Examples are not enough, learn to criticize! Criticism for Interpretability

Been Kim, Rajiv Khanna, Oluwasanmi Koyejo. NIPS-2016

Zoltán Szabó

Machine Learning Journal Club
CMAP, École Polytechnique

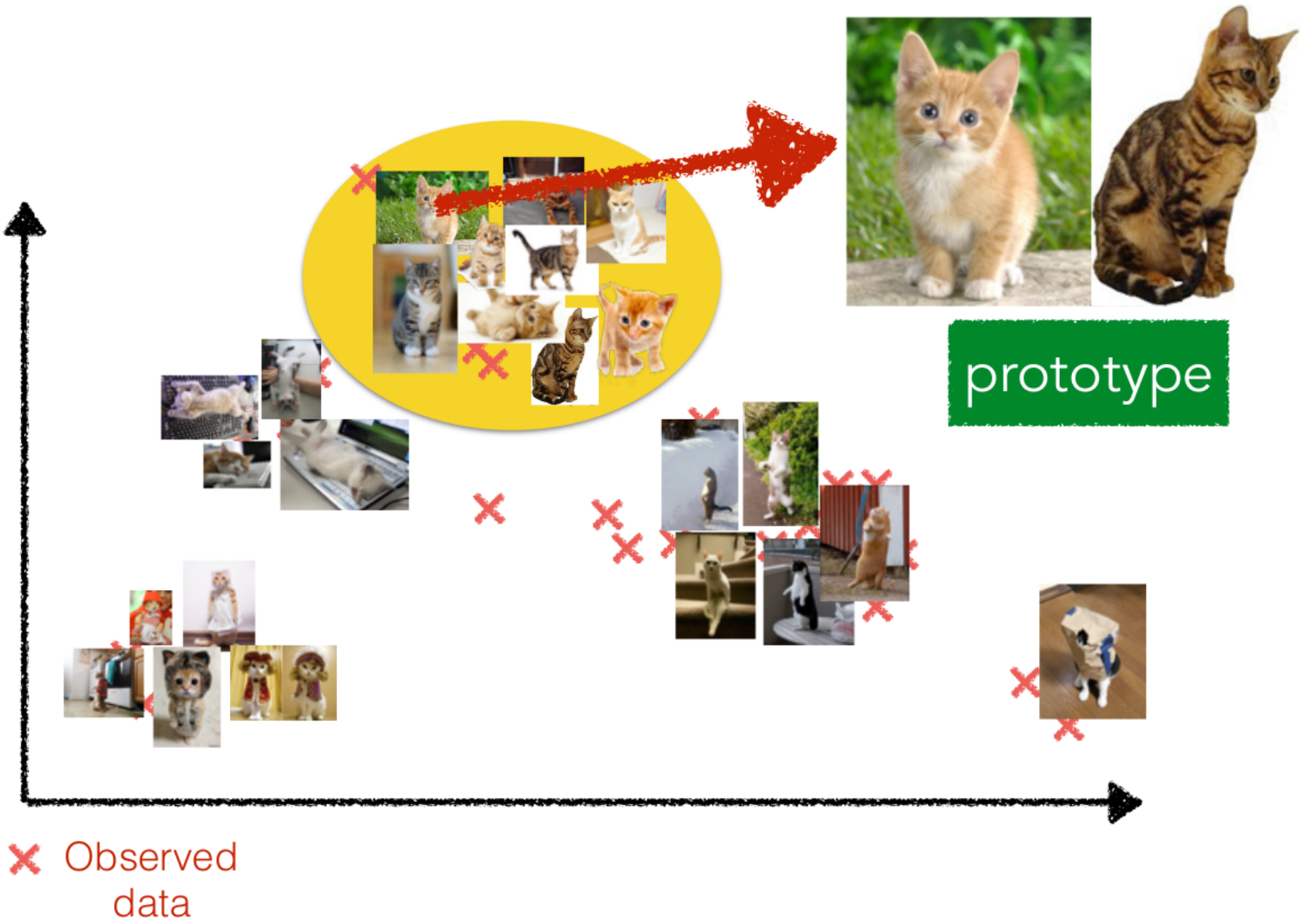
May 4, 2017

Motivation

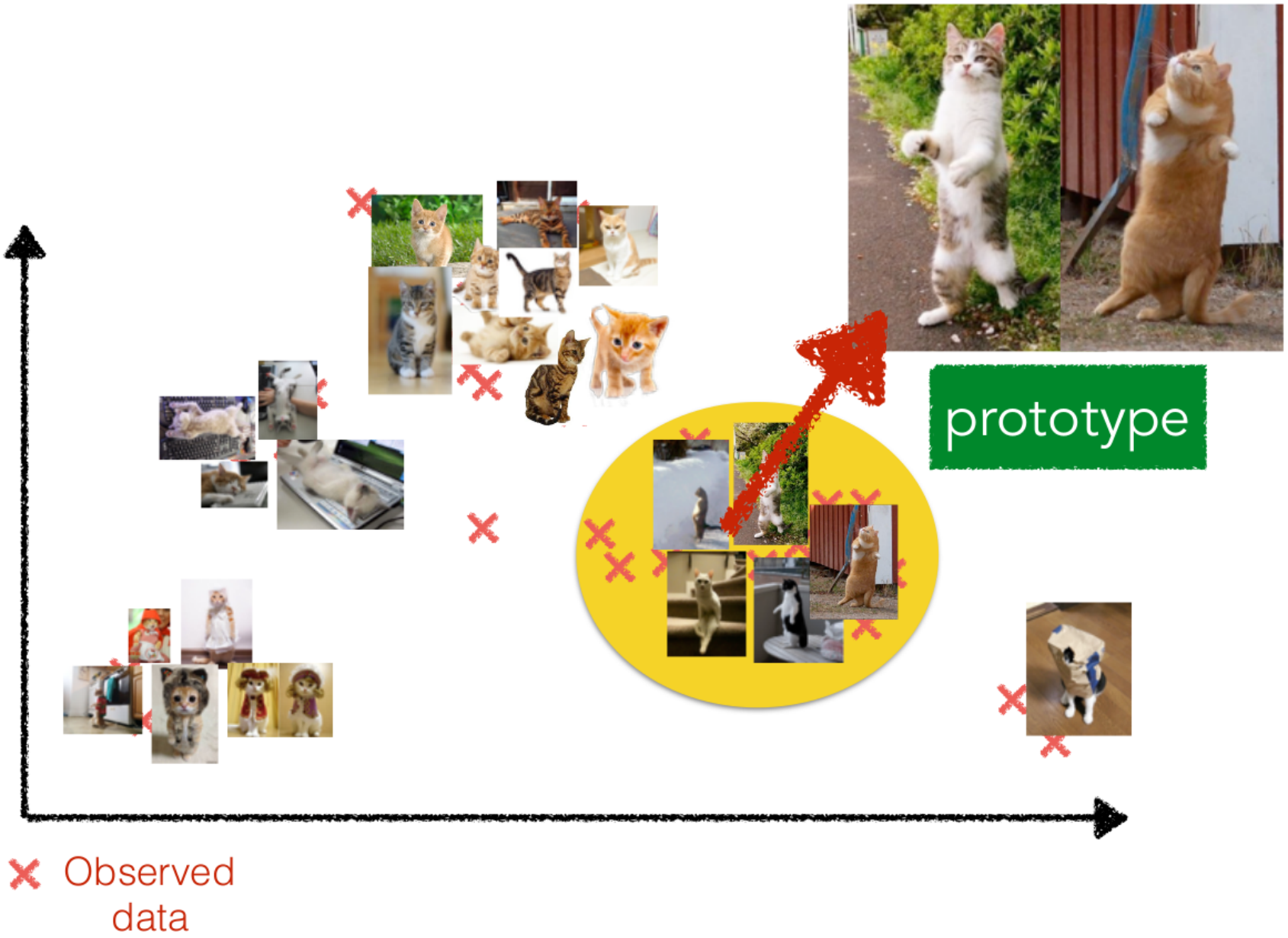
Motivation: dataset

†Illustrations are taken from Been Kim's slides.

