Theory of feature selection

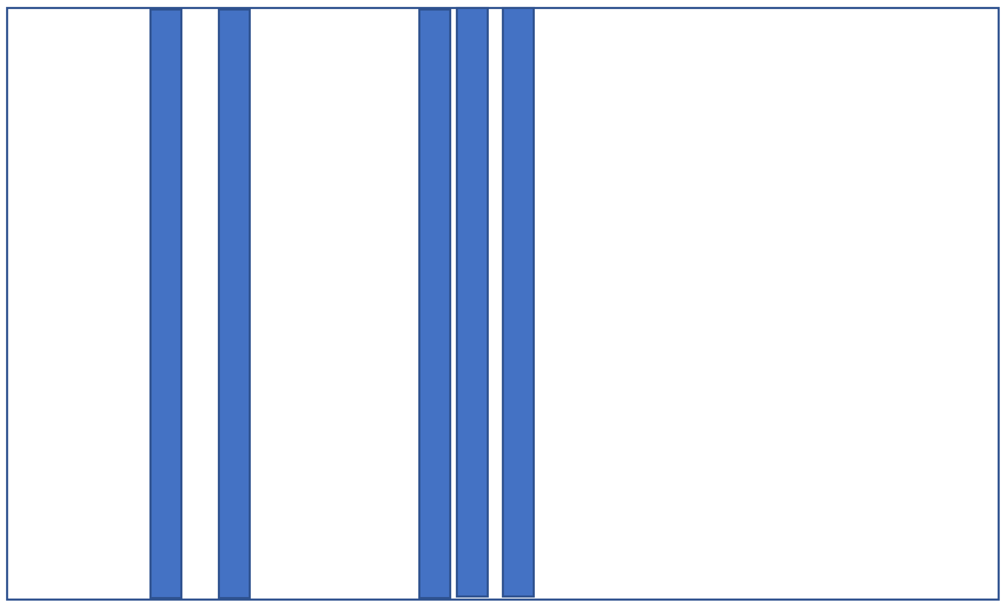
Rajiv Khanna
Purdue University

In collaboration with

Michael Mahoney (UC Berkeley), Alex Dimakis(UT Austin), Joydeep Ghosh (UT Austin), Oluwasanmi Koyejo (UIUC), Sahand Negaban (Yale), Michal Derezinski (UMich), Been Kim (Google Brain), Russell Poldrack (Stanford)

Feature Selection

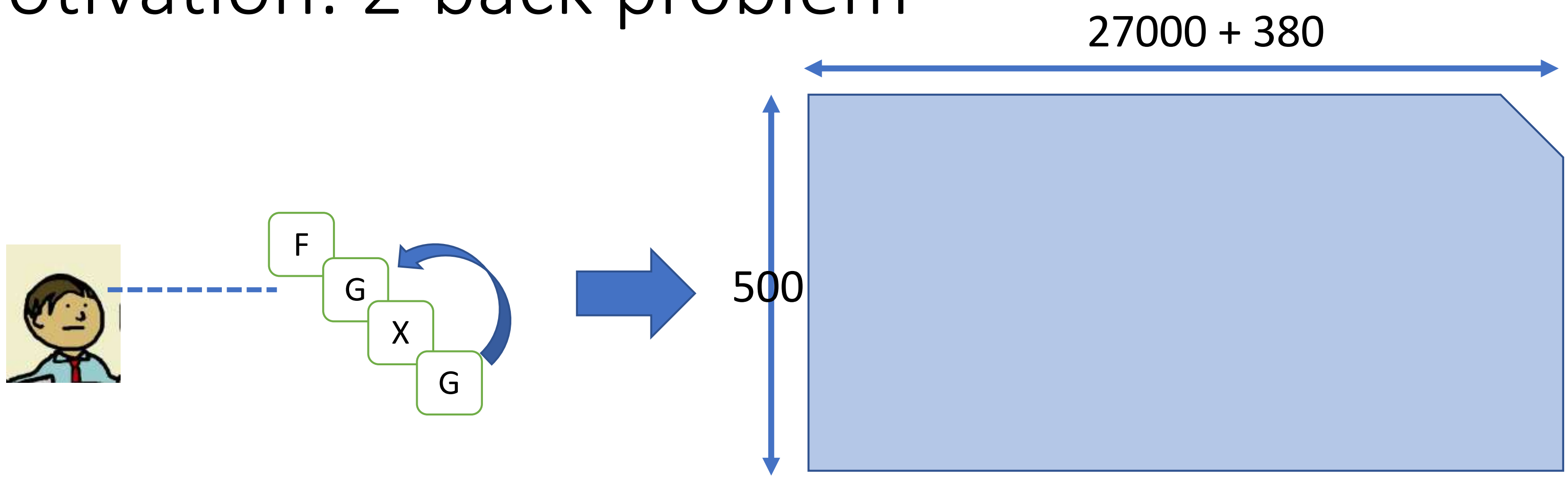
- Interpretability by modeling choice: Given a matrix, choose a few columns



Feature selection

- Interpretability by modeling choice – Feature Selection for sparsity
  - Provable guarantees for various greedy-variants
  - Cost of interpretability: Feature Selection vs SVD

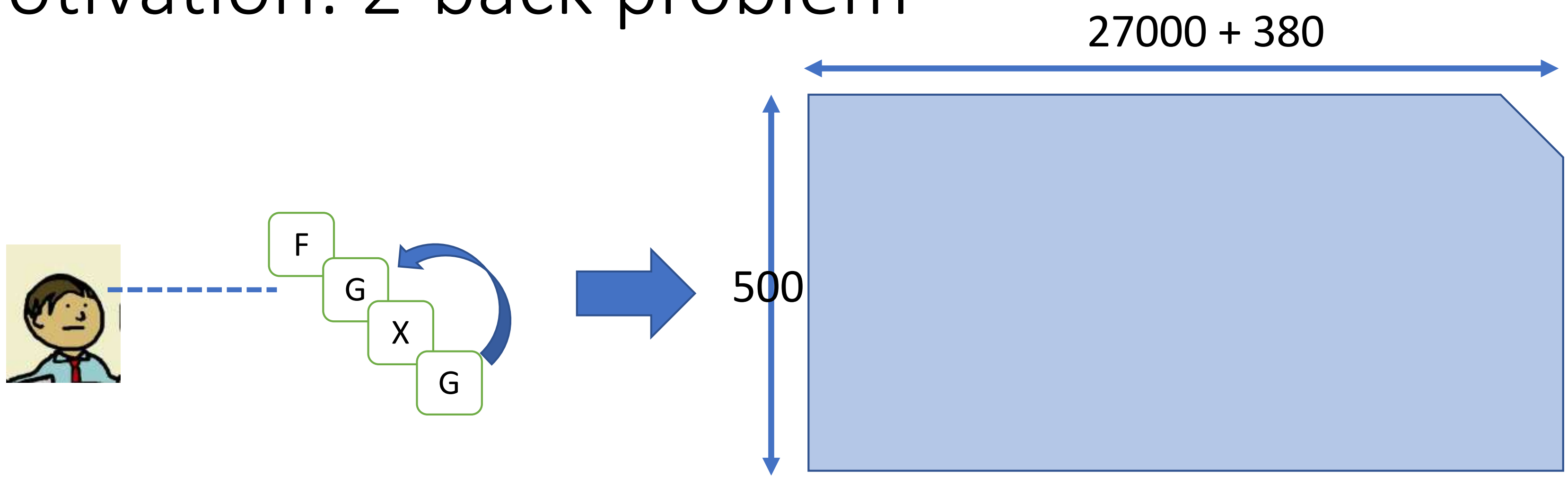
Motivation: 2-back problem



- Subject sees a stream of letters – “Does the current letter match the one 2 letters ago?”

