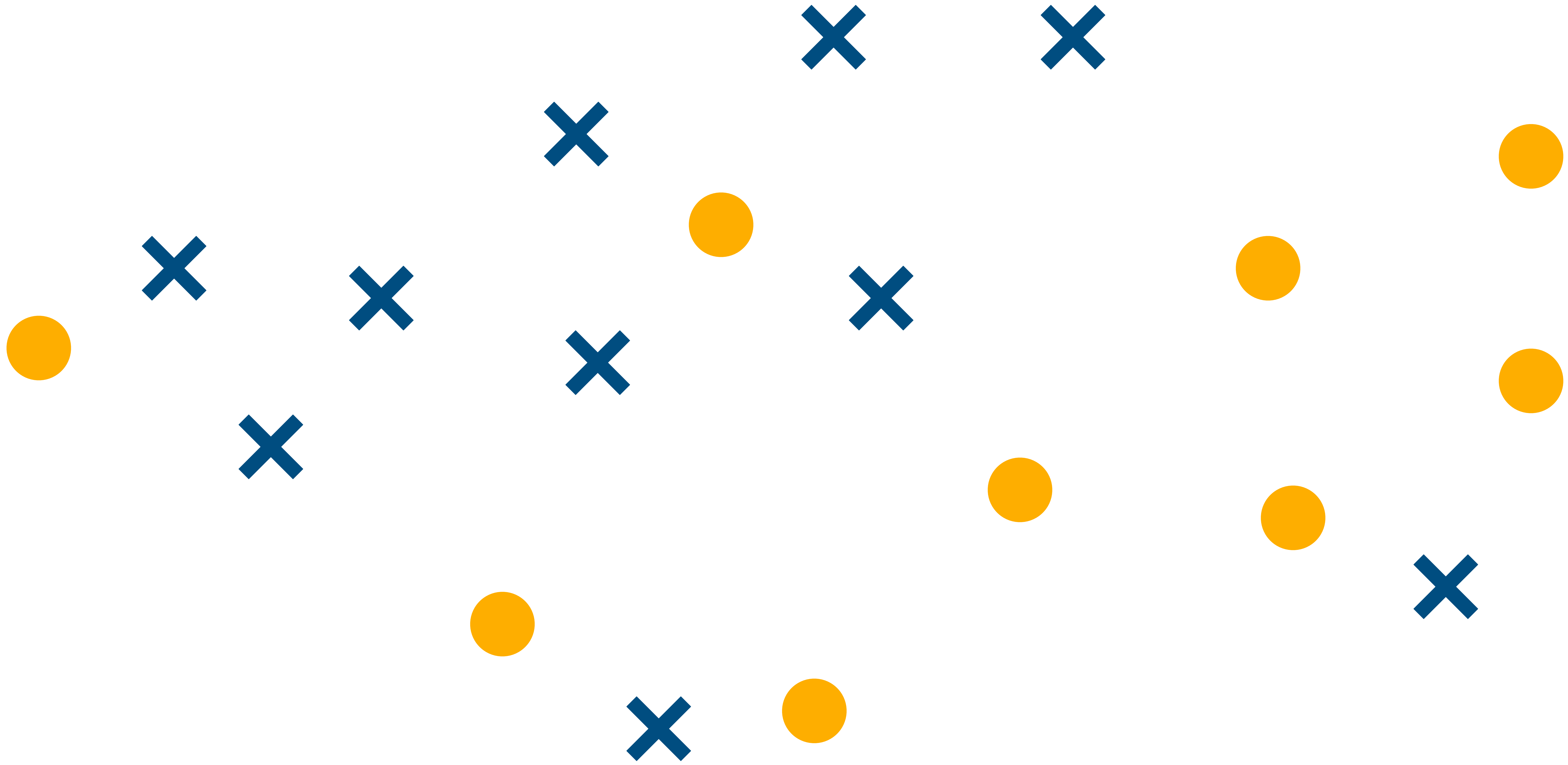
Consistency of the k_n-nearest neighbor rule under adaptive sampling