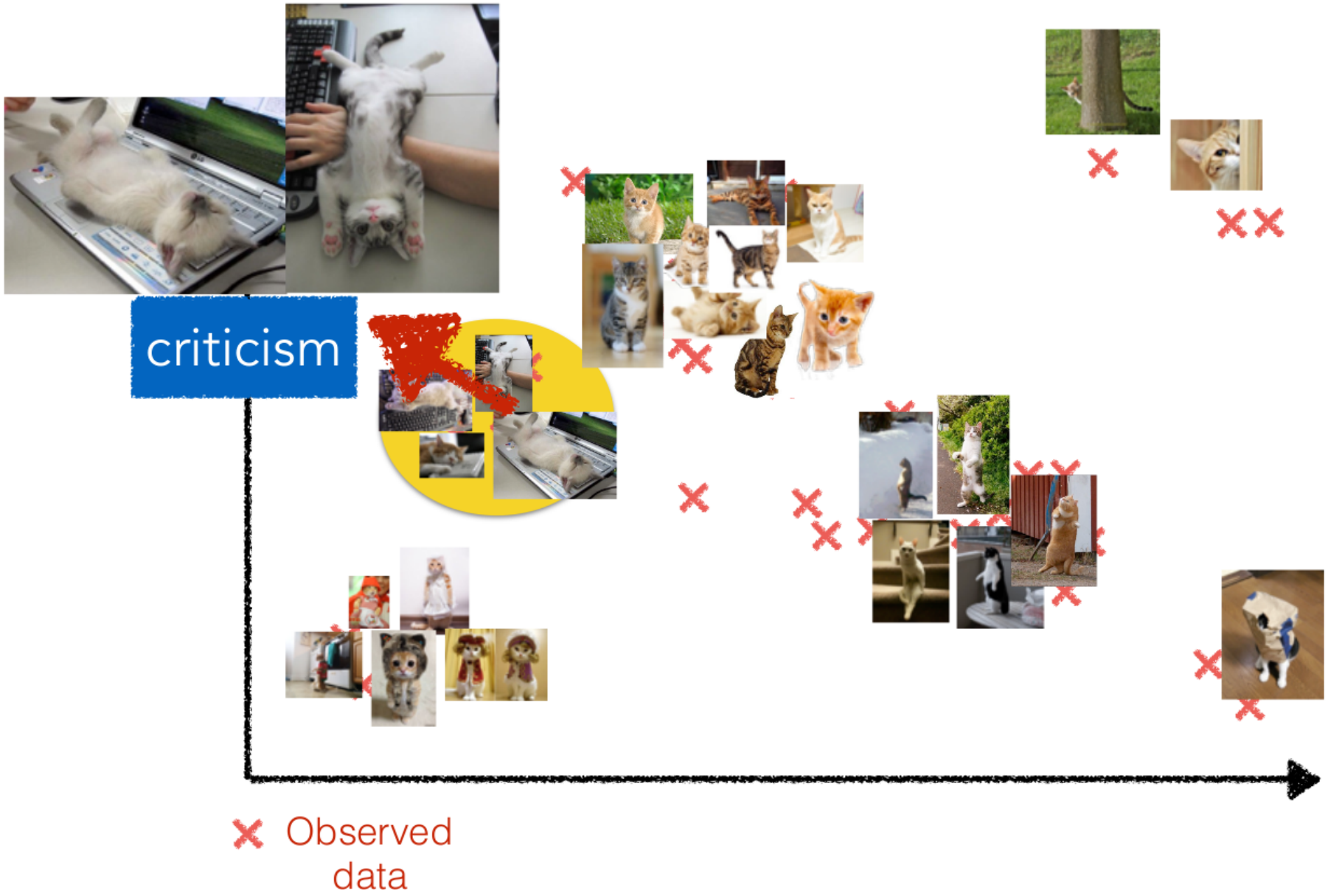
Motivation: prototypes



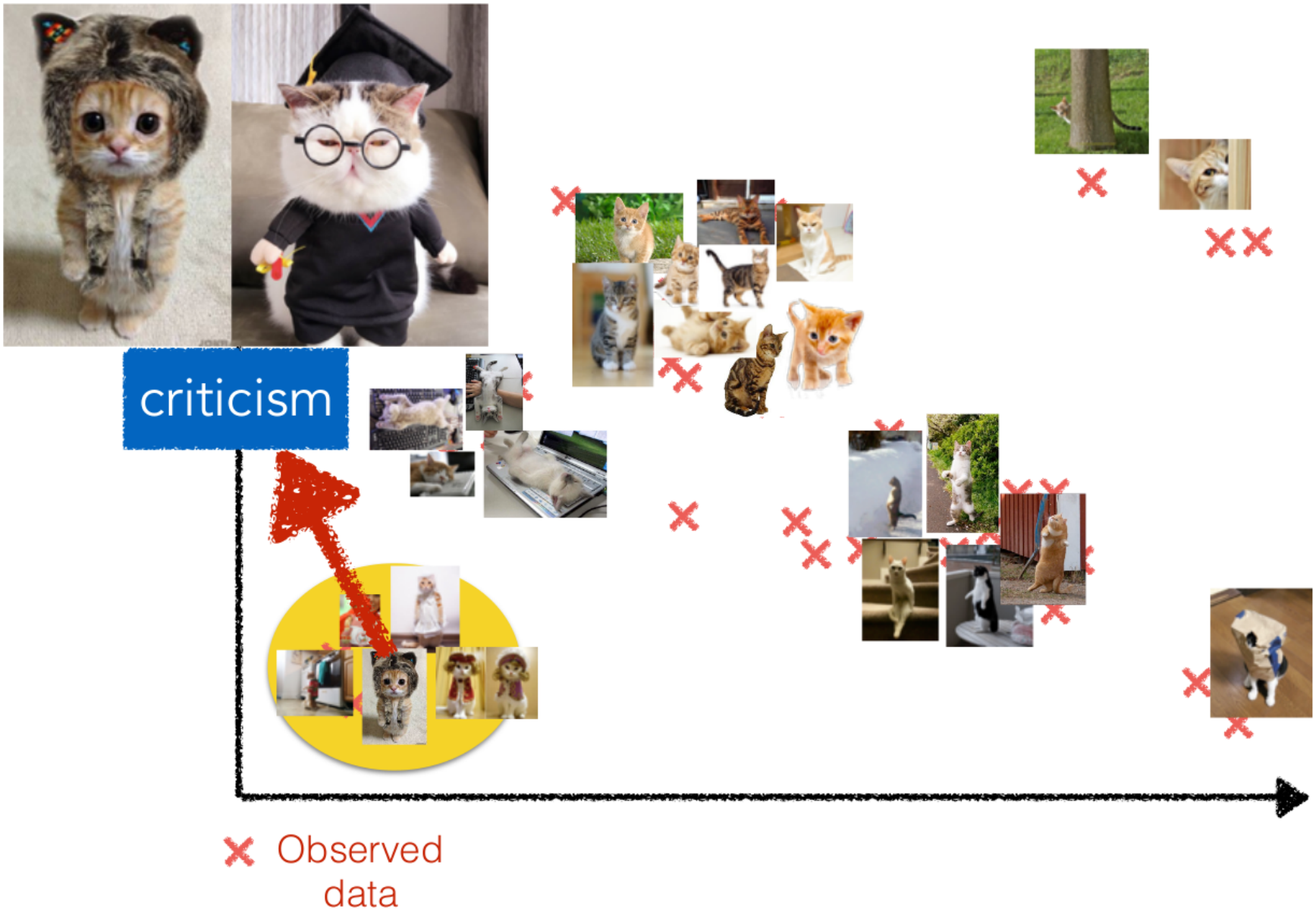
Motivation: prototypes



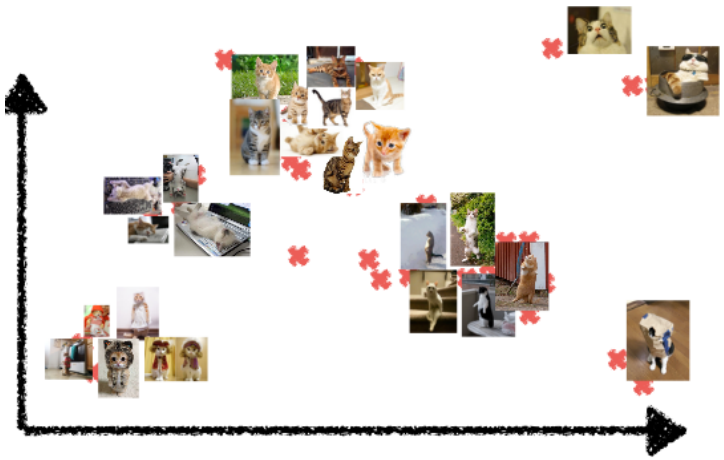
Motivation: criticisms



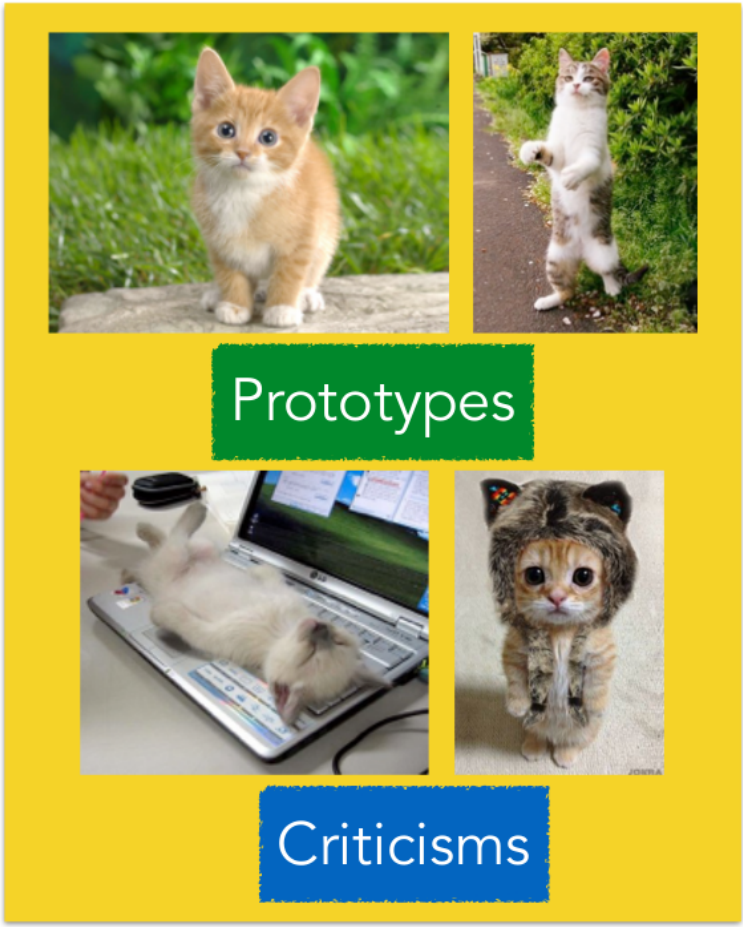
Motivation: criticisms



Motivation: high-level summary



MMD-critic



Task

- Given: dataset (say images).
- Goal:
  - summarize: prototypes (majorities),
  - criticize: atypical examples (minorities).
- Idea: use kernel $\rightarrow$ mean embedding $\rightarrow$ MMD.

Kernel, mean embedding, MMD

Kernel, RKHS definition(s)

Given: $\mathcal{X}$ set.

- Kernel:

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}, \quad \mathcal{F} : \text{some Hilbert space.}$$

Kernel, RKHS definition(s)

Given: $\mathcal{X}$ set.

- Kernel:

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}, \quad \mathcal{F} : \text{some Hilbert space.}$$

- Reproducing kernel of a $\mathcal{H}$(ilbert) $\subset \mathbb{R}^{\mathcal{X}}$:

$$k(\cdot, b) \in \mathcal{H}, \qquad \langle f, k(\cdot, b) \rangle_{\mathcal{H}} = f(b).$$

Kernel, RKHS definition(s)

Given: $\mathcal{X}$ set.

- Kernel:

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}, \quad \mathcal{F} : \text{some Hilbert space.}$$

- Reproducing kernel of a $\mathcal{H}$(ilbert) $\subset \mathbb{R}^{\mathcal{X}}$:

$$k(\cdot, b) \in \mathcal{H}, \qquad \langle f, k(\cdot, b) \rangle_{\mathcal{H}} = f(b).$$

$$\xrightarrow{\text{spec.}} k(a,b) = \langle k(\cdot, a), k(\cdot, b) \rangle_{\mathcal{H}}.$$

Kernel examples on $\mathcal{X} = \mathbb{R}^d$, $\theta > 0$

$$k_G(a,b) = e^{-\frac{\|a-b\|_2^2}{2\theta^2}}, \qquad k_e(a,b) = e^{-\frac{\|a-b\|_2}{2\theta^2}},$$

$$k_C(a,b) = \frac{1}{1+\frac{\|a-b\|_2^2}{\theta^2}}, \qquad k_t(a,b) = \frac{1}{1+\|a-b\|_2^\theta},$$