- Contrast response times (wrt 0-back) and brain map summary recorded

Motivation: 2-back problem



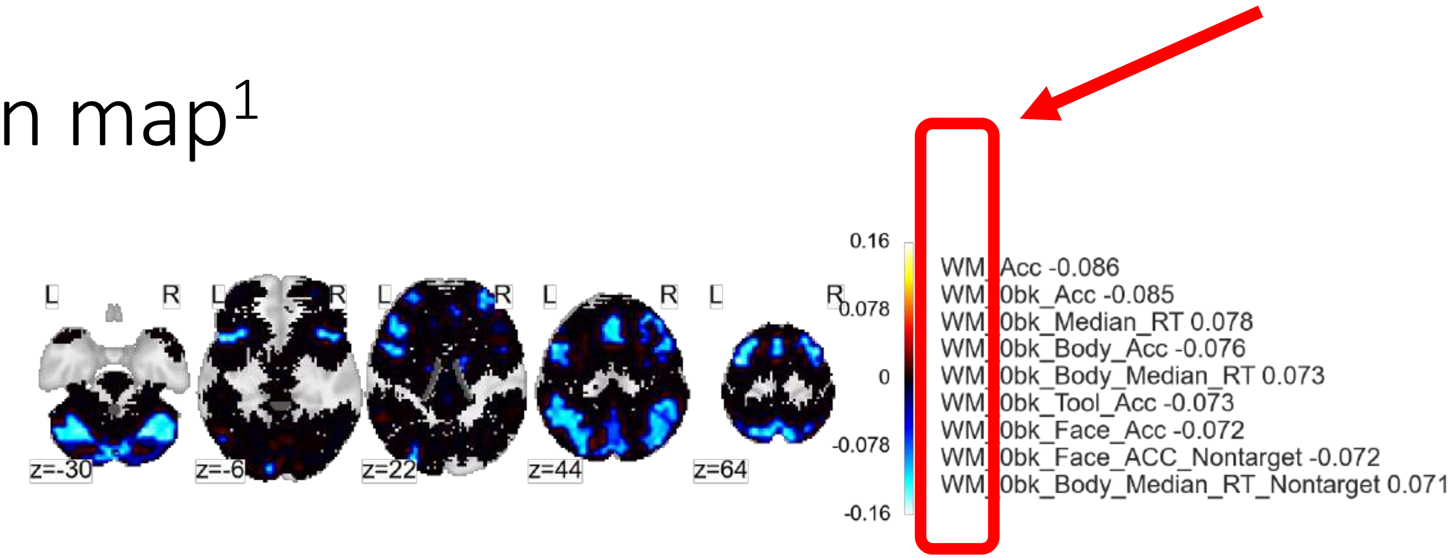
- Rows: 500 subjects
- Columns: 27000 brain voxels + 380 behavioural features
- Goal: Predict the response time using a <u>sparse</u> embedding of features

Modeling choices

- Sparse Canonical Correlation Analysis (Witten et. al. 2009) [SOTA]
  - Find the “best” Subspace -- Strong empirical performance
  - <u>Inconsistent</u> maps

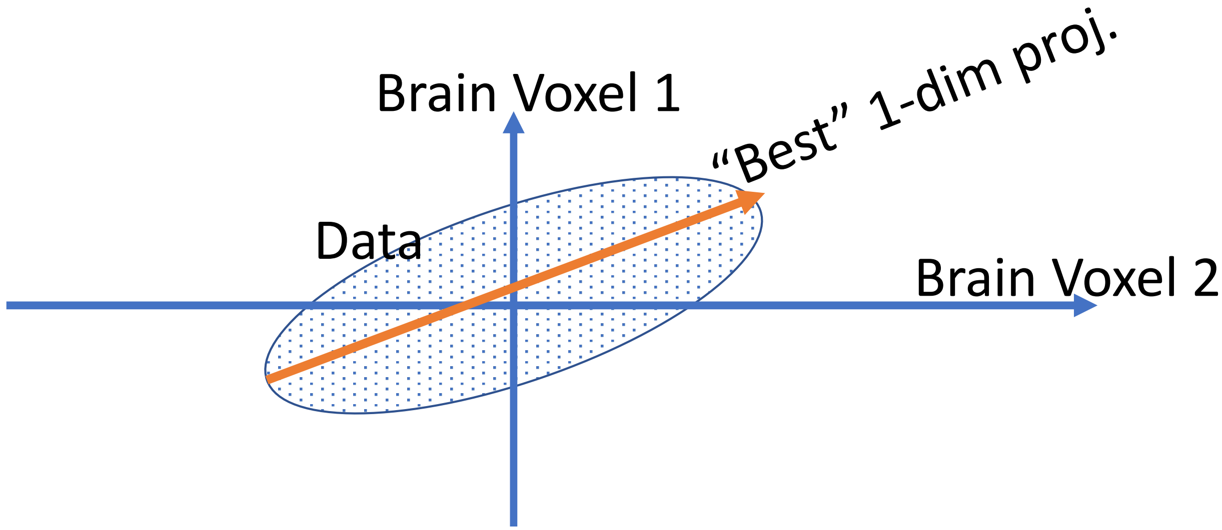
- Greedy Feature Selection with a Bayesian prior
  - Find the “best” subspace spanned by k columns - Strong empirical performance
  - <u>Interpretable</u> maps

Brain map[1]



- Neural support consistent with cognitive control systems akin to the task

- Selected behavioral features known to correlate with reaction time and accuracy

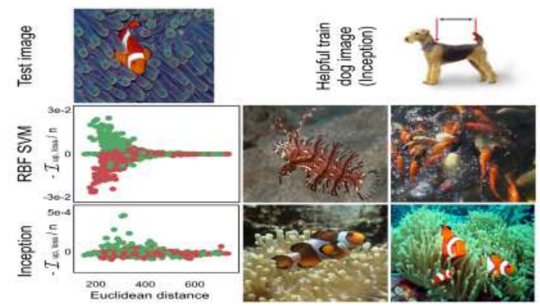


- Principal components give best quantitative performance but loses interpretability.
- Feature selection retain interpretability, but
  - Quantitatively worse, and
  - Harder – replace a poly-time solvable with a combinatorial one.

Greedy selection in practice

Document summarization.
[Vanderwende et. al. 2007]

Gene Analysis
[Paschou et. al. 2007]



Interpretability
[Koh et. al. 2017]

- Why does greedy work so well ?
- Why does feature selection perform well vs best rank-k approximation?