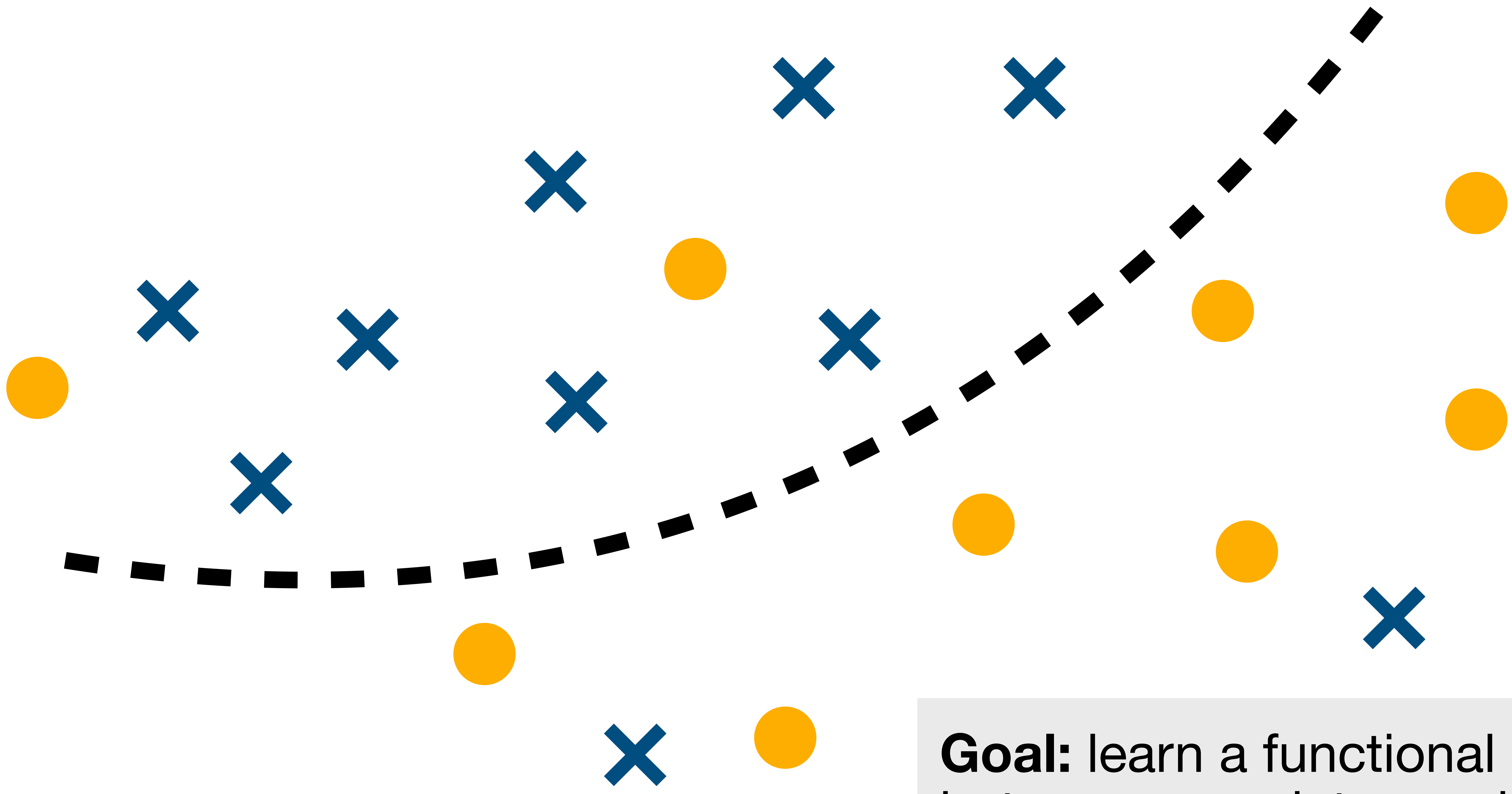
Robi Bhattacharjee, Sanjoy Dasgupta, Geelon So

University of Tübingen and Tübingen AI Center | University of California, San Diego

Binary Classification



Binary Classification



Goal: learn a functional relationship between covariates and responses.

Learning Setting

Let $\mathscr{X} \times \mathscr{Y}$ be a data space.

- The underlying relationship is defined by $\eta : \mathscr{X} \to [0,1]$, where:

$$\eta(x) \equiv \Pr\left(Y = 1 \,\middle|\, X = x\right).$$

- The Bayes-optimal prediction:

$$f^\star(x) = \mathbf{1}\left\{\eta(x) \geq 1/2\right\}.$$

Goal: make predictions that are consistent with the Bayes-optimal predictor.

I.I.D. Sampling

Most often in learning theory, data comes i.i.d. from a distribution:

$$X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} \mu \,.$$

Adaptive Sampling

What happens if X_n can depend on previously observed data?

- X_n is selected with knowledge of $(X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1})$.

Adaptive Sampling

Adaptive Sampling

?

