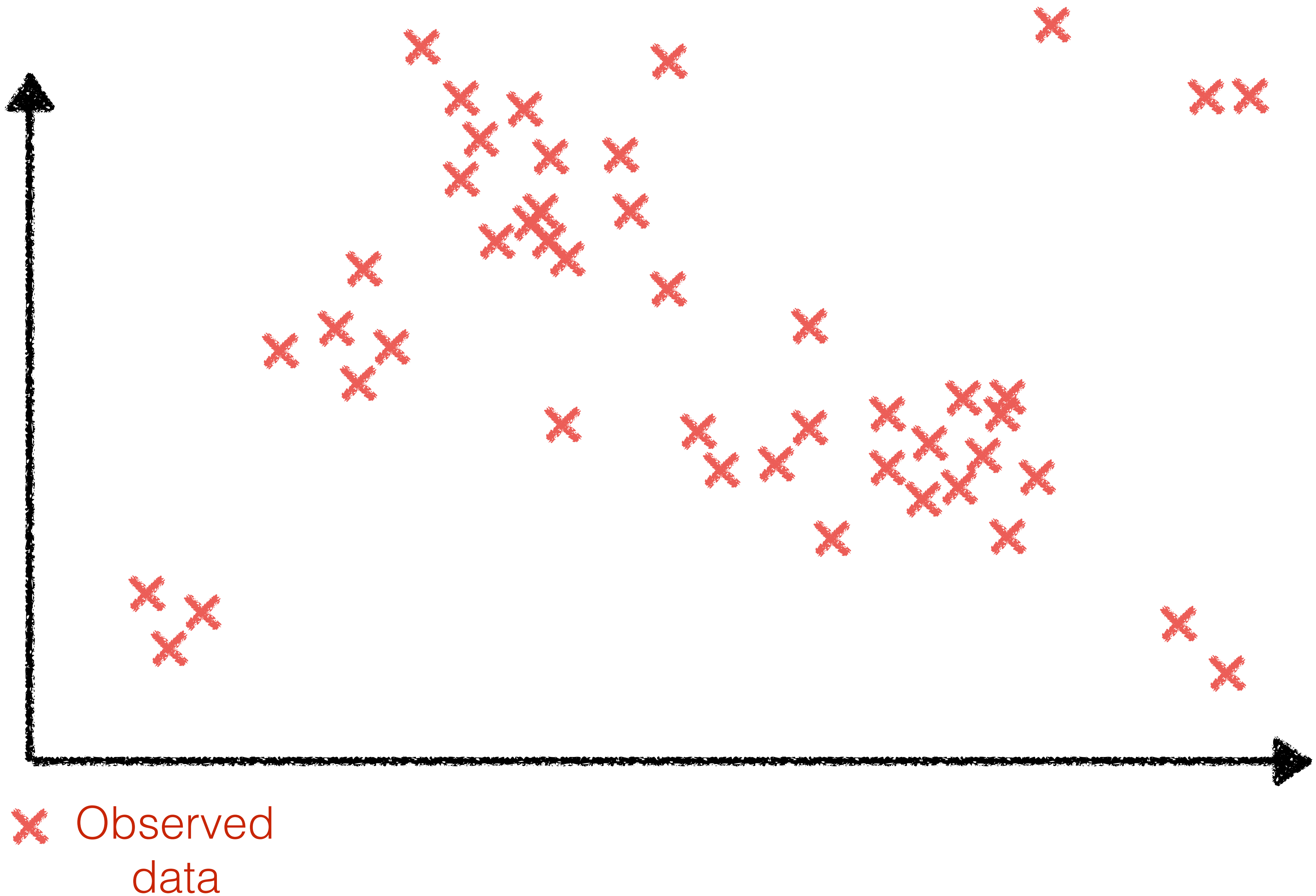
Examples are not Enough, Learn to Criticize! Criticism for Interpretability

Been Kim, Rajiv Khanna, Oluwasanmi Koyejo
*all authors contributed equally

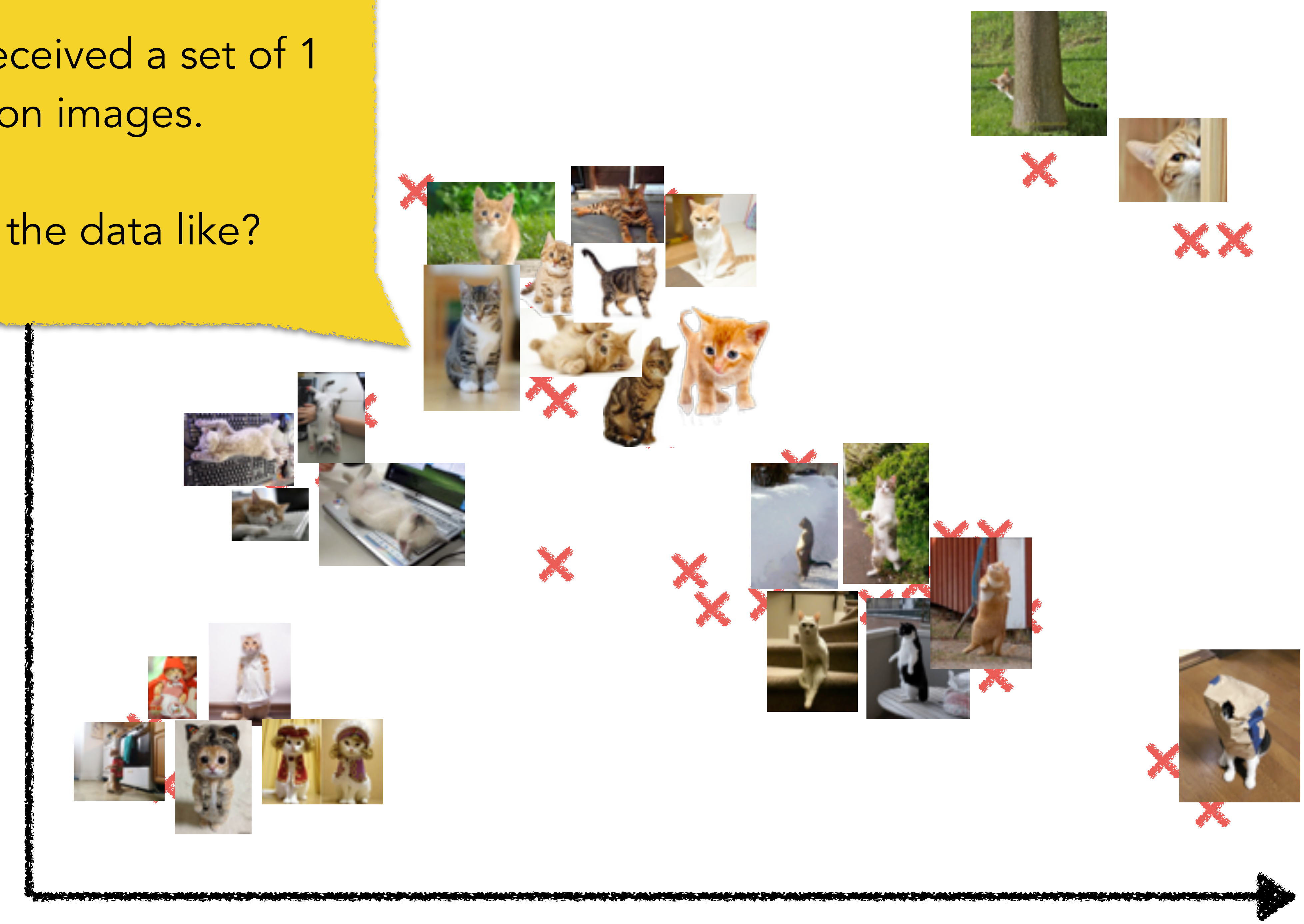
Understanding data through examples



Understanding data through examples

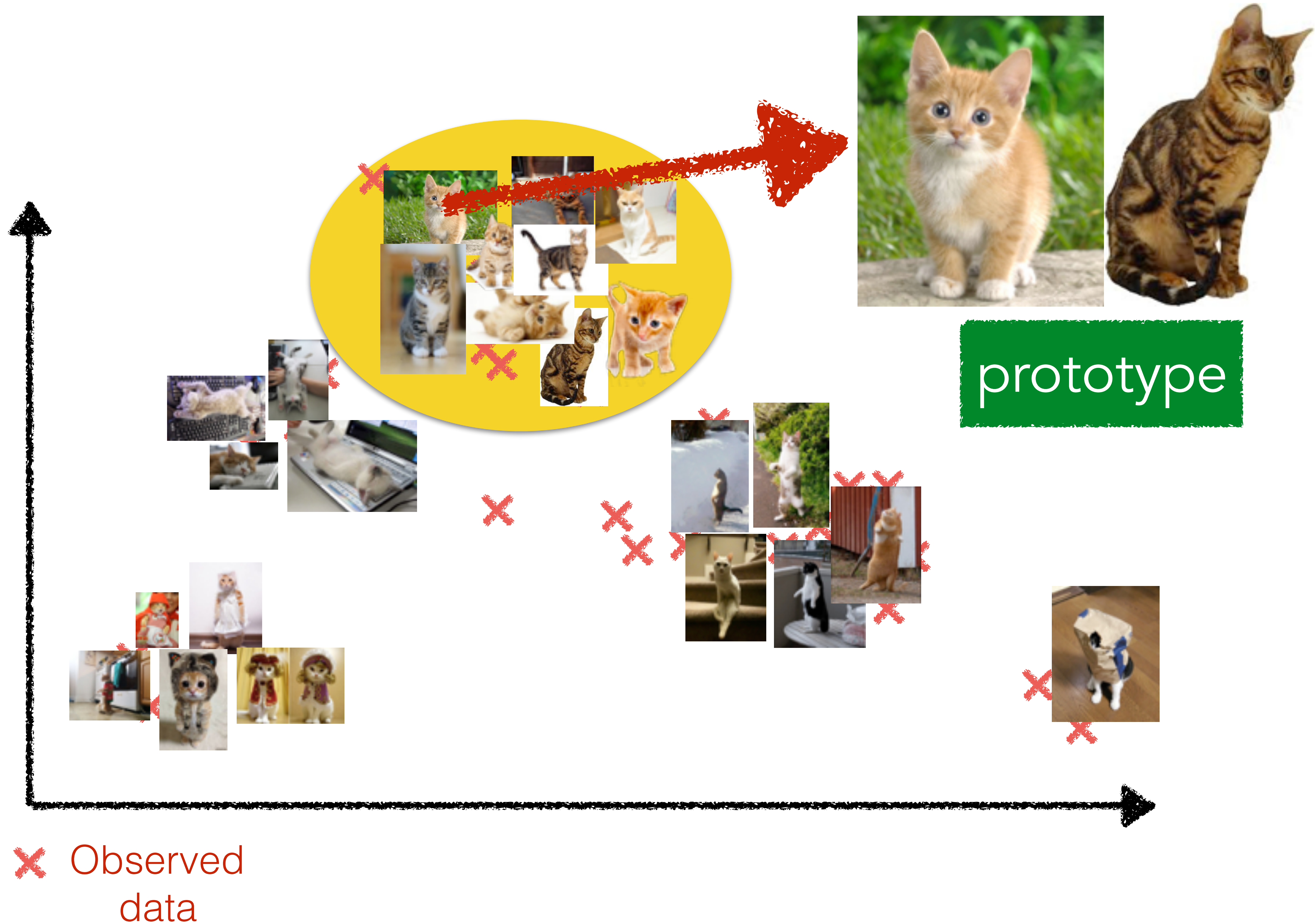
You just received a set of 1 billion images.

What's the data like?