$$k_p(a,b) = (\langle a,b\rangle + \theta)^p, \quad k_r(a,b) = 1 - \frac{\|a-b\|_2^2}{\|a-b\|_2^2+\theta},$$

$$k_i(a,b) = \frac{1}{\sqrt{\|a-b\|_2^2+\theta^2}},$$

$$k_{M,\frac{3}{2}}(a,b) = \left(1+\frac{\sqrt{3}\,\|a-b\|_2}{\theta}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\theta}}.$$

Mean embedding: kernel trick $\rightarrow$ mean trick

- Kernel: $k(x, x') = \left\langle k(\cdot, x), k(\cdot, x') \right\rangle_{\mathcal{H}_k}$, $\mu_{\mathbb{P}} := k(\cdot, x)$, $\mathbb{P} = \delta_x$.

Mean embedding: kernel trick $\rightarrow$ mean trick

- Kernel: $k(x, x') = \left\langle k(\cdot, x), k(\cdot, x') \right\rangle_{\mathcal{H}_k}$, $\mu_{\mathbb{P}} := k(\cdot, x)$, $\mathbb{P} = \delta_x$.
- Mean embedding (feature of $\mathbb{P}$):

$$\mathbb{P} = \sum_{i=1}^{N} w_i \delta_{x_i},$$

Mean embedding: kernel trick $\rightarrow$ mean trick

- Kernel: $k(x, x') = \left\langle k(\cdot, x), k(\cdot, x') \right\rangle_{\mathcal{H}_k}$, $\mu_{\mathbb{P}} := k(\cdot, x)$, $\mathbb{P} = \delta_x$.
- Mean embedding (feature of $\mathbb{P}$):

$$\mu_{\mathbb{P}} := \sum_{i=1}^{N} w_i k(\cdot, x_i) \in \mathcal{H}_k, \qquad \mathbb{P} = \sum_{i=1}^{N} w_i \delta_{x_i},$$

Mean embedding: kernel trick $\rightarrow$ mean trick

- Kernel: $k(x, x') = \left\langle k(\cdot, x), k(\cdot, x') \right\rangle_{\mathcal{H}_k}$, $\mu_{\mathbb{P}} := k(\cdot, x)$, $\mathbb{P} = \delta_x$.
- Mean embedding (feature of $\mathbb{P}$):

$$\mu_{\mathbb{P}} := \sum_{i=1}^{N} w_i k(\cdot, x_i) \in \mathcal{H}_k, \quad \mathbb{P} = \sum_{i=1}^{N} w_i \delta_{x_i},$$

$$\mu_{\mathbb{P}} := \mathbb{E}_{x \sim \mathbb{P}}[k(\cdot, x)] = \underbrace{\int k(\cdot, x) \mathrm{d}\mathbb{P}(x)}_{\text{Bochner integral}} \in \mathcal{H}_k.$$