Adaptive Sampling

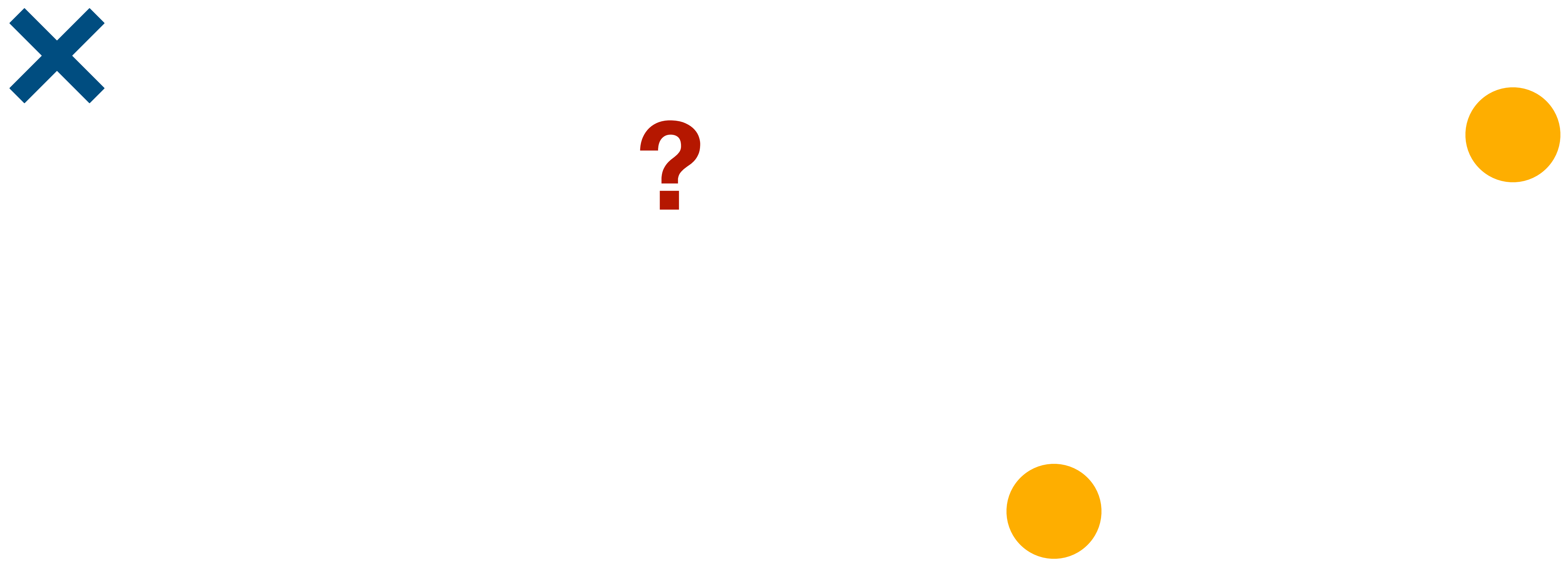
Adaptive Sampling

Adaptive Sampling

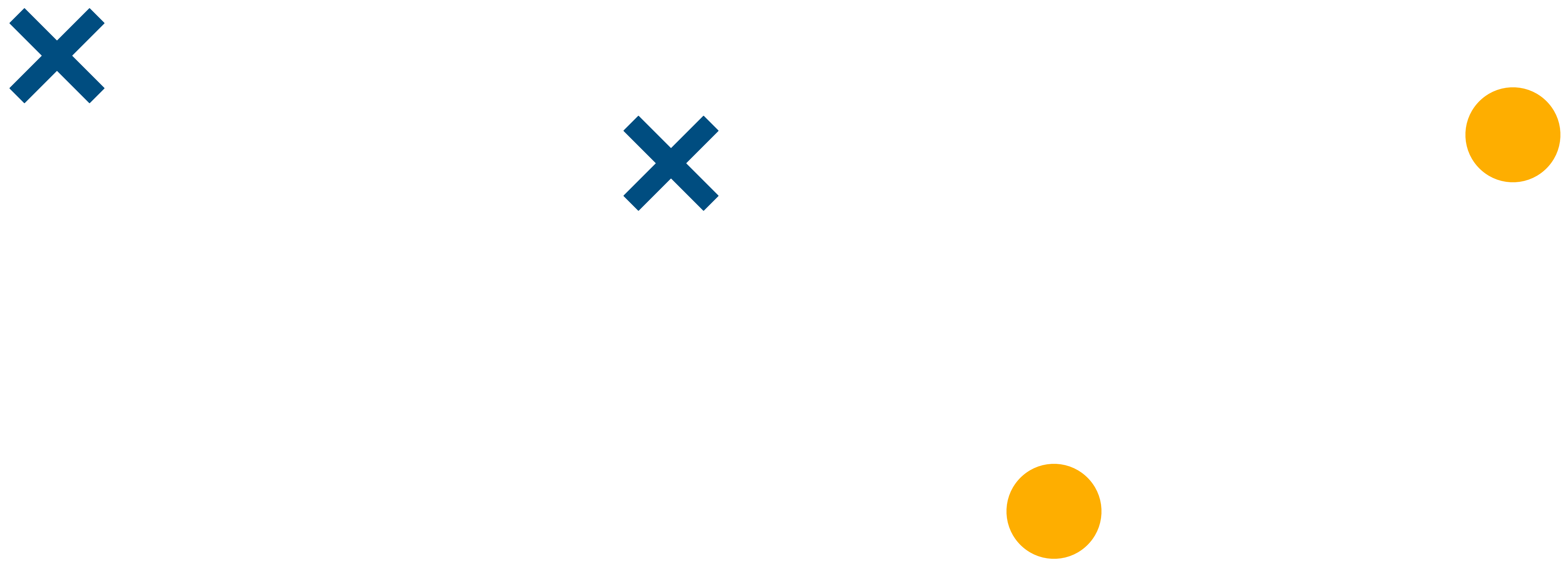
Adaptive Sampling

Adaptive Sampling

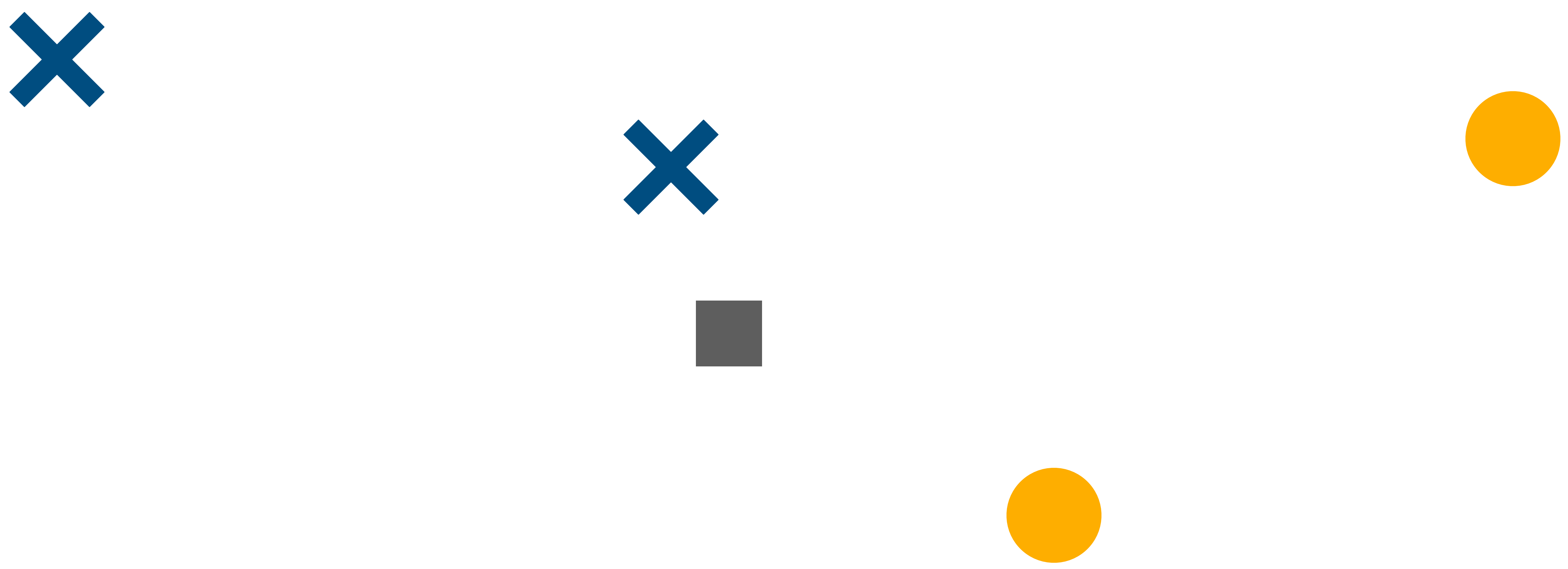
Adaptive Sampling