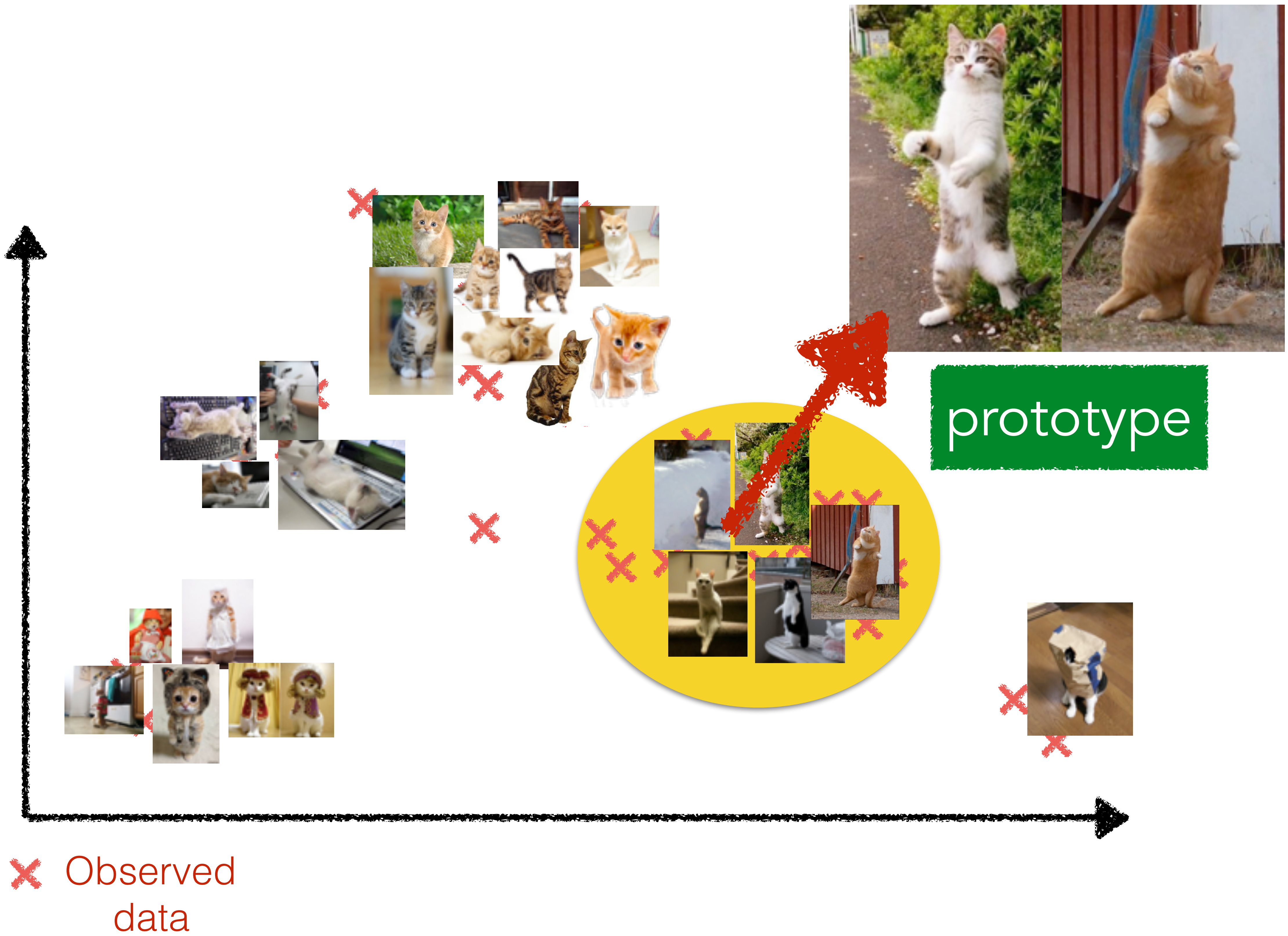


Observed data

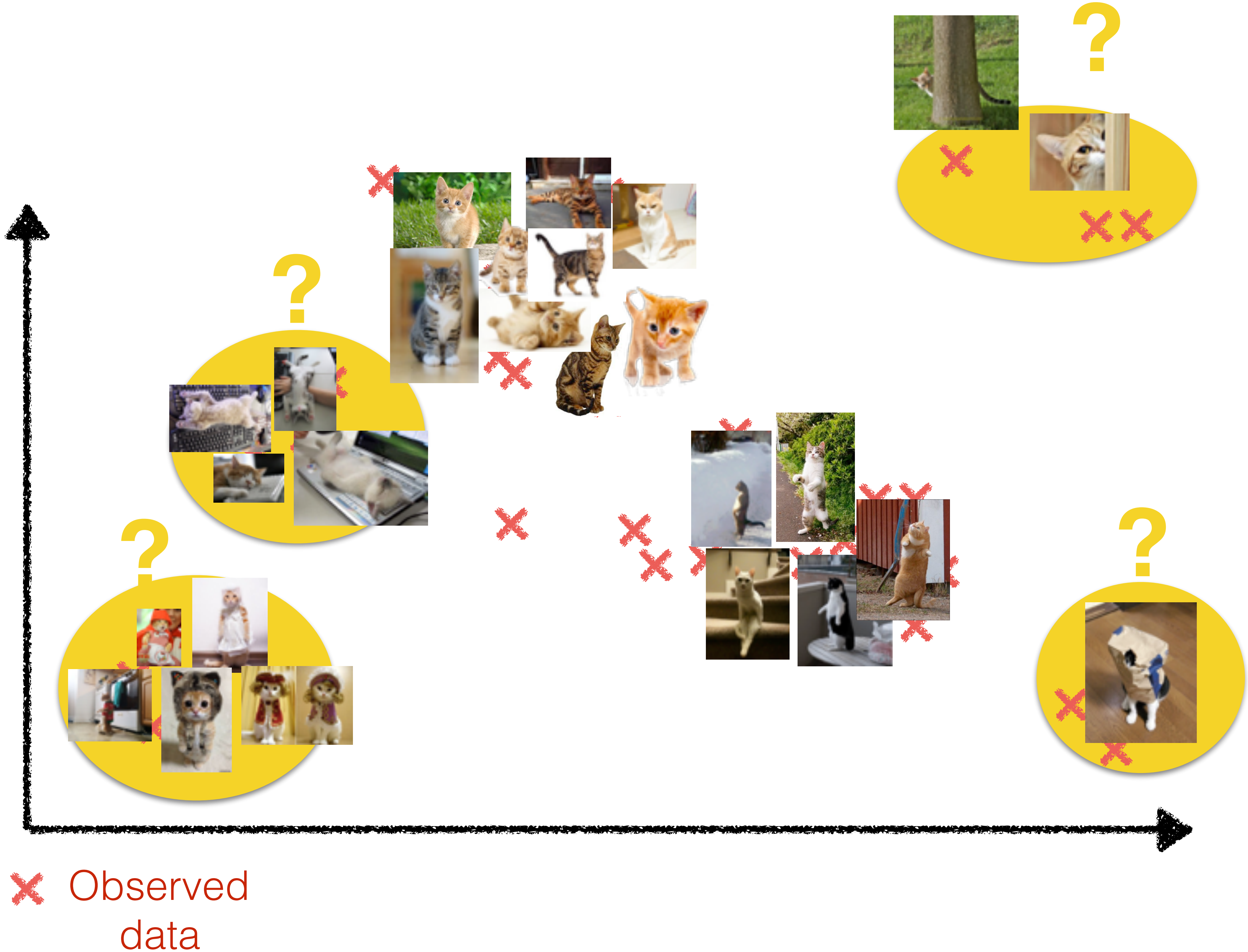
Understanding data through examples



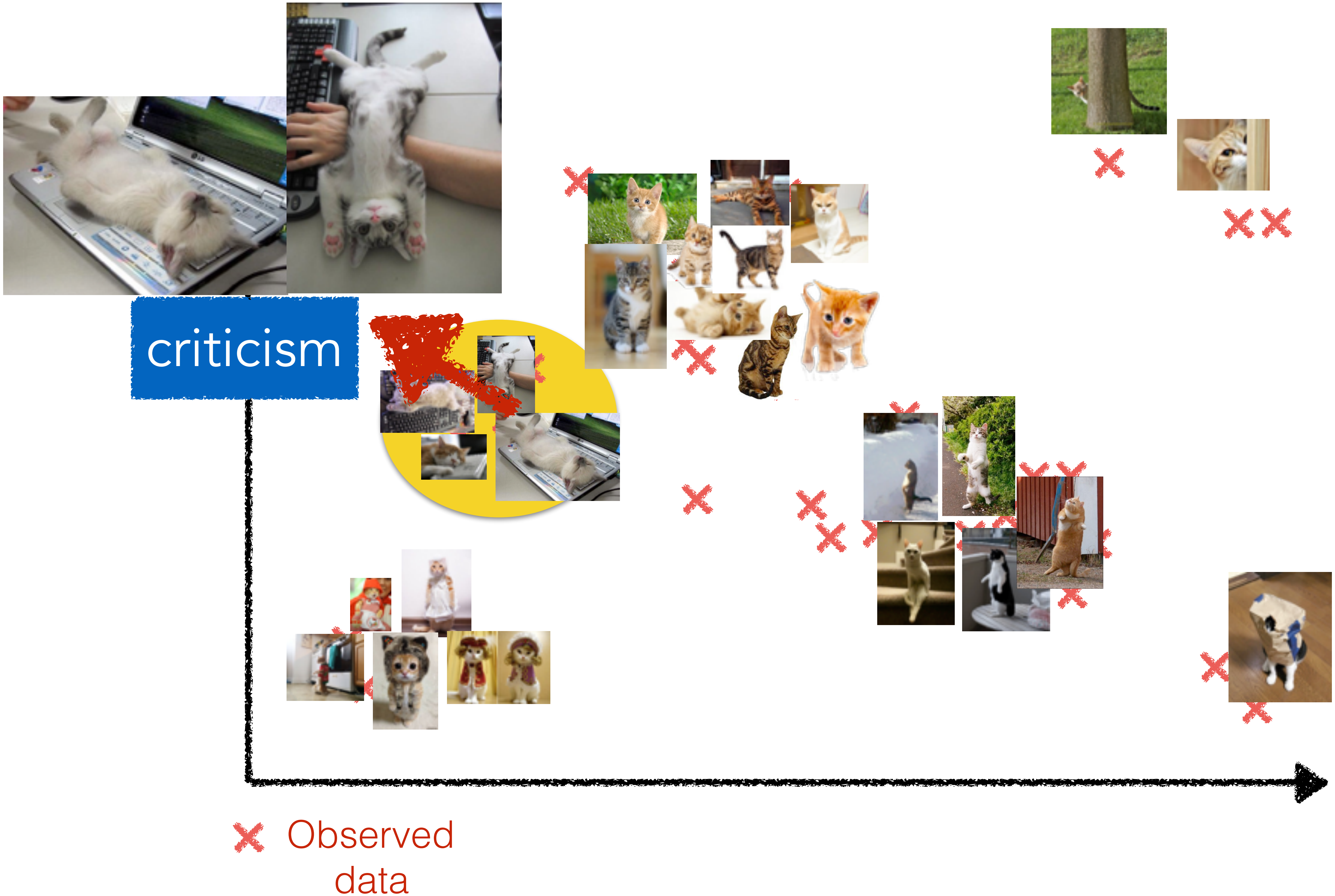
Understanding data through examples



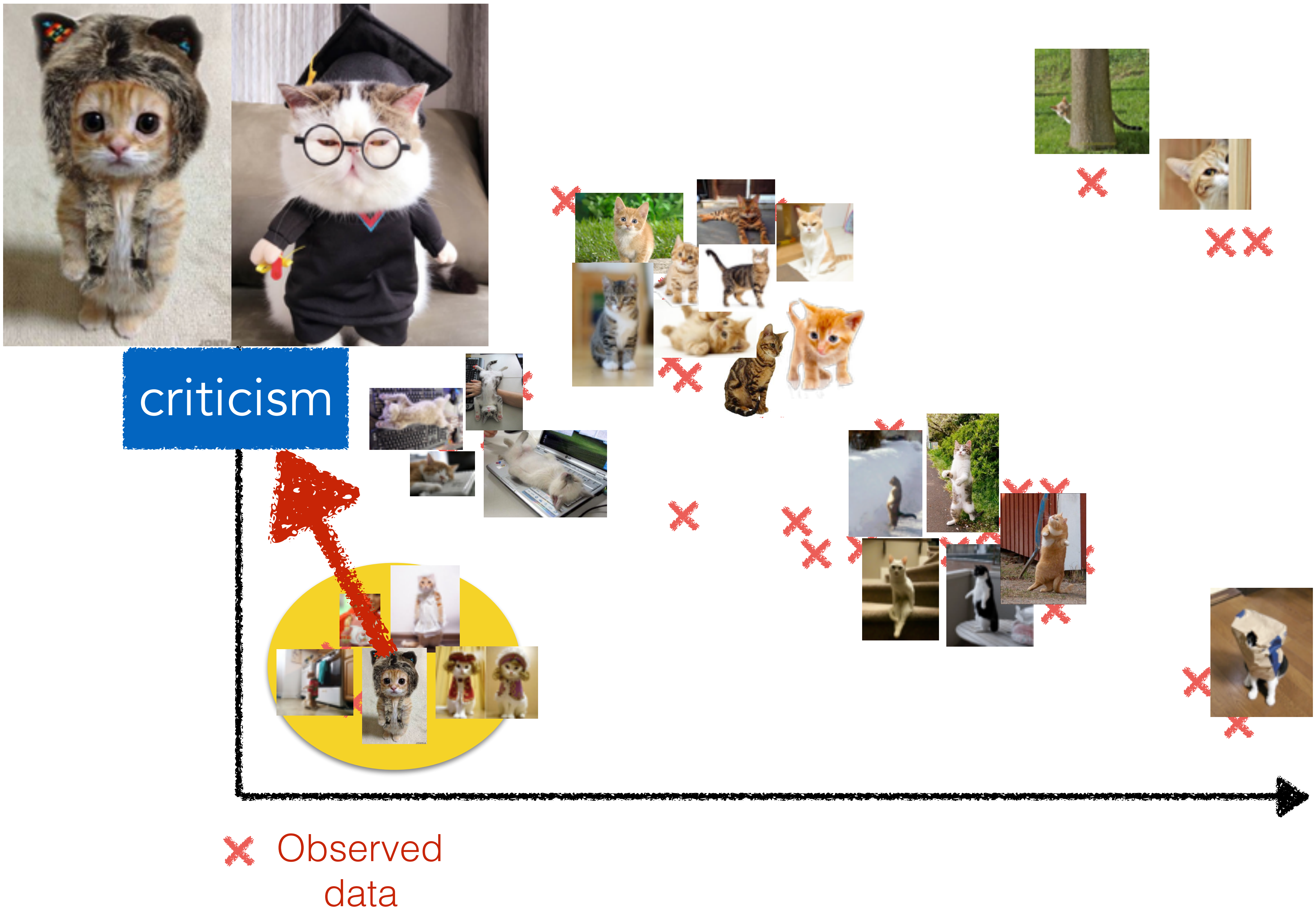
Understanding data through examples



Understanding data through examples



Understanding data through examples



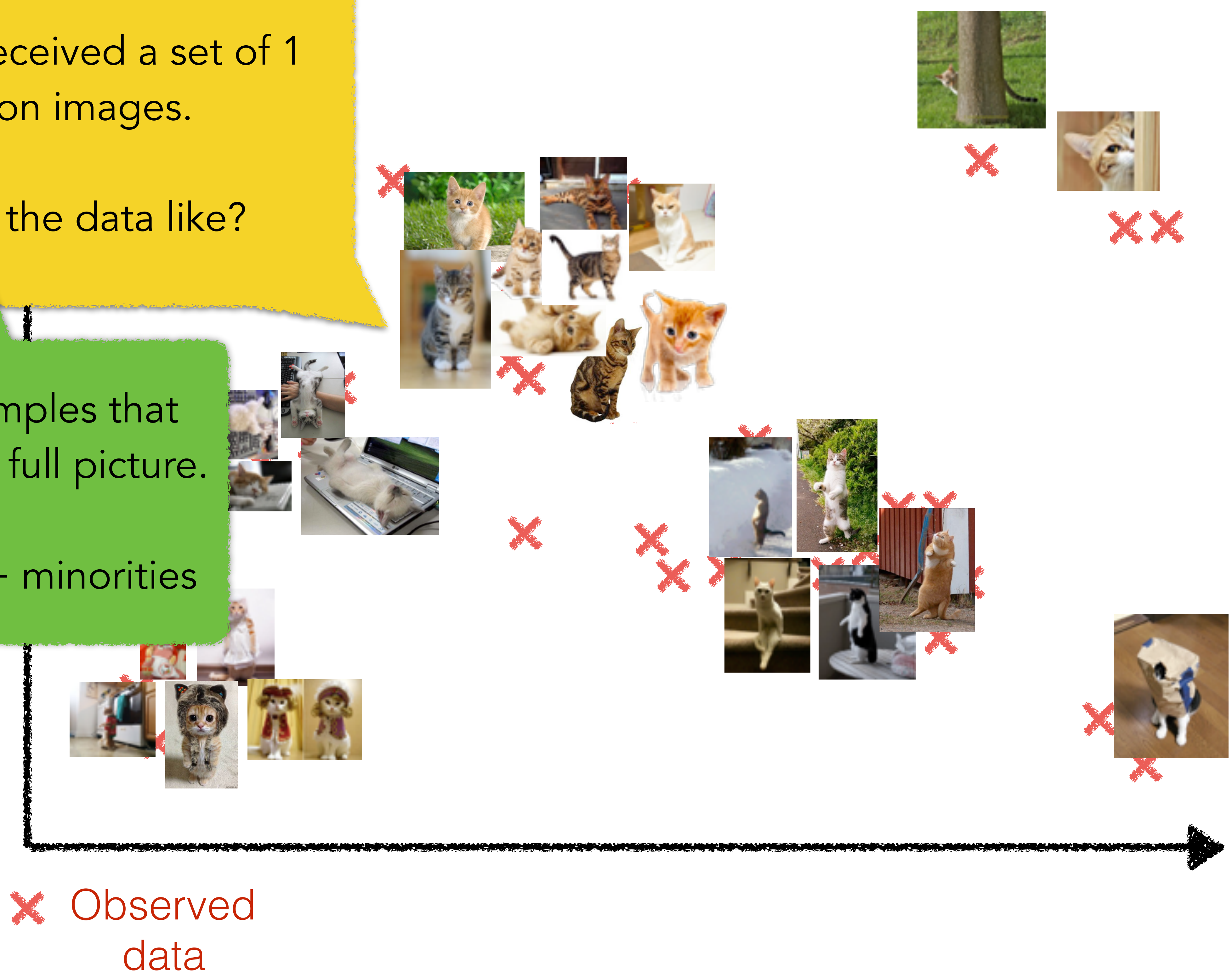
Understanding data through examples

You just received a set of 1 billion images.

