Learning from sketches: large-scale learning with the memory of a goldfish

Rémi Gribonval - ENS Lyon & Inria - DANTE team
remi.gribonval@inria.fr

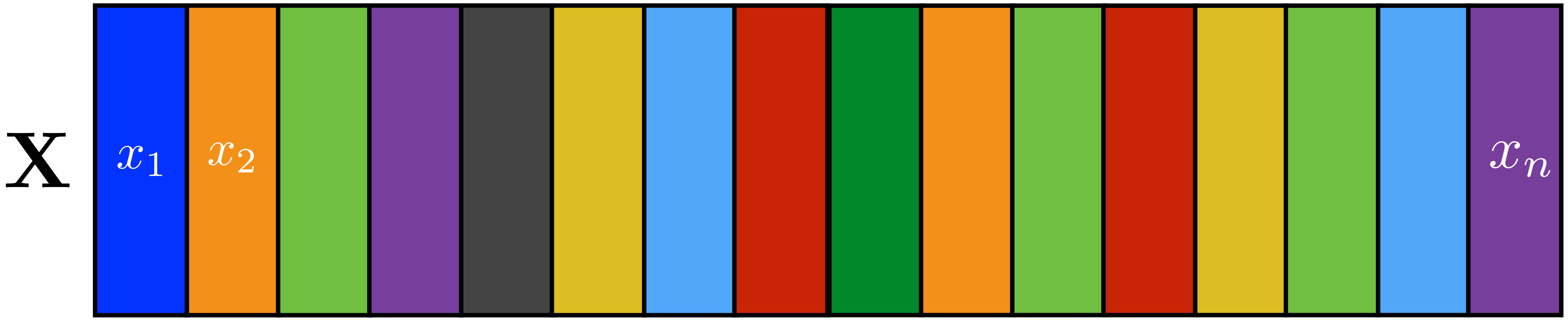
This project was supported in part by the AllegroAssai ANR project ANR-19-CHIA-0009

Contributors & Collaborators

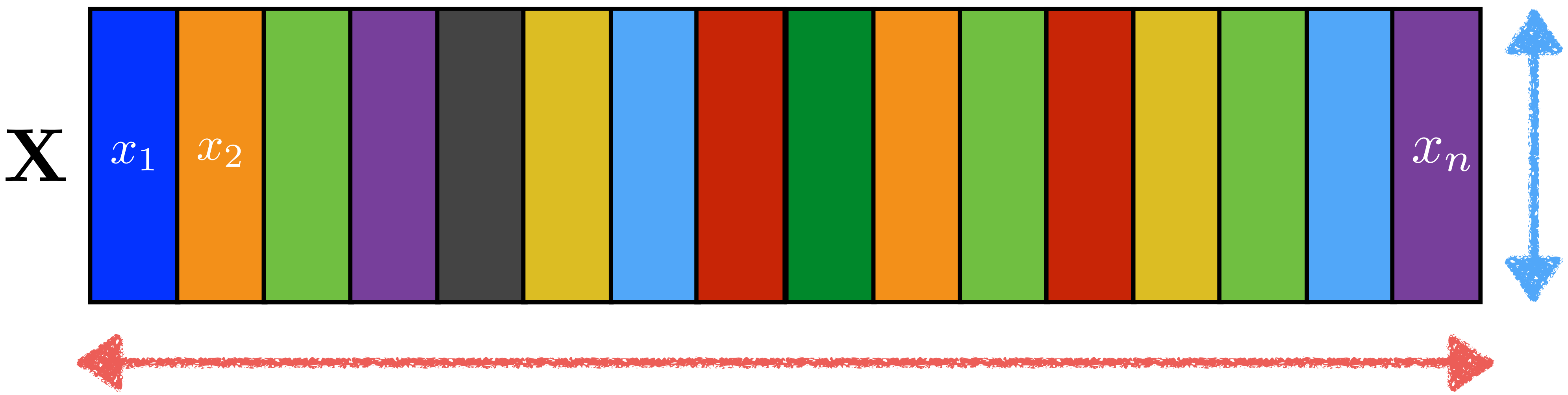
- Recently with **Titouan Vayer** and **Ayoub Belhadji**

- Building on a series of past work with
  - A. Bourrier, N. Keriven, A. Chatalic
  - G. Puy, N. Tremblay, Y. Traonmilin, C. Elvira, L. Giffon
  - P. Perez, M. Davies, G. Blanchard, P. Vandergheynst
  - L. Jacques, V. Schellekens, F. Houssiau, P. Schniter, E. Byrne, …

Large-scale learning

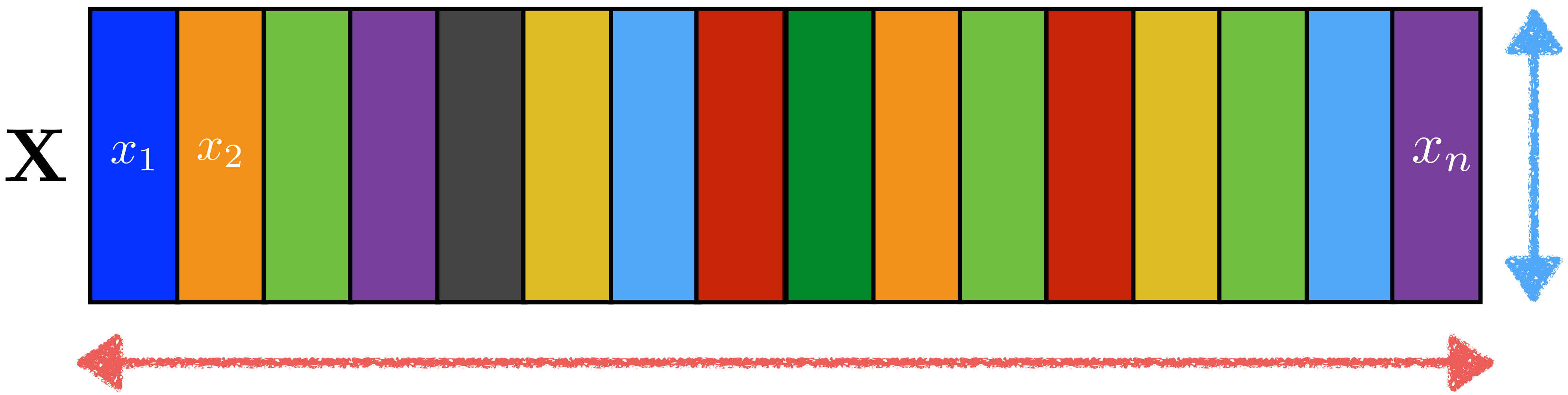


Large-scale learning



- High feature dimension d
- Large collection size n = “*volume*”

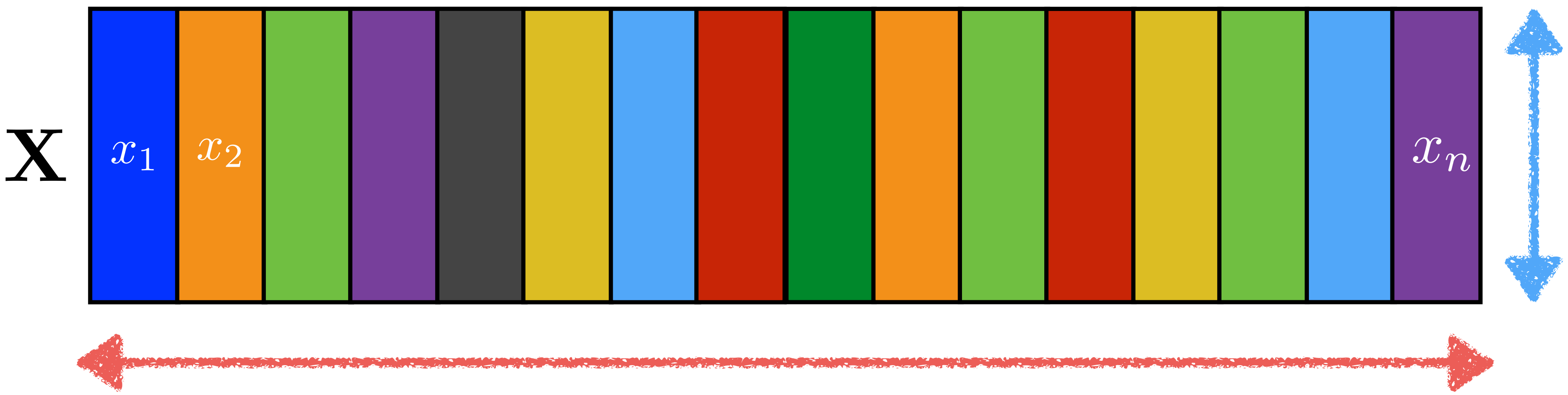
Large-scale learning



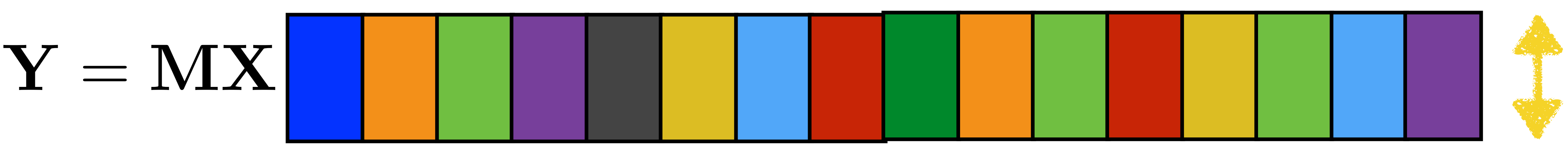
- High feature dimension d
- Large collection size n = *"volume"*

Challenge: compress $\mathcal{X}$ before learning ?

Compressive learning: three routes

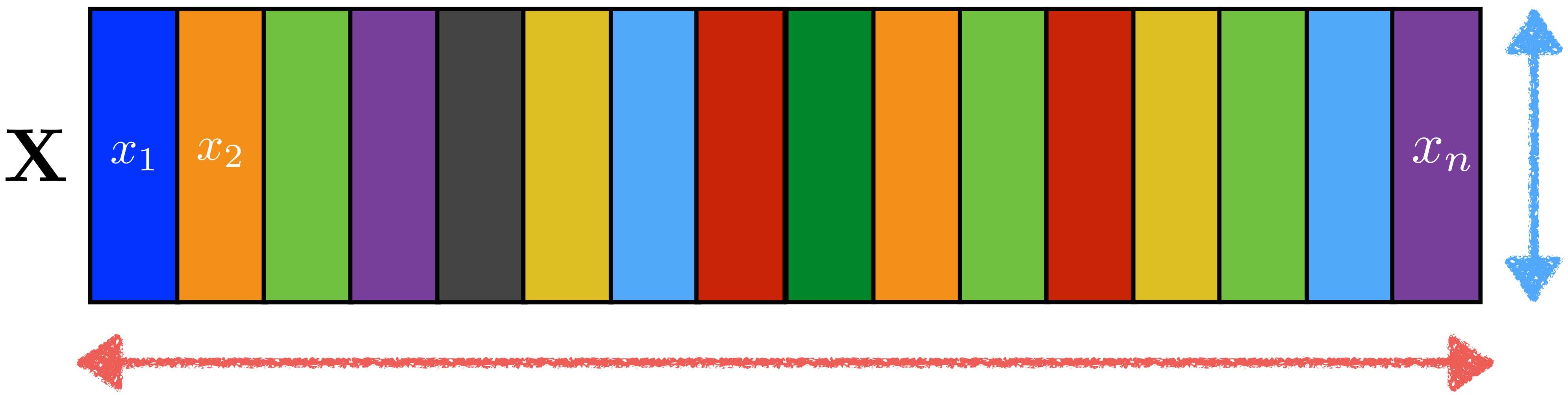


- dimension reduction
- subsampling
- sketching

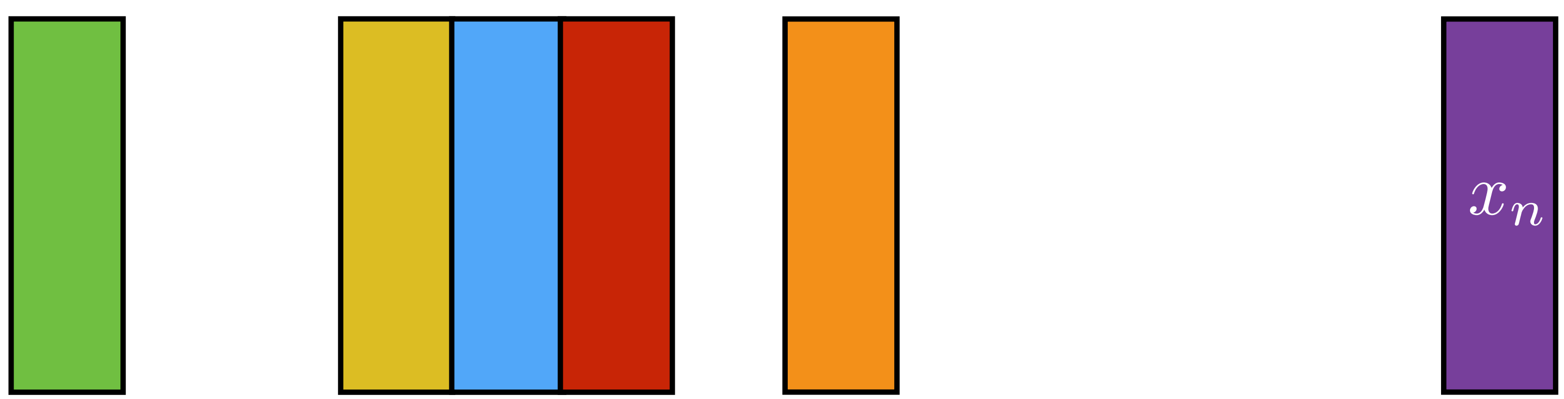


random projections - Johnson Lindenstrauss lemma
see e.g. [Calderbank & al 2009, Reboredo & al 2013]