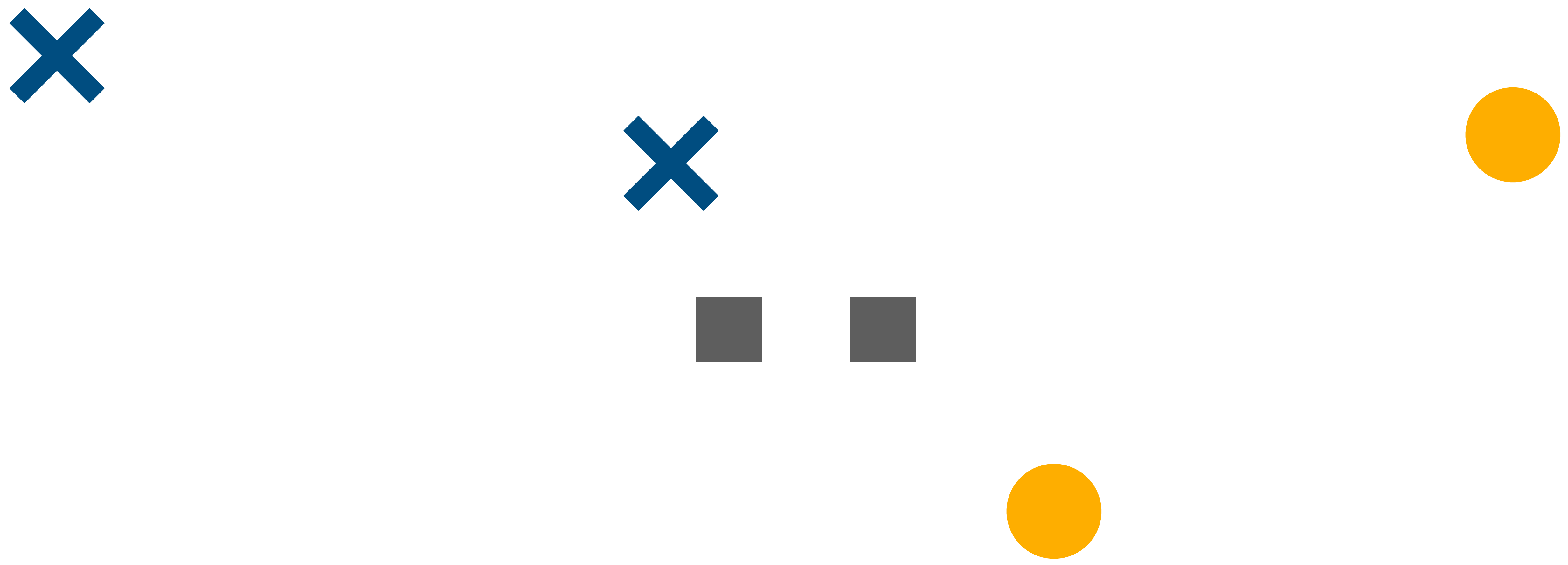
Adaptive Sampling



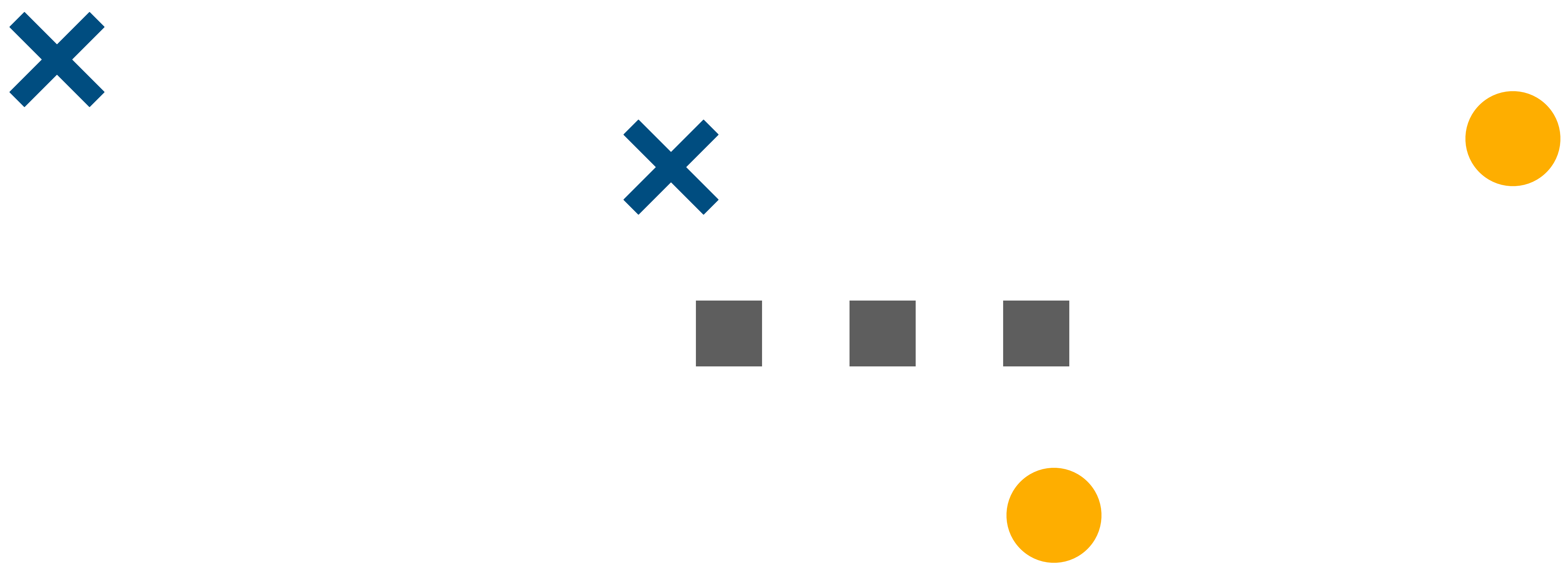
Adaptive Sampling



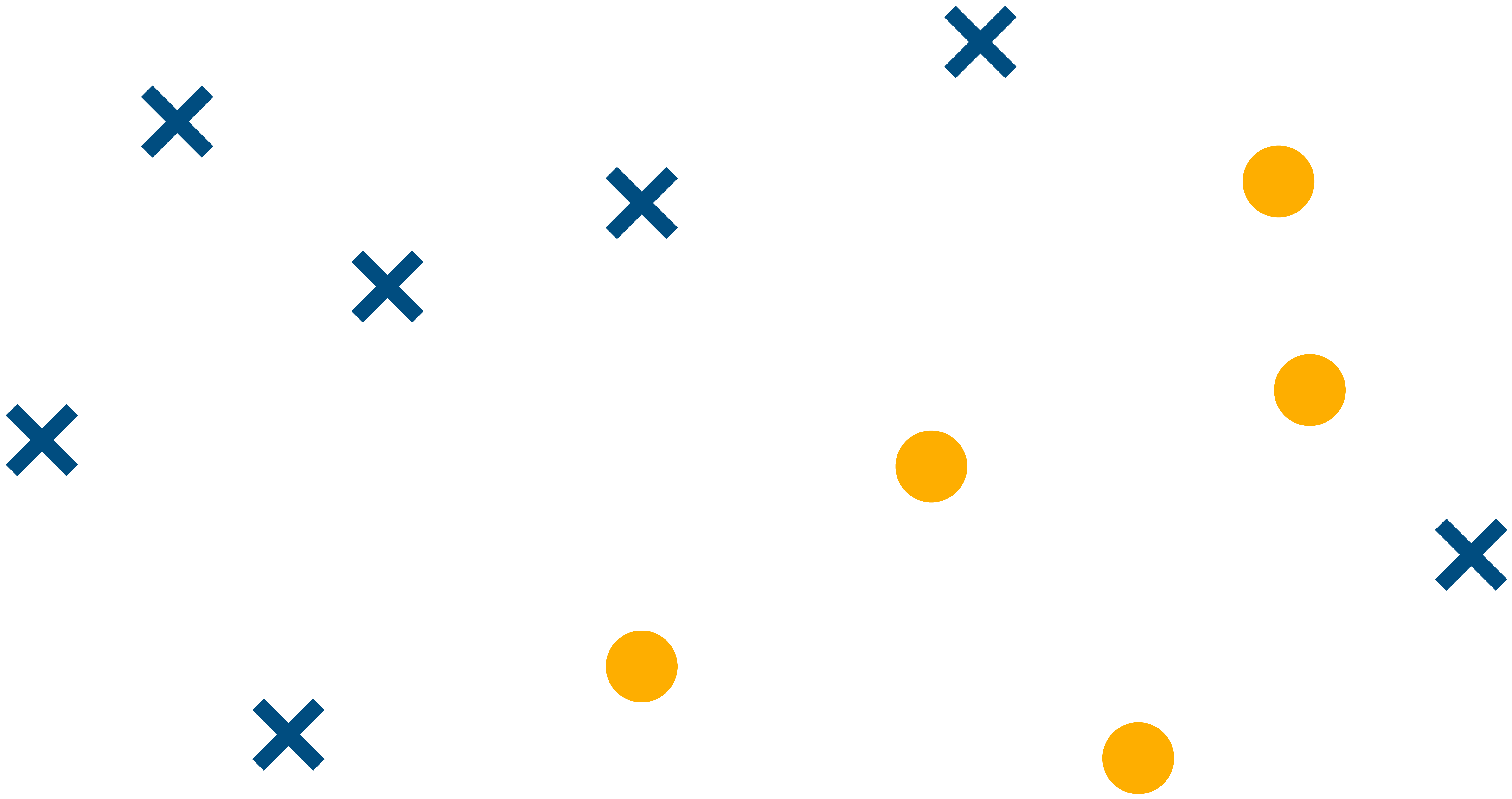
Adaptive Sampling



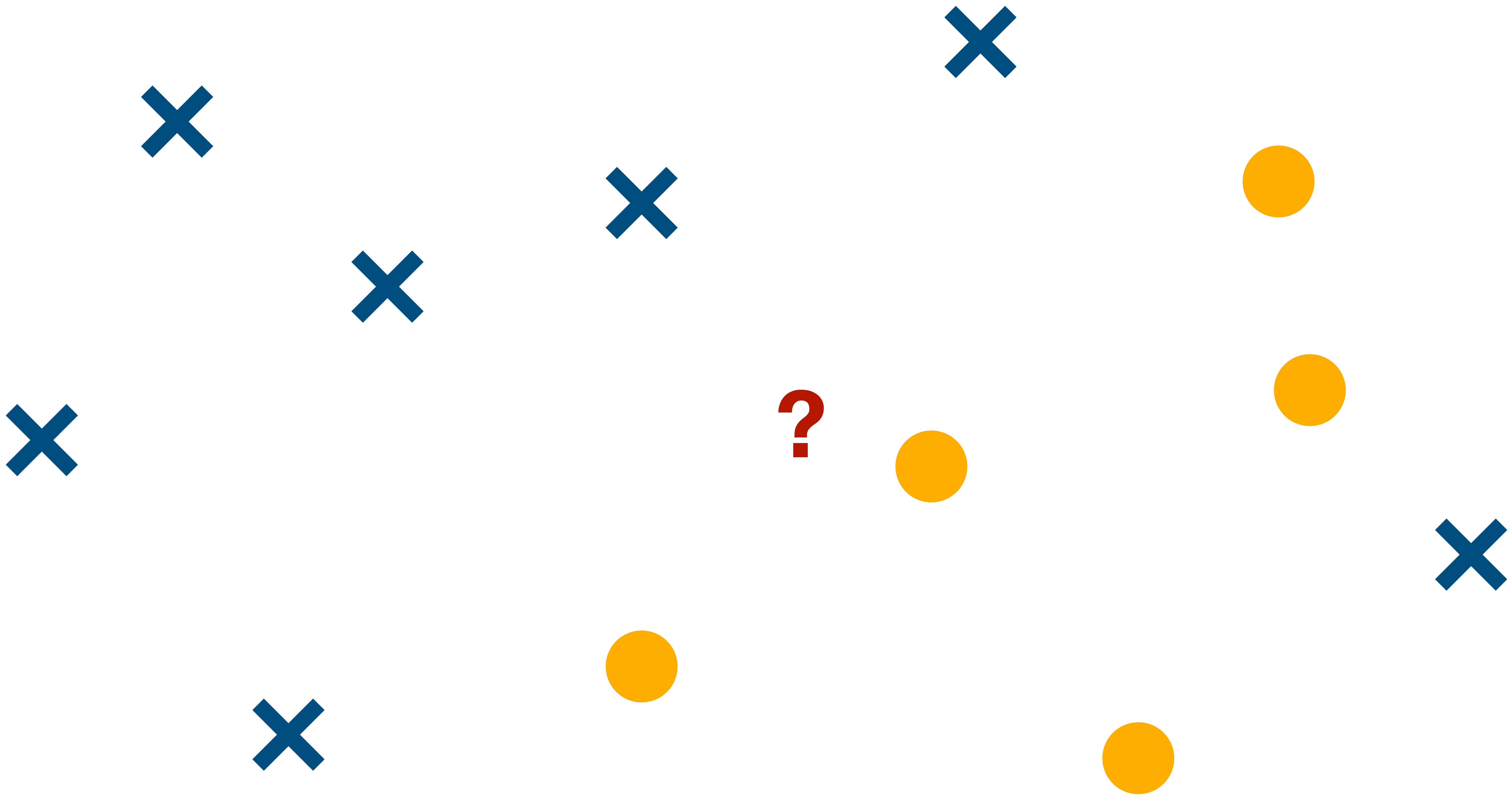
Adaptive Sampling



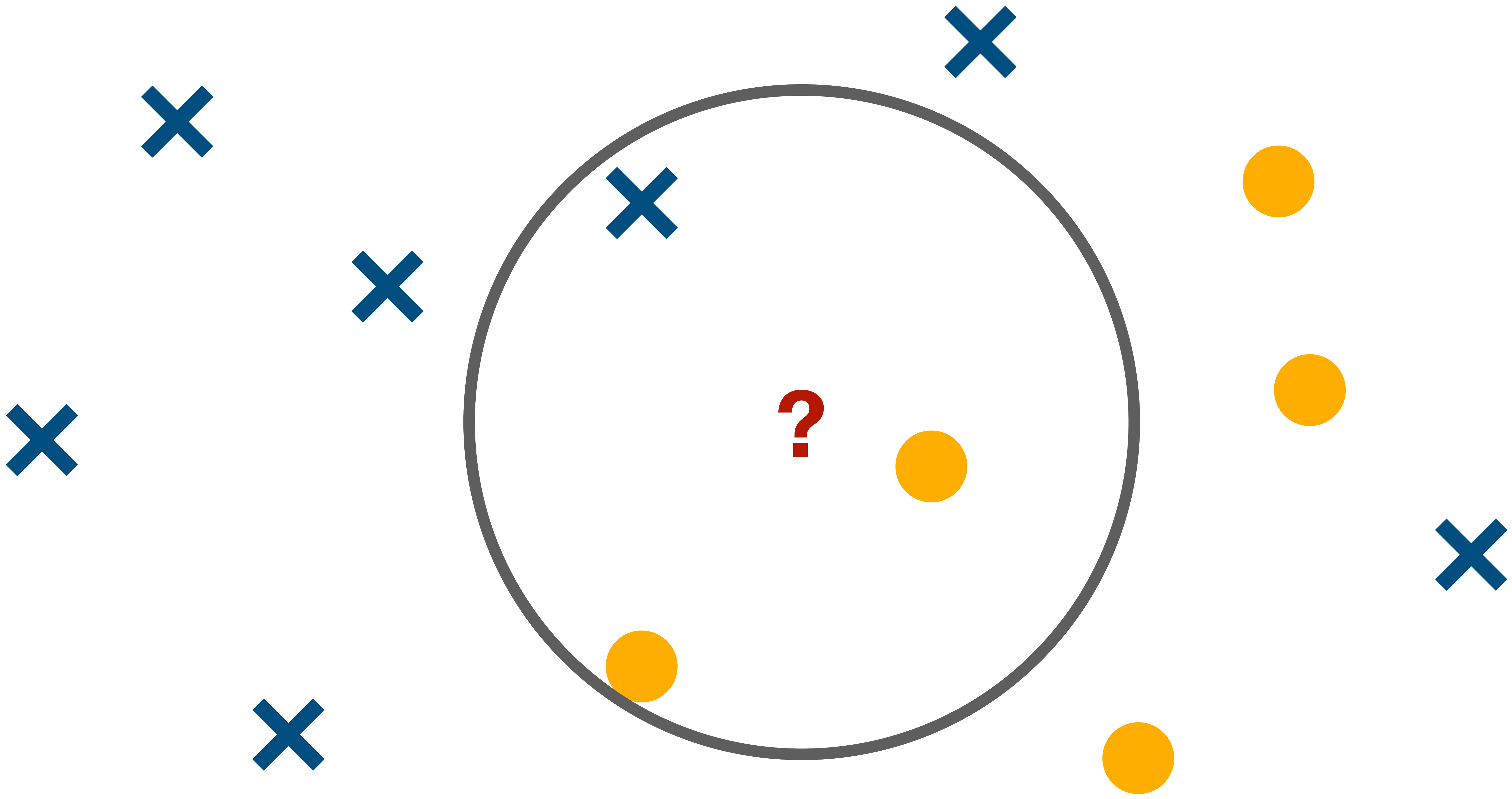