Mean embedding: kernel trick $\rightarrow$ mean trick

- Kernel: $k(x,x') = \left\langle k(\cdot,x), k(\cdot,x') \right\rangle_{\mathcal{H}_k}$, $\mu_{\mathbb{P}} := k(\cdot,x)$, $\mathbb{P} = \delta_x$.
- Mean embedding (feature of $\mathbb{P}$):

$$\mu_{\mathbb{P}} := \sum_{i=1}^{N} w_i k(\cdot,x_i) \in \mathcal{H}_k, \quad \mathbb{P} = \sum_{i=1}^{N} w_i \delta_{x_i},$$

$$\mu_{\mathbb{P}} := \mathbb{E}_{x\sim\mathbb{P}}[k(\cdot,x)] = \underbrace{\int k(\cdot,x)\mathrm{d}\mathbb{P}(x)}_{\text{Bochner integral}} \in \mathcal{H}_k.$$

- $\exists \mu_{\mathbb{P}} \Leftrightarrow \int \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}} \mathrm{d}\mathbb{P}(x) < \infty$. Example: bounded k.

Maximum mean discrepancy (MMD): specific IPM

$$MMD(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}$$

Maximum mean discrepancy (MMD): specific IPM

$$
\begin{aligned}
MMD(\mathbb{P},\mathbb{Q}) &= \|\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\|_{\mathcal{H}_k} \\
&= \sup_{f\in\mathcal{B}_k:=\left\{f:\|f\|_{\mathcal{H}_k}\le 1\right\}} \langle f, \mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\rangle_{\mathcal{H}_k}
\end{aligned}
$$

Maximum mean discrepancy (MMD): specific IPM

$$
\begin{aligned}
MMD(\mathbb{P},\mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} \\
&= \sup_{f\in\mathcal{B}_k := \left\{f : \|f\|_{\mathcal{H}_k} \le 1\right\}} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} \\
&= \sup_{f\in\mathcal{B}_k} \left[\mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{z\sim\mathbb{Q}} f(z)\right].
\end{aligned}
$$

MMD estimation

Using $\{x_i\}_{i=1}^n \sim \mathbb{P}, \{z_j\}_{j=1}^m \sim \mathbb{Q}$,

$$\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) = MMD^2\left(\hat{\mathbb{P}}, \hat{\mathbb{Q}}\right)$$

MMD estimation

Using $\{x_i\}_{i=1}^n \sim \mathbb{P}, \{z_j\}_{j=1}^m \sim \mathbb{Q}$,

$$\widehat{MMD^2}(\mathbb{P},\mathbb{Q}) = MMD^2\left(\hat{\mathbb{P}},\hat{\mathbb{Q}}\right) = \left\|\mu_{\hat{\mathbb{P}}} - \mu_{\hat{\mathbb{Q}}}\right\|_{\mathcal{H}_k}^2$$

MMD estimation

Using $\{x_i\}_{i=1}^n \sim \mathbb{P}, \{z_j\}_{j=1}^m \sim \mathbb{Q}$,

$$\widehat{MMD^2}(\mathbb{P},\mathbb{Q}) = MMD^2\left(\hat{\mathbb{P}},\hat{\mathbb{Q}}\right) = \left\|\mu_{\hat{\mathbb{P}}} - \mu_{\hat{\mathbb{Q}}}\right\|_{\mathcal{H}_k}^2$$

$$= \left\|\frac{1}{n}\sum_{i=1}^n k(\cdot,x_i) - \frac{1}{m}\sum_{j=1}^m k(\cdot,z_j)\right\|_{\mathcal{H}_k}^2$$

MMD estimation

Using $\{x_i\}_{i=1}^n \sim \mathbb{P}, \{z_j\}_{j=1}^m \sim \mathbb{Q}$,

$$\widehat{MMD^2}(\mathbb{P},\mathbb{Q}) = MMD^2\left(\hat{\mathbb{P}},\hat{\mathbb{Q}}\right) = \left\|\mu_{\hat{\mathbb{P}}} - \mu_{\hat{\mathbb{Q}}}\right\|_{\mathcal{H}_k}^2$$

$$= \left\|\frac{1}{n}\sum_{i=1}^n k(\cdot,x_i) - \frac{1}{m}\sum_{j=1}^m k(\cdot,z_j)\right\|_{\mathcal{H}_k}^2$$

$$= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i,x_j) + \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(z_i,z_j)$$

$$- \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m k(x_i,z_j).$$

IPM representation $\longrightarrow$ witness function of MMD

- Recall:

$$MMD(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{B}_k} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}.$$

IPM representation $\longrightarrow$ witness function of MMD

- Recall:

$$MMD(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{B}_k} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}.$$

- Witness function

$$f^*_{\mathbb{P},\mathbb{Q}} = \frac{\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}}{\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}} \propto \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}$$

IPM representation $\longrightarrow$ witness function of MMD

- Recall:

$$MMD(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{B}_k} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}.$$

- Witness function, empirical witness function:

$$f^*_{\mathbb{P},\mathbb{Q}} = \frac{\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}}{\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}} \propto \mu_{\mathbb{P}} - \mu_{\mathbb{Q}},$$

$$f^*_{\hat{\mathbb{P}},\hat{\mathbb{Q}}} = \frac{\mu_{\hat{\mathbb{P}}} - \mu_{\hat{\mathbb{Q}}}}{\|\mu_{\hat{\mathbb{P}}} - \mu_{\hat{\mathbb{Q}}}\|_{\mathcal{H}_k}} \propto \mu_{\hat{\mathbb{P}}} - \mu_{\hat{\mathbb{Q}}}.$$

Local summary

- Kernel:

$$k(x, x') = \left\langle \varphi(x), \varphi(x') \right\rangle_{\mathcal{H}},$$

  where $\varphi(x) := k(\cdot, x)$, $\mathcal{H} := \mathcal{H}_k$.

- Mean embedding, MMD, witness function:

$$\mu_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}}[k(\cdot, x)],$$
$$MMD(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k},$$
$$f^*_{\mathbb{P},\mathbb{Q}} \propto \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}.$$

Summarization objective

- Given:
  - $X = \{x_i\}_{i=1}^n$,
  - m^*: # of desired prototypes, $m^* < n$.

Summarization objective

- Given:
  - $X = \{x_i\}_{i=1}^n$,
  - m^*: # of desired prototypes, $m^* < n$.
- Choose subset S by

$$\min_{S \subset \{1,\dots,n\}, |S| \leq m^*} MMD^2\left(X, X_S\right).$$

Criticism objective

- Given:
  - S: selected prototypes.
  - c^*: # of desired criticism samples, $< n - m^*$.

