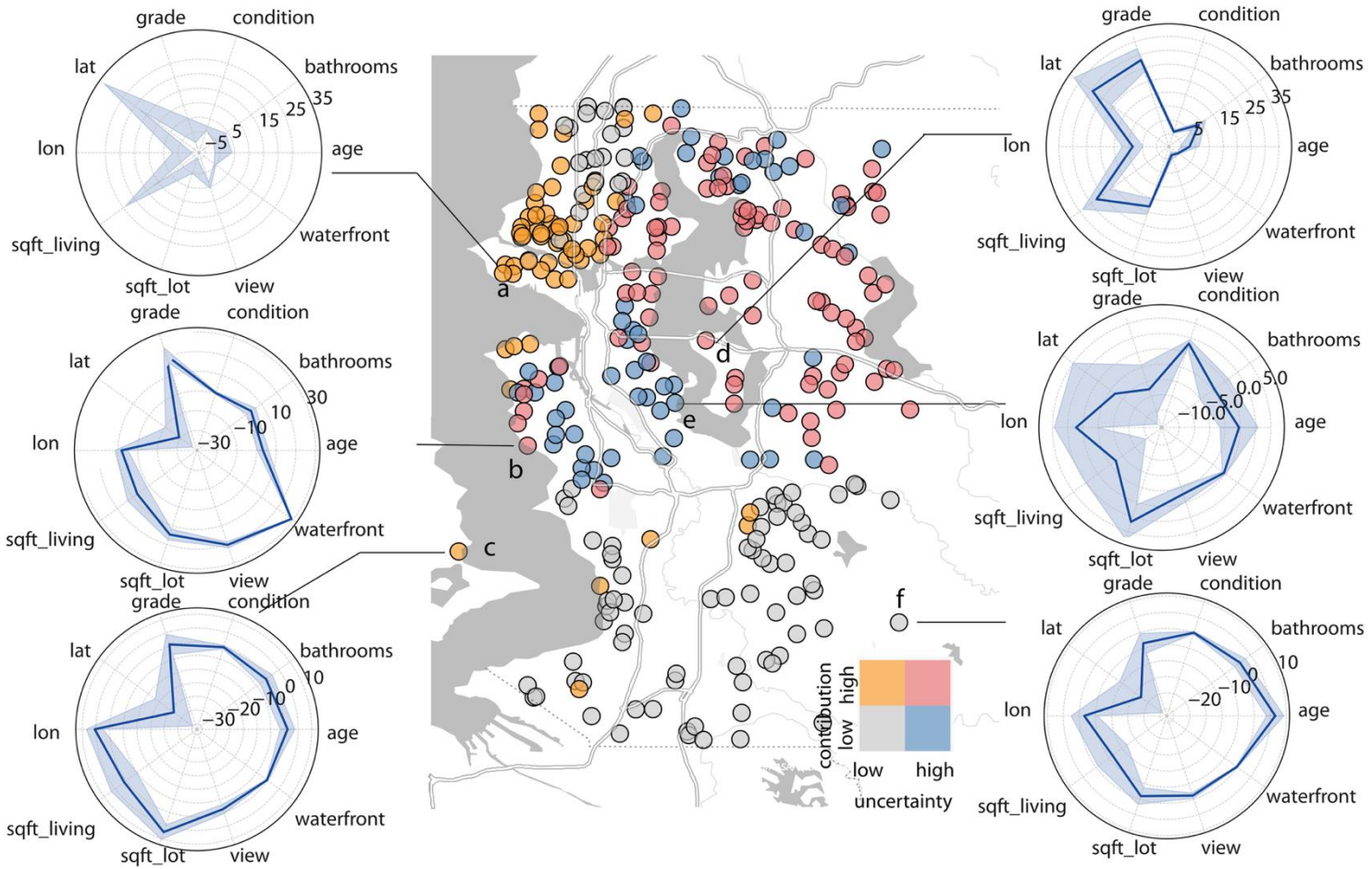
Home sales in Seattle as an example

3,000 home sales, 2014-15 · 8 explanatory variables + (lat, lon)

XGBoost · test $R^2 = 0.87$ · KernelSHAP wrapped by GeoXCP



Where to trust SHAP – and where not.



Each point = one home; radar callouts (a-f) show the per-variable SHAP profile and uncertainty band.