Adaptive Sampling



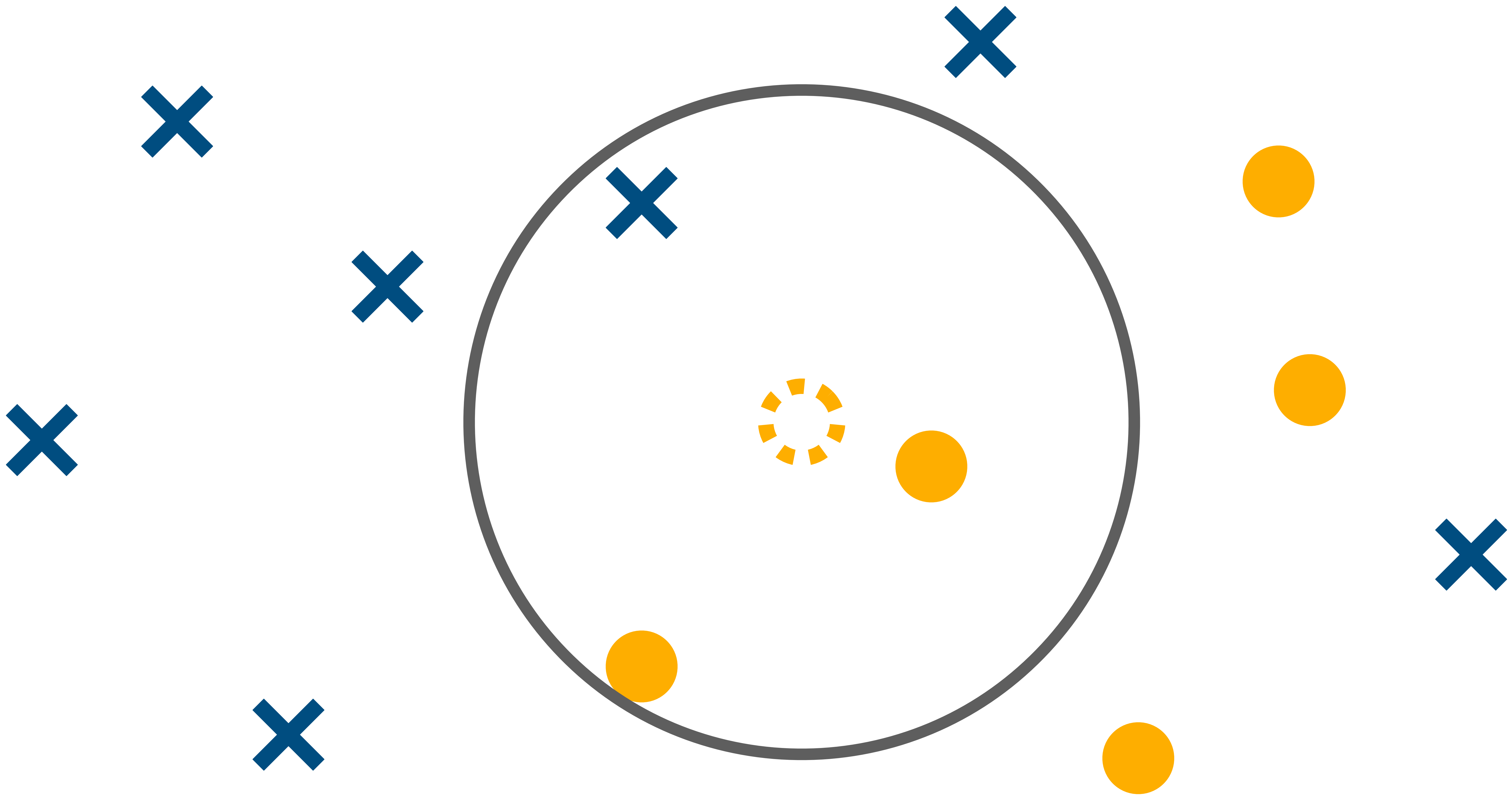
Prediction



The k_n-Nearest Neighbor Rule



The k_n-Nearest Neighbor Rule



The k_n-Nearest Neighbor Rule

For $n = 1,2,\ldots$

- Adaptive adversary selects X_n
- Learner predicts $\hat{Y}_n$
- Nature reveals label Y_n
    - It is drawn from $\mathrm{Ber}\big(\eta(X_n)\big)$

The k_n-Nearest Neighbor Rule

For $n = 1, 2, \ldots$

- Adaptive adversary selects X_n
- Learner predicts $\hat{Y}_n$
- Nature reveals label Y_n
  - It is drawn from $\mathrm{Ber}\big(\eta(X_n)\big)$

The k_n-Nearest Neighbor Rule

For $n = 1,2,\ldots$

- Adaptive adversary selects X_n
- Learner predicts $\hat{Y}_n$
- Nature reveals label Y_n
  - It is drawn from $\mathrm{Ber}\big(\eta(X_n)\big)$