Criticism objective

- Given:
  - S: selected prototypes.
  - c^*: # of desired criticism samples, $< n - m^*$.
- Witness function:

$$f^*_{X,X_S} \propto \mu_X - \mu_{X_S} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, x_i) - \frac{1}{|S|}\sum_{j\in S} k(\cdot, x_j).$$

Criticism objective

- Given:
  - S: selected prototypes.
  - c^*: # of desired criticism samples, $< n - m^*$.
- Witness function:

$$f^*_{X,X_S} \propto \mu_X - \mu_{X_S} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, x_i) - \frac{1}{|S|}\sum_{j\in S} k(\cdot, x_j).$$

- Objective function

$$\max_{C\subset\{1,\ldots,n\}\setminus S, |C|\leq c_*} \underbrace{\sum_{\ell\in C} \left|f^*_{X,X_S}(x_\ell)\right|}_{\text{criticise}}$$

Criticism objective

- Given:
  - S: selected prototypes.
  - c^*: # of desired criticism samples, $< n - m^*$.
- Witness function:

$$f^*_{X,X_S} \propto \mu_X - \mu_{X_S} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, x_i) - \frac{1}{|S|}\sum_{j\in S} k(\cdot, x_j).$$

- Objective function, $\mathbf{K} = [k(x_i, x_j)]$:

$$\max_{C\subset\{1,\dots,n\}\setminus S, |C|\le c_*} \underbrace{\sum_{\ell\in C} \left|f^*_{X,X_S}(x_\ell)\right|}_{\text{criticise}} + \underbrace{\log\det \mathbf{K}_{C,C}}_{\text{diversity of } X_C}.$$

Combinatorial optimization problems:

$$\max_{S\subset\{1,\dots,n\},|S|\leq m^*} F_1(S) := -MMD^2\left(X, X_S\right),$$

$$\max_{C\subset\{1,\dots,n\}\setminus S,|C|\leq c_*} F_2(C) := \sum_{\ell\in C}\left|f^*_{X,X_S}(x_\ell)\right| + \log\det \mathbf{K}_{C,C}.$$

Combinatorial optimization problems:

$$\max_{S\subset\{1,\dots,n\},|S|\le m^*} F_1(S) := -MMD^2(X, X_S),$$

$$\max_{C\subset\{1,\dots,n\}\setminus S,|C|\le c_*} F_2(C) := \sum_{\ell\in C} \left|f^*_{X,X_S}(x_\ell)\right| + \log\det \mathbf{K}_{C,C}.$$

Idea:

- greedy optimization of F

$$S_{t+1} := S_t \cup \operatorname*{arg\,max}_{u\in\{1,\dots,n\}\setminus S_t} F(S_t\cup\{u\})$$

gives $\underbrace{\left(1-\frac{1}{e}\right)}_{\approx 0.63}$-optimal S^*, if F is ...

. . . if F is

1. normalized: $F(\emptyset) = 0$.

. . . if F is

1. normalized: $F(\emptyset) = 0$.
2. monotone:

$$A \subseteq B \Rightarrow F(A) \leq F(B), \ \forall A, B.$$

. . . if F is

1. normalized: $F(\emptyset) = 0$.
2. monotone:

$$A \subseteq B \Rightarrow F(A) \leq F(B), \ \forall A, B.$$

3. submodular:

$$F(A \cup B) + F(A \cap B) \leq F(A) + F(B), \ \forall A, B.$$

Example:

- $F(A) = |A|$.

. . . if F is

1. normalized: $F(\emptyset) = 0$.
2. monotone:

$$A \subseteq B \Rightarrow F(A) \leq F(B), \ \forall A, B.$$

3. submodular:

$$F(A \cup B) + F(A \cap B) \leq F(A) + F(B), \ \forall A, B.$$

Example:

- $F(A) = |A|$.
- $F(A) = g(|A|)$ submodular $\Leftrightarrow$ g: concave.

Can we guarantee these conditions?

We focus on F_1. Recall:

$$F_1(S) = -MMD^2(X, X_S)$$
$$= -\frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) + \underbrace{\frac{2}{n|S|}\sum_{i=1}^{n}\sum_{j\in S} k(x_i, x_j) - \frac{1}{|S|^2}\sum_{i,j\in S} k(x_i, x_j)}_{=0 \text{ if } S=\emptyset}.$$

Can we guarantee these conditions?

We focus on F_1. Recall:

$$F_1(S) = -MMD^2(X, X_S)$$
$$= -\frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) + \underbrace{\frac{2}{n|S|}\sum_{i=1}^{n}\sum_{j\in S} k(x_i, x_j) - \frac{1}{|S|^2}\sum_{i,j\in S} k(x_i, x_j)}_{=0 \text{ if } S=\emptyset}.$$

$$F_1(S) := \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) - MMD^2(X, X_S).$$

$F_1 \leftarrow F_1 - F_1(\emptyset)$

The new F_1 is normalized: $F_1(\emptyset) = 0$.

Towards monotonicity and submodularity of F_1

$$F_1(S) = \frac{2}{n|S|} \sum_{i=1}^{n} \sum_{j \in S} k(x_i, x_j) - \frac{1}{|S|^2} \sum_{i,j \in S} k(x_i, x_j)$$

Observation: F_1 is linear in $\mathbf{K} = [K_{ij}] = [k(x_i, x_j)] \in \mathbb{R}^{n \times n}$

Towards monotonicity and submodularity of F_1

$$F_1(S) = \frac{2}{n|S|} \sum_{i=1}^{n} \sum_{j \in S} k(x_i, x_j) - \frac{1}{|S|^2} \sum_{i,j \in S} k(x_i, x_j) =: (*).$$

Observation: F_1 is linear in $\mathbf{K} = [K_{ij}] = [k(x_i, x_j)] \in \mathbb{R}^{n \times n}$,

$$(*) = \frac{2}{n|S|} \sum_{i=1}^{n} \sum_{j=1}^{n} I_{j \in S} k(x_i, x_j) - \frac{1}{|S|^2} \sum_{i,j=1}^{n} I_{i \in S} I_{j \in S} k(x_i, x_j)$$

Towards monotonicity and submodularity of F_1

$$F_1(S) = \frac{2}{n|S|}\sum_{i=1}^{n}\sum_{j\in S} k(x_i,x_j) - \frac{1}{|S|^2}\sum_{i,j\in S} k(x_i,x_j) =: (*).$$

Observation: F_1 is linear in $\mathbf{K} = [K_{ij}] = [k(x_i,x_j)] \in \mathbb{R}^{n\times n}$,

$$(*) = \frac{2}{n|S|}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{j\in S}\, k(x_i,x_j) - \frac{1}{|S|^2}\sum_{i,j=1}^{n} I_{i\in S} I_{j\in S}\, k(x_i,x_j)$$

$$= \sum_{i,j=1}^{n} \underbrace{\left(\frac{2}{n|S|} I_{j\in S} - \frac{1}{|S|^2} I_{i\in S} I_{j\in S}\right)}_{a_{ij}(S)} \underbrace{k(x_i,x_j)}_{K_{ij}}$$

Towards monotonicity and submodularity of F_1

$$F_1(S) = \frac{2}{n|S|}\sum_{i=1}^{n}\sum_{j\in S} k(x_i,x_j) - \frac{1}{|S|^2}\sum_{i,j\in S} k(x_i,x_j) =: (*).$$

Observation: F_1 is linear in $\mathbf{K} = [K_{ij}] = [k(x_i,x_j)] \in \mathbb{R}^{n\times n}$,

$$(*) = \frac{2}{n|S|}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{j\in S} k(x_i,x_j) - \frac{1}{|S|^2}\sum_{i,j=1}^{n} I_{i\in S} I_{j\in S} k(x_i,x_j)$$

$$= \sum_{i,j=1}^{n} \underbrace{\left(\frac{2}{n|S|} I_{j\in S} - \frac{1}{|S|^2} I_{i\in S} I_{j\in S}\right)}_{a_{ij}(S)} \underbrace{k(x_i,x_j)}_{K_{ij}} = \langle \mathbf{A}(S), \mathbf{K}\rangle .$$

Monotonicity & submodularity of $F(S, \mathbf{K}) = \langle \mathbf{A}(S), \mathbf{K} \rangle$

Theorem

- $\mathbf{K} = [K_{ij}] \in \mathbb{R}^{n \times n}$, $0 \leq K_{ij} \leq K_* := \max_{ij} K_{ij}$, $0 < K_*$.
- $\mathbf{E} := I_{\mathbf{K}=K_*} \in \{0,1\}^{n \times n}$ *('diagonal')*, $\mathbf{E}' = 1 - \mathbf{E}$.
- *Assume*

$$\alpha(n, m) = \frac{F(S \cup \{u\}, \mathbf{E}) - F(S, \mathbf{E})}{F(S, \mathbf{E}') > 0}.$$

If for $\forall \underbrace{(i,j) \text{ such that } (\mathbf{E}')_{ij} = 1}_{\text{'off-diagonal'}}$, $0 \leq \forall m \leq m^*$

$$K_{ij} \leq K_* \alpha(n, m), \quad [\mathbf{K} : \text{'diagonally dominant enough'}]$$

then $S \mapsto F(S, \mathbf{K})$ *is monotone.*

Note: Similar condition $\Rightarrow$ $S \mapsto F(S, \mathbf{K})$: submodular.