HOW TO READ IT

CONTRIBUTION →

UNCERTAINTY →

- **Act on it**. High signal, tight band.
- **Investigate**. Strong net effect, but unstable.
- **Distrust**. Net unclear–wide band, value could shift.
- **Net ~ 0**. Effects may cancel—inspect per-variable.

WHAT THE CALLOUTS REVEAL

a Downtown–**sqft_living** and **lat** drive the price, with relatively tight uncertainty band.

d East side of lake–**sqft_living**, **sqft_lot** and **grade** contribute strongly, but bands are wide.

e West side of lake—Net effect is close to zero, but very unstable.

f Rural south—summed contribution sits near zero and we can trust.

IF YOU **REMEMBER** ONE THING

A spatial explanation isn’t a number. It’s a distribution over places – and now we can measure it.

Lou, X., Luo, P., Li, Z., Gao, S. and Meng, L., 2025. GeoXCP: uncertainty quantification of spatial explanations in explainable AI. *International Journal of Geographical Information Science*, pp.1-31.

Thank you.

CONTACT

Xiayin Lou

Chair of Cartography and Visual Analytics

xiayin.lou@tum.de

Interactive version of GeoXCP

Paper

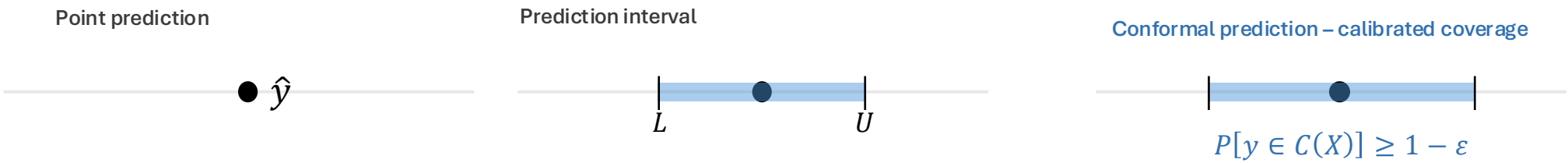
Code

Our pipeline forces three requirements – only conformal prediction meets them all.

WHAT OUR PIPELINE REQUIRES

- **A** Distribution free
- **B** Post-hoc – no retraining
- **C** Finite-sample guarantee

FIVE UQ FAMILIES	A	B	C
Bayesian UQ heaviest assumption · credible interval ≠ coverage guarantee	⊗	⊗	⊗
Ensemble UQ only ensemble-based GeoAI models apply	~	⊗	⊗
Bootstrapping UQ coverage holds only as N → ∞ · expensive on large datasets	✓	✓	⊗
Evidential Deep Learning needs a specific NN architecture · no formal coverage promise	⊗	⊗	⊗
Conformal Prediction exchangeability · exact 1 - ε coverage · wraps any model	✓	✓	✓



From a calibration set to a guarantee.

STEP 1 · SPLIT & TRAIN

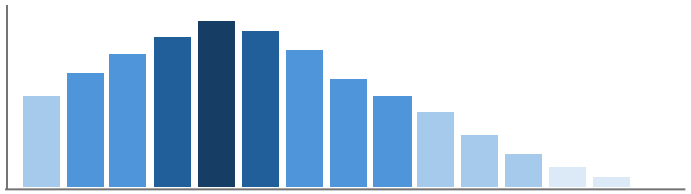
TRAINING	CALIBRATION	TEST
fit $\hat{f}$	held out n points	new X