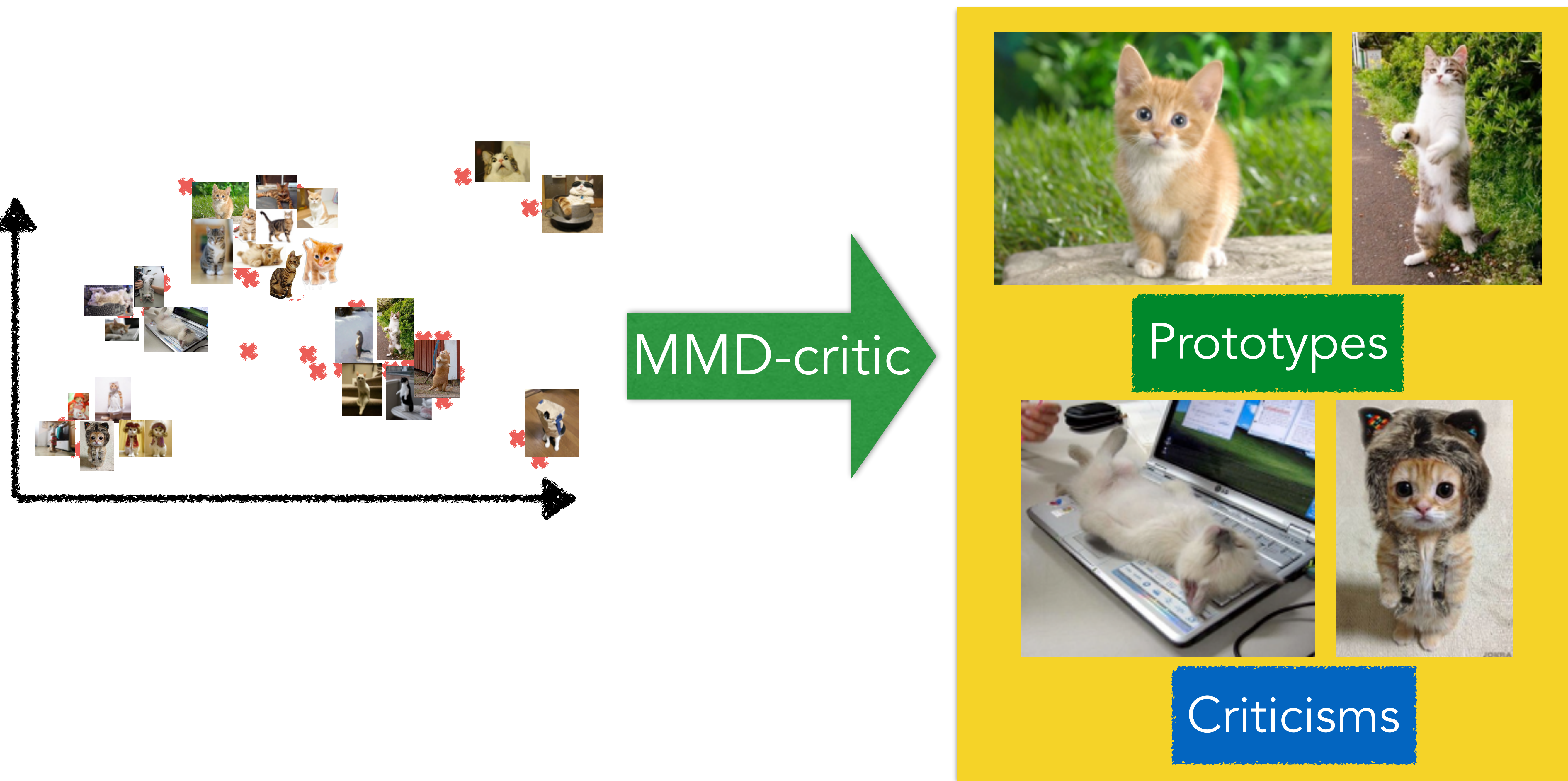
What's the data like?

Need examples that show us the full picture.

majorities + minorities



What this talk is about.



Insights from cognitive science

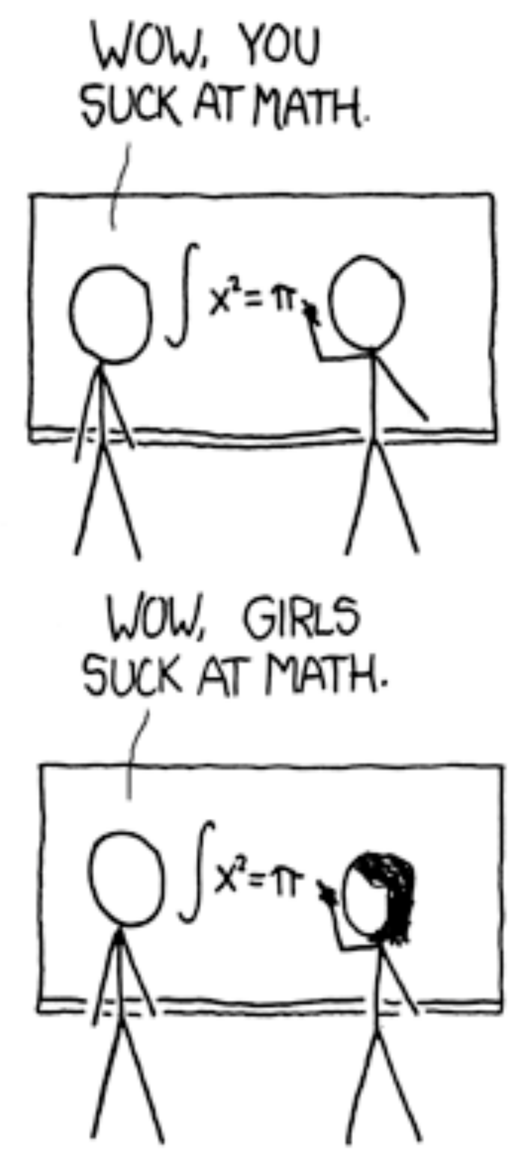
- Humans do exemplar-based reasoning for complex decisions [Cohen 96, Newell 72]
  - fire fighters [Klein 89]
- Mirror the way humans think: interpretability of data through examples.

However,
Humans tend to over-generalize

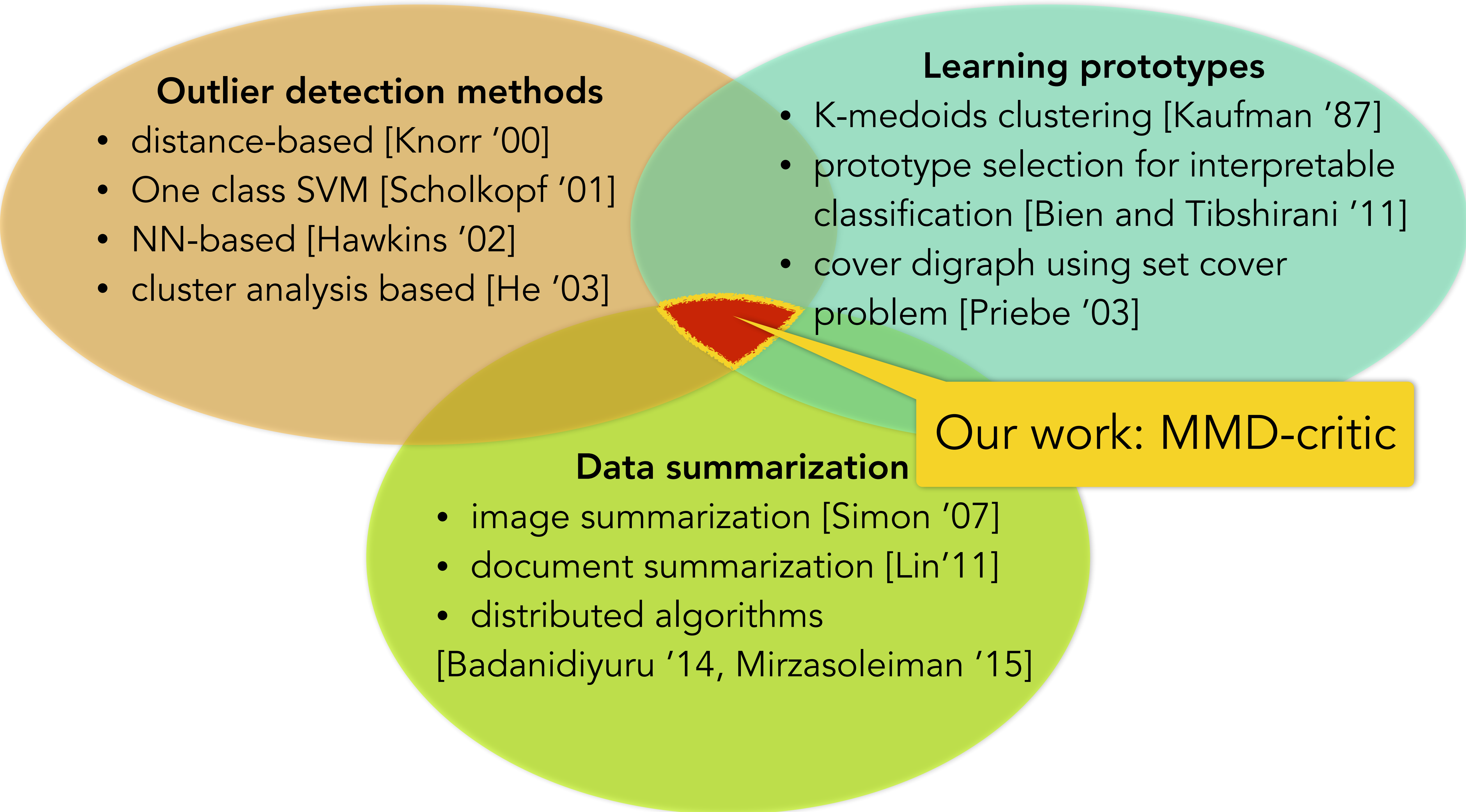
- Over-generalization is consistent with evolutionary theory [Zebrowitz '10, Schaller' 06]

→ algorithms can help against over-generalization

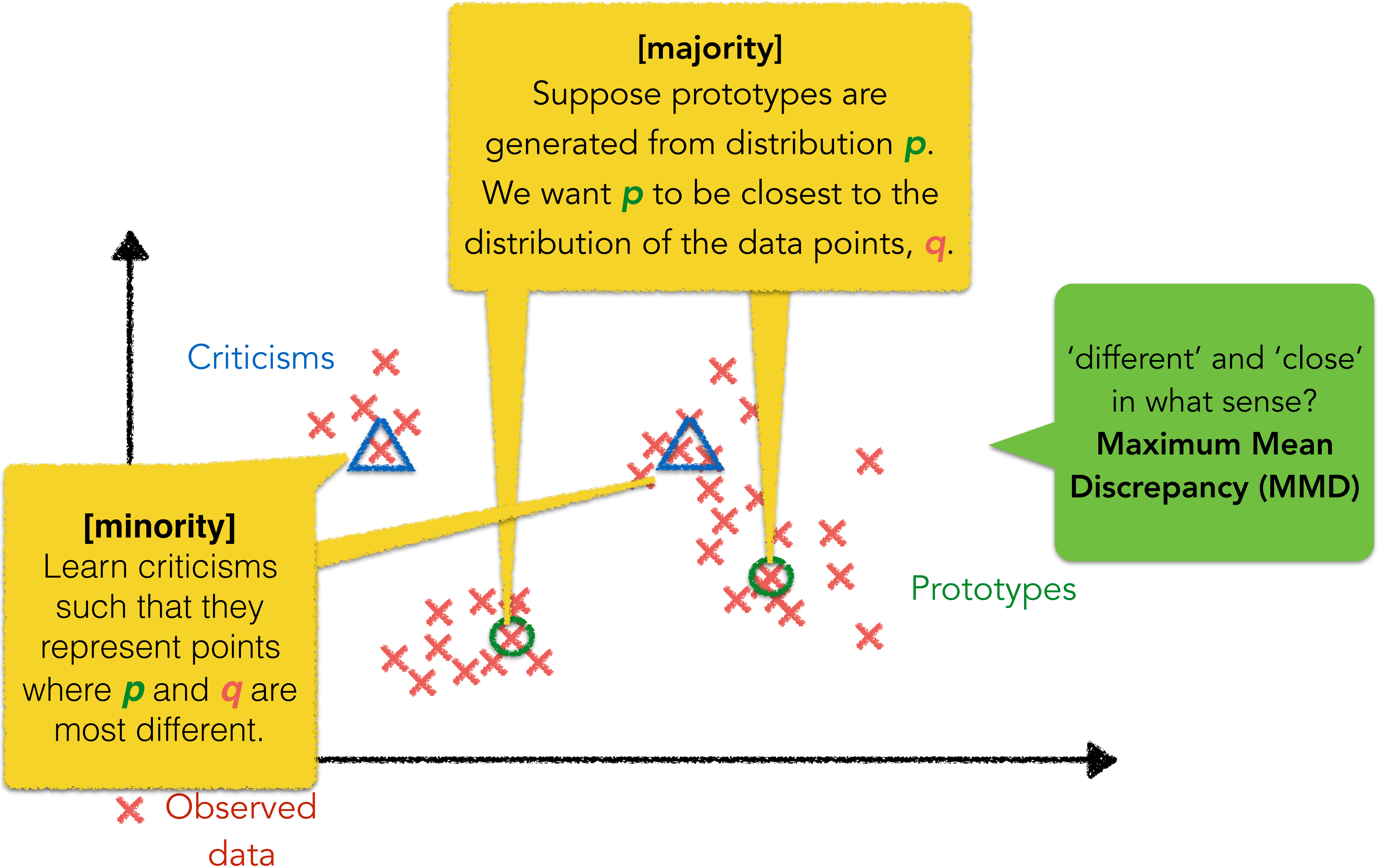
Our work:
Learn **prototypes + criticisms** to minimize over-generalization