Compressive learning: three routes

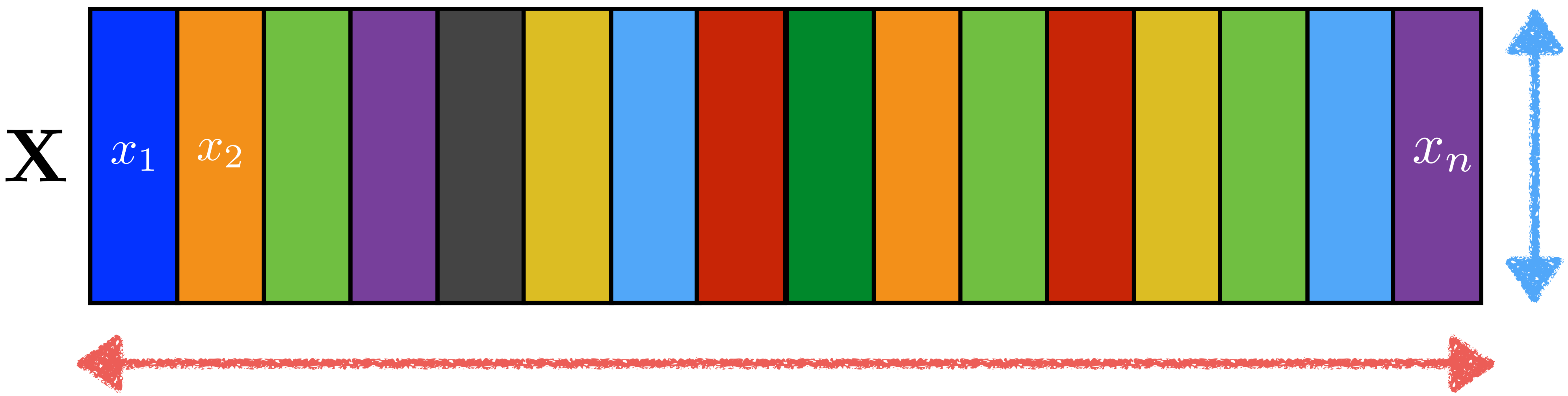


- dimension reduction
- subsampling
- **sketching**



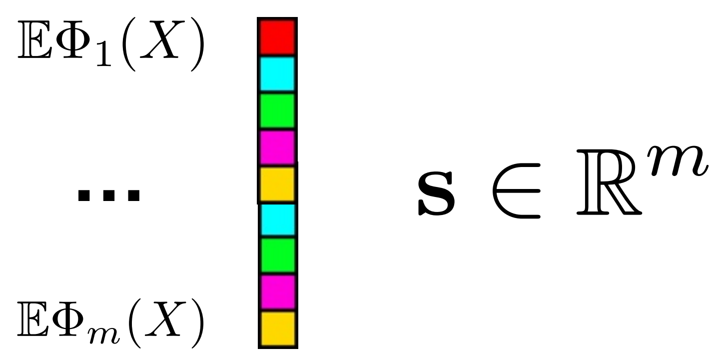
Nyström method & coresets
see e.g. [Williams&Seeger 2000, Agarwal & al 2003, Felman 2010]

Compressive learning: three routes



- dimension reduction
- subsampling
- **sketching**

  with random moments



Inspiration: *compressive sensing* *[Foucart & Rauhut 2013]*
sketching/hashing *[Thaper & al 2002, Cormode & al 2005]*
Connections with: *generalized method of moments* *[Hall 2005]*
kernel mean embeddings[Smola & al 2007, Sriperimbudur & al 2010]

- Principle of compressive learning
- Analogy with compressive sensing
- Inequalities between metrics on probability distribution
- Private sketching
- Take Home Messages

Statistical Learning

learning task →

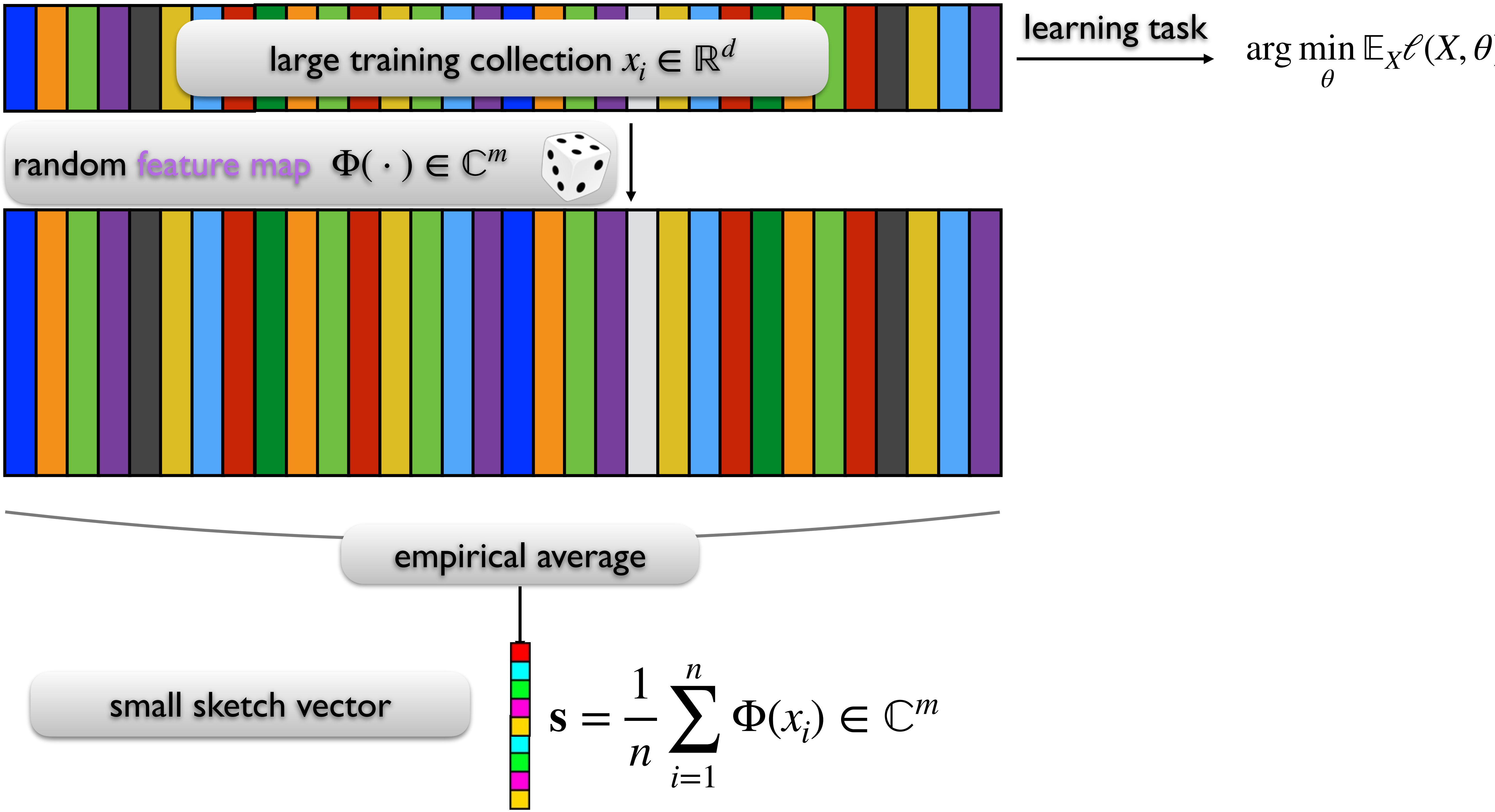
$$\arg\min_{\theta} \mathbb{E}_X \ell(X, \theta)$$

- Traditional approach:
  - (Convex) optimization
  - (Stochastic) gradient descent

$$\hat{\theta} \approx \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \theta)$$

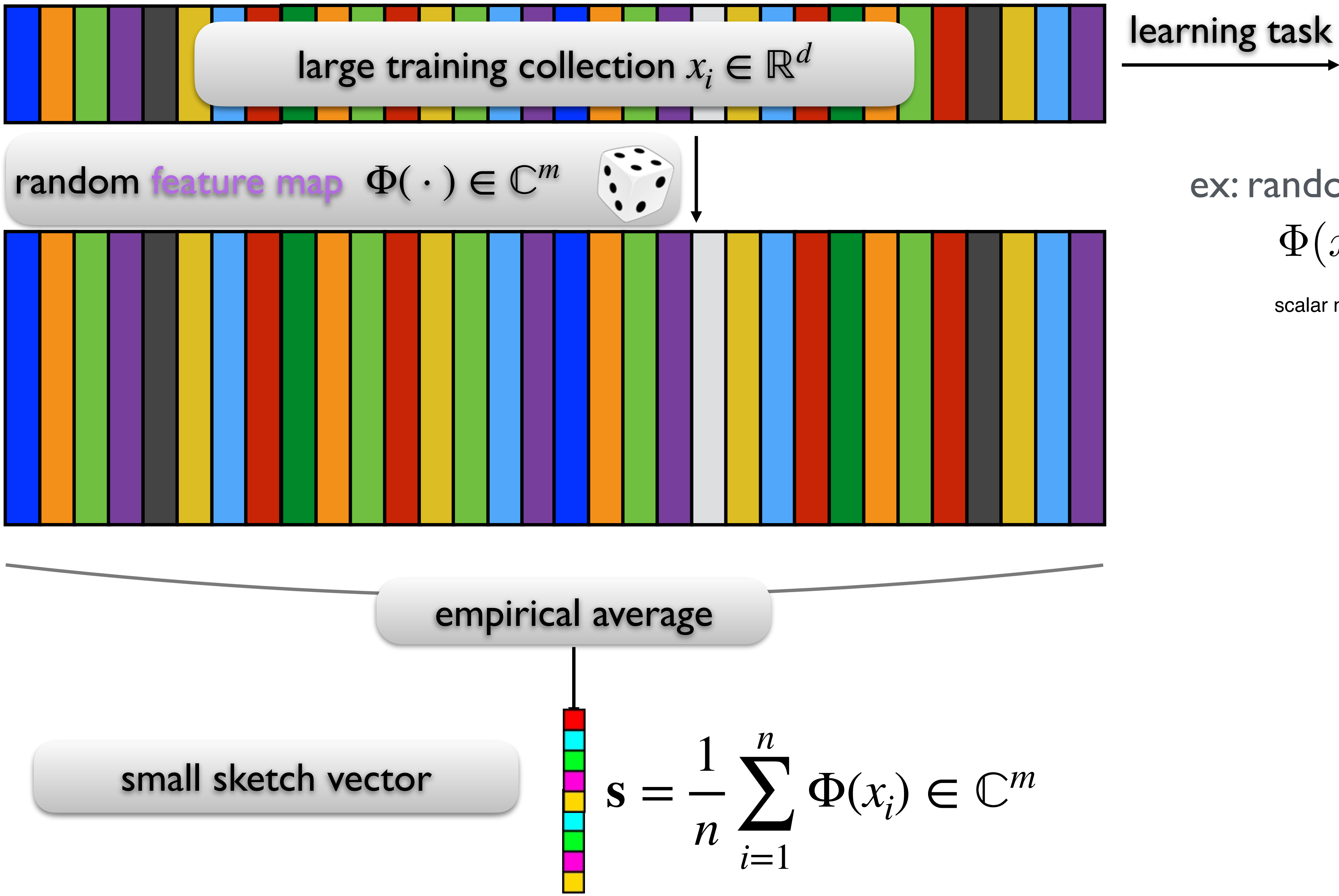
- Several passes on the training set
- Resource hungry at large scale