STEP 2 · NONCONFORMITY SCORE

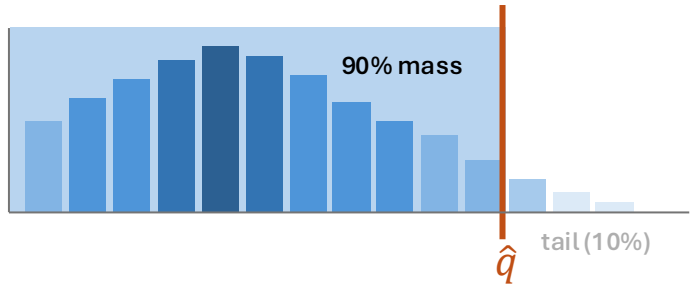
$$\alpha_i = |y_i - \hat{f}(X_i)|$$

How strange or unusual the model is by point i

Distribution of α_i on calibration

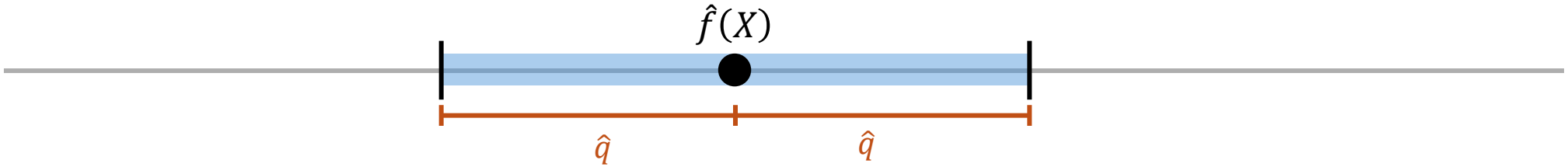


STEP 3 · TAKE THE $\lceil (n+1)(1-\varepsilon) \rceil / n$ QUANTILE

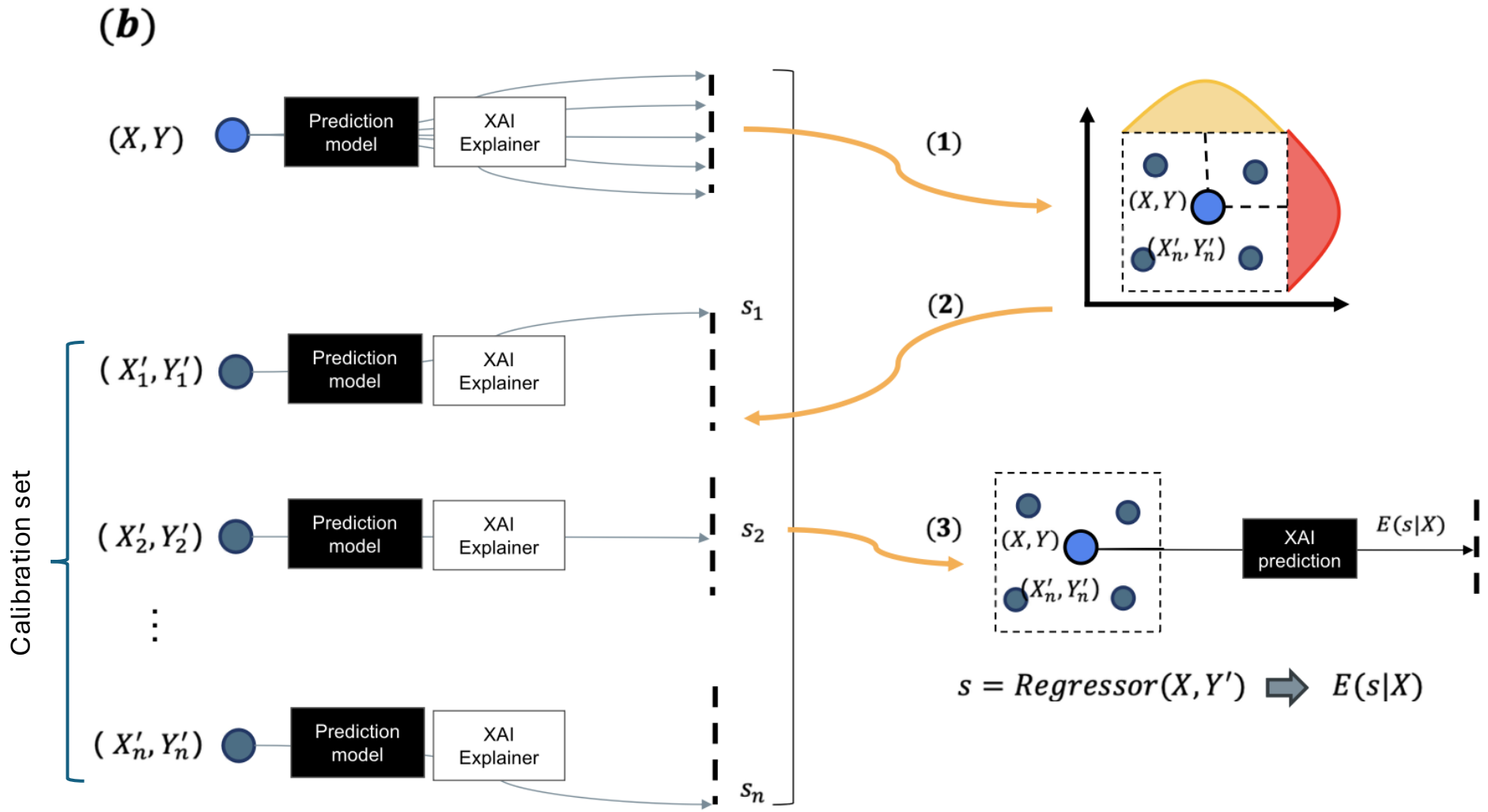


STEP 4 · BUILD THE INTERVAL AT TEST POINT

$$C(X) = \hat{f}(X) \pm \hat{q}$$



Solution #1: Train an XAI predictor to approximate the explainer