Submodular functions

- A set variate function *f* is submodular iff for any $S \subset T$:

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

- Diminishing returns property
- Examples: Entropy (e.g. log det), budget additive functions (e.g. facility location), rank function of matroids

Submodular functions

- Greedy algorithm is <u>provably</u> good.
- G_k is the set returned by the greedy algorithm
- S_k^* is the optimum solution to

$$\max_{\{|S| \leq k\}} f(S)$$

- If f is submodular:

$$f(G_k) \geq \left(1 - \frac{1}{e}\right) f(S_k^*)$$

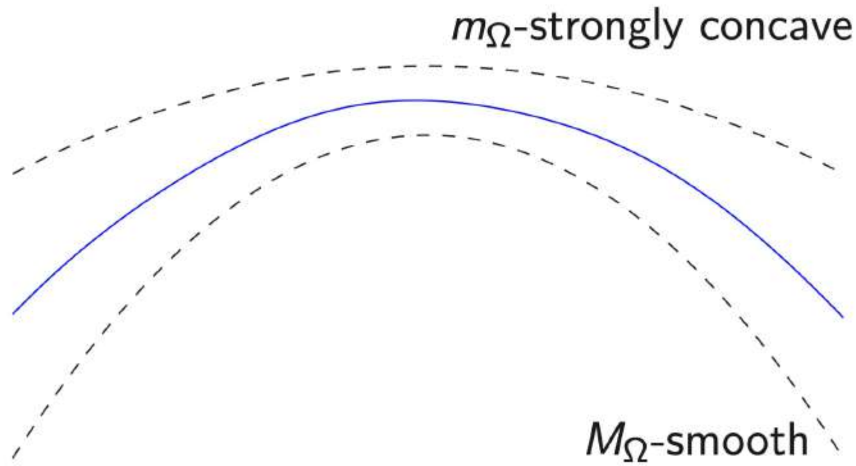
Weakly submodular functions[1]

- Submodularity sufficient, not necessary for greedy guarantees
- Weaker condition: based on submodularity ratio γ
- $\gamma_{\{L,S\}} = \frac{\sum_{\{j\ \in S\}}[f(L \cup \{\ i\ \}) - f(L)\ ]}{f(L \cup S) - f(L)}$
- $\gamma \geq 1\ \Rightarrow$ submodularity
- $\gamma > 0$ suffices for $(1\ -\ \frac{1}{e^{\gamma}})$ guarantees

[1] Elenberg, **Khanna**, Dimakis, Negahban. Annals of Stats 2018

Restricted strong concavity/smoothness (RSC/RSM)

- To bound the submodularity ratio for general functions, I will make use of RSC/RSM



$$-\frac{m_\Omega}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \geq l(\mathbf{y}) - l(\mathbf{x}) - \langle \nabla l(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \geq -\frac{M_\Omega}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$$

First result: RSC implies weak submodularity

- Goal:

$$\max_{\mathsf{S}:|\mathsf{S}|\leq k} f(\mathsf{S}) \Leftrightarrow \max_{\substack{\boldsymbol{\beta}:\boldsymbol{\beta}_{\mathsf{S}^c}=0 \\ |\mathsf{S}|\leq k}} l(\boldsymbol{\beta}) - l(\mathbf{0})$$

- If $l()$ is:
  - m-Restricted strong concave on a certain subdomain, and
  - M-smooth on another subdomain,

  then,

$$\boxed{\gamma \geq \frac{m}{M} \Longrightarrow f(G_k) \geq \left(1 - \frac{1}{e\text{^}\{m/M\}}\right) f(S_k^*)}$$

Recovery based bounds

- For any β^s of size k, greedy/MP based algorithm recovers β^r

$$\|\widehat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^s\|_2^2 \leq \frac{4}{m_{s+r}^2}\|\nabla l(\boldsymbol{\beta}^s)\|_{2,(s+r)}^2 + \frac{4}{m_{s+r}}(1 - C_{s,r})\big[l(\boldsymbol{\beta}^s) - l(\mathbf{0})\big]$$

Matching pursuit based selection

- Greedy choice

$$s \leftarrow \arg\max_{j \in [p] \setminus \mathsf{S}_{i-1}} f(\mathsf{S}_{i-1}^G \cup \{j\}) - f(\mathsf{S}_{i-1}^G)$$

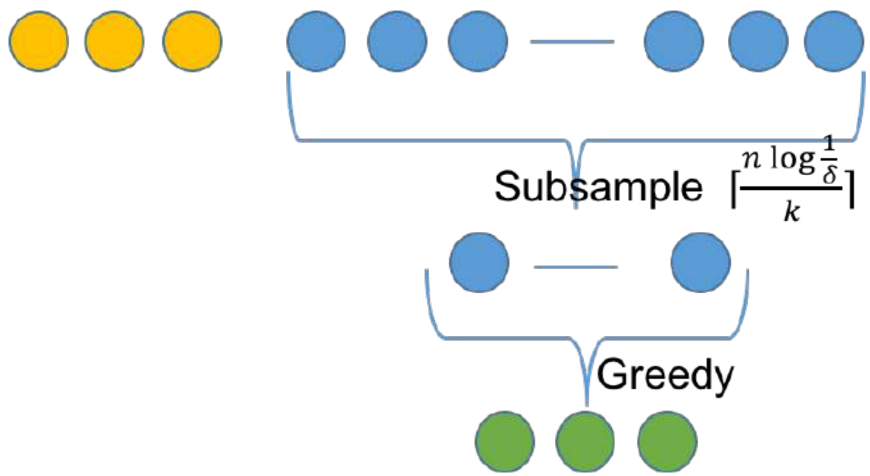
- Matching pursuit choice

$$s \leftarrow argmax_j \; \left| \langle e_j, \nabla l(\beta^S) \rangle \right|$$

- Same bounds hold!

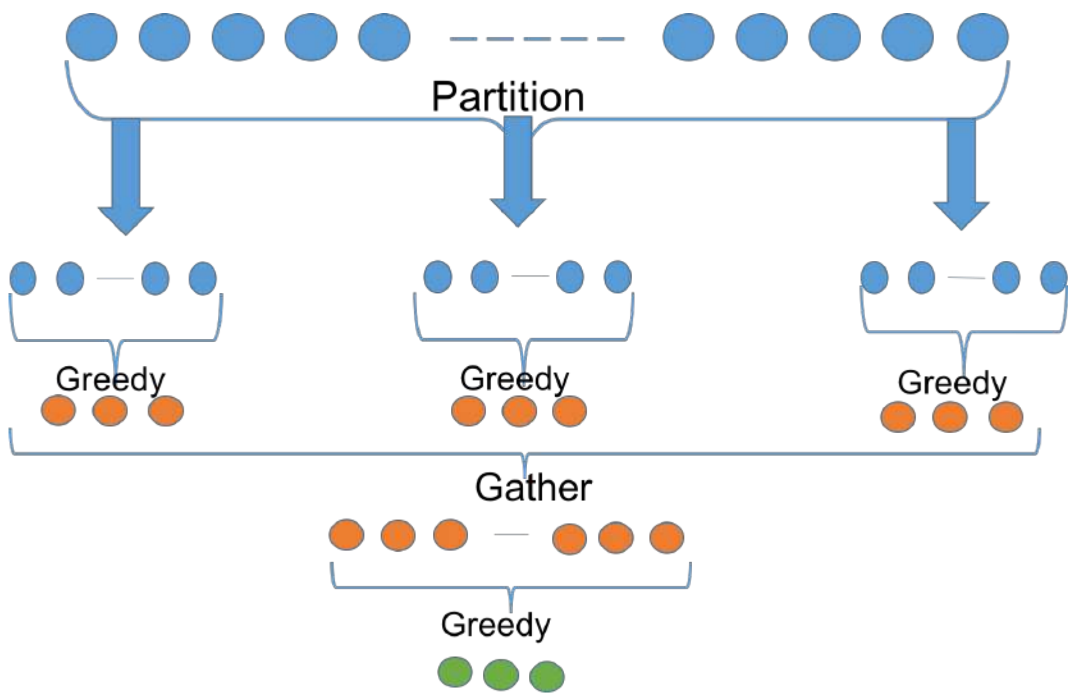
Stochastic selection

- Subsample before every greedy step



$$\mathbb{E}[f(\mathsf{G}_{sg})] \geq (1 - \frac{1}{e^{\gamma}} - \delta) f(\mathsf{A}^{\star})$$