Related work



Our approach: MMD-critic



Maximum Mean Discrepancy (MMD)

- MMD is a measure of the difference between distributions *P* and *Q* [Borgwardt '06, Gretton '07]

$$\mathrm{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)])$$

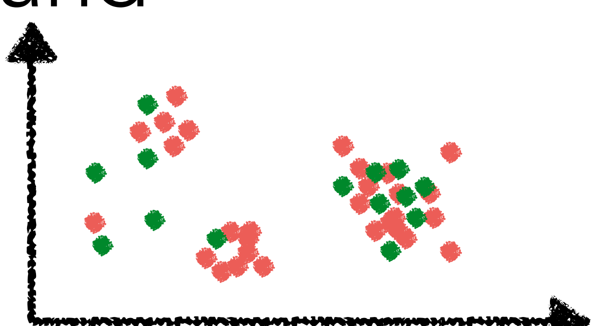
reproducing kernel Hilbert space with kernel function *k*

witness function
gives analytic solution

- Empirically can be measured using samples:

$$\mathrm{MMD}^2[\mathcal{F}, p, q] := \frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(y_i, y_j)$$

- Used for Bayesian model criticism [Lloyd '15] and two-sample tests [Gretton '07]



MMD-critic:
learning prototypes and criticisms

1. Choose the number of prototypes and criticisms
2. Select prototypes using greedy search
3. Select criticisms using greedy search

Submodular functions

Let X be a finite set. A function $f : 2^X \to \mathbb{R}$ is sub modular if for all subsets $S \subset T \subset X$ and all $x \in X / T$

$$f(S \cup \{x\}) - f(S) \geq f(T \cup \{x\}) - f(T)$$

then greedy method guarantees at least $(1 - \frac{1}{e})$ of the optimal solution

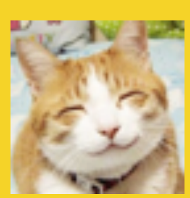
MMD-critic: learning prototypes and criticisms

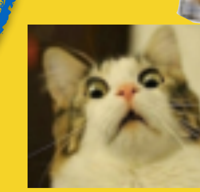
1. Choose the number of prototypes and criticisms
2. Select prototypes using greedy search
3. Select criticisms using greedy search

Submodular functions

Let X be a finite set. A function $f : 2^X \rightarrow \mathbb{R}$ is sub modular if for all subsets $S \subset T \subset X$ and all $x \in X / T$

$f($  $\cup \{$  $\}) - f(S) \geq f($  $\cup \{$  $\}) - f(T)$



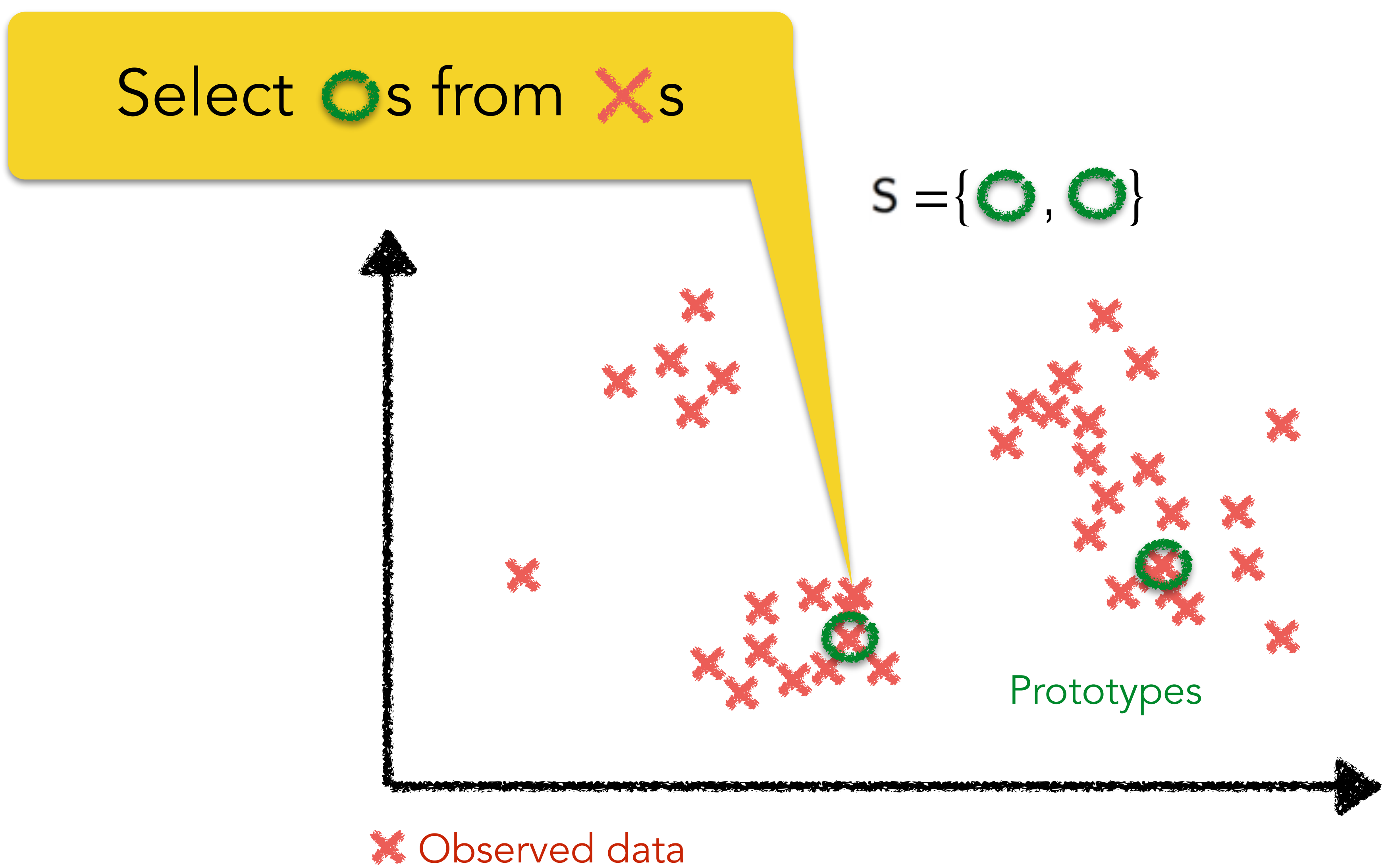


then greedy method guarantees at least $(1 - \frac{1}{e})$ of the optimal solution

MMD-critic:
learning prototypes and criticisms

1. Selecting prototypes by **minimizing** MMD

$S \in 2^{[n]}, |S| \le m_*$

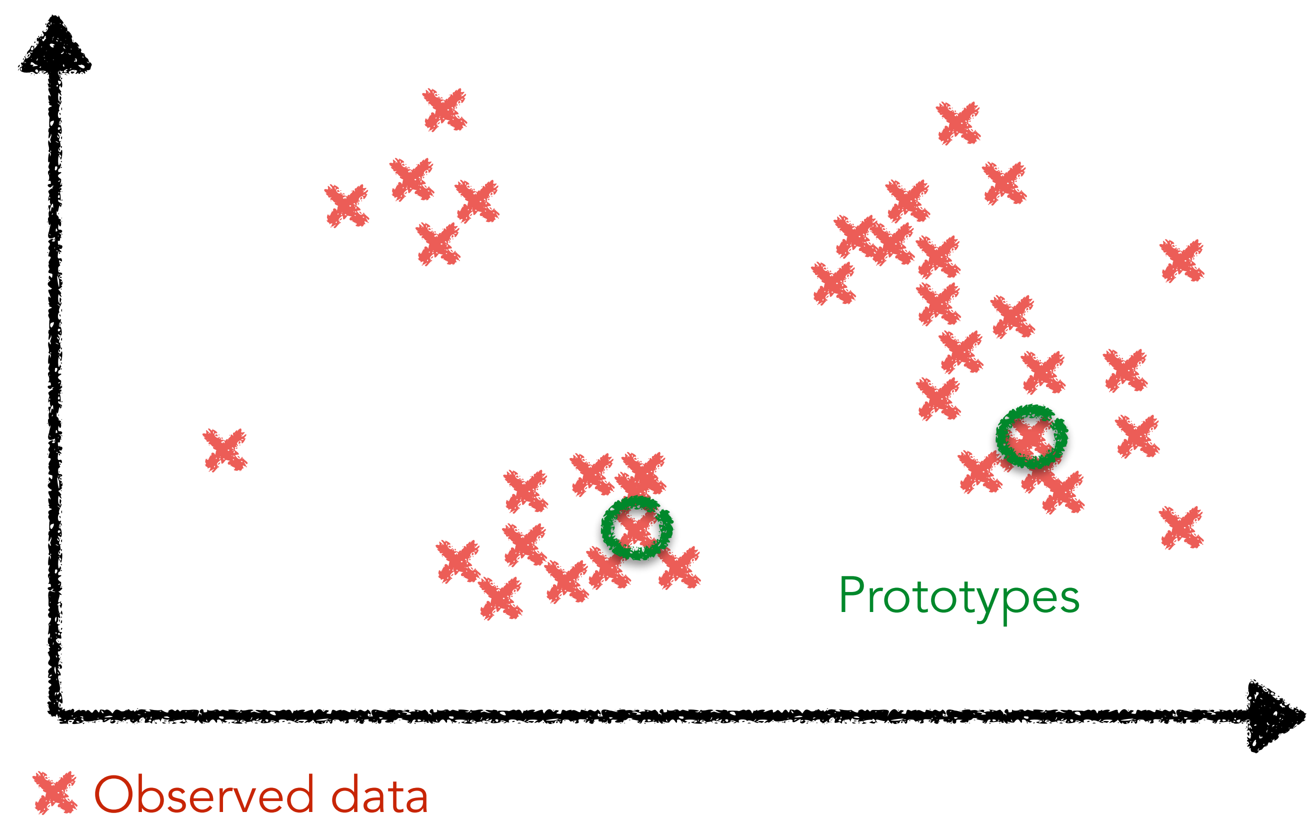


MMD-critic:
learning prototypes and criticisms

1. Selecting prototypes by **minimizing** MMD

$$\max_{S \in 2^{[n]}, |S| \leq m_*} J_b(S) = -\,\mathrm{MMD}^2(\mathcal{F}, X, X_S)$$

Suppose prototypes (◯) are generated from distribution ***p***. We want ***p*** to be closest to the distribution of the data points (✕), ***q***.



MMD-critic: learning prototypes and criticisms

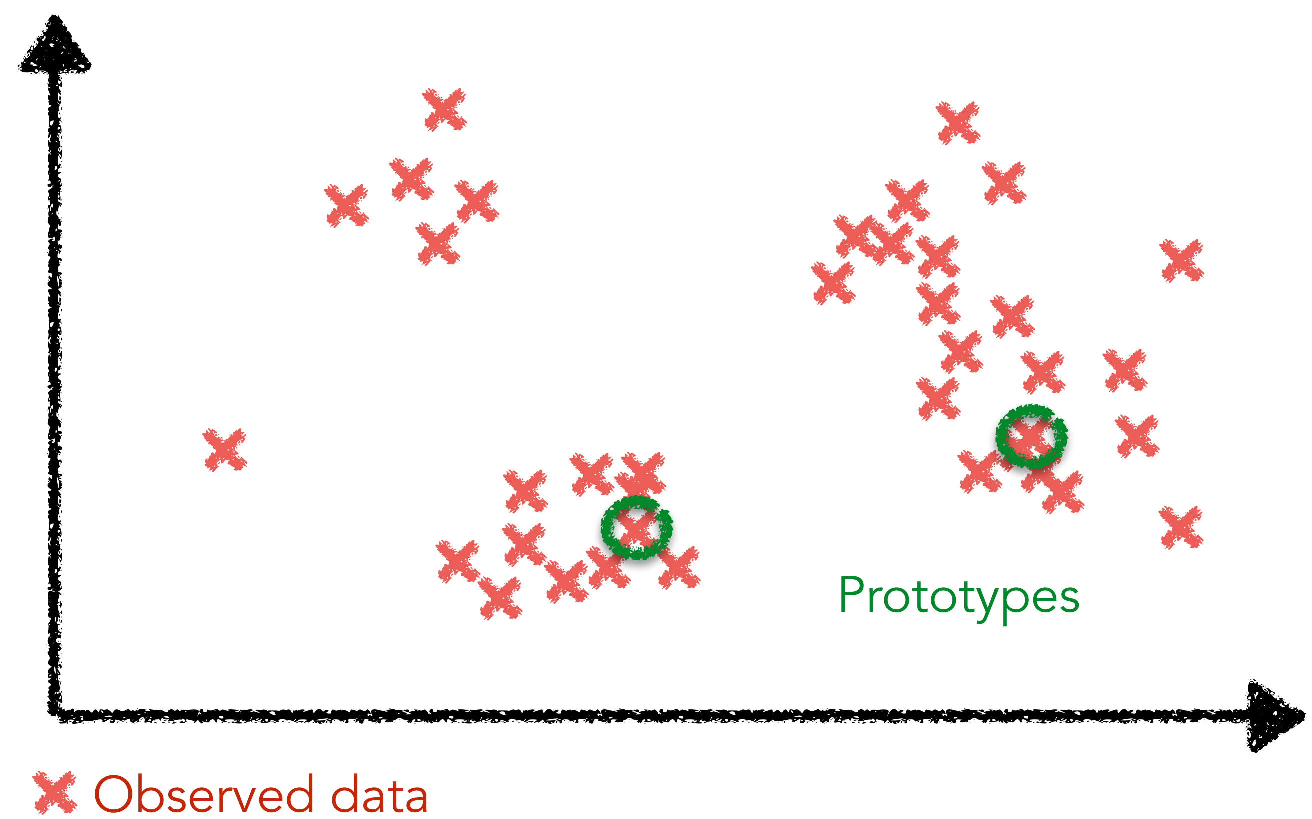
1. Selecting prototypes by **minimizing** MMD

$$\max_{S \in 2^{[n]}, |S| \leq m_*} J_b(S) = -\mathrm{MMD}^2(\mathcal{F}, X, X_S) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j)$$

submodular if the kernel matrix is diagonally dominant:

$$0 \leq k_{i,j} \leq \frac{k^*}{n^3 + 2n^2 - 2n - 3}$$

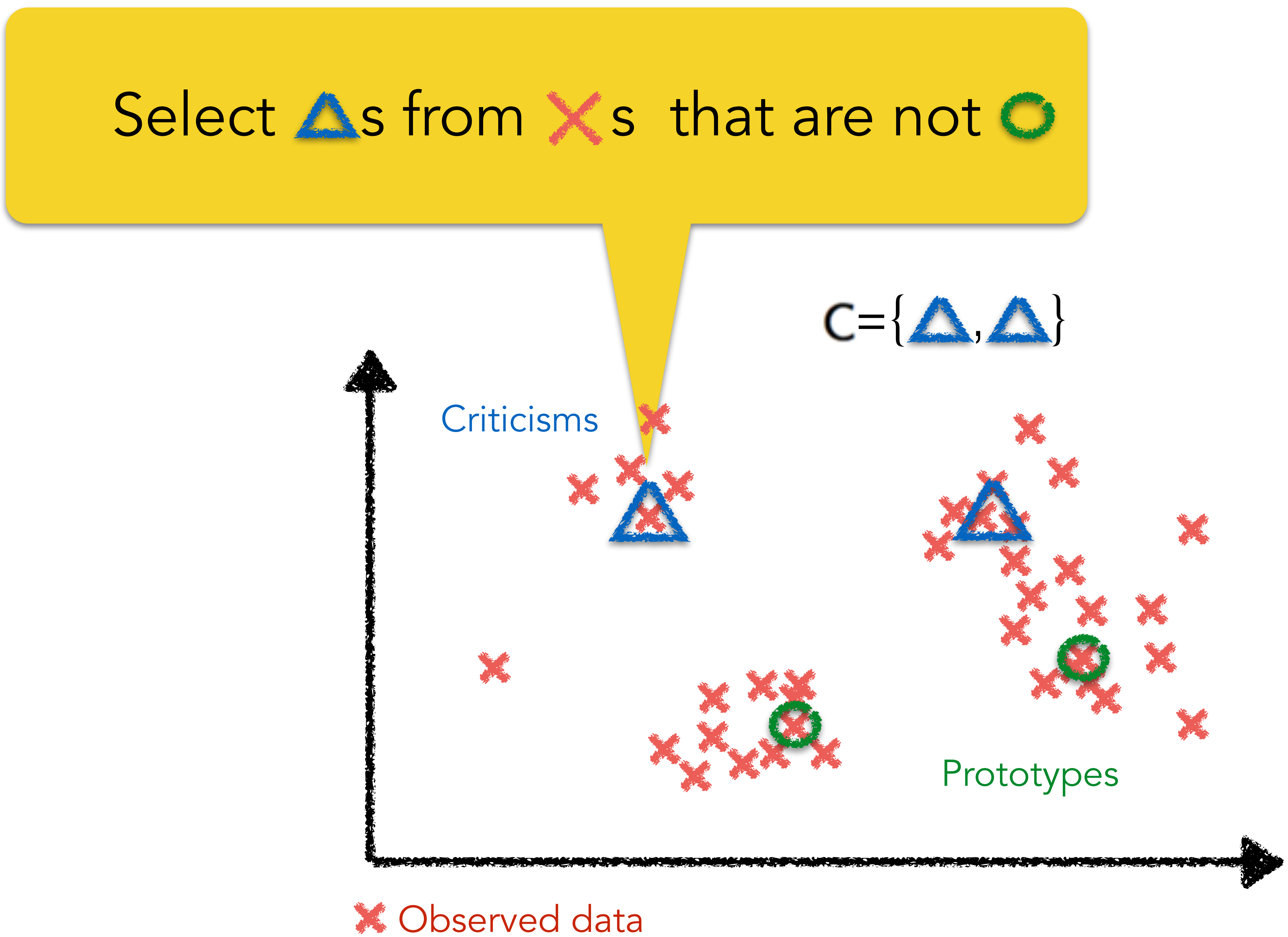
(Detailed proofs in the paper)



MMD-critic:
learning prototypes and criticisms

2. Selecting criticisms by **maximizing** - finding peaks in witness function

$C \subseteq [n] \setminus S, |C| \leq c_*$



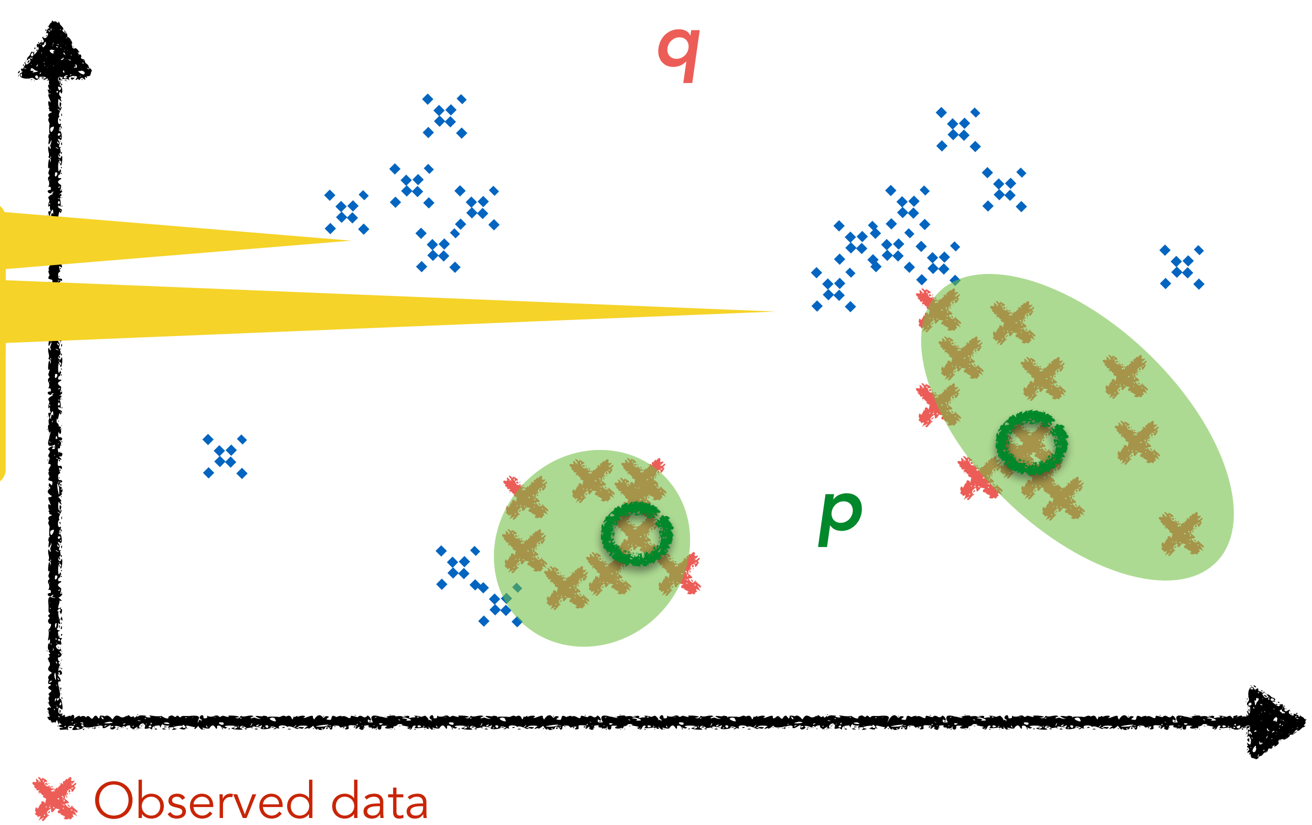
MMD-critic: learning prototypes and criticisms

2. Selecting criticisms by **maximizing -** finding peaks in witness function

Learn criticisms such that they represent where prototype distribution (***p***) and data distribution (***q***) are most different

$$\max_{\mathsf{C} \subseteq [n] \setminus \mathsf{S}, |\mathsf{C}| \leq c_*} L(\mathsf{C})$$

'peaks' in the witness function (analytical solution to MMD)



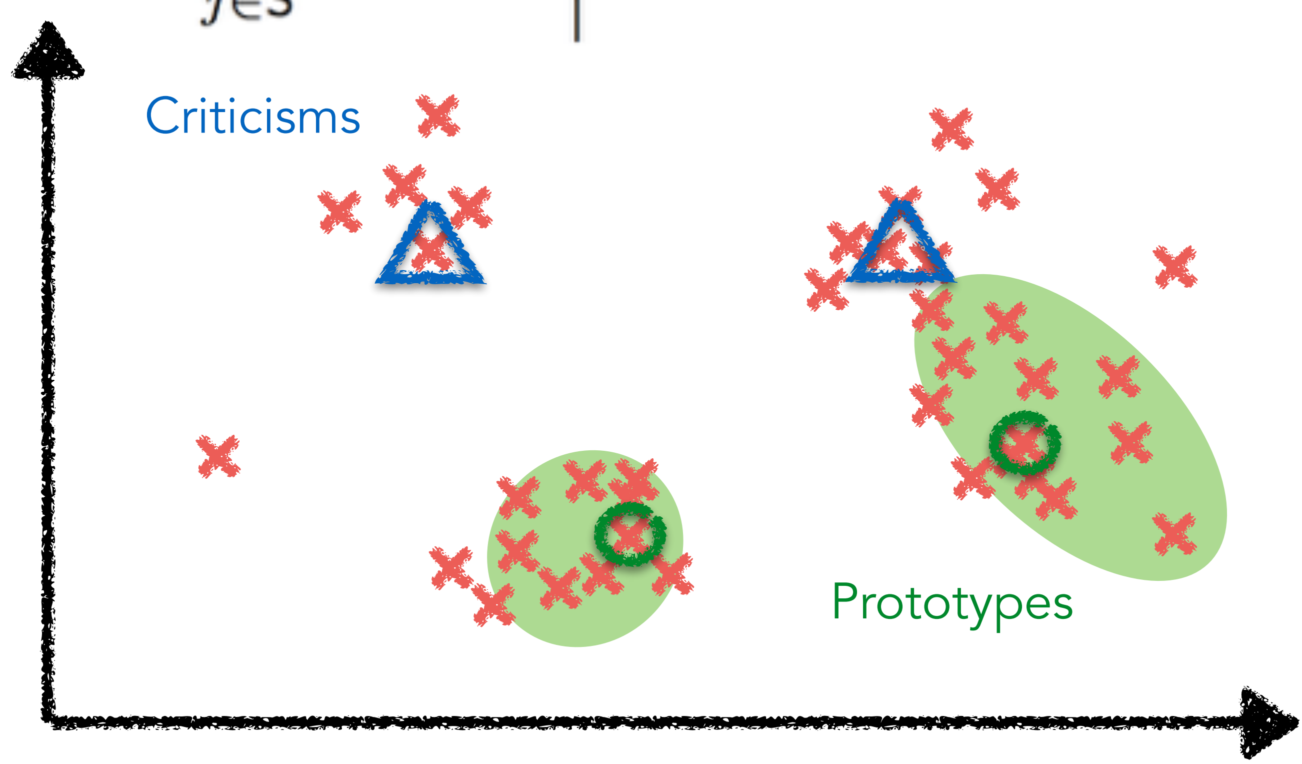
MMD-critic: learning prototypes and criticisms

2. Selecting criticisms by **maximizing** - finding peaks in witness function

Learn criticisms such that they represent where prototype distribution (***p***) and data distribution (***q***) are most different

$$\max_{C\subseteq[n]\setminus S,\,|C|\le c_*} L(C) = \sum_{l\in C}\left|\frac{1}{n}\sum_{i\in[n]} k(x_i, x_l) - \frac{1}{m}\sum_{j\in S} k(x_j, x_l)\right|$$

also submodular



MMD-critic: learning prototypes and criticisms

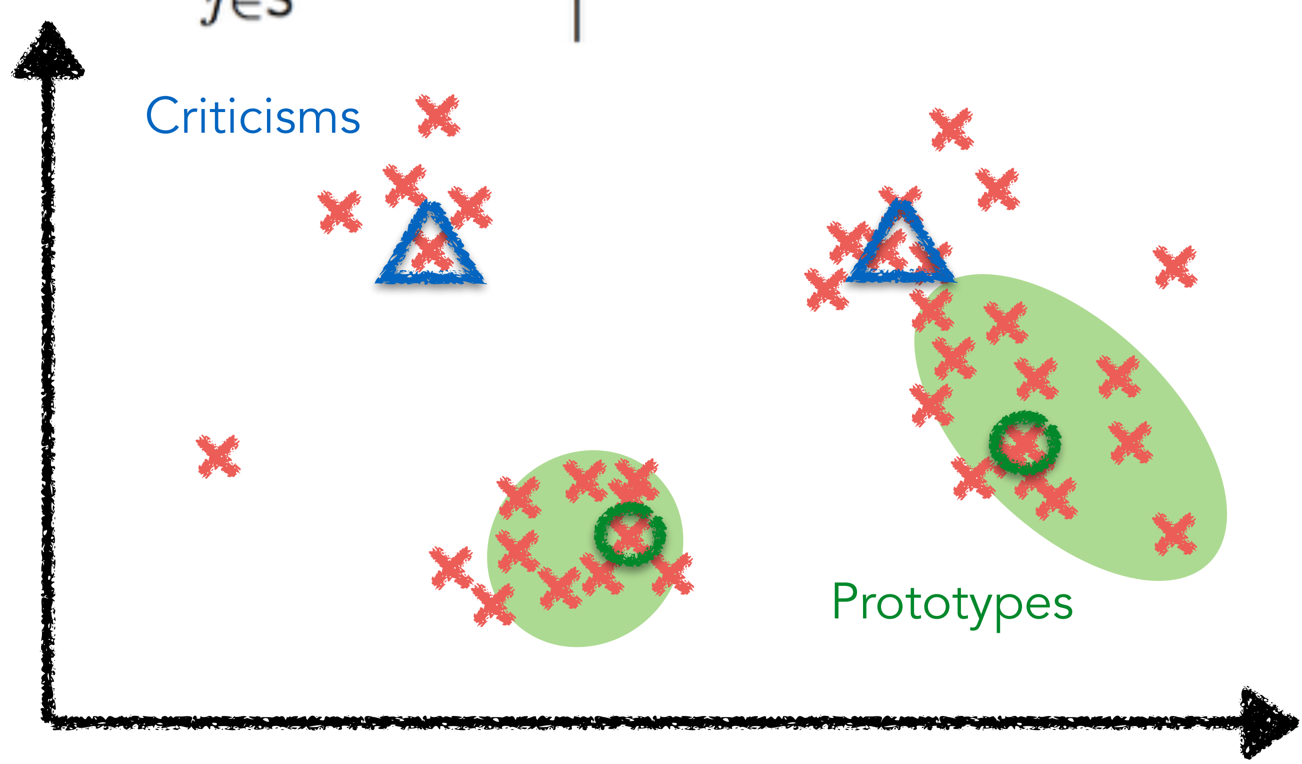
2. Selecting criticisms by **maximizing** - finding peaks in witness function

Learn criticisms such that they represent where prototypes and data distribution (q) are most different

want them to be diverse [Krause '08]
$r(\mathbf{K}, C) = \log \det \mathbf{K}_{C,C}$