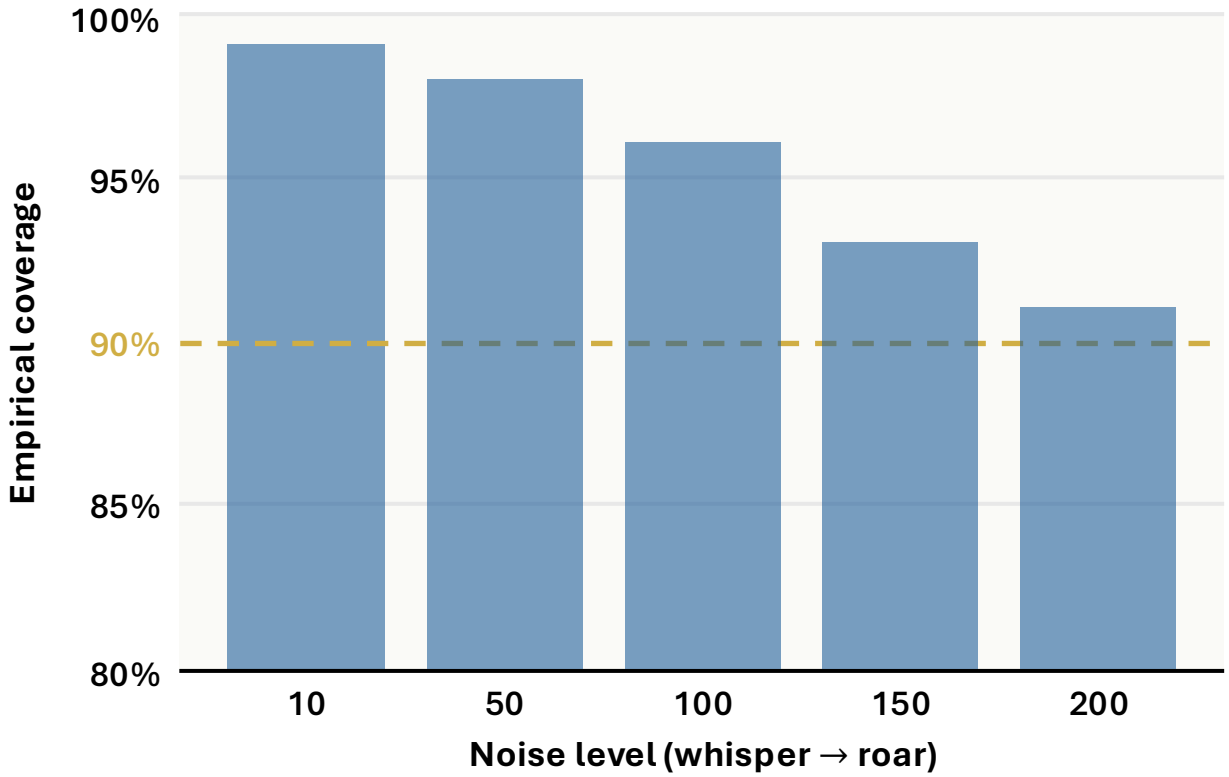
Yes – at every noise level, GeoXCP kept its 90% promise.

HOW WE TESTED IT

1 Build a world
two variables vary in space, two stay constant – every true SHAP value known

2 Form the noise
magnitude 10 → 200, Gaussian + Uniform – from a whisper to a roar

3 Count the coverage
how often did the truth land inside our interval? promise ≥ 90%



EVEN AT THE LOUDEST NOISE

91%

of the time, the true SHAP value sat inside our interval – promise kept.