Distributed selection



Distributed selection results

- With ν_k as subadditivity constant

$$\nu_{\mathsf{S}} := \min_{\substack{\mathsf{A} \cup \mathsf{B} = \mathsf{S} \\ \mathsf{A} \cap \mathsf{B} = \phi}} \frac{f(\mathsf{A}) + f(\mathsf{B})}{f(\mathsf{S})}. \quad \nu_k := \min_{\mathsf{S}:|\mathsf{S}|=k} \nu_{\mathsf{S}}$$

- Can show: RSC implies weak subadditivity

$$\mathbb{E}\left[f(\mathsf{G}_{dg})\right] \geq \frac{\nu_k}{2}(1 - \frac{1}{e^{\gamma}})f(\mathsf{A}^\star)$$

Extension to adaptive sequencing

- Goal: reduce the number of oracle calls (bottleneck in greedy algorithms)
- Also, extension to beyond k-sparsity constraint through p-systems
- Intuition: Generate a ``good enough" random subset of features, choose several of them in every oracle call.
- Oracle calls: $O(\epsilon^{-2} r \log n)$

$$\frac{\mathbb{E}\left[\, f(\mathsf{S}^*)\,\right]}{\mathtt{OPT}} \geq \frac{1}{1+p}\left(1 - \exp\left\{-(1-\varepsilon)^2 \frac{m^3}{M^3}\right\}\right)$$

Other extensions

- Greedy low rank approximation
- Kernel herding

Cost of interpretability

- Why does feature selection retain quantitative performance vs best rank-k approximation?[1]

- Goal: Interpretable dimensionality reduction for an arbitrary matrix A.

[1] Derezinski, **Khanna**, Mahoney. NeurIPS 2020 (best paper award)

Cost of Interpretability

- Optimal rank-k approximation

$$\mathrm{OPT}_k := \min_{\mathbf{B}:\, \mathrm{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_F^2$$

- Choose a set *S* of *k* columns

$$\mathrm{Er}_{\mathbf{A}}(S) := \|\mathbf{A} - \mathbf{P}_S \mathbf{A}\|_F^2$$

- What can we say about the cost of interpretability

$$\frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_k} = ?$$

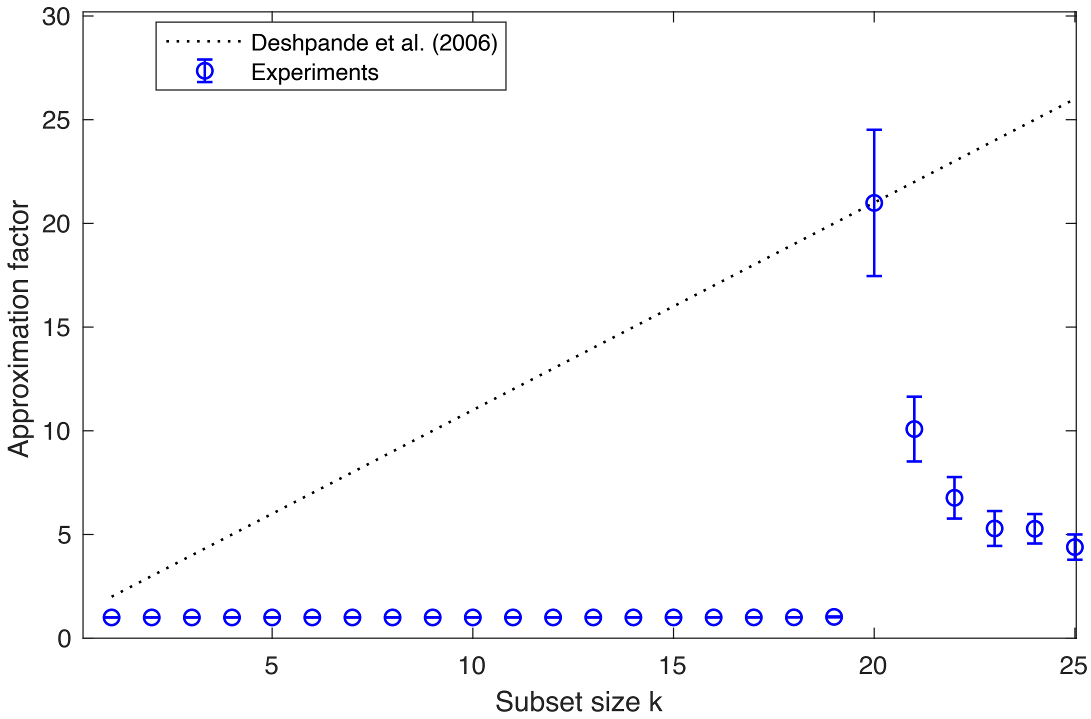
Prior Art – worst case analysis

- Deshpande et. al. 2006
- There exists an algorithm with O(k) guarantees for $|S| \leq k$

“Tight” bounds

- There exist matrices for which better than O(k) not possible

(Worst-case) Theory vs Practice



The sampling algorithm used is k-DPP (Determinantal Point Processes)

Image courtesy Michal Derezinski

Main results[1]

- Beyond worst-case analysis based on spectrum of the matrix
  - Previous bound:

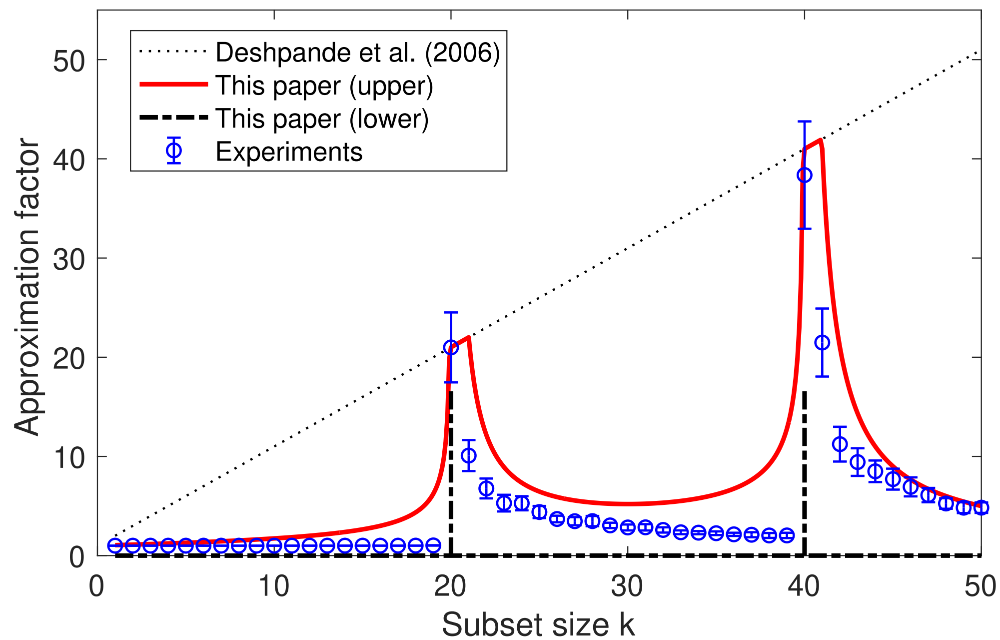
$$\frac{\mathrm{E}[\mathrm{Er}_{\mathbf{A}}(S)]}{\mathrm{OPT}_k} \leq (k+1)$$

  - New bound: For s<rank(A) and a favorable t_s:

$$\frac{\mathrm{E}[\mathrm{Er}_{\mathbf{A}}(S)]}{\mathrm{OPT}_k} \leq \left(1+\frac{s}{k-s}\right)\sqrt{1+\frac{2(k-s)}{t_s-k}}$$

[1] Derezinski, **Khanna**, Mahoney. NeurIPS 2020 (best paper award)

Summary of results



“Bad” cases only when there is a sudden drop in spectrum of the matrix.