The k_n-Nearest Neighbor Rule

For $n = 1,2,\ldots$

- Adaptive adversary selects X_n
- Learner predicts $\hat{Y}_n$
- Nature reveals label Y_n
  - It is drawn from $\mathrm{Ber}\big(\eta(X_n)\big)$

The k_n-Nearest Neighbor Rule

For $n = 1,2,\ldots$

- Adaptive adversary selects X_n
- Learner predicts $\hat{Y}_n$
- Nature reveals label Y_n
  - It is drawn from $\mathrm{Ber}\big(\eta(X_n)\big)$

The k_n-Nearest Neighbor Rule

For $n = 1,2,\ldots$

- Adaptive adversary selects X_n
- Learner predicts $\hat{Y}_n$
- Nature reveals label Y_n
  - It is drawn from $\mathrm{Ber}\big(\eta(X_n)\big)$

Is the k_n-nearest neighbor rule consistent?

The k_n-Nearest Neighbor Rule

For $n = 1,2,\ldots$

- Adaptive adversary selects X_n
- Learner predicts $\hat{Y}_n$
- Nature reveals label Y_n
  - It is drawn from $\mathrm{Ber}\big(\eta(X_n)\big)$

Is the k_n-nearest neighbor rule consistent?

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\left\{\hat{Y}_n \neq f^{\star}(X_n)\right\}$$

mistake rate at time N

The k_n-Nearest Neighbor Rule

For $n = 1,2,\ldots$

- Adaptive adversary selects X_n
- Learner predicts $\hat{Y}_n$
- Nature reveals label Y_n
  - It is drawn from $\mathrm{Ber}\big(\eta(X_n)\big)$

Is the k_n-nearest neighbor rule consistent?

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\Big\{\hat{Y}_n \neq f^\star(X_n)\Big\} = 0$$

asymptotic mistake rate converges to zero

Worst-Case Setting

No. The k_n-nearest neighbor rule is **not consistent** in the **worst-case setting.**

Counterintuitive Behavior

$\mathscr{X} = \mathbb{R}$

Counterintuitive Behavior

$$\eta(x) = \frac{1}{2}$$

Labels are generated by a unbiased coin flip.

Counterintuitive Behavior

What happens if we sample X_n using the binary search algorithm?

$$\eta(x) = \frac{1}{2}$$

Labels are generated by a unbiased coin flip.

Counterintuitive Behavior

Counterintuitive Behavior

Counterintuitive Behavior

Counterintuitive Behavior

Counterintuitive Behavior

Counterintuitive Behavior

Counterintuitive Behavior

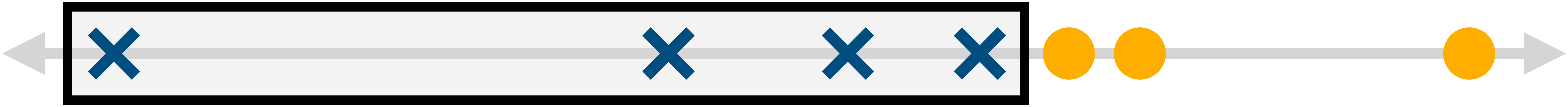
We end up with a dataset that looks linearly separable!

Counterintuitive Behavior

We end up with a dataset that looks linearly separable!

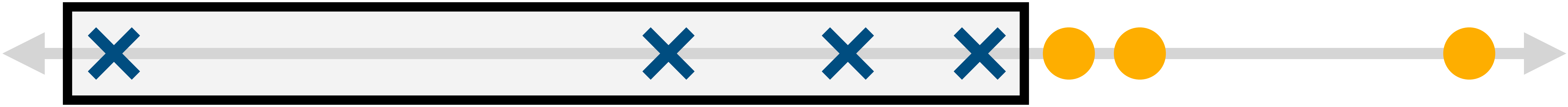
But this pattern will not generalize to future data.

Counterintuitive Behavior



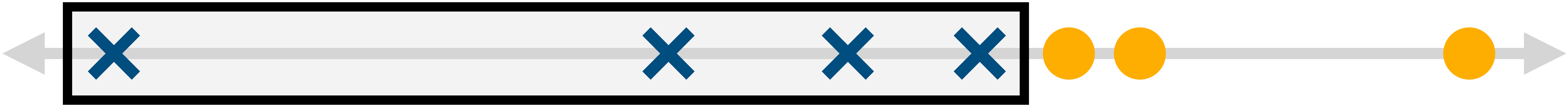
This interval has many points.

Counterintuitive Behavior



This interval has many points. But, its average label is very far from the “expected” 1/2.

Counterintuitive Behavior



This interval has many points. But, its average label is very far from the “expected” 1/2.

- **Uniform law of large numbers** from I.I.D. setting does not apply.

How “realistic” or likely is such worst-case data?

How “realistic” or likely is such worst-case data?

- These examples are actually very pathological.

How “realistic” or likely is such worst-case data?

- These examples are actually very pathological.

In some sense, they will **almost never appear** in the wild.

Smoothed Online Learning

By imposing **extremely mild** constraints on the adaptive sampling mechanism:

Smoothed Online Learning

By imposing **extremely mild** constraints on the adaptive sampling mechanism:

- **We can recover a uniform law of large numbers for dependent data.**

Smoothed Online Learning

By imposing **extremely mild** constraints on the adaptive sampling mechanism:

- **We can recover a uniform law of large numbers for dependent data.**
- **We can show that the k_n-nearest neighbor rule is asymptotically consistent.**

Takeaway

- Dependent data is everywhere and can behave counterintuitively.
- Worst-case adaptivity can also be quite unrealistic.
- Theoretical and practical opportunities in average-case adaptivity.

Thanks!