And the interval widens honestly as noise grows – GeoXCP doesn't just stay calibrated, it tells you when to be more cautious.

The conformal guarantee isn't just theory. It survives contact with data.

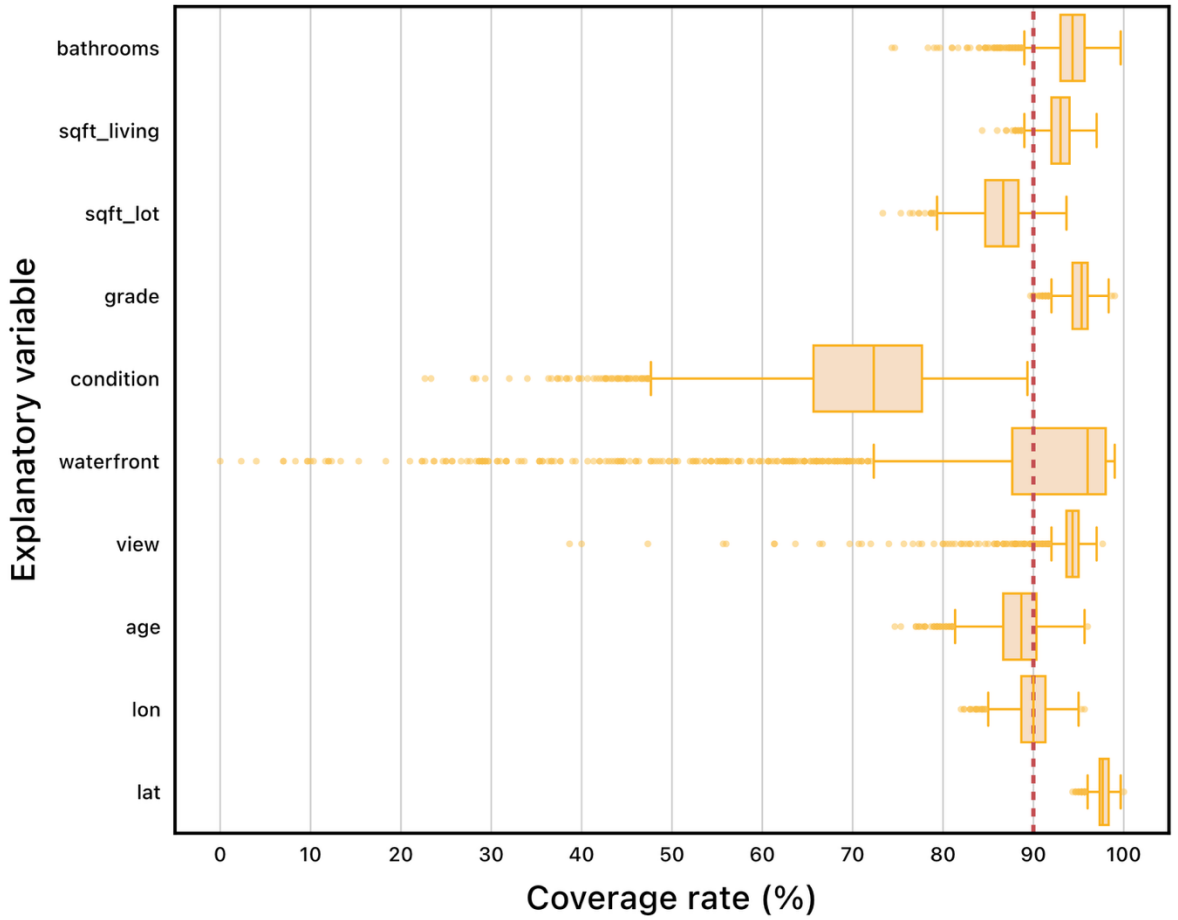
Intervals are calibrated for most variables.

Resample & retrain 2,000×. For every location, check what fraction of bootstrap SHAP values fall inside the GeoXCP interval from a single fit.

≥90%

Median coverage for most variables

Condition falls short – uncertainty is under-estimated for that variable.