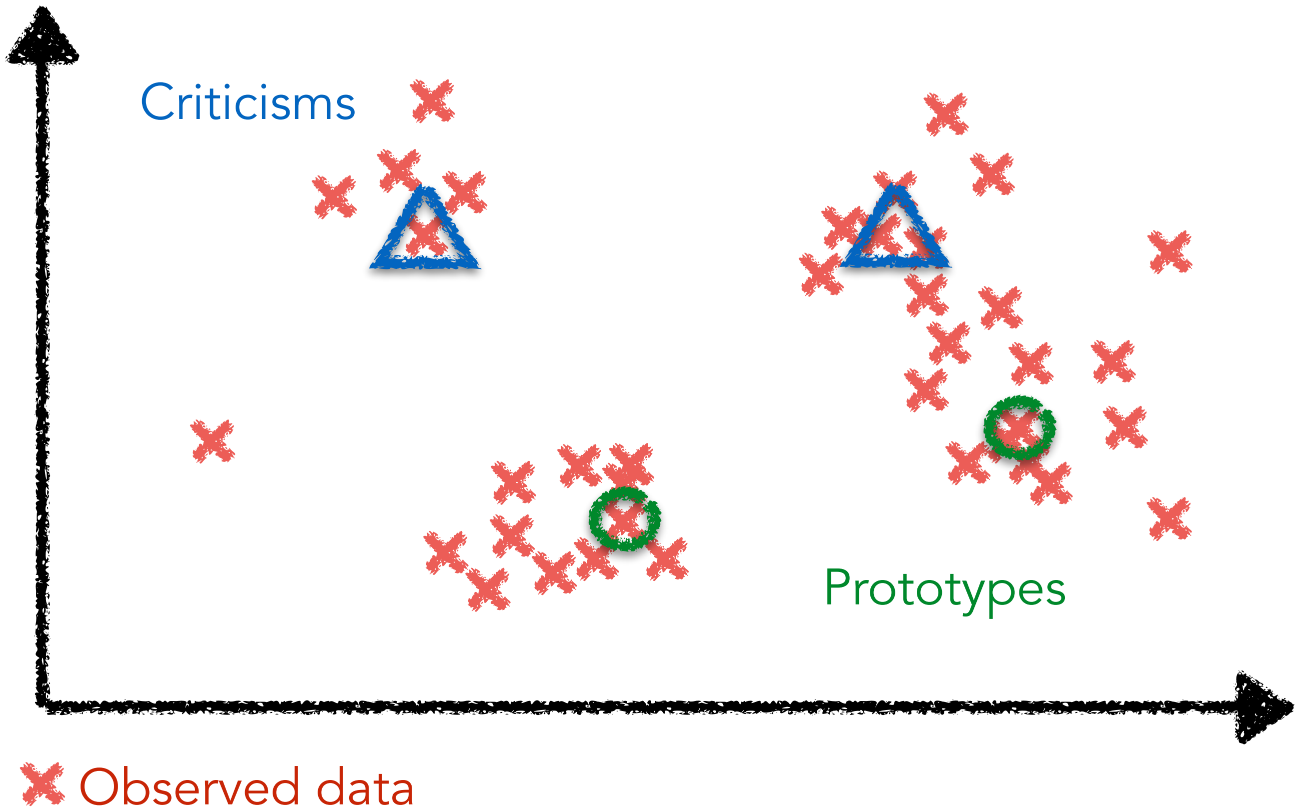
$$\max_{C \subseteq [n] \setminus S, |C| \leq c_*} L(C) = \sum_{l \in C} \left| \frac{1}{n} \sum_{i \in [n]} k(x_i, x_l) - \frac{1}{m} \sum_{j \in S} k(x_j, x_l) \right| + r(\mathbf{K}, C)$$

still submodular



Results

- Eval1: [quantitative] prototype-based classification
- Eval2: [qualitative] prototypes and criticisms across various data sets
- Eval3: [quantitative] Pilot study with human subjects

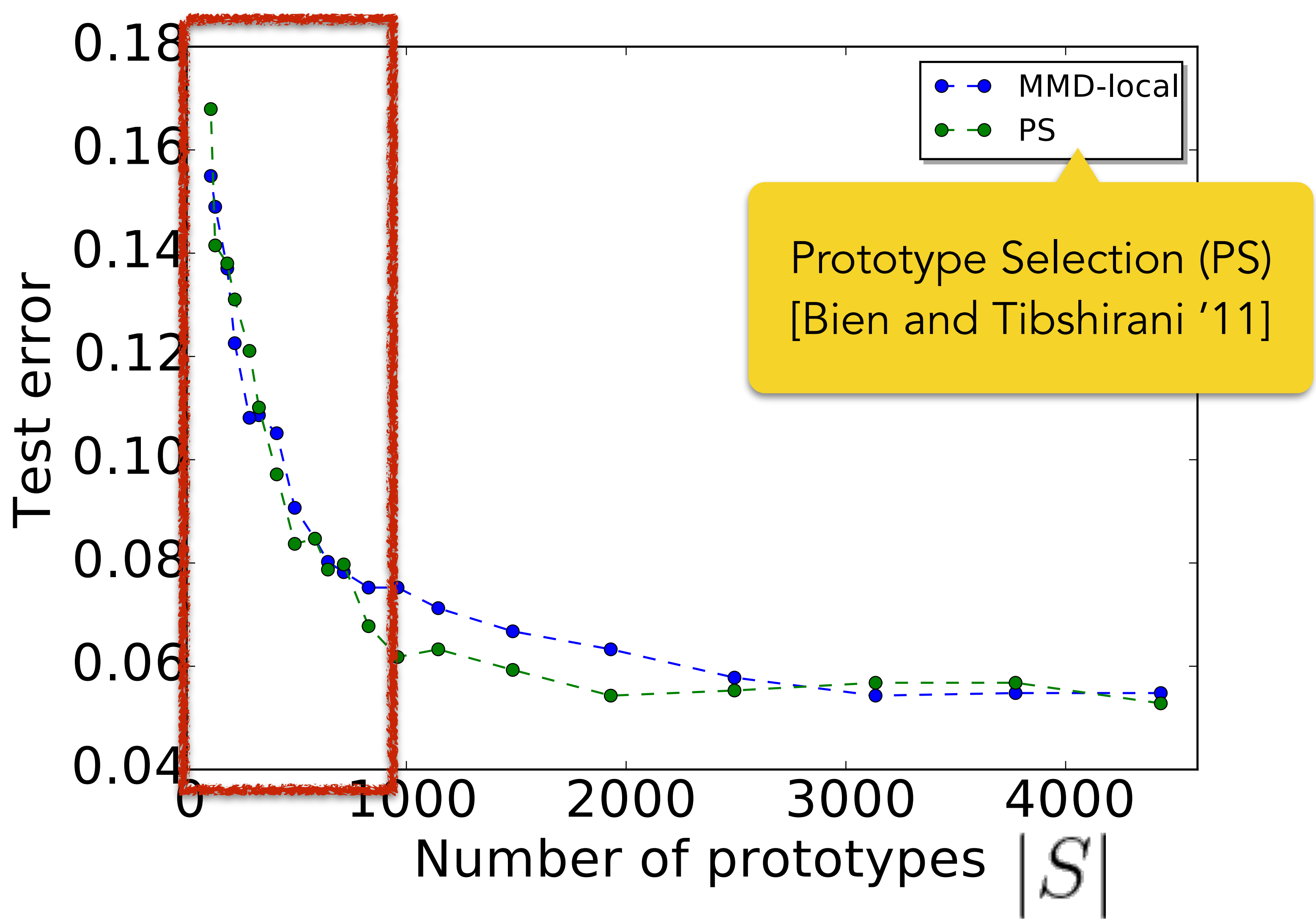


Eval1

Prototype-based classification

- Use the learned prototypes for classification (nearest-neighbor)

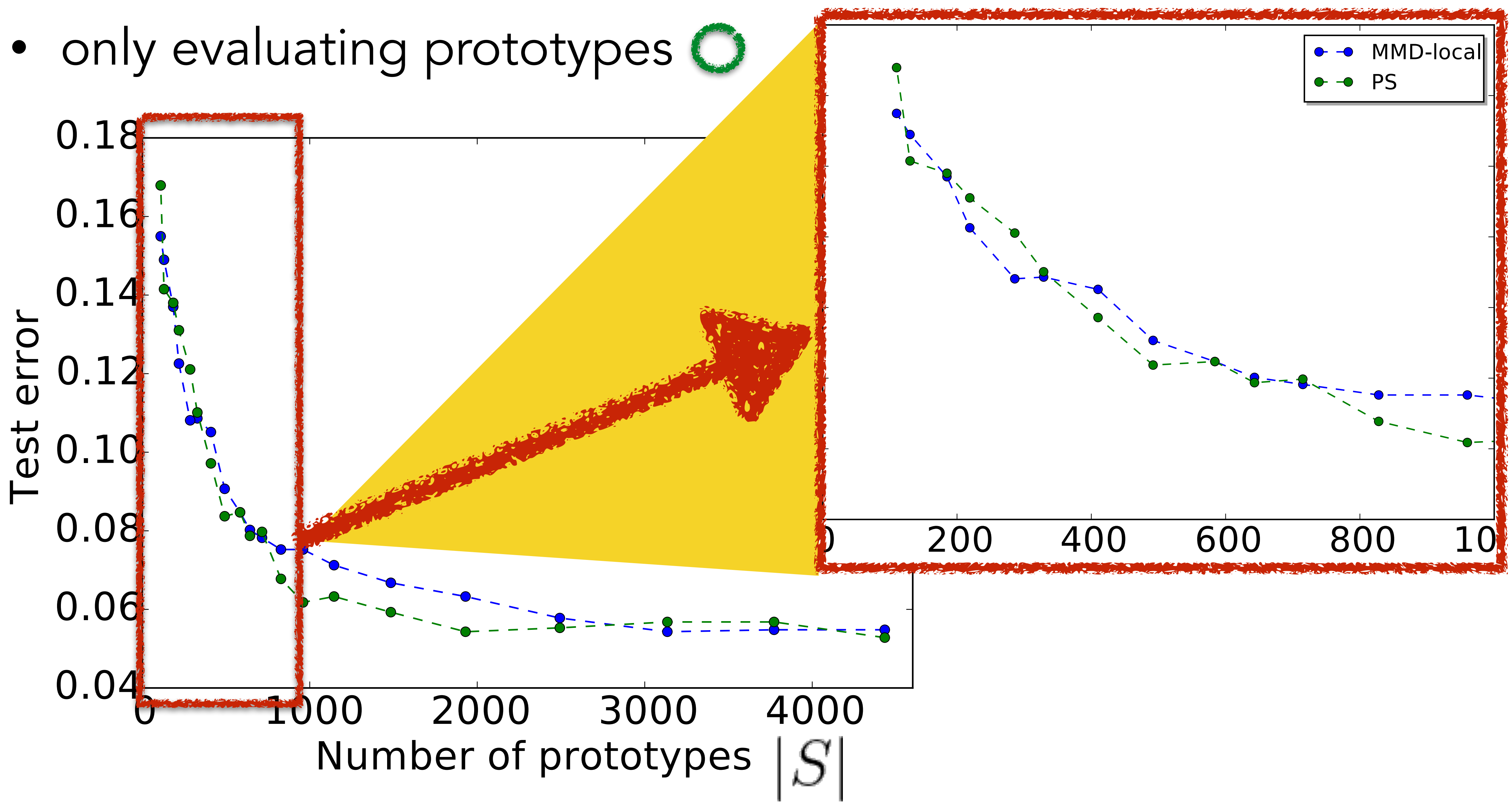
- only evaluating prototypes



Eval1

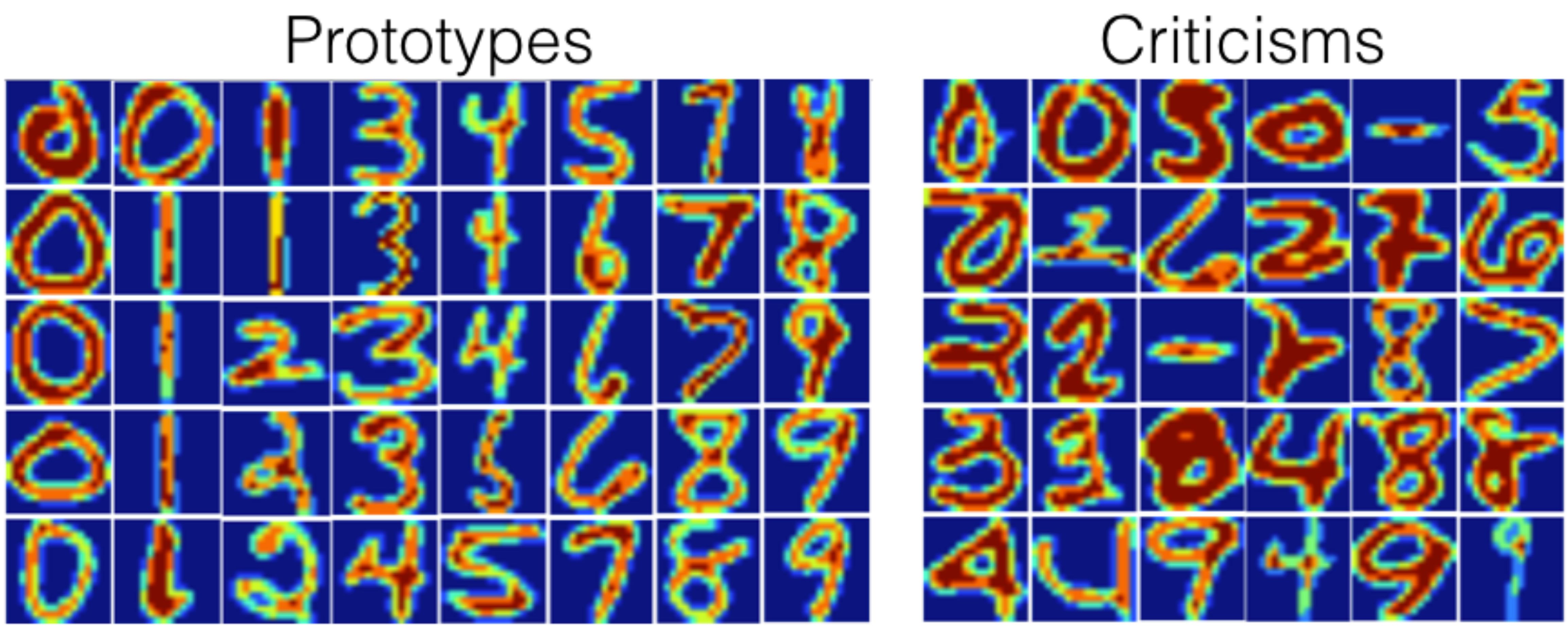
Prototype-based classification

- Use the learned prototypes for classification (nearest-neighbor)
- only evaluating prototypes