Example of a sharp transition

- Let sr(A) be stable rank of A. There exists a matrix A such that if

$sr(A) - 1 < k < sr(A)$, then

$$\frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_k} \geq 0.9k$$

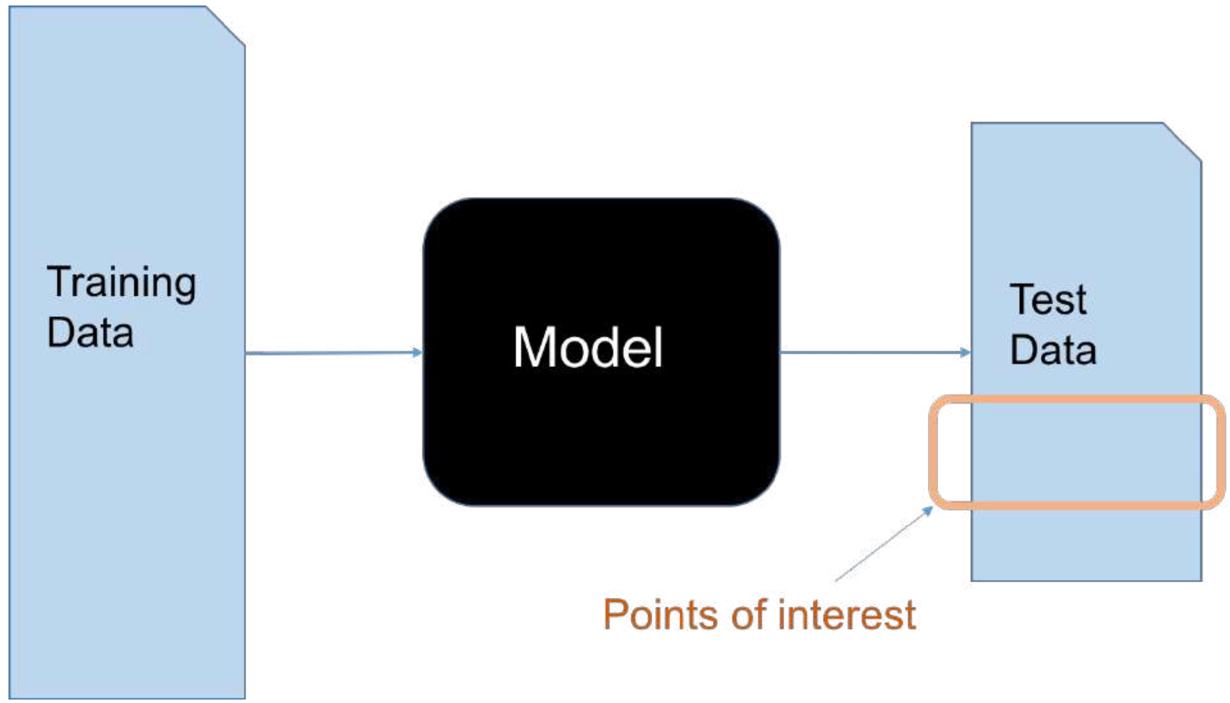
Example: smooth decay (polynomial)

- No sharp drop in spectrum ⇒ even better guarantees!
- If $c_1 i^{-p} \leq \lambda_i \leq c_2 i^{-p}$, then the upper bound:

$$\frac{\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)]}{\mathrm{OPT}_k} \leq \frac{c_2}{c_1} cp$$

- Example: Matern kernel

Black-box interpretability



Which training data points are "most" responsible for predicting on points of interest ?

Working idea

- Approximate the test data distribution using samples from the training data[1]

- Challenges:
    - Not the same support
    - How to incorporate the model ?
    - How to reliably choose training data samples ?
    - Scale up on both time and space?

Working idea

- Approximate the test data distribution using samples from the training data[1]

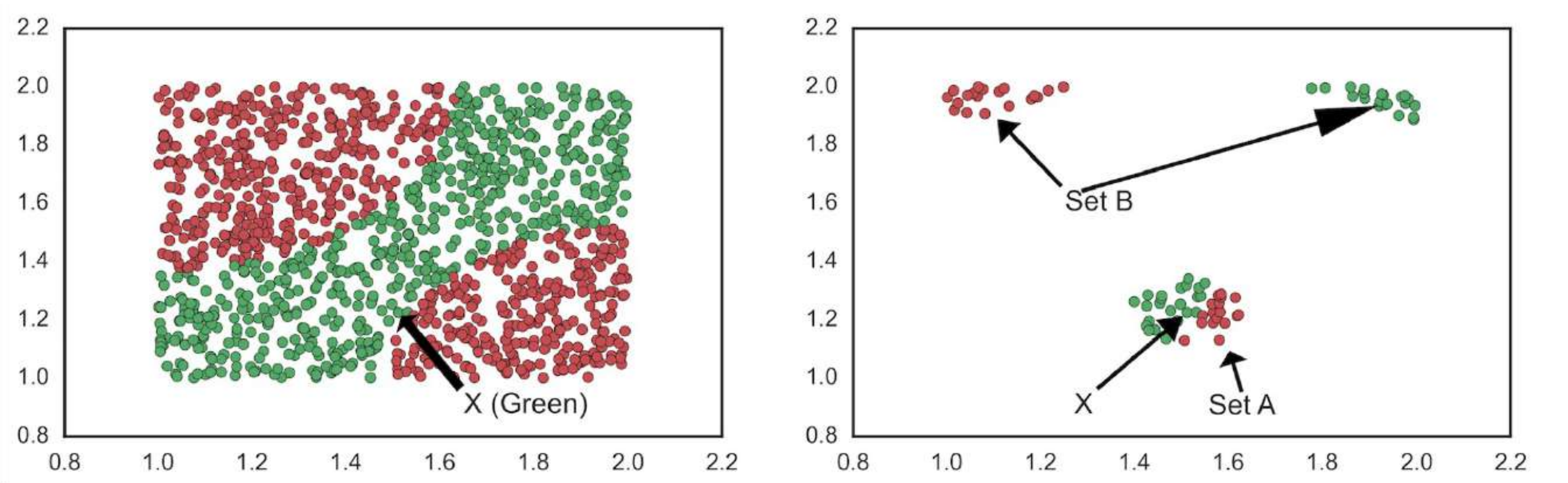
- Challenges:
  - Not the same support -- Use a smoothing prior GP(0,k)
  - How to incorporate the model ? – Fisher Kernels
  - How to reliably choose training data samples ? – Weak submodularity/DPP
  - Scale up on both time and space? – Weak submodularity/DPP

Fisher Kernels

- Slight perturbations in the neighborhood of fitted parameters would impact the fit of two similar objects similarly
- Two similar objects will have similar gradients in the parameter of the model