Compressive Statistical Learning



$$\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i) \in \mathbb{C}^m$$

Compressive Statistical Learning

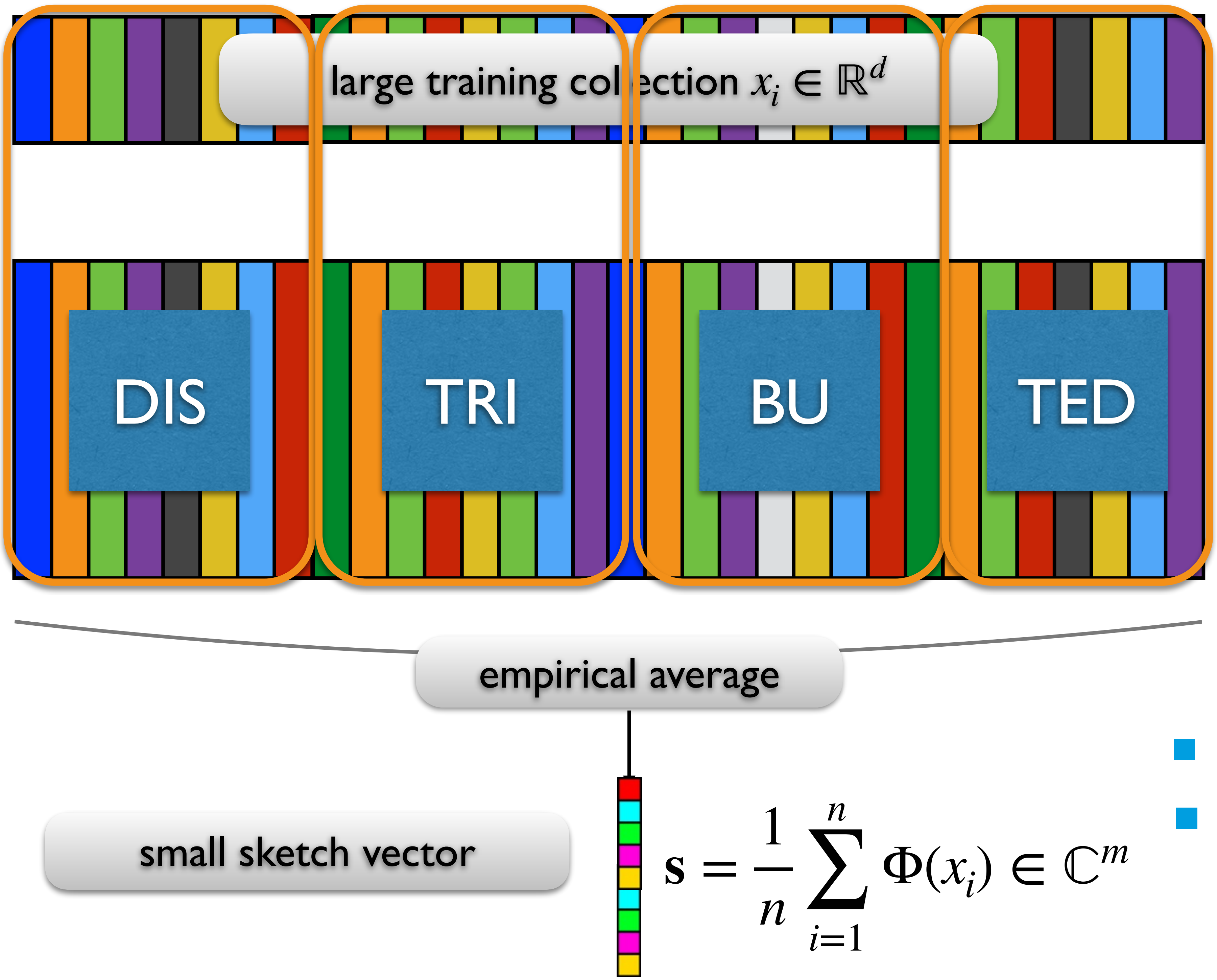


ex: random Fourier features

$$\Phi(x) = \rho(\mathbf{W}x)$$

scalar nonlinearity
random projection

Compressive Statistical Learning

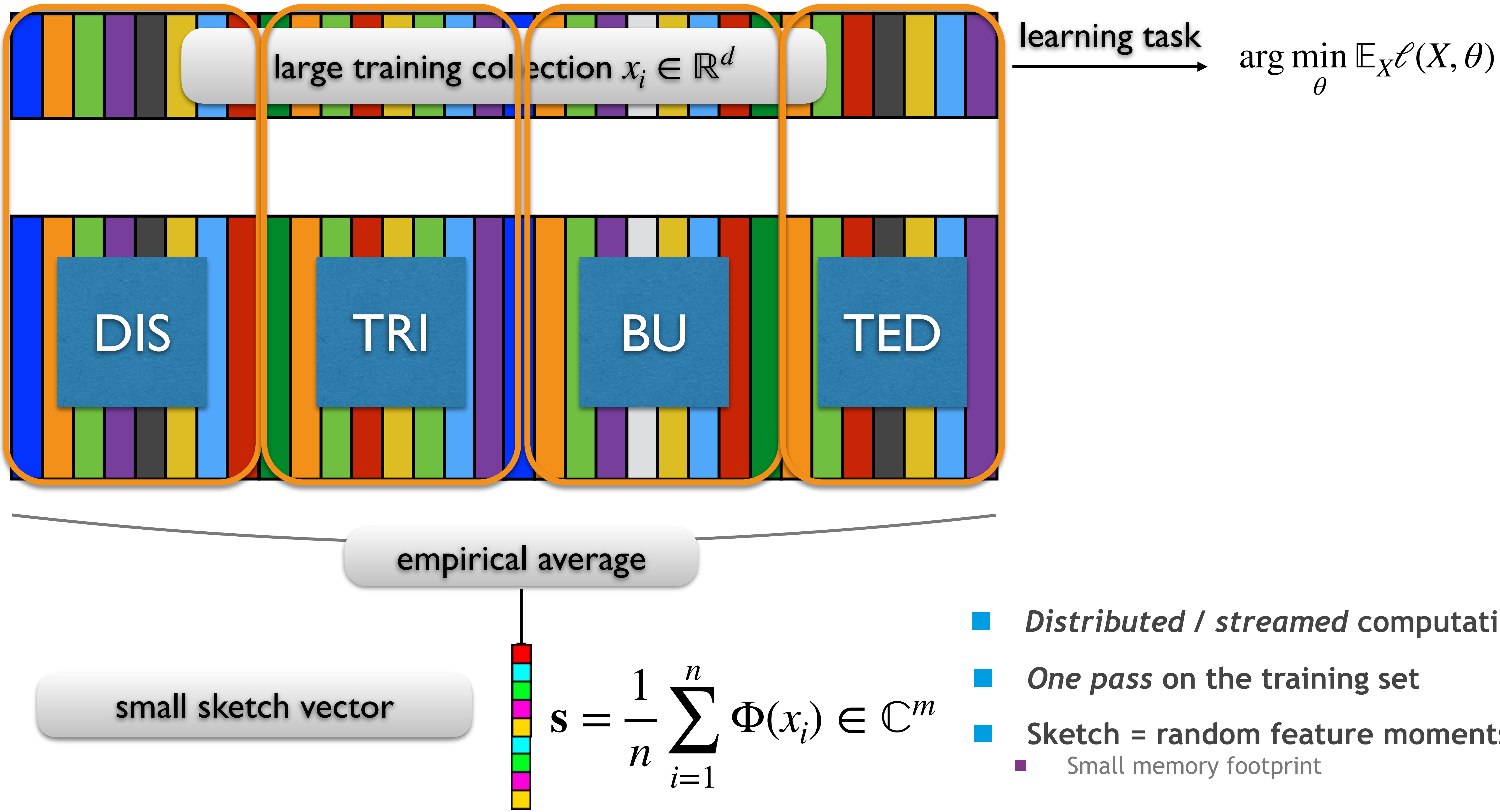


learning task → $\arg\min_{\theta} \mathbb{E}_X \ell(X, \theta)$

$$\mathbf{s} = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \in \mathbb{C}^m$$

- *Distributed / streamed* computation
- *One pass* on the training set

Compressive Statistical Learning



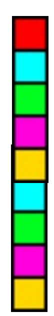
$$\mathbf{s} = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \in \mathbb{C}^m$$

- *Distributed / streamed* computation
- *One pass* on the training set
- Sketch = random feature moments
  - Small memory footprint

Compressive Statistical Learning

learning task → $\arg\min_{\theta} \mathbb{E}_X \ell(X, \theta)$

small sketch vector



$$\mathbf{s} = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \in \mathbb{C}^m$$

- *Distributed / streamed* computation
- *One pass* on the training set
- Sketch = random feature moments
  - Small memory footprint
  - Privacy

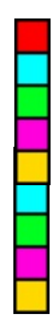
Compressive Statistical Learning

learning task → $\arg\min_{\theta} \mathbb{E}_X \ell(X, \theta)$

Learning from a sketch ?
Optimize a surrogate

$$\tilde{\theta} \approx \arg\min_{\theta} \; C(\theta \,|\, \mathbf{s})$$

small sketch vector

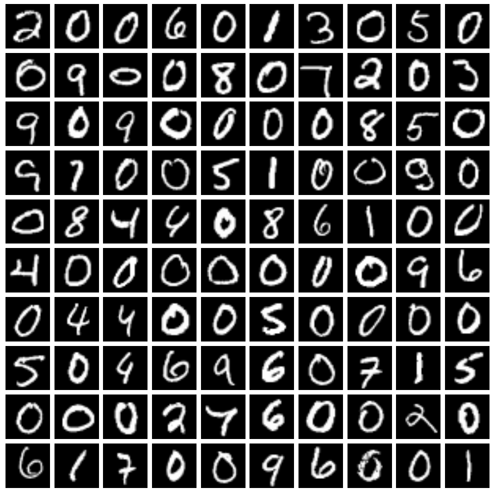


$$\mathbf{s} = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \in \mathbb{C}^m$$

- *Distributed / streamed* computation
- *One pass* on the training set
- Sketch = random feature moments
  - Small memory footprint
  - Privacy

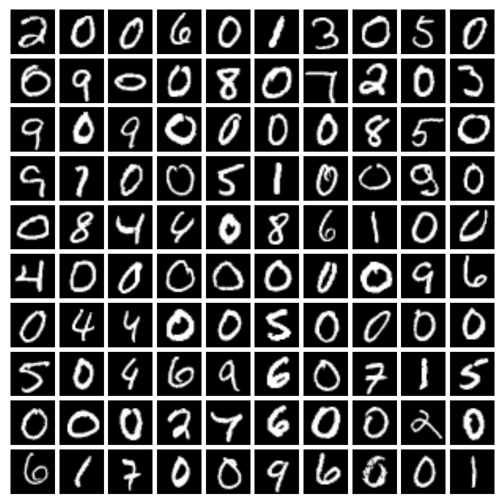
Example: clustering MNIST

Handwritten digits



Example: clustering MNIST

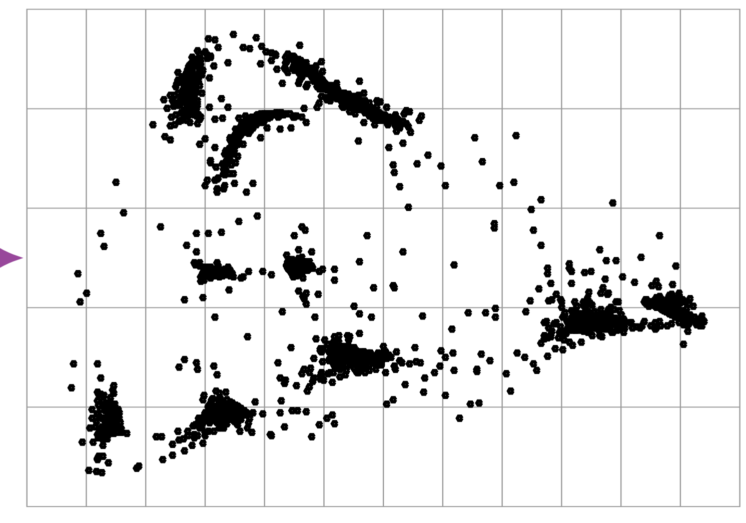
Handwritten digits



Pre-processing

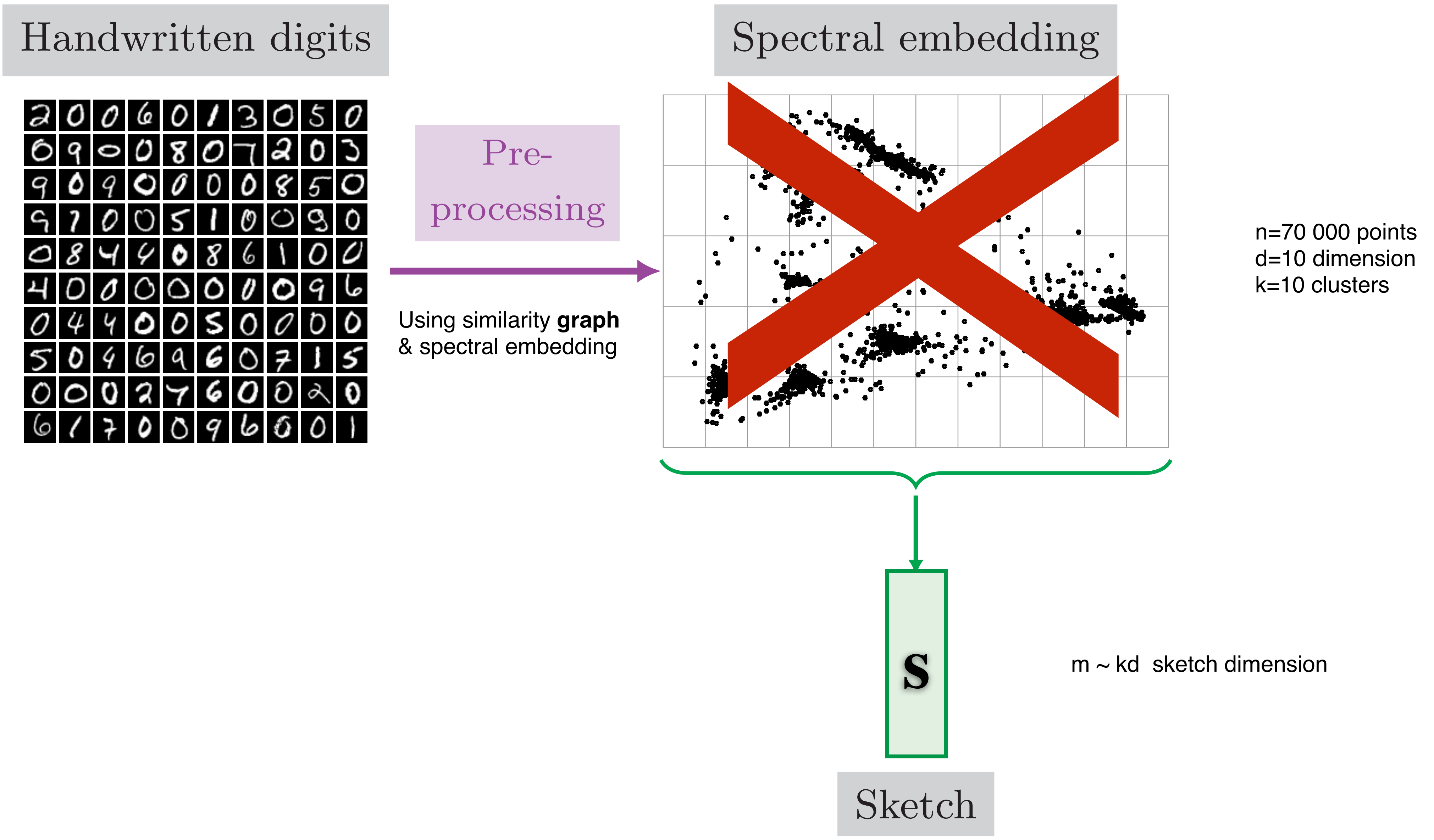
Using similarity **graph**
& spectral embedding