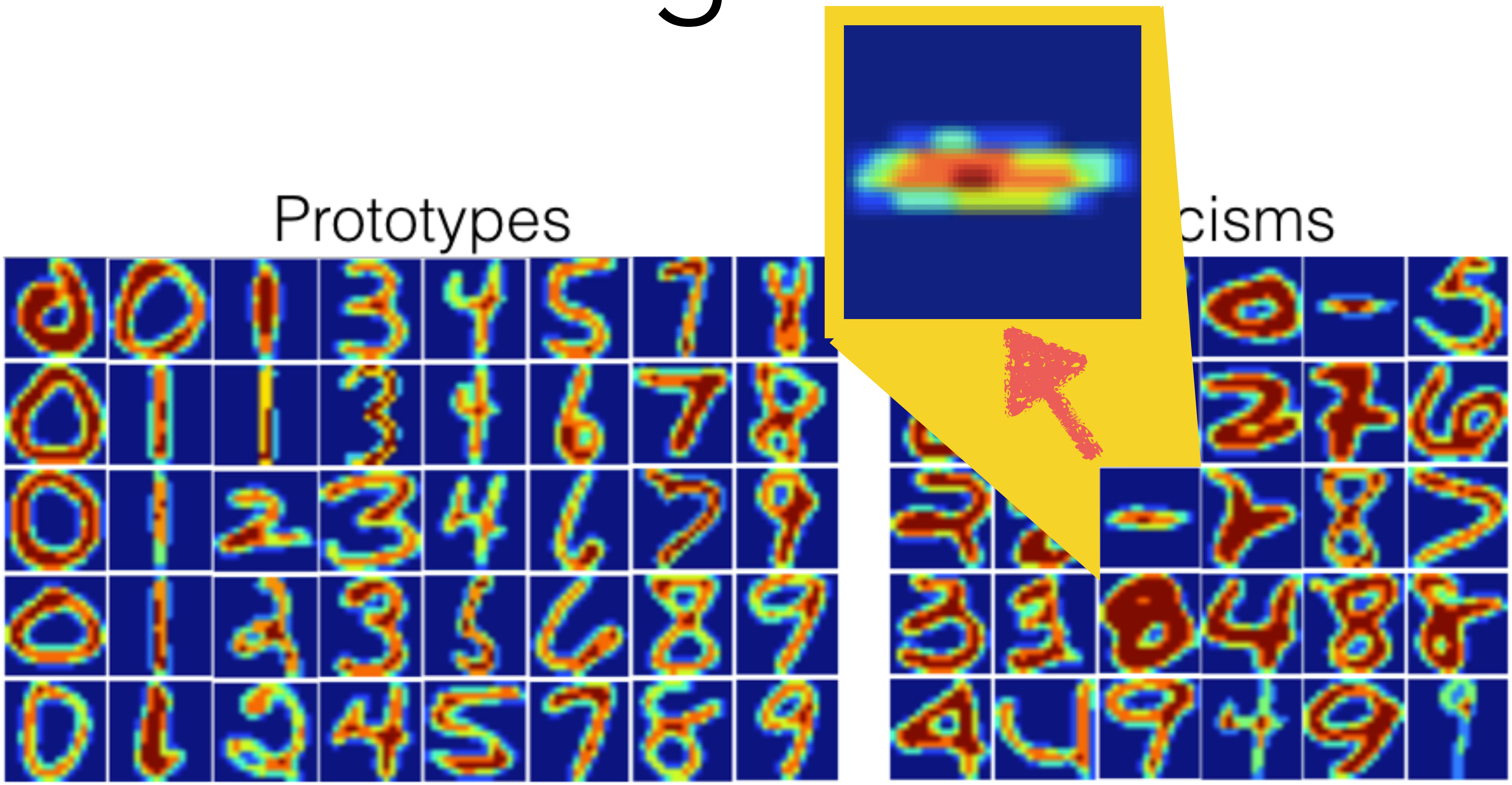
Eval2

USPS digits dataset



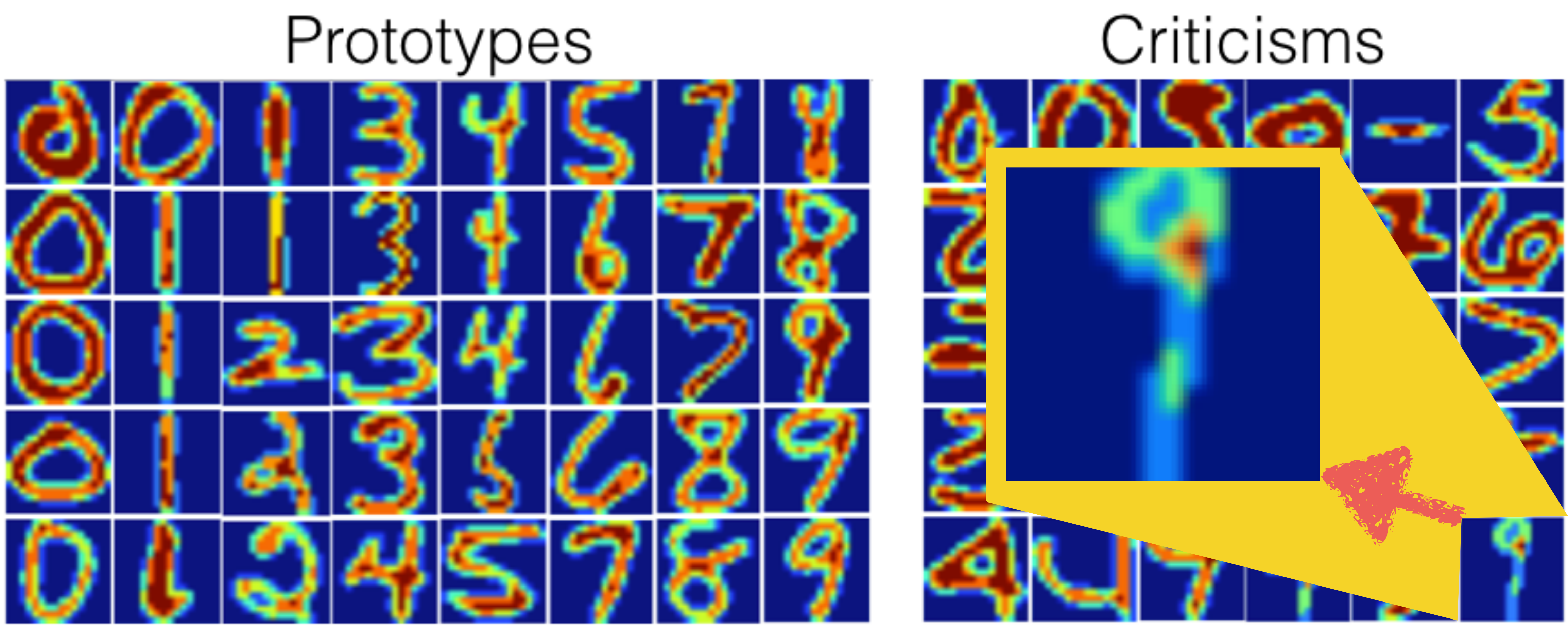
Eval2

USPS digits dataset



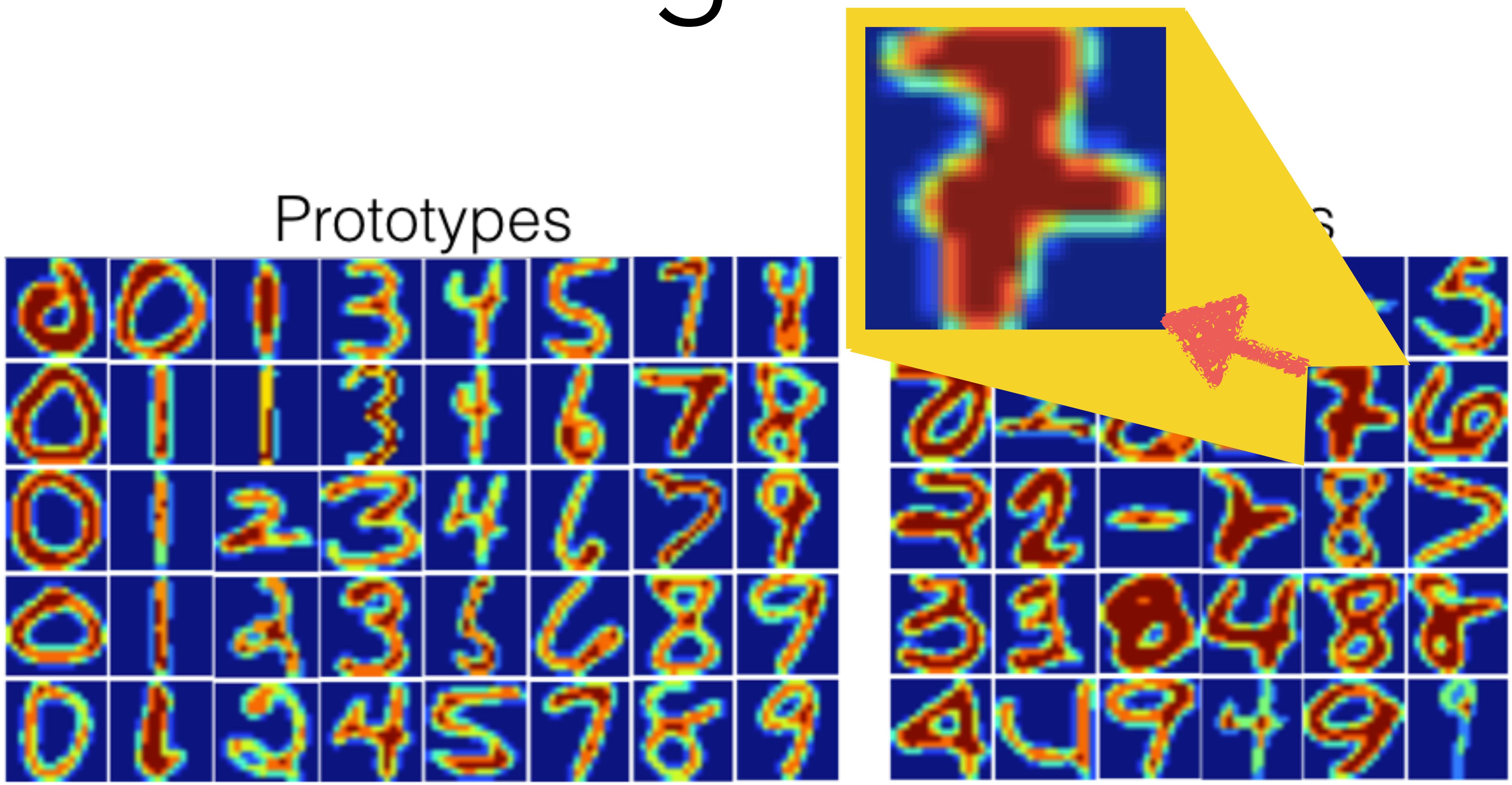
Eval2

USPS digits dataset



Eval2

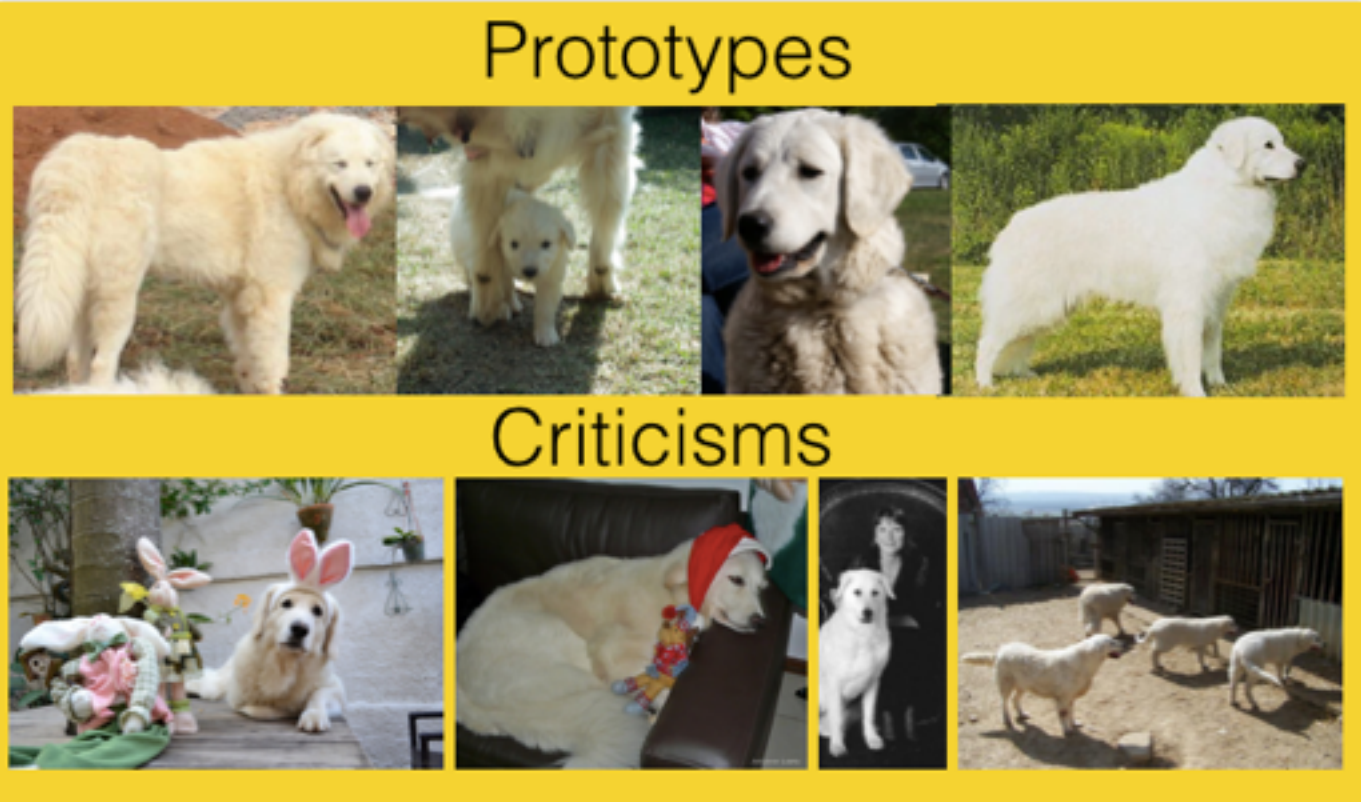
USPS digits dataset



Eval2

ImageNet dataset

- ImageNet dataset [Russakovsky et al '15] using image embeddings from [He '15]



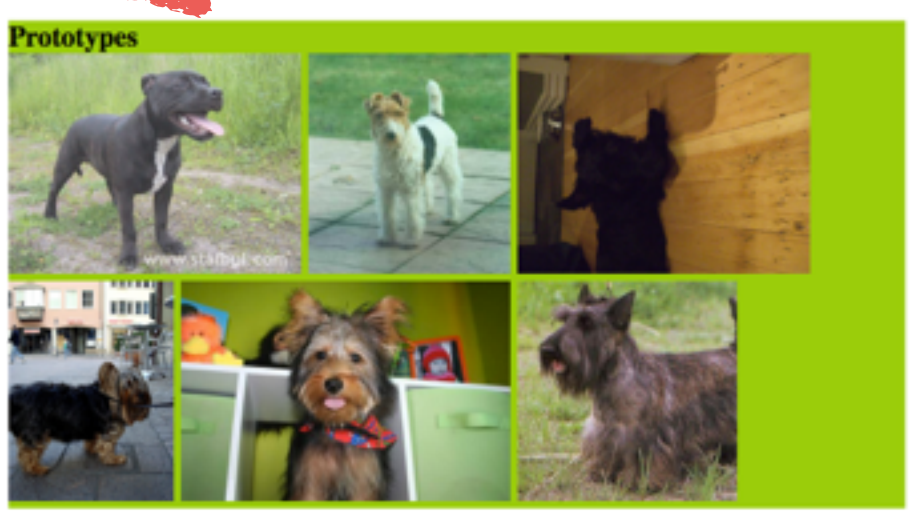
Eval3

Pilot study with human subjects

- Definition of interpretability: A method is interpretable if a user can correctly and efficiently predict the method's results.