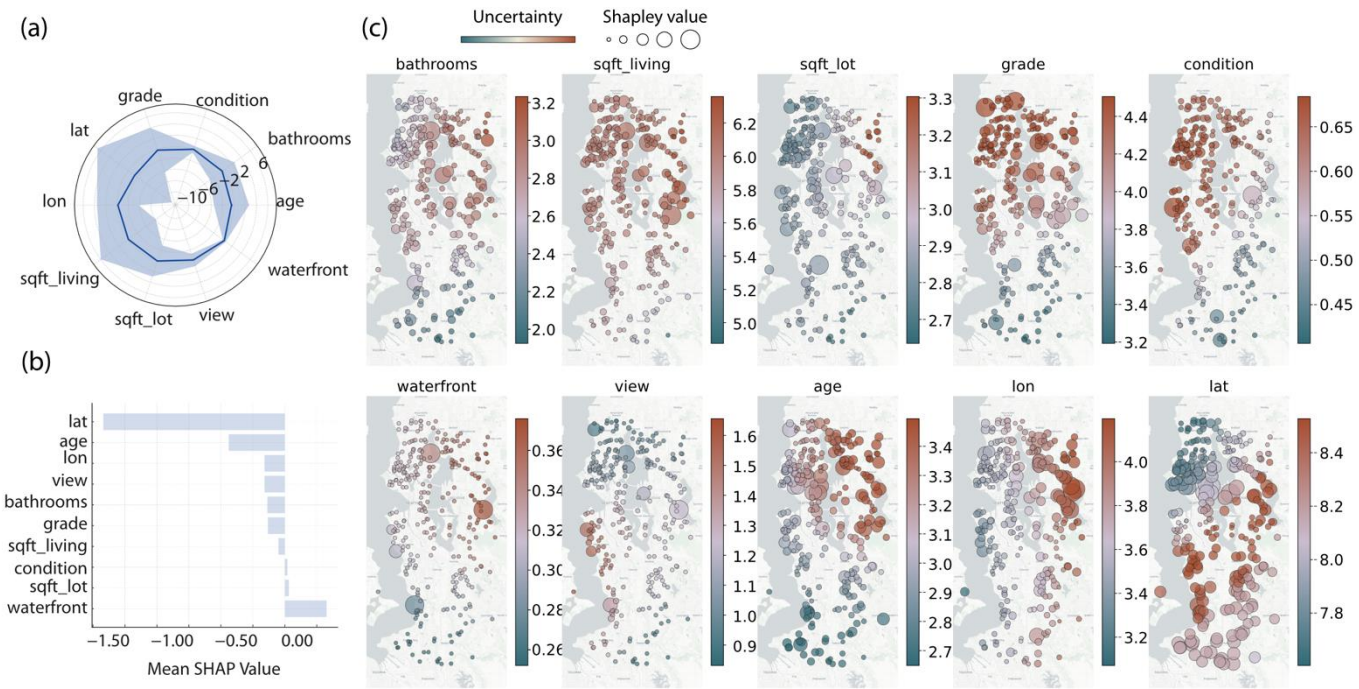


Results

Where to trust SHAP – and where not.

Seattle, WA · 3,000 home sales, 2014-15 · 8 explanatory variables + (lat, lon)

XGBoost · test $R^2 = 0.87$ · KernelSHAP wrapped by GeoXCP



For each of 8 variables: point size = |Shapley value| · color = uncertainty (red high, teal low)

WHAT THE MAPS REVEAL

- **sqft_living** – high uncertainty almost everywhere.
- **sqft_lot** – stable across the region.
- **grade & condition** – trustworthy in the south, unreliable in the north.
- **view**– unstable along the western coast.

A flat SHAP map hides all of this.

Next: can we see contribution and uncertainty together?

What GeoXCP doesn’t do – yet.

01 · ASSUMPTION

Weighting uses geographic distance only.

GeoXCP assumes local exchangeability holds along geographic proximity – yet distant places with similar feature configurations can share explanatory regimes. Pure geo-weighting misses that channel.

02 · MODEL

The XAI predictor itself adds noise.

Without ground-truth explanations, $\hat{g}$ stands in for the conditional expectation – a useful proxy that carries its own model-included error. Simulations show coverage stays above 75% even under heavy predictor noise.

03 · COST

Geographic weighting isn’t free.

Each test point requires a weighted quantile over the full calibration set – $O(n)$ per query – and KernelSHAP itself is sampling-based. Large datasets compound both costs.