Spectral embedding

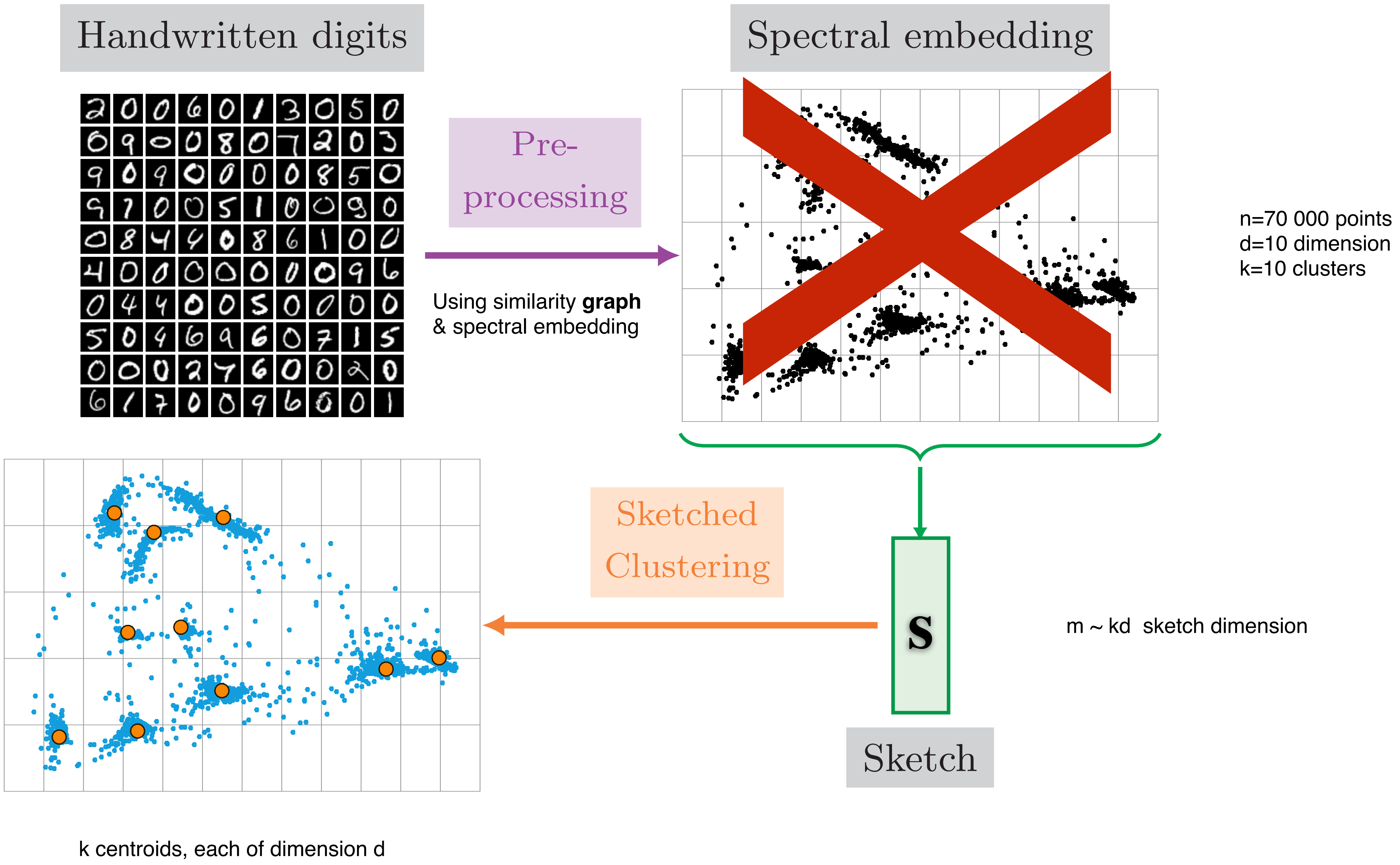


n=70 000 points
d=10 dimension
k=10 clusters

Example: clustering MNIST



Example: clustering MNIST



Comparison with traditional learning

Traditional approach

- ideal goal: minimize **risk**

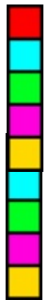
$$\mathcal{R}(p,\theta) = \mathbb{E}_{X\sim p}\ell(X,\theta)$$

- empirical risk minimization

$$\hat{\theta} \approx \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\ell(x_i,\theta)$$

Compressive learning

- sketch the training data



$$\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i) \in \mathbb{R}^m$$

- optimize a **surrogate**

$$\tilde{\theta} \approx \arg\min_{\theta} \; C(\theta|\mathbf{s})$$

Comparison with traditional learning

Traditional approach

- ideal goal: minimize **risk**

$$\mathcal{R}(p,\theta) = \mathbb{E}_{X\sim p}\ell(X,\theta)$$

- empirical risk minimization

$$\hat{\theta} \approx \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \ell(x_i,\theta)$$

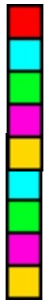
- need access to training samples

Computationally expensive.
High energy consumption.
Multiple passes on the data.

Sensitive data
(e.g. emails, medical data).

Compressive learning

- sketch the training data



$$\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i) \in \mathbb{R}^m$$

- optimize a **surrogate**

$$\tilde{\theta} \approx \arg\min_{\theta} \; C(\theta|\mathbf{s})$$

Comparison with traditional learning

Traditional approach

- ideal goal: minimize **risk**

$$\mathcal{R}(p,\theta) = \mathbb{E}_{X\sim p}\ell(X,\theta)$$

- empirical risk minimization

$$\hat{\theta} \approx \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\ell(x_i,\theta)$$

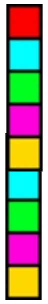
- need access to training samples

Computationally expensive.
High energy consumption.
Multiple passes on the data.

Sensitive data
(e.g. emails, medical data).

Compressive learning

- sketch the training data



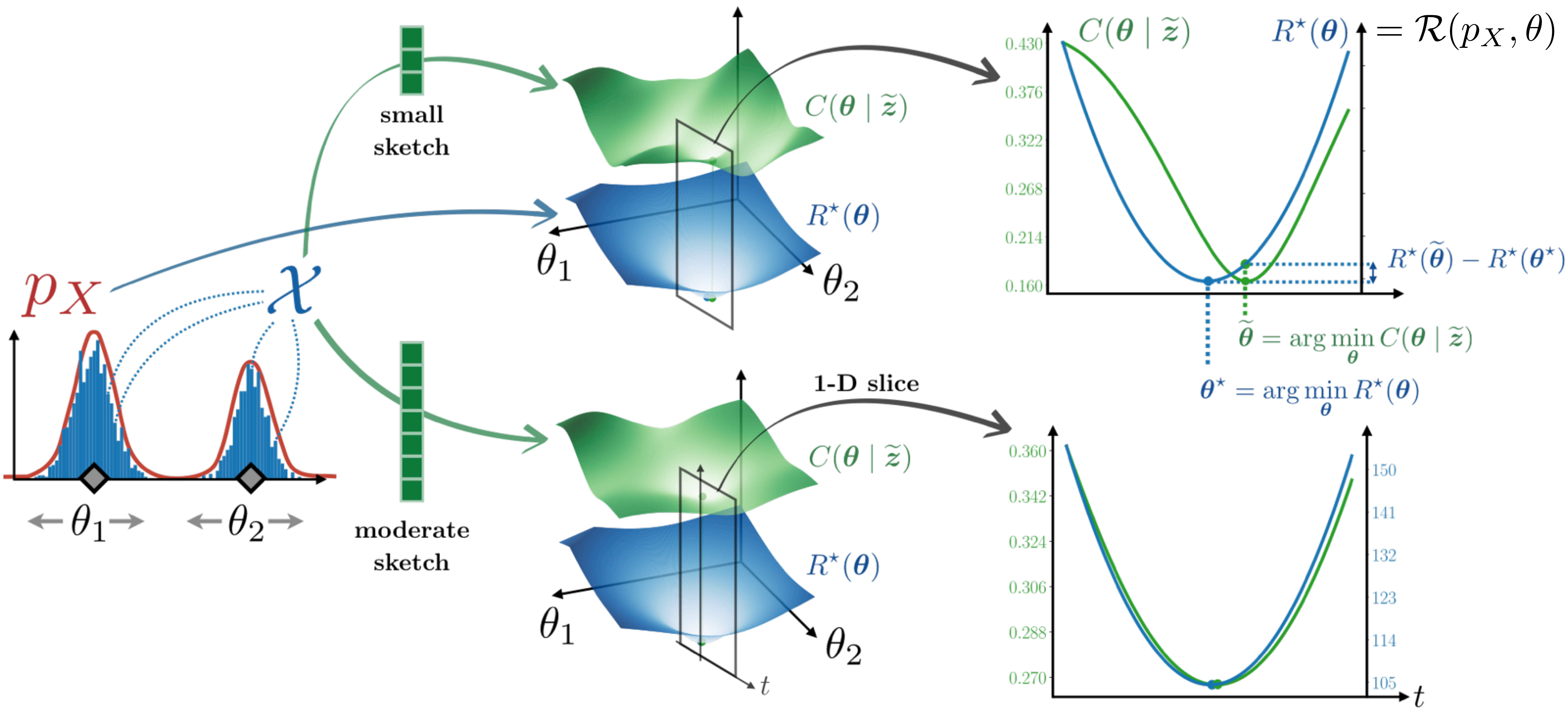
$$\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i) \in \mathbb{R}^m$$

- optimize a **surrogate**

$$\tilde{\theta} \approx \arg\min_{\theta} \; C(\theta|\mathbf{s})$$

- can « forget » training samples
  - complexity **independent of n**
  - slight variant **differentially private**
- **good surrogate ??**

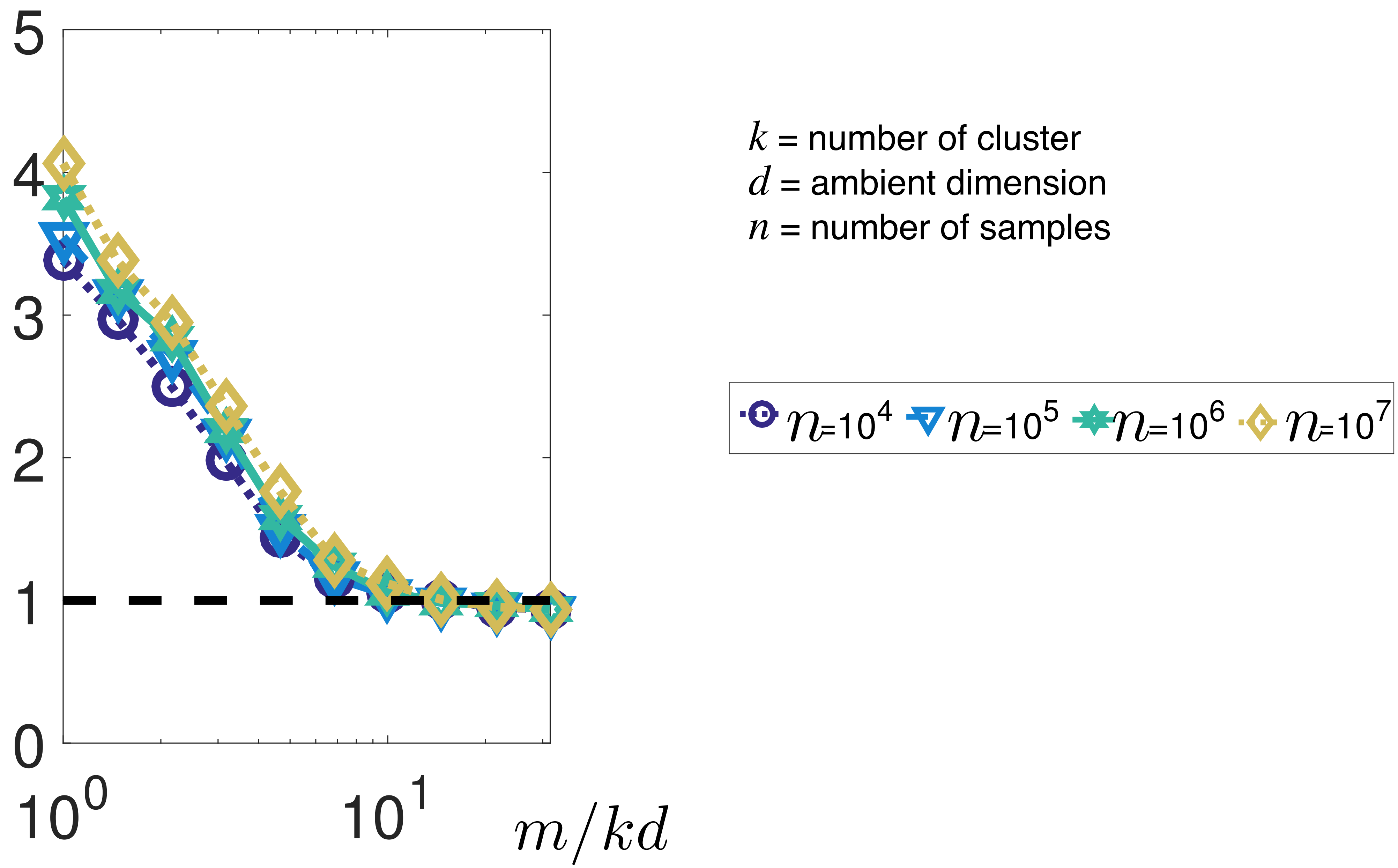
Optimization landscapes
toy example - clustering



Surrogate $C(\theta \,|\, \mathbf{s})$ = distance between sketch of model 2-mixture and empirical sketch

Effect of sketch size *m* (clustering - planted model)

Relative loss $\frac{\mathcal{R}(p,\hat{\theta})}{\mathcal{R}(p,\theta^\star)}$



k = number of cluster
d = ambient dimension
n = number of samples

- Principle of compressive learning
- Analogy with compressive sensing
- Inequalities between metrics on probability distribution
- Private sketching
- Take Home Messages

Moments & kernel mean embeddings

- **Data distribution** $X \sim p(x)$

- **Sketch = vector of *generalized moments***

$$\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i) \approx \mathbb{E}\,\Phi(X) = \int \Phi(x)p(x)dx$$