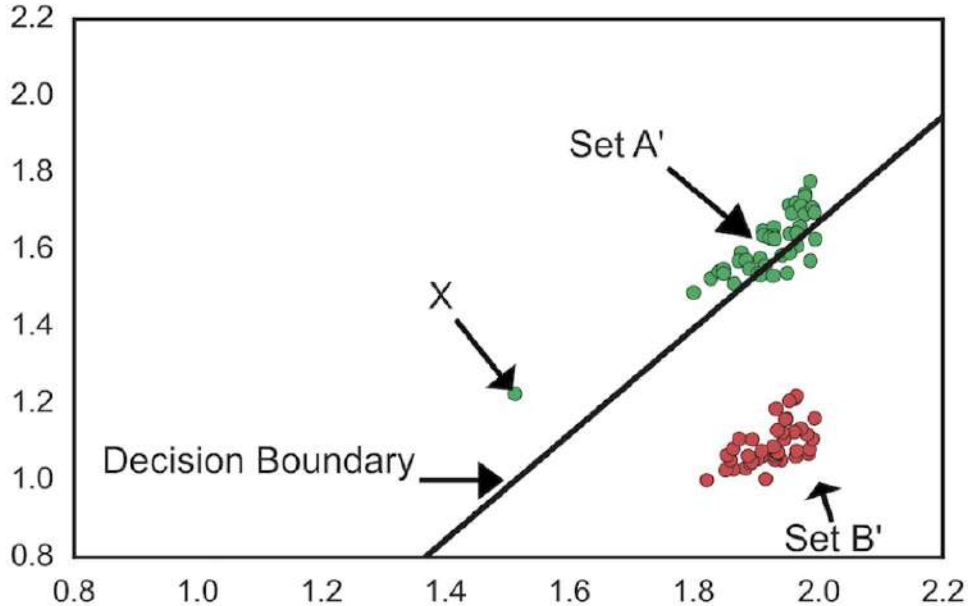
- $K(x, y) = \langle \nabla_\theta l(x), I^{-1} \nabla_\theta l(y) \rangle$. I is the Fisher information matrix.

Fisher kernel: Intuition



- Set A: Fifty closest points in RBF similarity
- Set B: Fifty farthest points

Fisher kernel: intuition



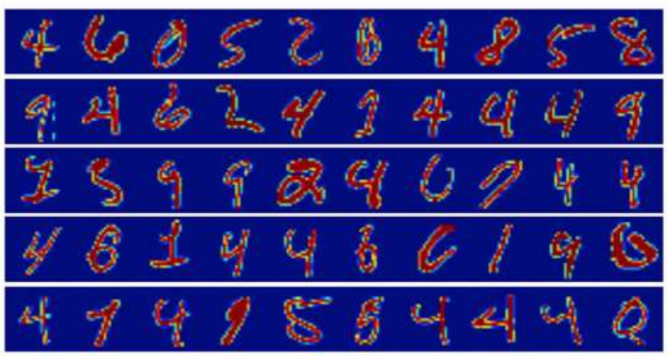
- Set A’: Fifty closest points in Fisher similarity
- Set B’: Fifty farthest points

Implications

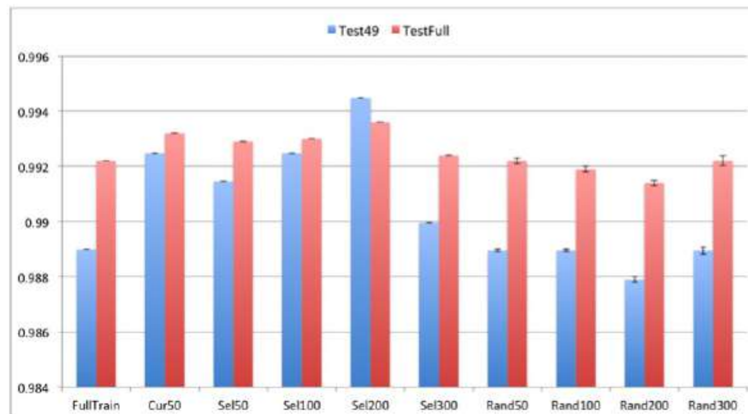
- Influence functions: Up-weighing which training point impacts the prediction of a given test point the most ? [Koh/Liang ICML 2017]
- Strict generalization of influence functions, provide a probabilistic foundation
  - Lasso ⟺ MAP solution of Bayesian Linear regression with Laplace priors
  - K-Means ⟺ EM algorithm to optimize for parameters of mixture of gaussians
- Greedily selecting influential training data points is weakly submodular
- Empirical evidence for data summarization and data cleaning

Experiments

Removing “bad” points from MNIST. Identify which data points are responsible for misclassification on 4s vs 9s



(a) A subset of selected prototypes responsible for misclassifying 4s and 9s in the test set



(b) Accuracy fractions on test data 4s and 9s (Test49), and the full test set after removing random (Rand), algorithm selected (Sel), or Curated (Cur) prototypes.

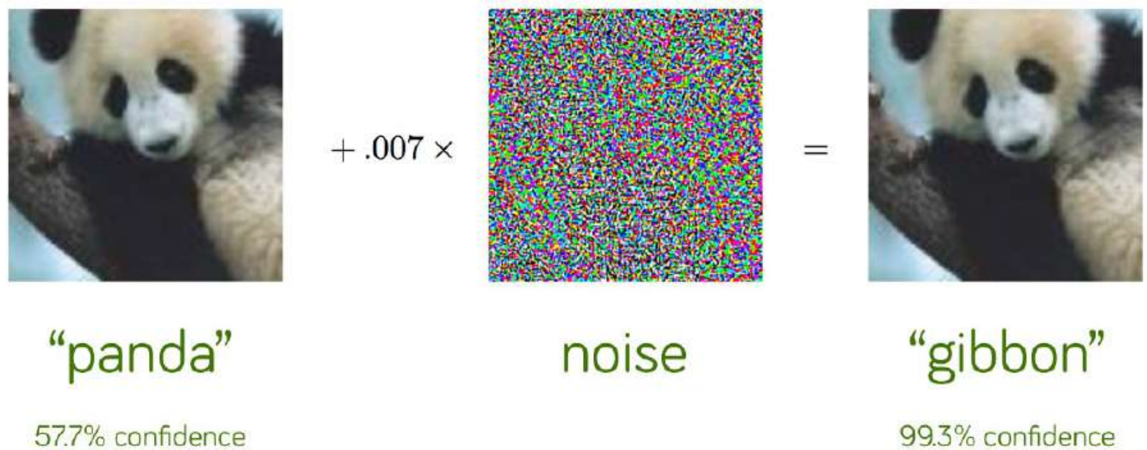
Thank you!

Take-aways:

- Greed is ***<u>provably</u>*** good.

- For feature selection, cost of interpretability is not high.

Neural networks are brittle



Can "fool" a NN by humanly-imperceptible injecting adversarial noise
(Image from Goodfellow et. al.2015)