- Task: Assign a new data point to one of the groups using 1) all images 2) prototypes 3) prototypes and criticisms 4) small set of randomly selected images

a new data point



group 1

group 2

Eval3

Pilot study with human subjects

- Definition of interpretability: A method is interpretable if a user can correctly and efficiently predict the method's results.

- Task: Assign a new data point to one of the groups using 1) all images 2) prototypes 3) prototypes and criticisms 4) small set of randomly selected images

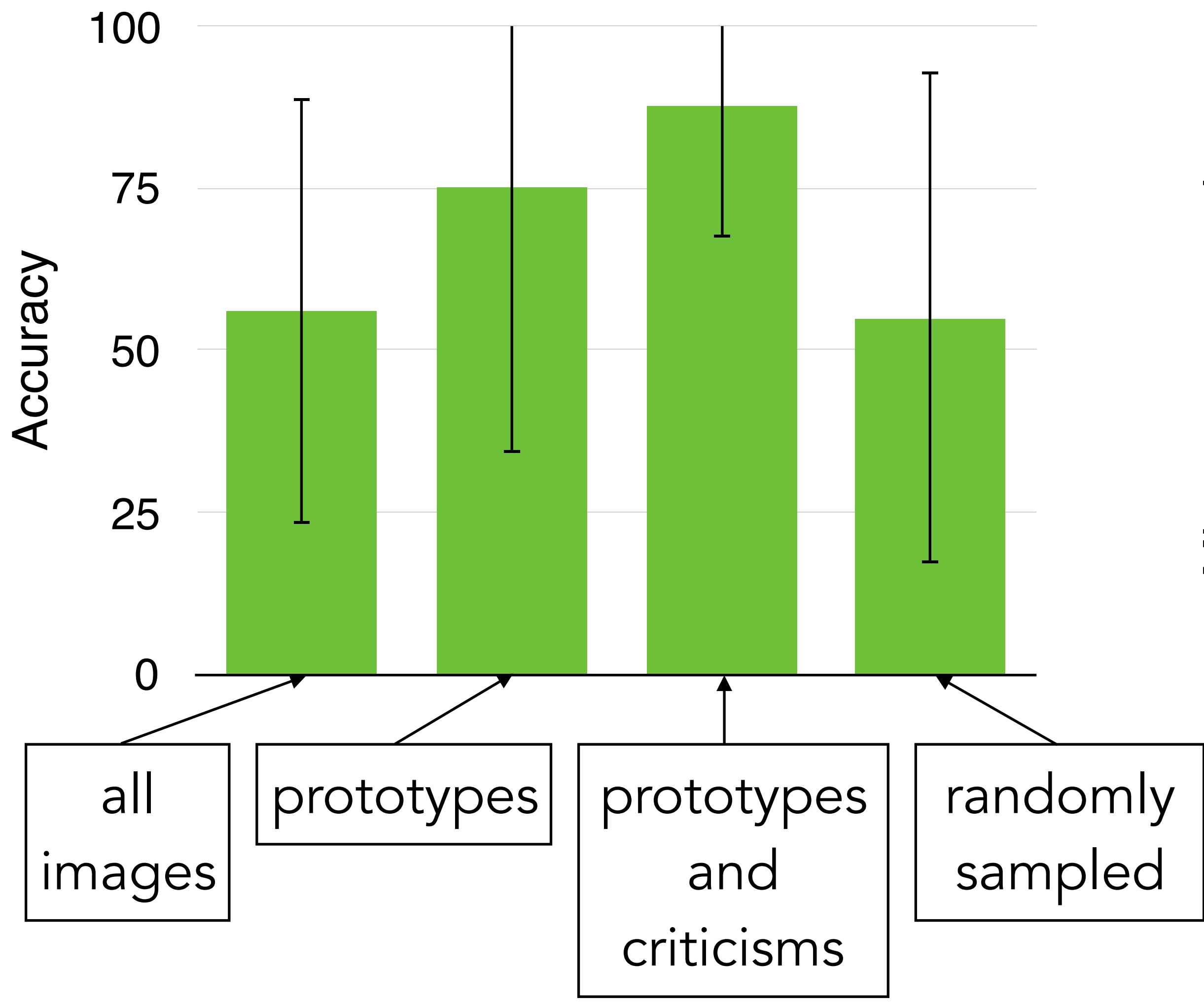
a new data point

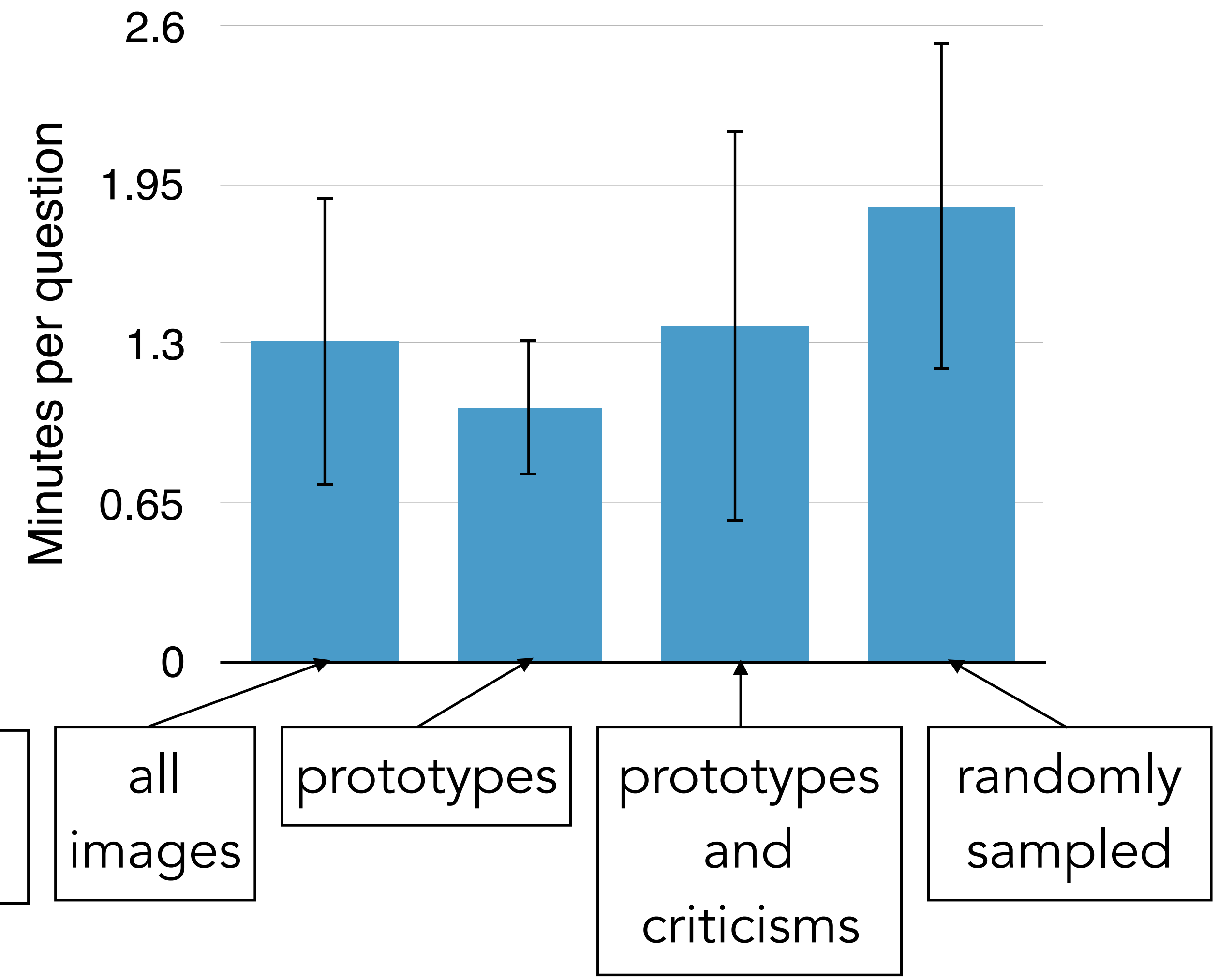
group 1

group 2

Eval3

Pilot study with human subjects





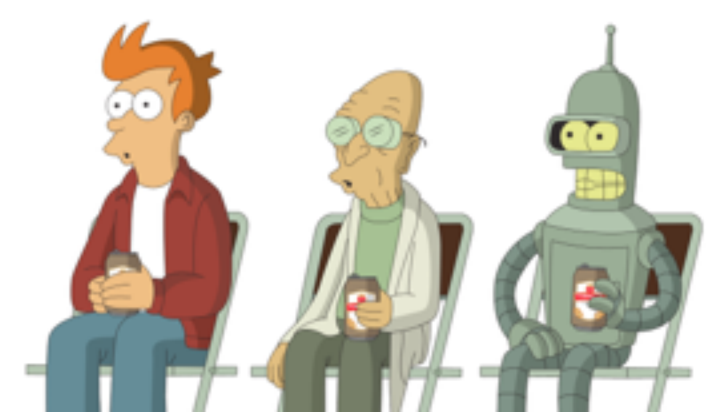
Comment:

"[Proto and Criticism Condition resulted in] less confusion from trying to discover hidden patterns in a ton of images, more clues indicating what features are important"

n = 3
21 questions each

Future work

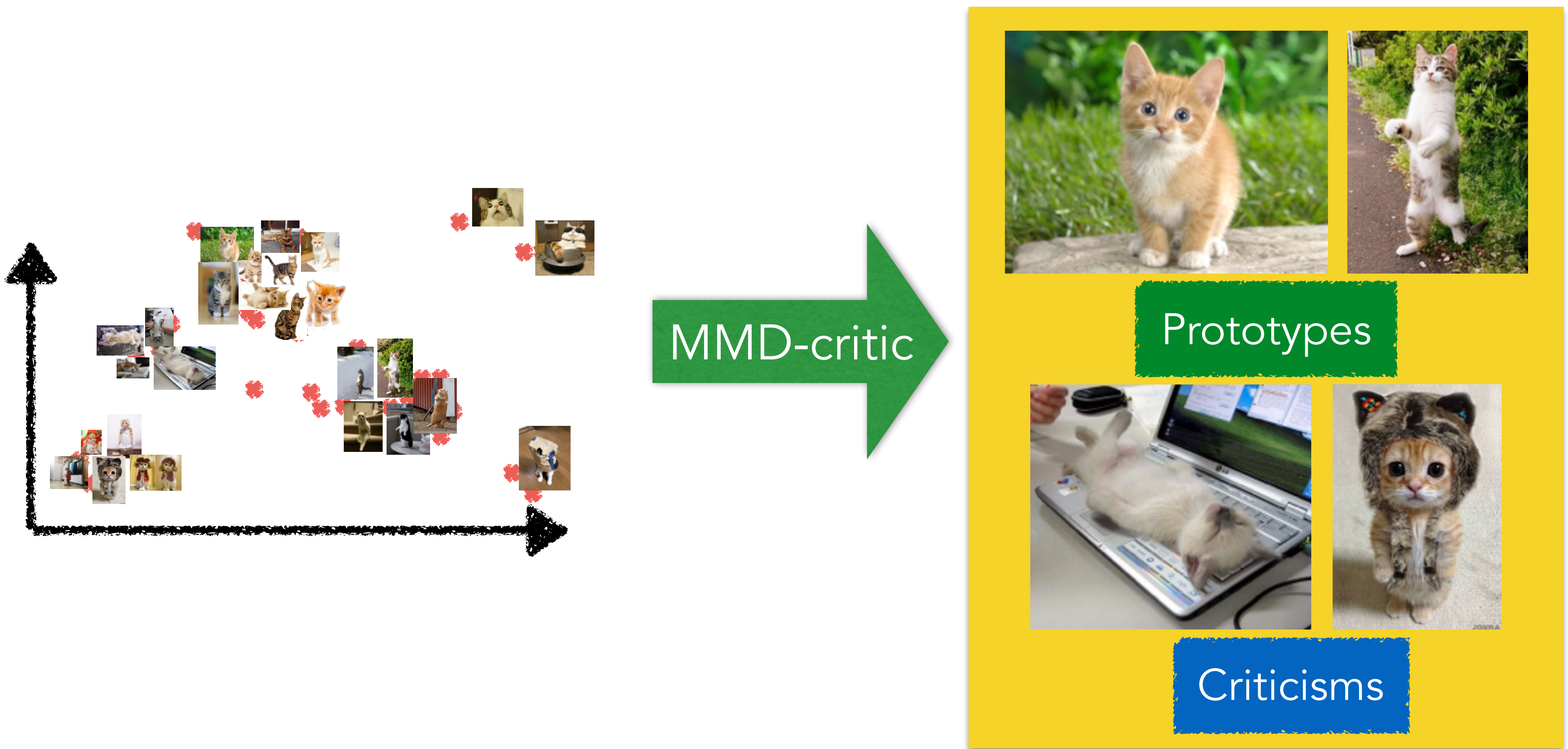
- explore to use for evaluating ML models by using model dependent kernels (e.g., Fisher kernel)
- heuristics to select the number of prototypes and criticisms
- human experiments to compare with outlier methods
- better understand the effect of the choice of kernel



Conclusion

MMD-critic learns **prototypes + criticisms that** highlight aspects of data that are overlooked by prototypes.

code: https://github.com/BeenKim/MMD-critic



Questions?

MMD-critic learns **prototypes + criticisms that** highlight aspects of data that are overlooked by prototypes.

code: https://github.com/BeenKim/MMD-critic