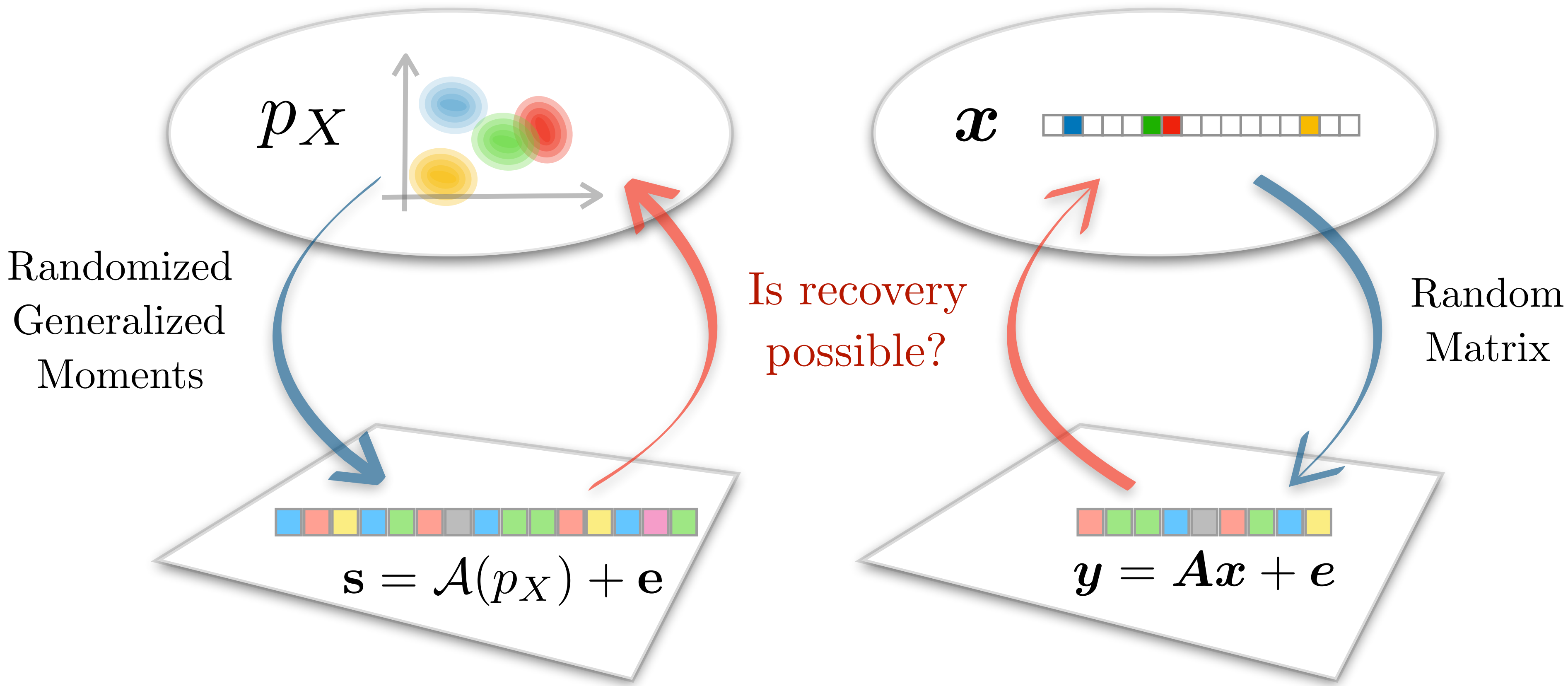
  - ***nonlinear*** in the feature vectors
  - ***linear*** in the distribution *p(x)*

  - finite-dimensional **Mean Map Embedding**, *[cf Smola & al 2007, Sriperumbudur & al 2010]*

$$\mathcal{A}(p) := \mathbb{E}_{X\sim p}\Phi(X)$$

Utility guarantees ?

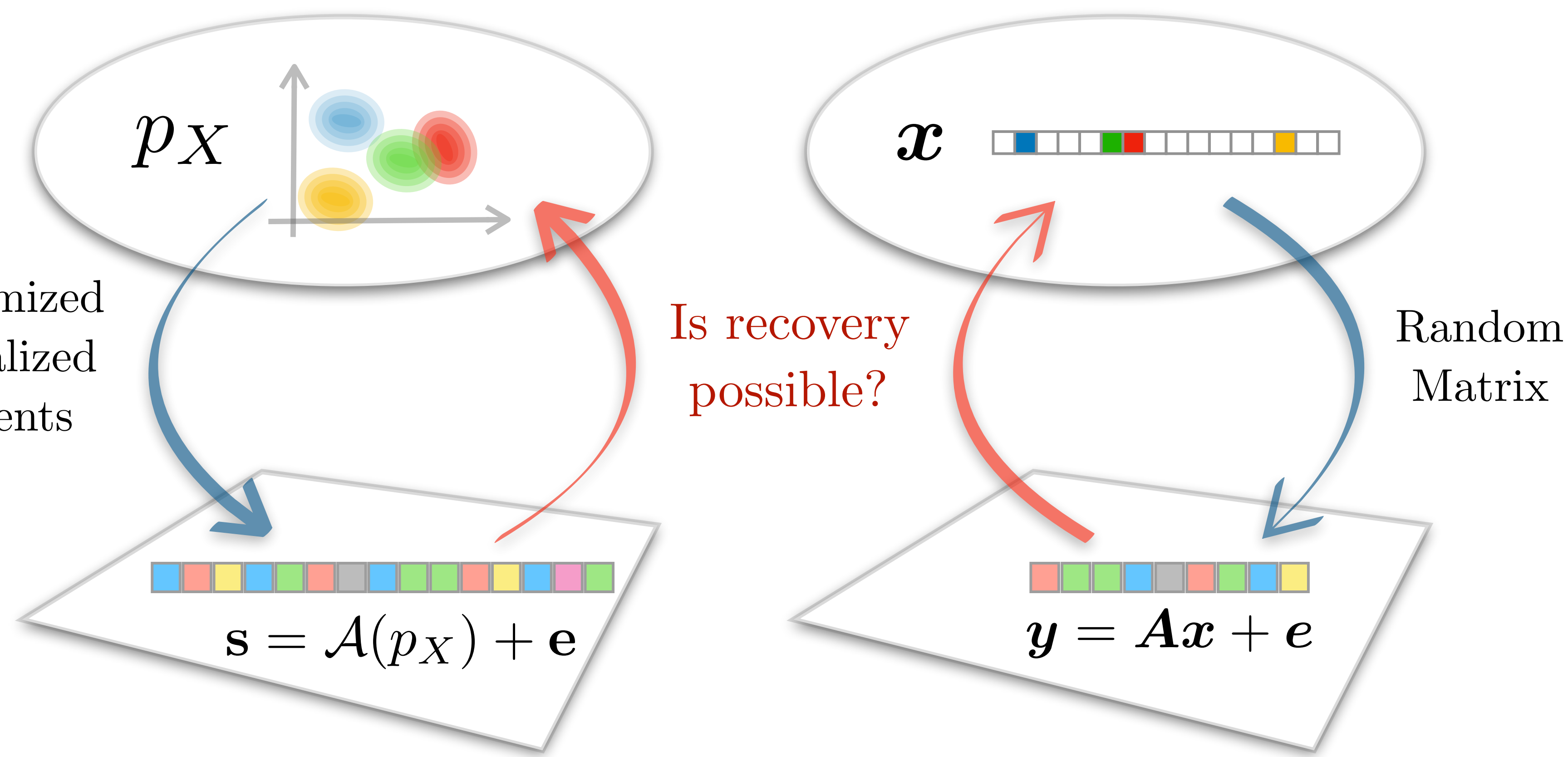
- **Compressive learning**
  - ex: mixtures of k Gaussians
- **Compressive sensing**
  - ex: k-sparse vectors



[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*]

Utility guarantees ?

- Compressive learning
  - ex: mixtures of k Gaussians
- Compressive sensing
  - ex: k-sparse vectors



Statistical guarantees = control of excess risk

Key ingredients

Task-based metric on distributions

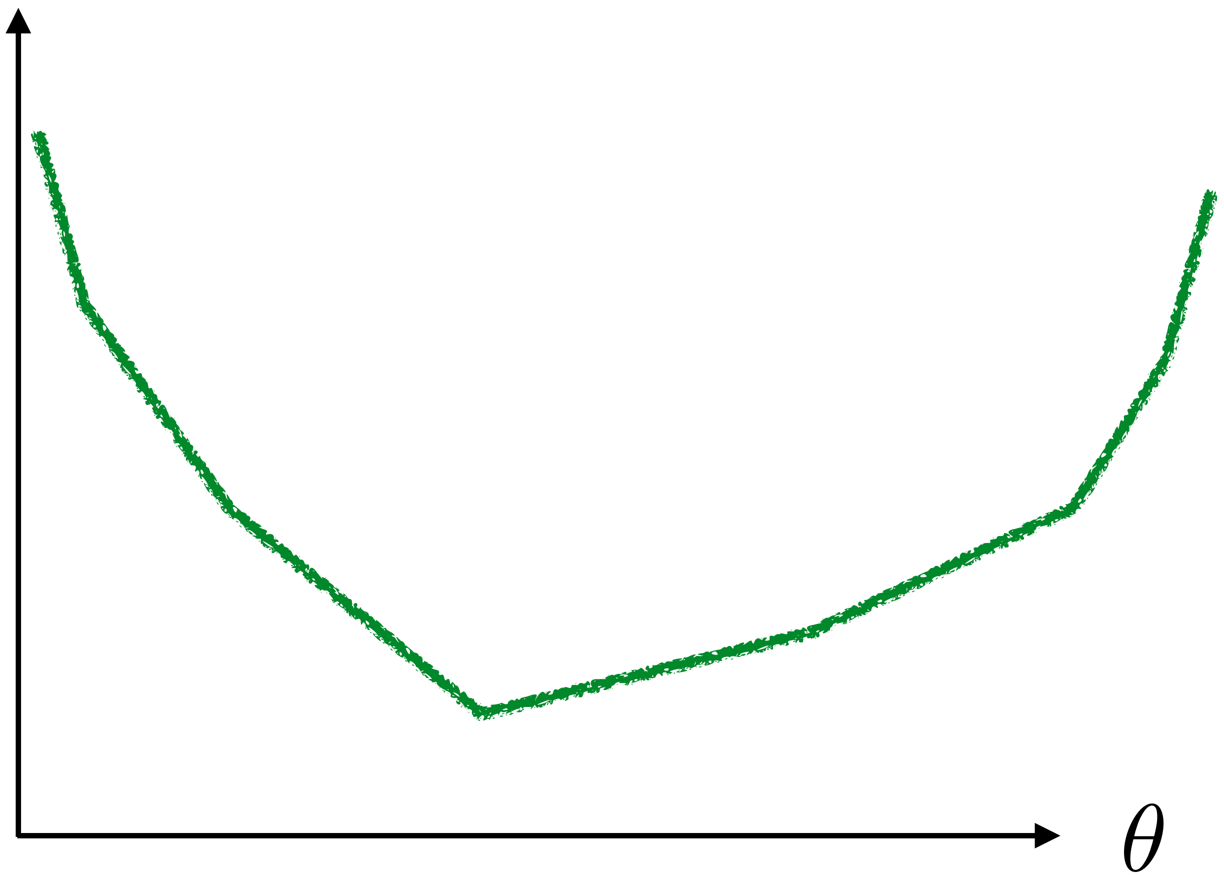
Low-dim model set (task-dependent)

[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*]

Risk & task-dependent metric

- Statistical risk

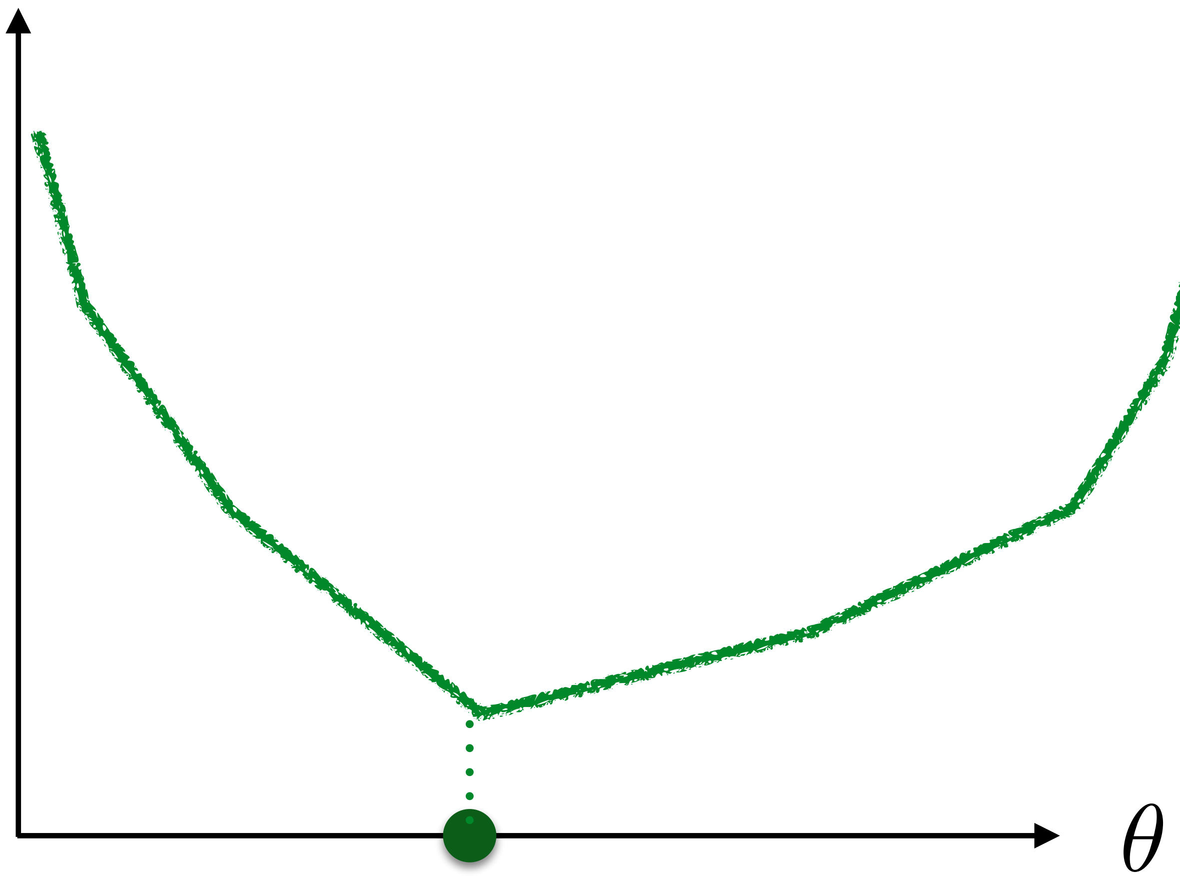
$$\mathcal{R}(p,\theta) := \mathbb{E}_{X\sim p}\ell(X,\theta)$$