Adversarial Training

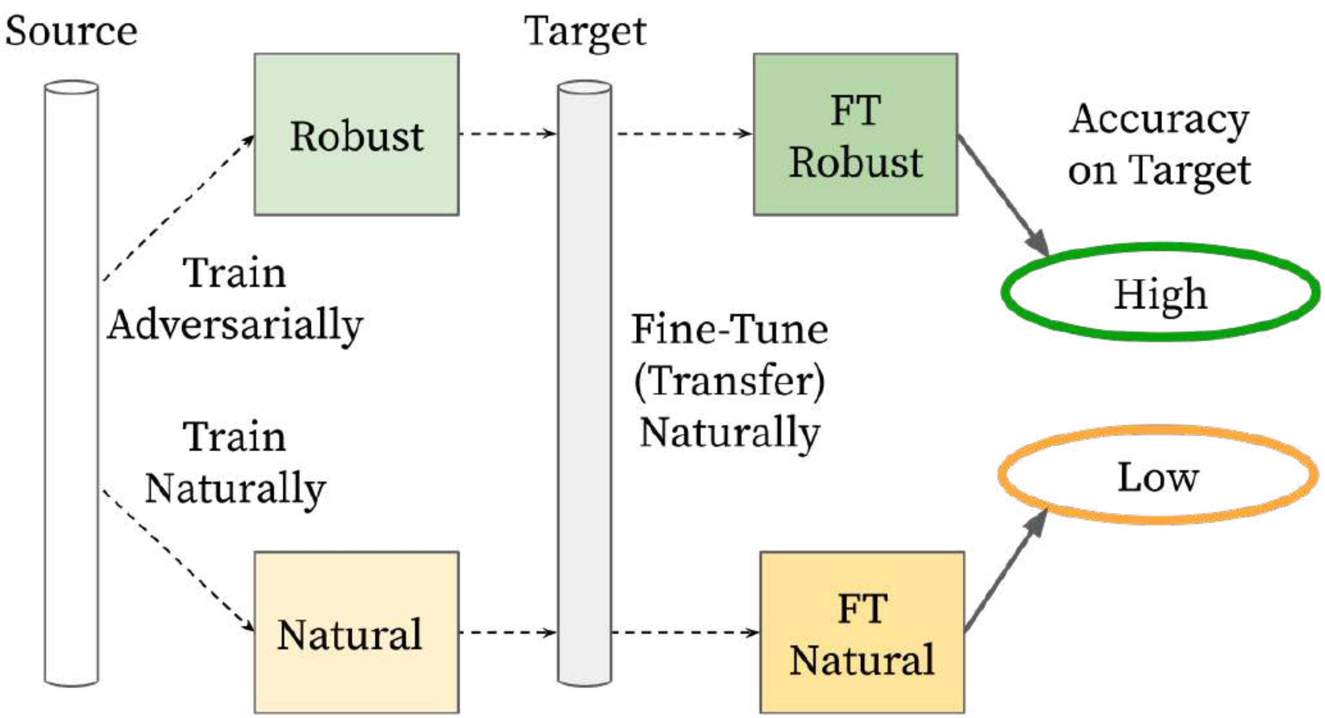
- Natural Training: $\min_{\theta} \sum_i L_\theta(x_i, y_i)$

- Adversarial Training $\min_{\theta} \sum_i \max_{\|\delta_i\| \le \epsilon} L_\theta(x_i + \delta_i, y_i)$

- There is an unintended side effect[1]

1 Urtera, Kravitz, Erichson, **Khanna**, Mahoney. ICLR 2021

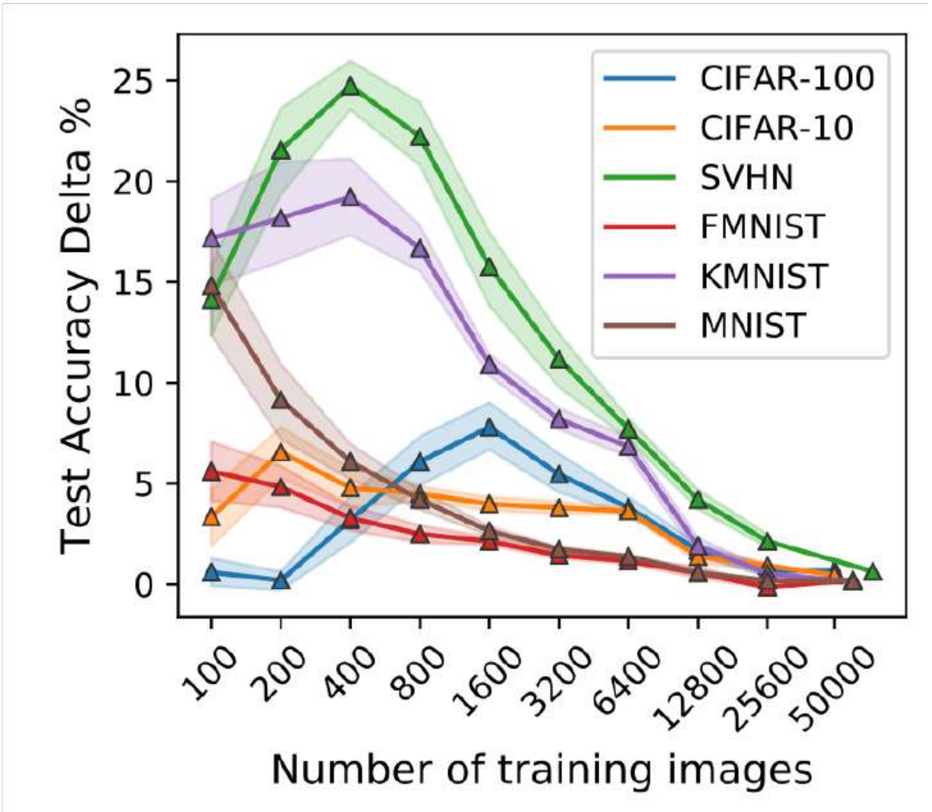
Setup



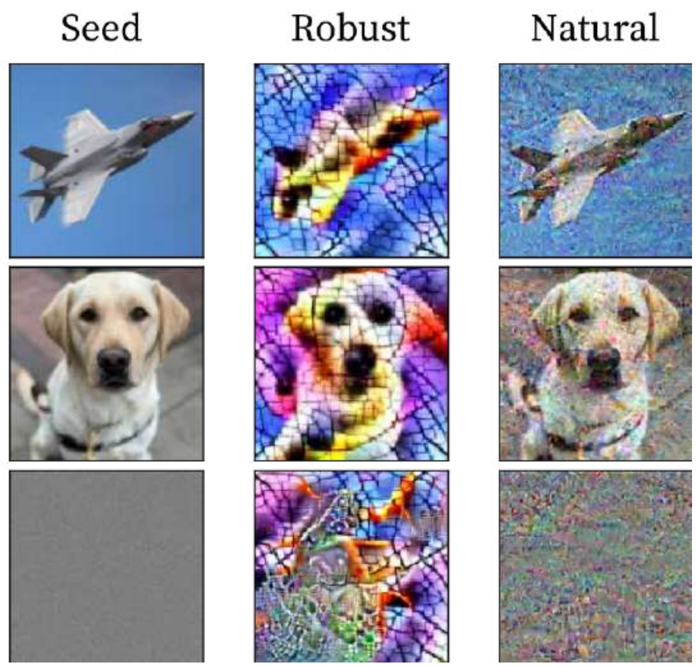
Source: ImageNet

Target: CIFAR-10, CIFAR-100, SVHN, FMNIST, KMNIST, MNIST

Adversarial Training transfers better!



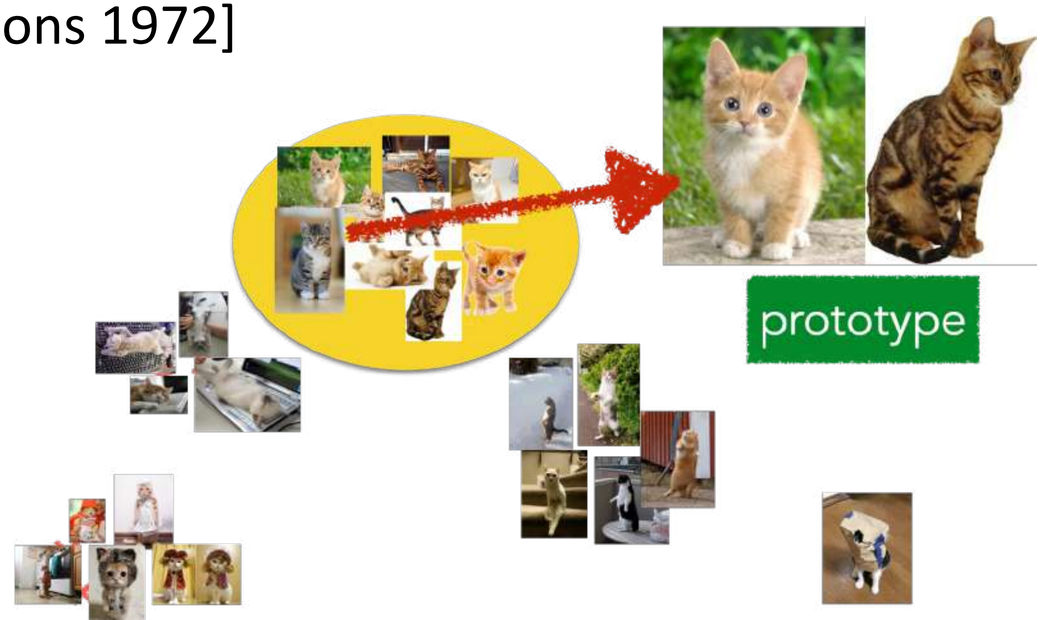
Feature visualization



Robust training seems to retain more humanly identifiable information. How do we test this?