Monotonicity & submodularity of F_1

Theorem

Let $\mathbf{K} = [K_{ij}]$. *If*

1. *Non-negative entries:* $0 \leq K_{ij}$ $(\forall i, j)$.

Monotonicity & submodularity of F_1

Theorem

Let $\mathbf{K} = [K_{ij}]$. *If*

1. *Non-negative entries:* $0 \leq K_{ij}$ $(\forall i, j)$.
2. *Equal, positive diagonal:* $0 < K_* := K_{ii}$ $(\forall i)$.

Monotonicity & submodularity of F_1

Theorem

Let $\mathbf{K} = [K_{ij}]$. *If*

1. *Non-negative entries:* $0 \leq K_{ij}$ $(\forall i, j)$.
2. *Equal, positive diagonal:* $0 < K_* := K_{ii}$ $(\forall i)$.
3. *Diagonally dominance:* $\sum_{j \neq i} K_{ij} < K_{ii}$ $(\forall i)$.

Monotonicity & submodularity of F_1

Theorem

Let $\mathbf{K} = [K_{ij}]$. *If*

1. *Non-negative entries:* $0 \leq K_{ij}$ $(\forall i, j)$.
2. *Equal, positive diagonal:* $0 < K_* := K_{ii}$ $(\forall i)$.
3. *Diagonally dominance:* $\sum_{j \neq i} K_{ij} < K_{ii}$ $(\forall i)$.
4. *Off-diagonal entries are small enough:*

$$K_{ij} \leq \frac{K_*}{n^3 + 2n^2 - 2n - 3} \quad (\forall i \neq j).$$

then $S \mapsto F_1(S, \mathbf{K})$ *is monotone and submodular.*

Example

- $X = \{\mathbf{x}_i\}_{i=1}^n$: disjunct points. $k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2}$, $\gamma > 0$.

Example

- $X = \{\mathbf{x}_i\}_{i=1}^n$: disjunct points. $k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2}$, $\gamma > 0$.
- In this case:
  - $k(\mathbf{x}_i, \mathbf{x}_j) > 0$: non-negative ✓

Example

- $X = \{\mathbf{x}_i\}_{i=1}^n$: disjunct points. $k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2}$, $\gamma > 0$.
- In this case:
  - $k(\mathbf{x}_i, \mathbf{x}_j) > 0$: non-negative ✓
  - Diagonal $\equiv 1$, > 0.

Example

- $X = \{\mathbf{x}_i\}_{i=1}^n$: disjunct points. $k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2}$, $\gamma > 0$.
- In this case:
  - $k(\mathbf{x}_i, \mathbf{x}_j) > 0$: non-negative ✓
  - Diagonal $\equiv 1$, > 0.
  - Diagonal dominance: disjunct $\mathbf{x}_i$-s.

Example

- $X = \{\mathbf{x}_i\}_{i=1}^n$: disjunct points. $k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2}$, $\gamma > 0$.
- In this case:
  - $k(\mathbf{x}_i, \mathbf{x}_j) > 0$: non-negative ✓
  - Diagonal $\equiv 1$, > 0.
  - Diagonal dominance: disjunct $\mathbf{x}_i$-s.
  - Small enough off-diagonal entries: disjunct $\mathbf{x}_i$-s $\Rightarrow$

$$K_{ij}(\gamma) \xrightarrow{\gamma \to \infty} 0, < \frac{1}{n^3 + 2n^2 - 2n - 3}.$$

Proof: general theorem – monotonicity

- Goal: $F(S \cup \{u\}) - F(S) \geq 0$, $\forall S, \forall u$.
- Idea:
  - Construct F bounds.
  - F is linear in $\mathbf{K} \Rightarrow \mathbf{K}$ bounds are enough.

Proof: monotonicity

- $\mathbf{K}$ bounds: using $0 \leq K_{ij}$

$$K_* \mathbf{E} \leq \mathbf{K} \leq \underbrace{K_* \mathbf{E}}_{\max_{\mathbf{E}} \mathbf{K}} + \underbrace{\nu \mathbf{E}'}_{\max_{\mathbf{E}'} \mathbf{K}}, \quad \nu := \max_{(i,j) \in E'} K_{ij}.$$

Proof: monotonicity

- $\mathbf{K}$ bounds: using $0 \leq K_{ij}$

$$K_* \mathbf{E} \leq \mathbf{K} \leq \underbrace{K_* \mathbf{E}}_{\max_{\mathbf{E}} \mathbf{K}} + \underbrace{\nu \mathbf{E}'}_{\max_{\mathbf{E}'} \mathbf{K}}, \quad \nu := \max_{(i,j) \in E'} K_{ij}.$$

- Propagation to F bounds: by linearity of $\mathbf{K} \mapsto F(S, \mathbf{K}) = \langle \mathbf{A}(S), \mathbf{K} \rangle$

$$\langle \mathbf{A}(S), K_* \mathbf{E} \rangle \leq \langle \mathbf{A}(S), \mathbf{K} \rangle \leq \left\langle \mathbf{A}(S), K_* \mathbf{E} + \nu \mathbf{E}' \right\rangle, ???$$

.

Proof: monotonicity

- $\mathbf{K}$ bounds: using $0 \leq K_{ij}$

$$K_* \mathbf{E} \leq \mathbf{K} \leq \underbrace{K_* \mathbf{E}}_{\max_{\mathbf{E}} \mathbf{K}} + \underbrace{\nu \mathbf{E}'}_{\max_{\mathbf{E}'} \mathbf{K}}, \quad \nu := \max_{(i,j) \in E'} K_{ij}.$$

- Propagation to F bounds: by linearity of $\mathbf{K} \mapsto F(S, \mathbf{K}) = \langle \mathbf{A}(S), \mathbf{K} \rangle$

$$\langle \mathbf{A}(S), K_* \mathbf{E} \rangle \leq \langle \mathbf{A}(S), \mathbf{K} \rangle \leq \left\langle \mathbf{A}(S), K_* \mathbf{E} + \nu \mathbf{E}' \right\rangle, \; ???$$

$$\underbrace{F(S, K_* \mathbf{E})}_{K_* F(S, \mathbf{E})} \leq F(S, \mathbf{K}) \leq \underbrace{F(S, K_* \mathbf{E} + \nu \mathbf{E}')}_{K_* F(S, \mathbf{E}) + \nu F(S, \mathbf{E}')}.$$

Proof: monotonicity

- $\mathbf{K}$ bounds: using $0 \leq K_{ij}$

$$K_* \mathbf{E} \leq \mathbf{K} \leq \underbrace{K_* \mathbf{E}}_{\max_{\mathbf{E}} \mathbf{K}} + \underbrace{\nu \mathbf{E}'}_{\max_{\mathbf{E}'} \mathbf{K}}, \quad \nu := \max_{(i,j) \in E'} K_{ij}.$$

- Propagation to F bounds: by linearity of $\mathbf{K} \mapsto F(S, \mathbf{K}) = \langle \mathbf{A}(S), \mathbf{K} \rangle$

$$\langle \mathbf{A}(S), K_* \mathbf{E} \rangle \leq \langle \mathbf{A}(S), \mathbf{K} \rangle \leq \left\langle \mathbf{A}(S), K_* \mathbf{E} + \nu \mathbf{E}' \right\rangle, \; ???$$

$$\underbrace{F(S, K_* \mathbf{E})}_{K_* F(S, \mathbf{E})} \leq F(S, \mathbf{K}) \leq \underbrace{F(S, K_* \mathbf{E} + \nu \mathbf{E}')}_{K_* F(S, \mathbf{E}) + \nu F(S, \mathbf{E}')}.$$