Risk & task-dependent metric

- Statistical risk

$$\mathcal{R}(p,\theta) := \mathbb{E}_{X\sim p}\ell(X,\theta)$$

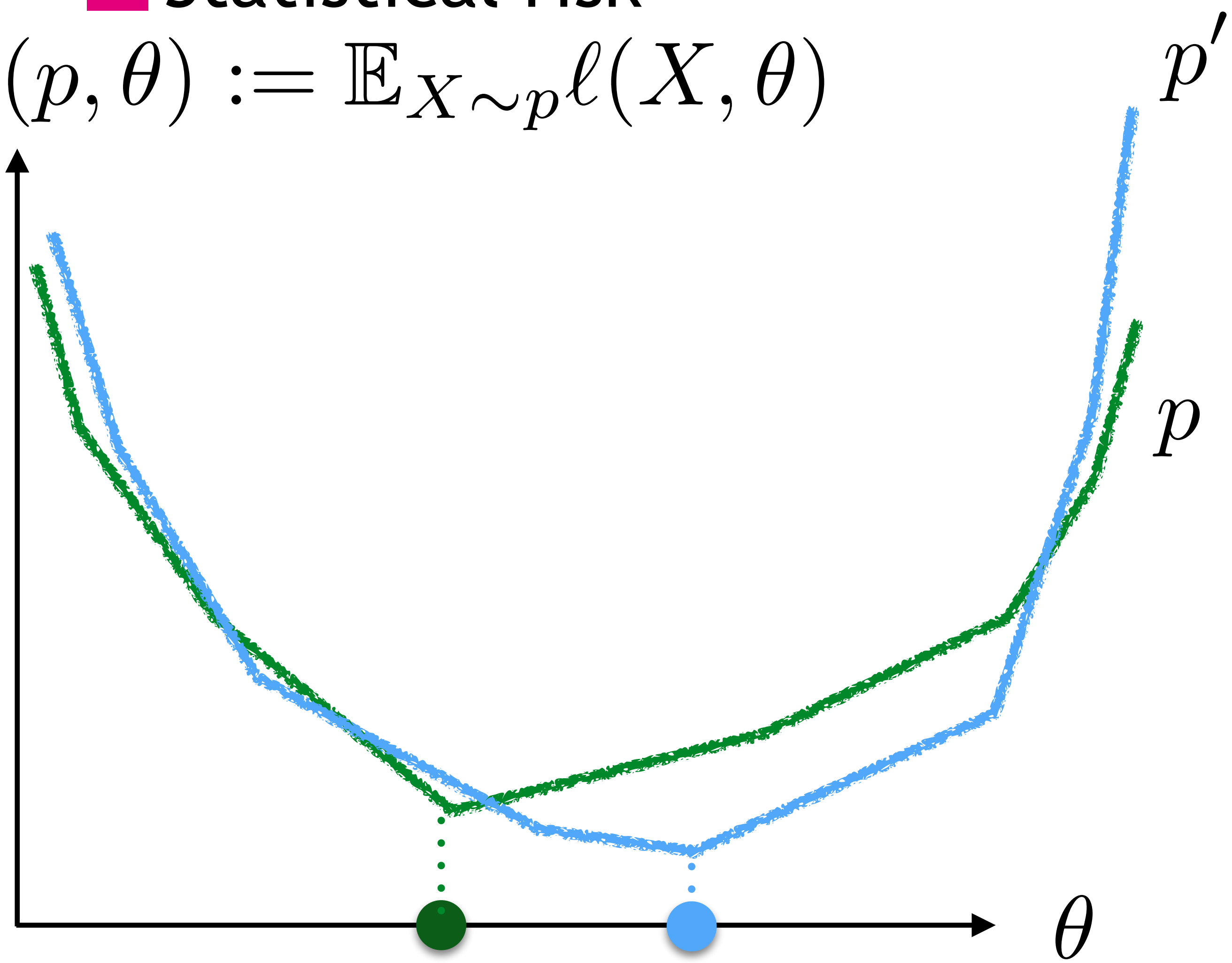


$\theta^\star \in \arg\min_\theta \mathcal{R}(p,\theta)$

Risk & task-dependent metric

- Statistical risk

$$\mathcal{R}(p,\theta) := \mathbb{E}_{X\sim p}\ell(X,\theta)$$



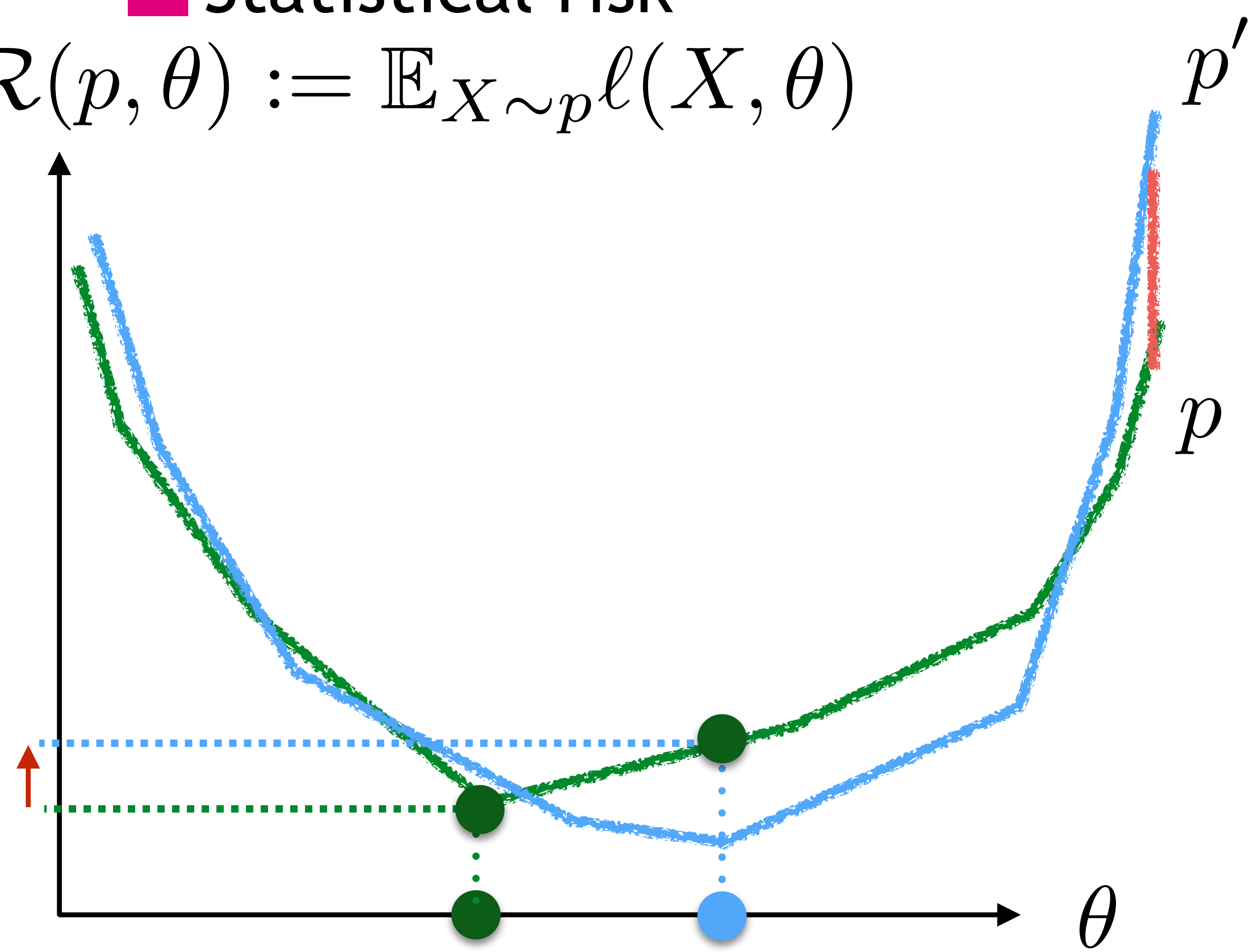
$\theta^\star \in \arg\min_\theta \mathcal{R}(p,\theta)$ $\theta' \in \arg\min_\theta \mathcal{R}(p',\theta)$

Traditional approach: use *empirical* distribution $p' = \hat{p}_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$

Risk & task-dependent metric

- Statistical risk

$$\mathcal{R}(p,\theta) := \mathbb{E}_{X\sim p}\ell(X,\theta)$$



$\theta^\star \in \arg\min_\theta \mathcal{R}(p,\theta)$ $\quad$ $\theta' \in \arg\min_\theta \mathcal{R}(p',\theta)$

- Bound on excess risk

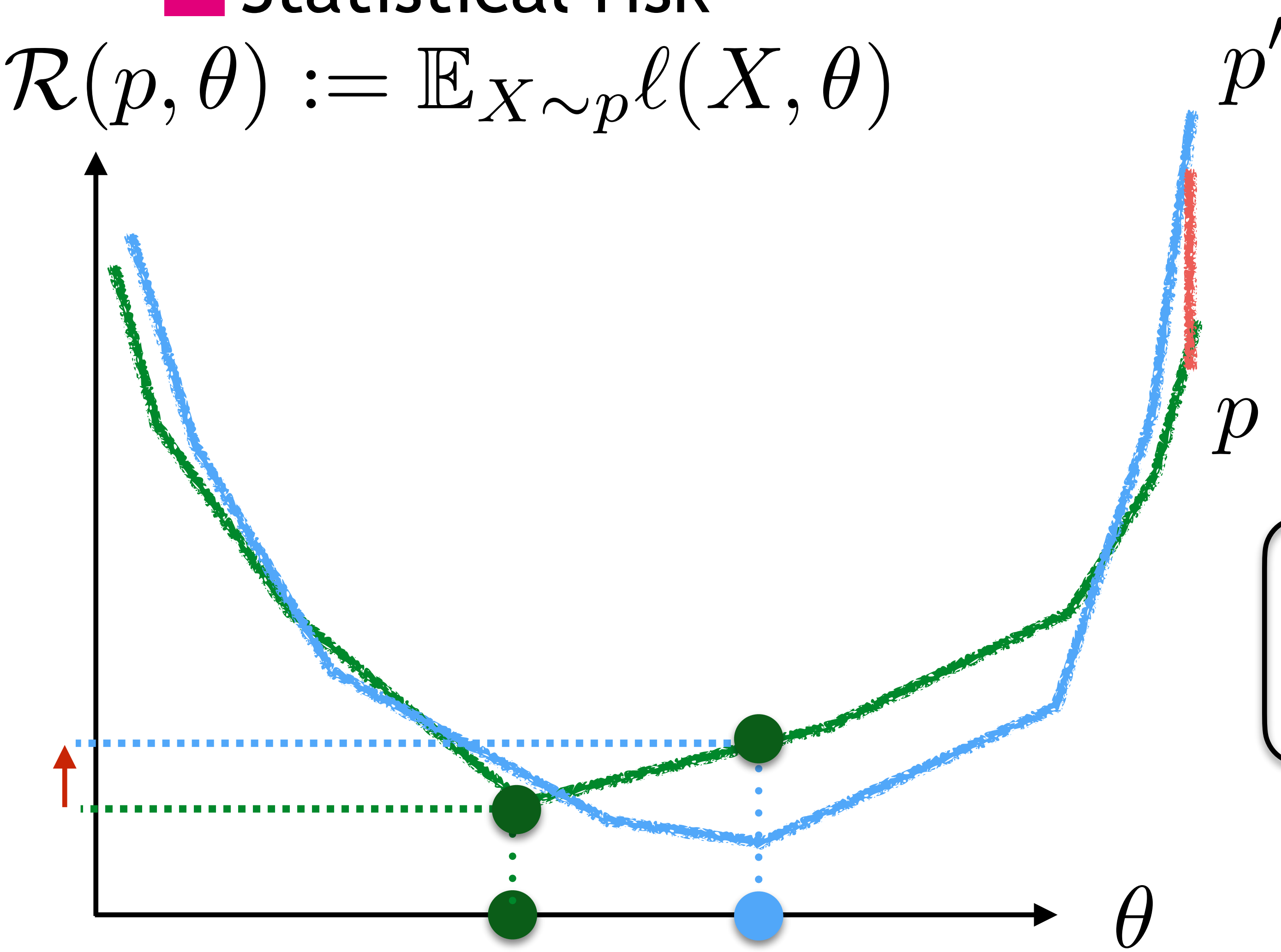
$$\mathcal{R}(p,\theta') - \mathcal{R}(p,\theta^\star) \leq 2 \sup_\theta |\mathcal{R}(p',\theta) - \mathcal{R}(p,\theta)|$$

Traditional approach: use *empirical* distribution $\quad p' = \hat{p}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

Risk & task-dependent metric

- **Statistical risk**

$$\mathcal{R}(p,\theta) := \mathbb{E}_{X\sim p}\ell(X,\theta)$$



$\theta^\star \in \arg\min_\theta \mathcal{R}(p,\theta)$ $\quad$ $\theta' \in \arg\min_\theta \mathcal{R}(p',\theta)$

- **Bound on excess risk**

$$\mathcal{R}(p,\theta') - \mathcal{R}(p,\theta^\star) \leq 2\sup_\theta |\mathcal{R}(p',\theta) - \mathcal{R}(p,\theta)|$$

- **suggests task-based metric**
  - *with abstract hypothesis class $\mathscr{H}$*

$$\mathrm{TaskMetric}(p,p') := \sup_{\theta\in\mathcal{H}} |\mathcal{R}(p,h) - \mathcal{R}(p',h)|$$

Traditional approach: use *empirical* distribution $\quad p' = \hat{p}_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$

A key metric inequality

[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*]

Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$$

A key metric inequality

[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*]

Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$$

Model set, e.g. set of (separated) *k*-mixtures of Diracs or Gaussians

A key metric inequality

[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*]

Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \ \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$$

Model set, e.g. set of (separated) *k*-mixtures of Diracs or Gaussians

$$\Downarrow$$

Statistical Guarantees $\underline{\forall \pi \in \mathcal{P}(\mathcal{X})}$

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^\circ(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2$$

A key metric inequality

[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*]

Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \ \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$$

Model set, e.g. set of (separated) *k*-mixtures of Diracs or Gaussians

$$\Downarrow$$

Statistical Guarantees $\forall \pi \in \mathcal{P}(\mathcal{X})$

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^\circ(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2$$