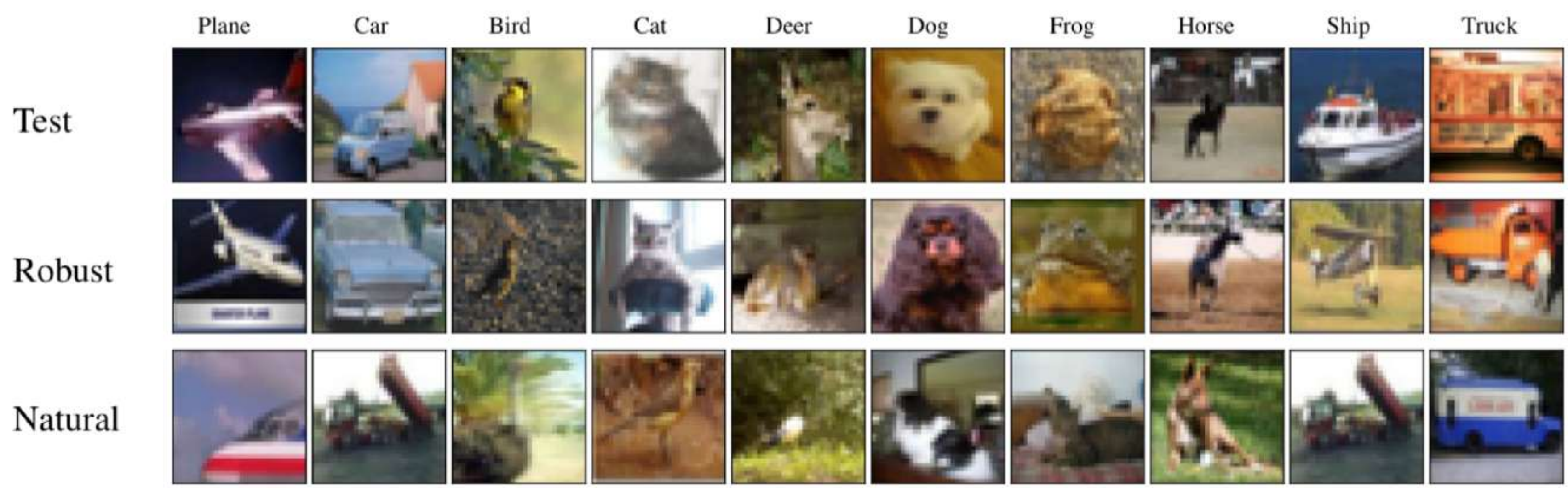
Example based learning

- Human beings learn through examples/prototypes, but can over-generalize[1]
- Build a mental model of the concept based on prototypes [Newell/Simons 1972]



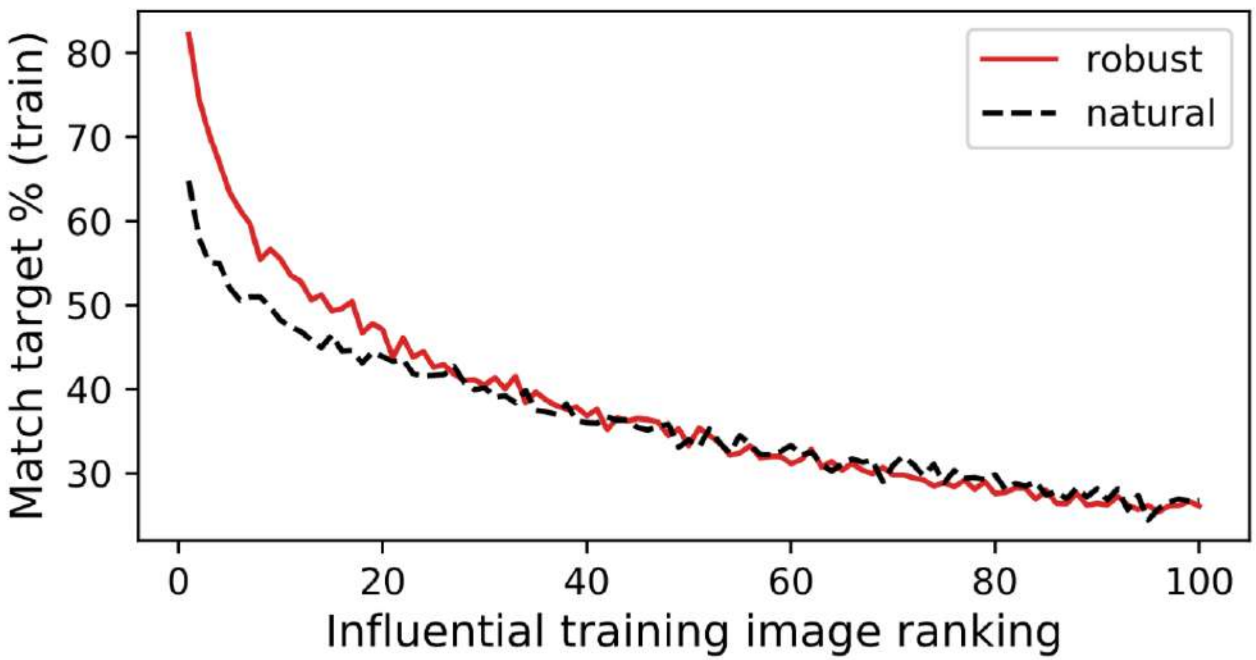
1 **Khanna***, Kim*, Koyejo* NeurIPS 2016

Using Fisher Kernels for Interpretation



Most influential images

Using Fisher Kernels for interpretation



- Takeaway: Similar images are more influential for robust than natural
- Throwback: Theoretical guarantees for the greedy selection are vital here

Closing the loop: Human insights for classification

- Non-adversarially trained models are biased towards retaining texture info. [Geirhos et. al. 2019]
- Human beings use shape more than textures when classifying objects [Landau et. al. 1988]
- Robustly trained neural networks are biased towards retaining shape information (Additional experiments on stylized ImageNet)

Thanks to my wonderful collaborators!

- Optimization: Michael Mahoney (UC Berkeley), Anastasios Kyrillidis (Rice), Martin Jaggi (ETH Zurich)
- Theory: Sahand Negahban (Yale)
- Machine Learning: Joydeep Ghosh (UT Austin), Alex Dimakis (UT Austin)
- Bayesian learning: Oluwasanmi Koyejo (UIUC)
- Neuroscience: Russell Poldrack (Stanford)
- Interpretability: Been Kim (Google) ... several others
- Other + ongoing works: Optimization guarantees for greedy-like algorithms, Bayesian Coresets, Algorithmic generalization guarantees.