- Applying the F bounds, and using $F(S, \mathbf{E}') > 0$:

$$F(S \cup \{u\}, \mathbf{K}) - F(S, \mathbf{K}) \geq K_* F(S \cup \{u\}, \mathbf{E}) - K_* F(S, \mathbf{E}) - \nu F(S, \mathbf{E}')$$
$$\geq 0? \Leftrightarrow$$

Proof: monotonicity

- $\mathbf{K}$ bounds: using $0 \leq K_{ij}$

$$K_* \mathbf{E} \leq \mathbf{K} \leq \underbrace{K_* \mathbf{E}}_{\max_{\mathbf{E}} \mathbf{K}} + \underbrace{\nu \mathbf{E}'}_{\max_{\mathbf{E}'} \mathbf{K}}, \quad \nu := \max_{(i,j) \in E'} K_{ij}.$$

- Propagation to F bounds: by linearity of $\mathbf{K} \mapsto F(S, \mathbf{K}) = \langle \mathbf{A}(S), \mathbf{K} \rangle$

$$\langle \mathbf{A}(S), K_* \mathbf{E} \rangle \leq \langle \mathbf{A}(S), \mathbf{K} \rangle \leq \left\langle \mathbf{A}(S), K_* \mathbf{E} + \nu \mathbf{E}' \right\rangle, ???$$

$$\underbrace{F(S, K_* \mathbf{E})}_{K_* F(S, \mathbf{E})} \leq F(S, \mathbf{K}) \leq \underbrace{F(S, K_* \mathbf{E} + \nu \mathbf{E}')}_{K_* F(S, \mathbf{E}) + \nu F(S, \mathbf{E}')}.$$

- Applying the F bounds, and using $F(S, \mathbf{E}') > 0$:

$$F(S \cup \{u\}, \mathbf{K}) - F(S, \mathbf{K}) \geq K_* F(S \cup \{u\}, \mathbf{E}) - K_* F(S, \mathbf{E}) - \nu F(S, \mathbf{E}')$$
$$\geq 0? \Leftrightarrow$$
$$\nu \leq K_* \frac{F(S \cup \{u\}, \mathbf{E}) - F(S, \mathbf{E})}{F(S, \mathbf{E}')}.$$

Consequence: idea

- $\mathbf{K}$: diagonal dominant $\Rightarrow \mathbf{E} = \mathbf{I}$, $\mathbf{E}' = 1 - \mathbf{I}$.

Consequence: idea

- $\mathbf{K}$: diagonal dominant $\Rightarrow \mathbf{E} = \mathbf{I}$, $\mathbf{E}' = 1 - \mathbf{I}$.
- Recall: $\mathbf{A}(S) = [a_{ij}(S)]$, $a_{ij}(S) = \frac{2}{n|S|} I_{j \in S} - \frac{1}{|S|^2} I_{i \in S} I_{j \in S}$.

Consequence: idea

- $\mathbf{K}$: diagonal dominant $\Rightarrow \mathbf{E} = \mathbf{I}$, $\mathbf{E}' = 1 - \mathbf{I}$.
- Recall: $\mathbf{A}(S) = [a_{ij}(S)]$, $a_{ij}(S) = \frac{2}{n|S|} I_{j \in S} - \frac{1}{|S|^2} I_{i \in S} I_{j \in S}$.
- Relevant quantities:

$$\langle \mathbf{A}(S), \mathbf{E} \rangle = \frac{2}{n} - \frac{1}{m}, \quad \langle \mathbf{A}(S), \mathbf{E}' \rangle = \frac{2(n-1)}{n} - \frac{m-1}{m},$$

Consequence: idea

- $\mathbf{K}$: diagonal dominant $\Rightarrow \mathbf{E} = \mathbf{I}$, $\mathbf{E}' = 1 - \mathbf{I}$.
- Recall: $\mathbf{A}(S) = [a_{ij}(S)]$, $a_{ij}(S) = \frac{2}{n|S|} I_{j \in S} - \frac{1}{|S|^2} I_{i \in S} I_{j \in S}$.
- Relevant quantities:

$$\langle \mathbf{A}(S), \mathbf{E} \rangle = \frac{2}{n} - \frac{1}{m}, \quad \langle \mathbf{A}(S), \mathbf{E}' \rangle = \frac{2(n-1)}{n} - \frac{m-1}{m},$$

$$\alpha(n, m) = \frac{n}{(m+1)\,[m(n-2)+n]} \quad \text{(monotonicity)},$$

Consequence: idea

- $\mathbf{K}$: diagonal dominant $\Rightarrow \mathbf{E} = \mathbf{I}$, $\mathbf{E}' = 1 - \mathbf{I}$.
- Recall: $\mathbf{A}(S) = [a_{ij}(S)]$, $a_{ij}(S) = \frac{2}{n|S|} I_{j \in S} - \frac{1}{|S|^2} I_{i \in S} I_{j \in S}$.
- Relevant quantities:

$$\langle \mathbf{A}(S), \mathbf{E} \rangle = \frac{2}{n} - \frac{1}{m}, \quad \langle \mathbf{A}(S), \mathbf{E}' \rangle = \frac{2(n-1)}{n} - \frac{m-1}{m},$$

$$\alpha(n, m) = \frac{n}{(m+1)\left[m(n-2)+n\right]} \quad \text{(monotonicity)},$$

$$\beta(n, m) = \frac{n}{(m+1)\left[n\left(m^2+3m+1\right)-2\left(m^2+2m\right)\right]} \quad \text{(submodularity)}$$

Consequence: idea

- $\mathbf{K}$: diagonal dominant $\Rightarrow \mathbf{E} = \mathbf{I}$, $\mathbf{E}' = 1 - \mathbf{I}$.
- Recall: $\mathbf{A}(S) = [a_{ij}(S)]$, $a_{ij}(S) = \frac{2}{n|S|} I_{j \in S} - \frac{1}{|S|^2} I_{i \in S} I_{j \in S}$.
- Relevant quantities:

$$\langle \mathbf{A}(S), \mathbf{E} \rangle = \frac{2}{n} - \frac{1}{m}, \quad \langle \mathbf{A}(S), \mathbf{E}' \rangle = \frac{2(n-1)}{n} - \frac{m-1}{m},$$

$$\alpha(n, m) = \frac{n}{(m+1)\left[m(n-2)+n\right]} \quad \text{(monotonicity)},$$

$$\beta(n, m) = \frac{n}{(m+1)\left[n\left(m^2+3m+1\right)-2\left(m^2+2m\right)\right]} \quad \text{(submodularity)}$$

- $m \mapsto \alpha(n, m)$ decreasing, $\alpha(n, n) = \frac{1}{n^2-1}$.

Consequence: idea

- $\mathbf{K}$: diagonal dominant $\Rightarrow \mathbf{E} = \mathbf{I}$, $\mathbf{E}' = 1 - \mathbf{I}$.
- Recall: $\mathbf{A}(S) = [a_{ij}(S)]$, $a_{ij}(S) = \frac{2}{n|S|} I_{j \in S} - \frac{1}{|S|^2} I_{i \in S} I_{j \in S}$.
- Relevant quantities:

$$\langle \mathbf{A}(S), \mathbf{E} \rangle = \frac{2}{n} - \frac{1}{m}, \quad \langle \mathbf{A}(S), \mathbf{E}' \rangle = \frac{2(n-1)}{n} - \frac{m-1}{m},$$

$$\alpha(n, m) = \frac{n}{(m+1)\left[m(n-2)+n\right]} \quad \text{(monotonicity)},$$

$$\beta(n, m) = \frac{n}{(m+1)\left[n\left(m^2+3m+1\right)-2\left(m^2+2m\right)\right]} \quad \text{(submodularity)}$$

- $m \mapsto \alpha(n, m)$ decreasing, $\alpha(n, n) = \frac{1}{n^2-1}$.
- $m \mapsto \beta(n, m)$ decreasing, $\beta(n, n) = \frac{1}{n^3+2n^2-2n-3}$.