Excess-risk
$\hat{h}$ minimizing surrogate
$h^\star$ optimum hypothesis

A key metric inequality

[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*]

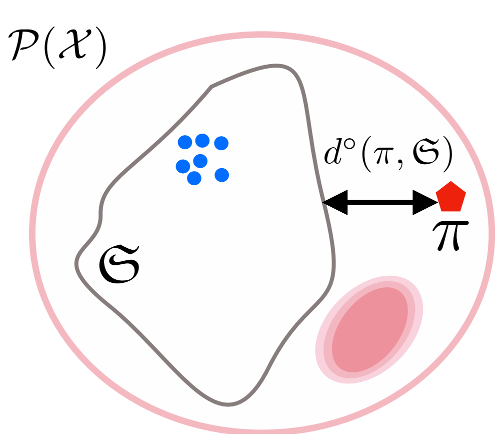
Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$$

$$\Downarrow$$

Statistical Guarantees $\underline{\forall \pi \in \mathcal{P}(\mathcal{X})}$

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^\circ(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2$$



Notion of **distance** to the model set

Bias term: vanishes when the true distrib. in the model

$$\pi \in \mathfrak{S} \implies d^\circ(\pi, \mathfrak{S}) = 0$$

$\approx$ **Approximation error in traditional statistical learning**

A key metric inequality - and a variant

[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*] **[Vayer & Gribonval, 2023, *JMLR*]**

Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$$

$$\Downarrow$$

Statistical Guarantees $\underline{\forall \pi \in \mathcal{P}(\mathcal{X})}$

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^\circ(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2$$

$\mathbf{s} = \mathcal{A}(\pi_n)$

Distance between the empirical and true sketch

Basically converges to zero in $\mathcal{O}(n^{-1/2})$

$\approx$ **Estimation error in traditional statistical learning**

A key metric inequality - and a variant

[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*] **[Vayer & Gribonval, 2023, *JMLR*]**

Lower Restricted Isometric Property (LRIP) **Hölder LRIP with $0 < \delta \leq 1$**

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^{\delta}$$

$$\Downarrow$$

Statistical Guarantees $\underline{\forall \pi \in \mathcal{P}(\mathcal{X})}$

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^\circ(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2^{\delta}$$

$\mathbf{s} = \mathcal{A}(\pi_n)$

Distance between the empirical and true sketch

Basically converges to zero in $\mathcal{O}(n^{-1/2})$ $\rightarrow \mathcal{O}(n^{-\delta/2})$

$\approx$ **Estimation error in traditional statistical learning**

- Principle of compressive learning
- Analogy with compressive sensing
- Inequalities between metrics on probability distribution
- Private sketching
- Take Home Messages

Goal = establish the Hölder LRIP

Lower Restricted Isometric Property (LRIP)

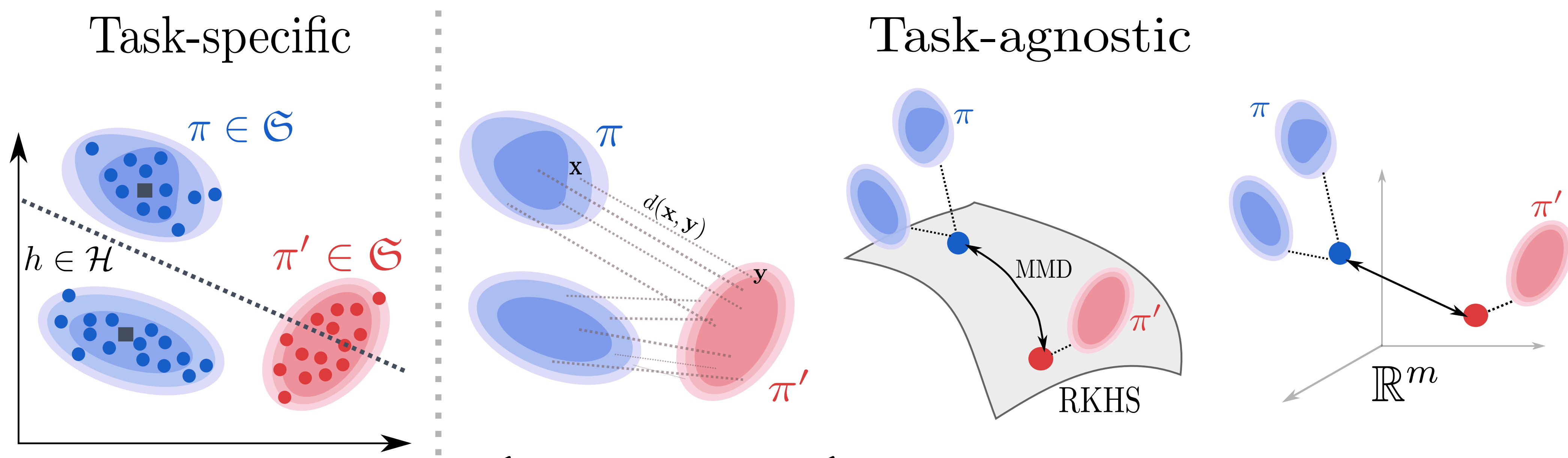
$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^{\delta}$$

Goal = establish the Hölder LRIP

Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^{\delta}$$

Roadmap



$$\mathrm{TaskMetric}(\pi, \pi') \lesssim W_q(\pi, \pi') \lesssim \mathrm{MMD}^{\delta}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^{\delta}$$