Consequence: idea

- $\mathbf{K}$: diagonal dominant $\Rightarrow \mathbf{E} = \mathbf{I}$, $\mathbf{E}' = 1 - \mathbf{I}$.
- Recall: $\mathbf{A}(S) = [a_{ij}(S)]$, $a_{ij}(S) = \frac{2}{n|S|} I_{j \in S} - \frac{1}{|S|^2} I_{i \in S} I_{j \in S}$.
- Relevant quantities:

$$\langle \mathbf{A}(S), \mathbf{E} \rangle = \frac{2}{n} - \frac{1}{m}, \quad \langle \mathbf{A}(S), \mathbf{E}' \rangle = \frac{2(n-1)}{n} - \frac{m-1}{m},$$

$$\alpha(n,m) = \frac{n}{(m+1)\left[m(n-2)+n\right]} \quad \text{(monotonicity)},$$

$$\beta(n,m) = \frac{n}{(m+1)\left[n\left(m^2+3m+1\right)-2\left(m^2+2m\right)\right]} \quad \text{(submodularity)}$$

- $m \mapsto \alpha(n,m)$ decreasing, $\alpha(n,n) = \frac{1}{n^2-1}$.
- $m \mapsto \beta(n,m)$ decreasing, $\beta(n,n) = \frac{1}{n^3+2n^2-2n-3}$.
- $\beta(n,n) \leq \alpha(n,n)$ so $K_{ij} \leq K_* \beta(n,n)$ is sufficient.

F_2 quickly: criticism samples

$$\max_{C\subset\{1,\dots,n\}\setminus S,|C|\le c_*} F_2(C) := \underbrace{\sum_{\ell\in C}\left|f^*_{X,X_S}(x_\ell)\right|}_{\text{additive}} + \underbrace{\log\det \mathbf{K}_{C,C}}_{\text{submodular}}.$$

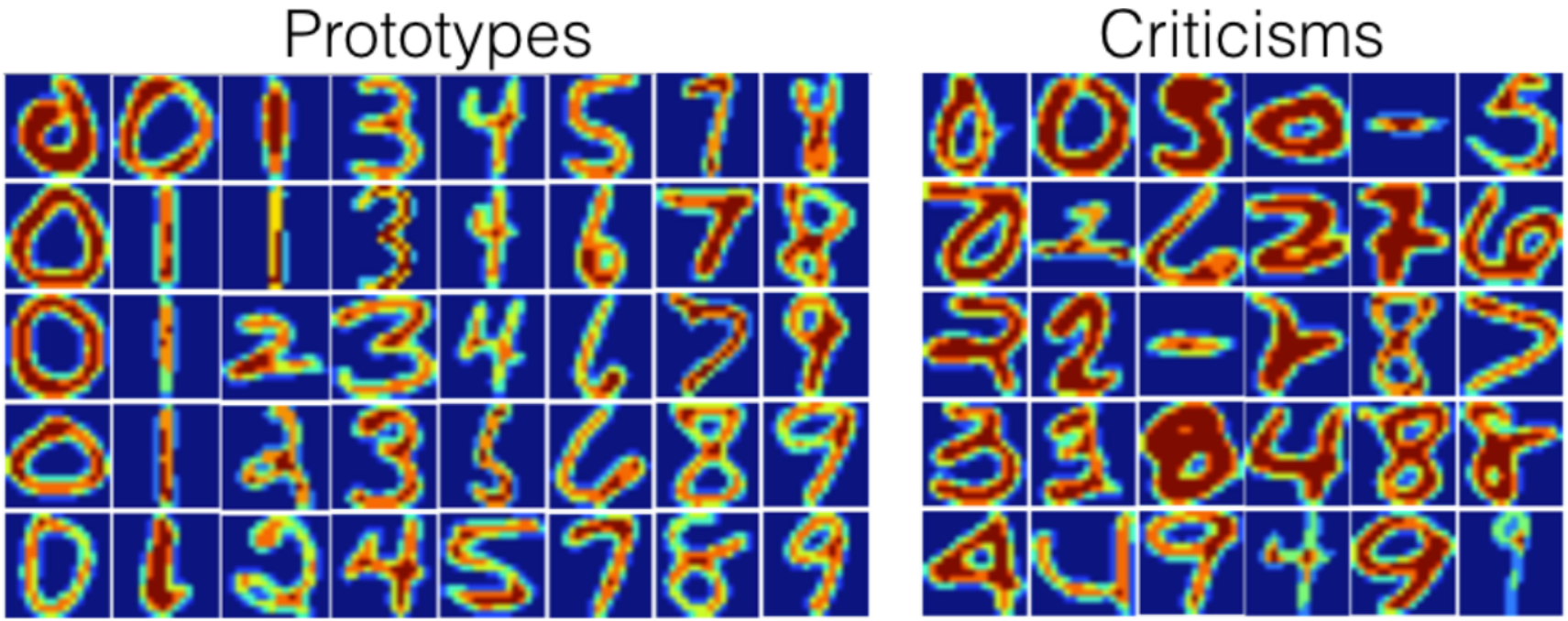
Notes: F_2 is submodular

- additive $\Rightarrow$ submodular,
- submodular + submodular = submodular.

Numerical illustrations-1

Goodness of S (prototypes): nearest prototype classifier.

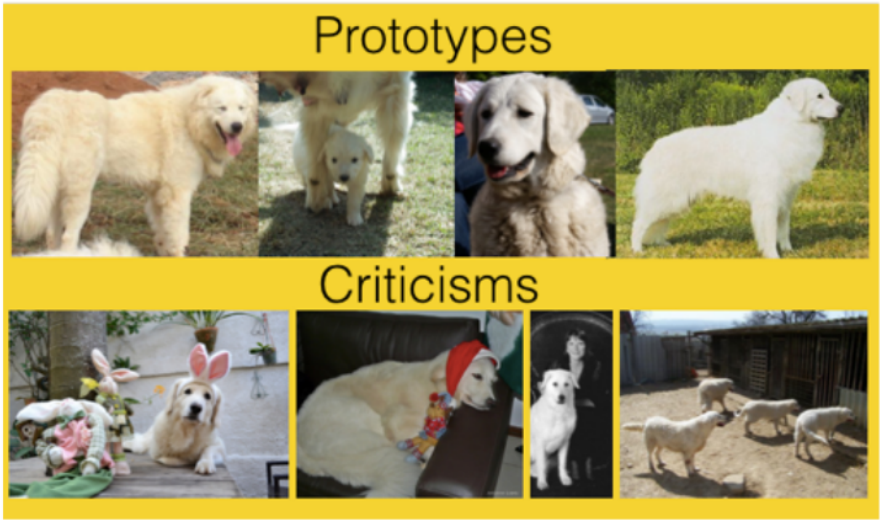
- Data: USPS handwritten digits, raw pixels.
- It compares nicely to the state-of-the-art.
- Especially good when m^* is small (≤ 1000).



Numerical illustrations-2

Goodness of S and C. ImageNet dataset:

- Features: 2048-dimensional vector embedding [He et al. 2015].
- Demo:



- Human pilot study: best performance with S & C.

Summary

- Goal: find prototypes and criticisms.
- Objective:
  - relied on mean embedding.
  - submodularity $\rightarrow$ greedy optimization.

Summary

- Goal: find prototypes and criticisms.
- Objective:
  - relied on mean embedding.
  - submodularity $\rightarrow$ greedy optimization.
- Practice:
  1. good compression: 1NN prototype classifier.
  2. useful for humans.

Summary

- Goal: find prototypes and criticisms.
- Objective:
  - relied on mean embedding.
  - submodularity $\rightarrow$ greedy optimization.
- Practice:
  1. good compression: 1NN prototype classifier.
  2. useful for humans.