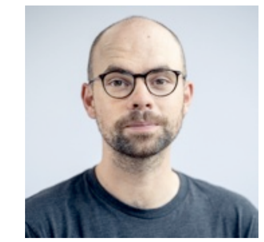
Wasserstein Learnability

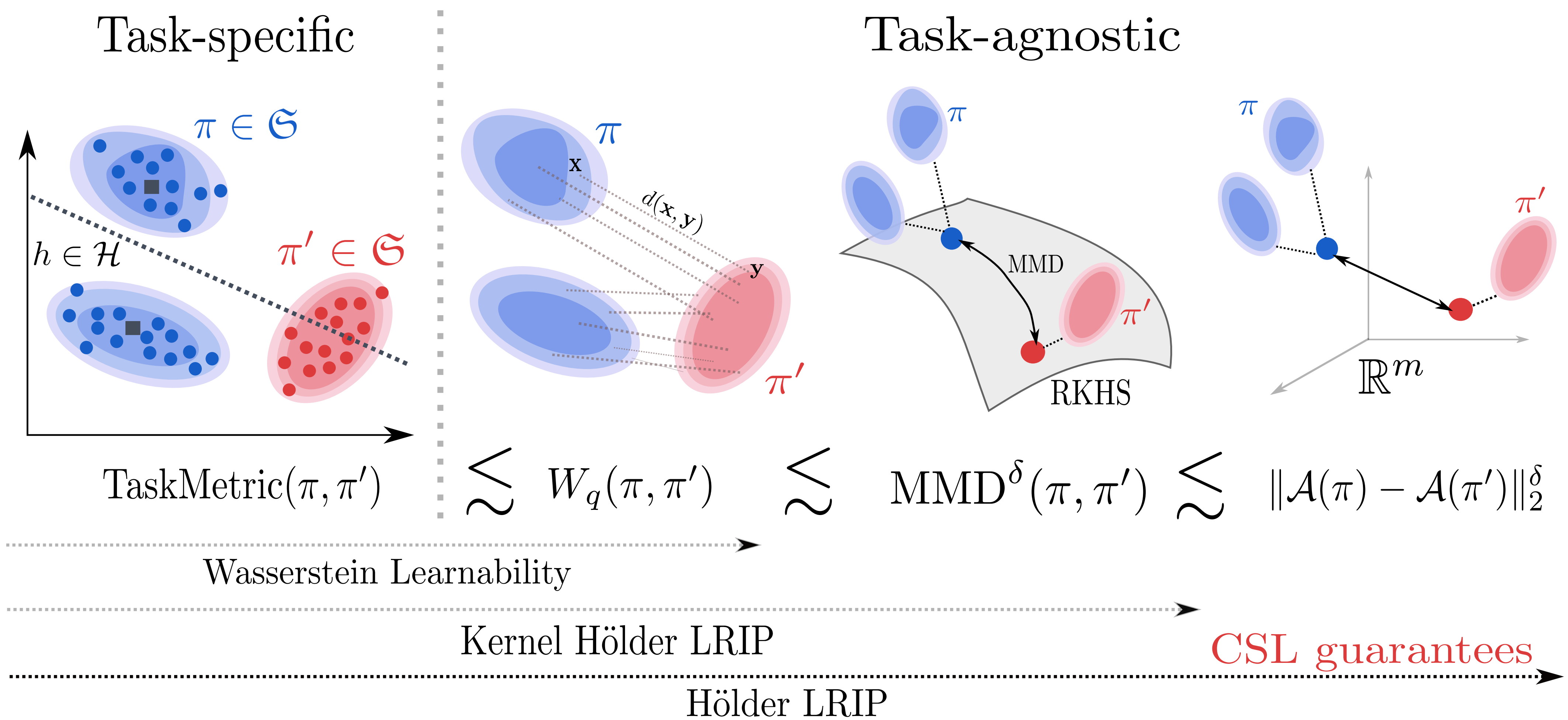
Kernel Hölder LRIP

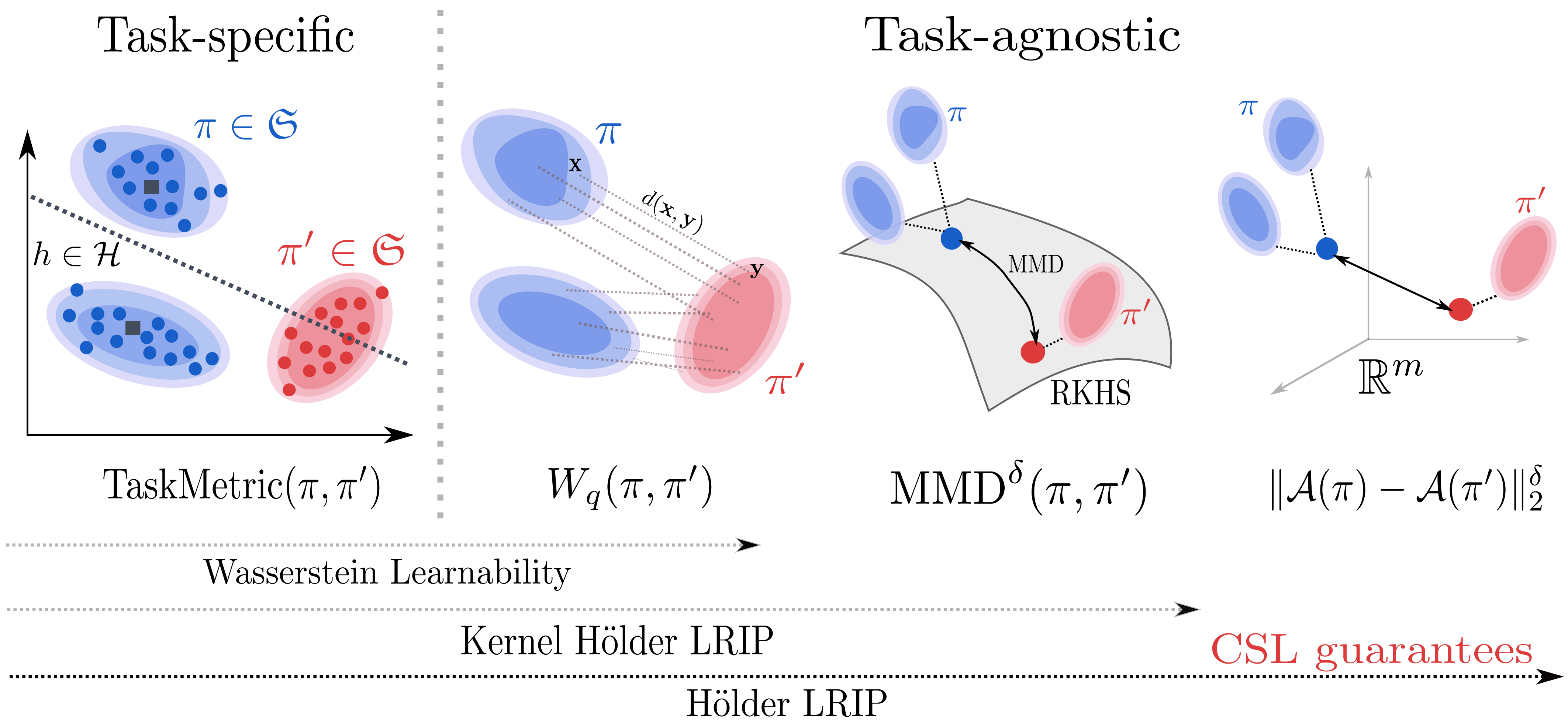
CSL guarantees

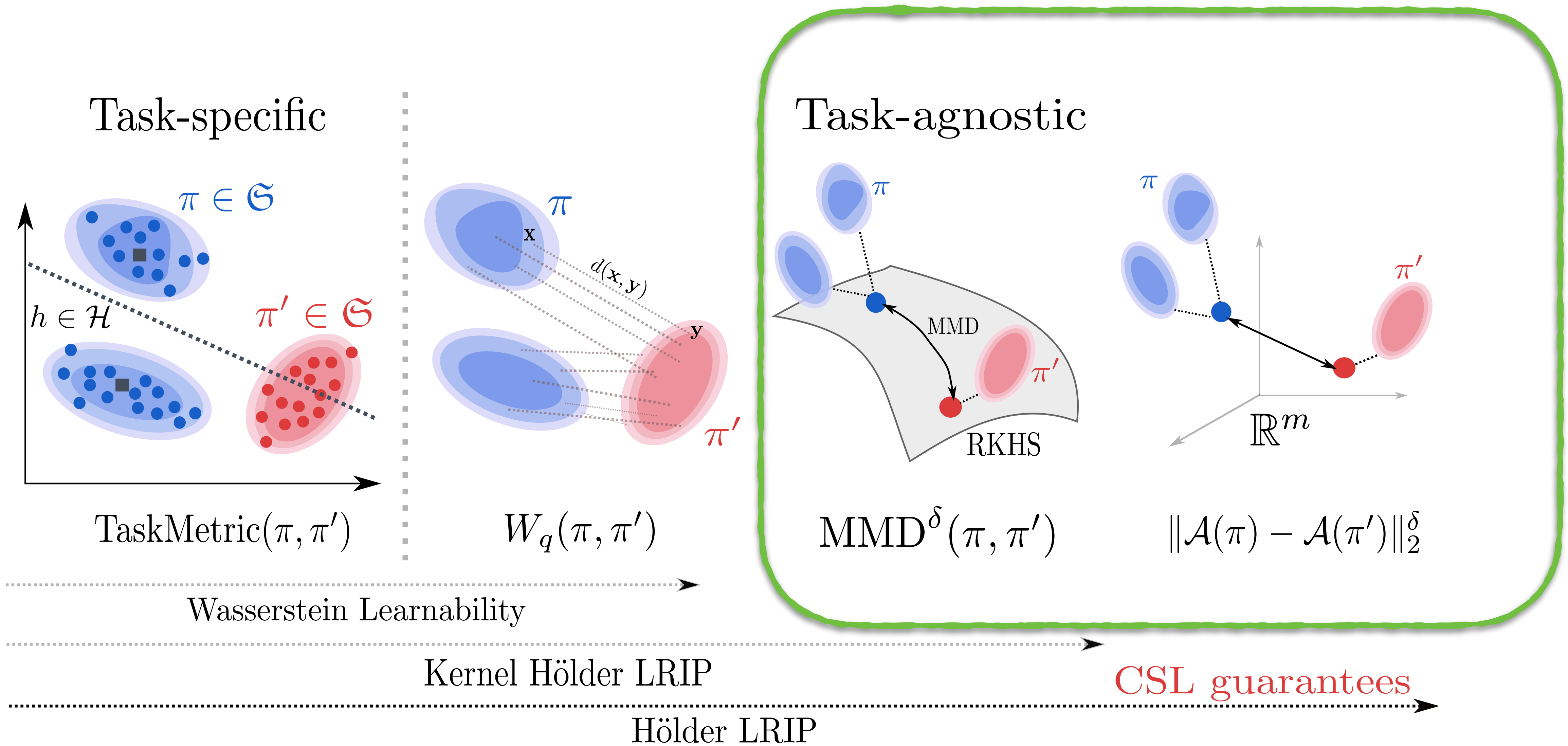
Hölder LRIP

[Vayer & Gribonval, 2023, *Controlling Wasserstein Distances by Kernel Norms with Application to Compressive Statistical Learning, JMLR*]






Task-specific
Task-agnostic
$\pi \in \mathfrak{S}$
$\pi' \in \mathfrak{S}$
$h \in \mathcal{H}$
π
x
y
$d(\mathbf{x}, \mathbf{y})$
π'
MMD
RKHS
$\mathbb{R}^m$
$\mathrm{TaskMetric}(\pi, \pi')$
$W_q(\pi, \pi')$
$\mathrm{MMD}^\delta(\pi, \pi')$
$\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^\delta$
Wasserstein Learnability
Kernel Hölder LRIP
CSL guarantees
Hölder LRIP


Task-specific
Task-agnostic
$\pi \in \mathfrak{S}$
$h \in \mathcal{H}$
$\pi' \in \mathfrak{S}$
π
x
$d(\mathbf{x}, \mathbf{y})$
y
π'
π
MMD
π'
RKHS
π
π'
$\mathbb{R}^m$
$\mathrm{TaskMetric}(\pi, \pi')$
$W_q(\pi, \pi')$
$\mathrm{MMD}^{\delta}(\pi, \pi')$
$\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^{\delta}$
Wasserstein Learnability
Kernel Hölder LRIP
CSL guarantees
Hölder LRIP

Key ideas to achieve sketches of small size :

JL lemma and Compressive Sensing

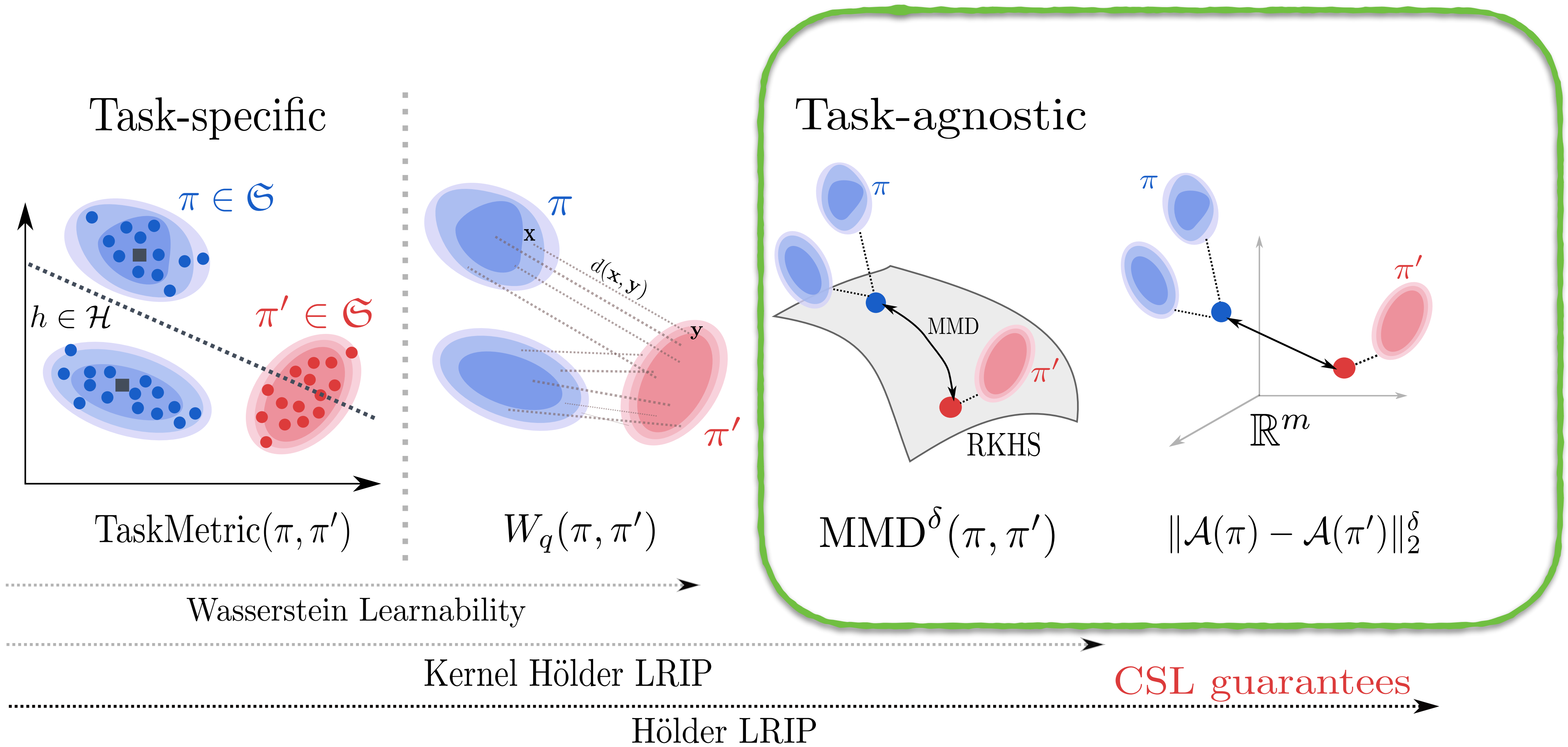
RIP for Mean Map Embedding (~ extensions of JL lemma)

Random kernel approximations [Rahimi & Recht 07, Bach 15]

Covering dim of « secant set », cf also monograph [Robinson 2010]

[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*]

for k-mixtures in $\mathbb{R}^d$ $m = \mathcal{O}(k^2 d)$



Key ideas to achieve sketches of small size :

JL lemma and Compressive Sensing

RIP for Mean Map Embedding (~ extensions of JL lemma)

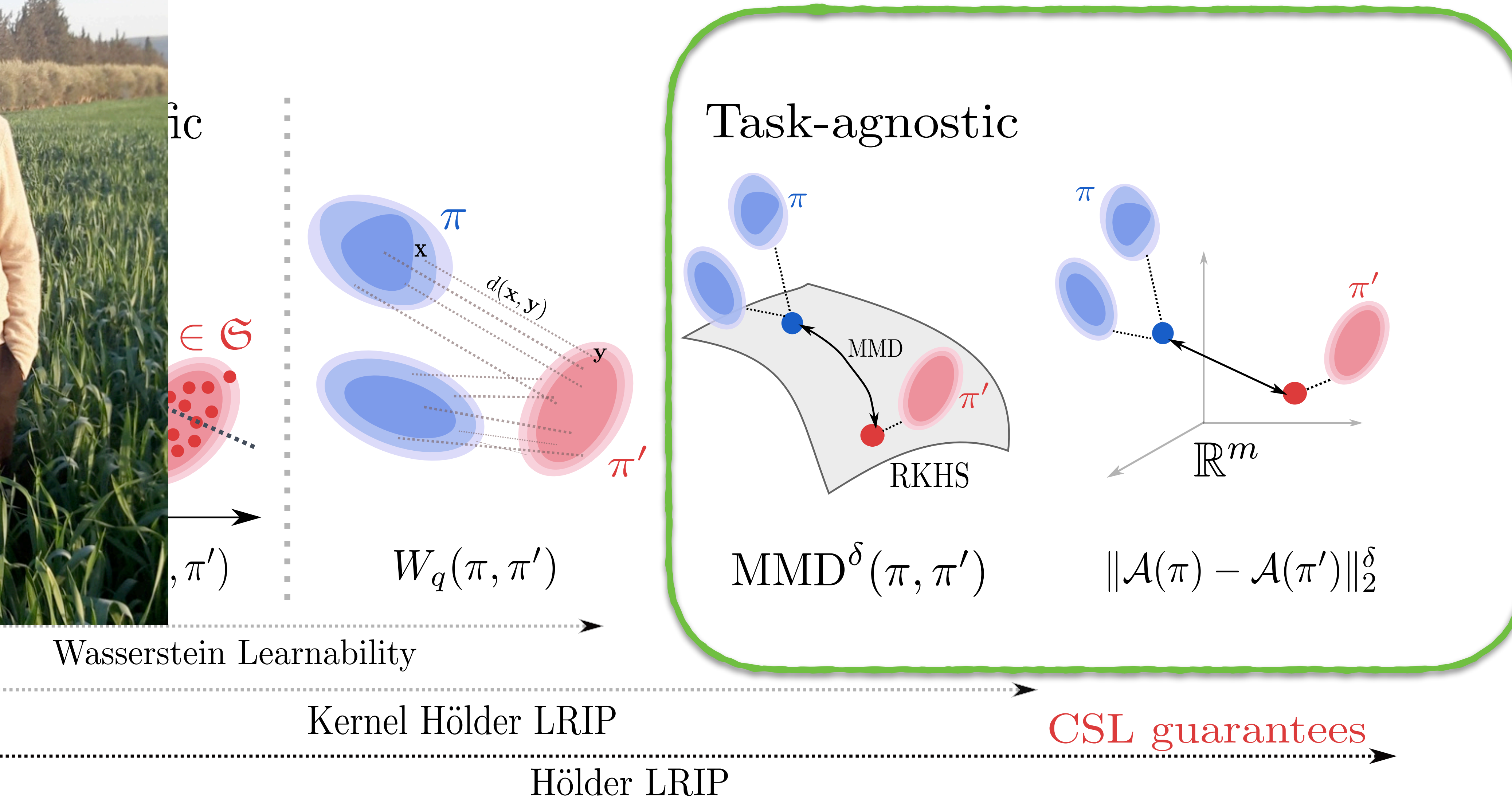
Random kernel approximations [Rahimi & Recht 07, Bach 15]

Covering dim of « secant set », cf also monograph [Robinson 2010]

[Gribonval, Blanchard, Keriven & Traonmilin 2021, in *Mathematics of Statistics and Learning*] for k-mixtures in $\mathbb{R}^d$ $m = \mathcal{O}(k^2 d)$

Belhadji & G.,*Revisiting RIP Guarantees for Sketching Operators on Mixture Models*, JMLR 2024 $m = \mathcal{O}(kd)$

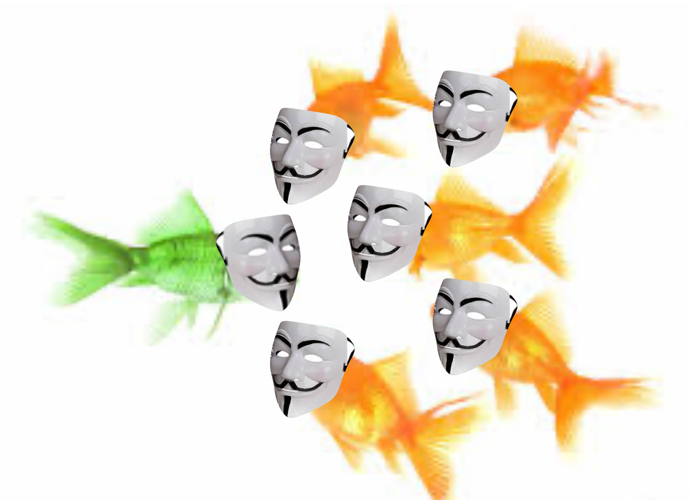
+[Belhadji & G., Sketch and shift: a robust decoder for compressive clustering, TMLR 2024]
also reduces *empirically* needed sketch size via better learning from sketch



Learning with *limited memory*

- Memory = limited resource
  - Compressive Learning:
    - Goal = handle large-scale collections
    - « **enough information** » for learning should be captured
- Privacy = desirable target
  - Differential privacy
    - Goal = learn without memorizing individual information
    - « **no more information than needed** » should be captured

Learning with *limited memory*



- Memory = limited resource
  - Compressive Learning:
    - Goal = handle large-scale collections
    - « **enough information** » for learning should be captured
- Privacy = desirable target
  - Differential privacy
    - Goal = learn without memorizing individual information
    - « **no more information than needed** » should be captured

Private sketched learning ?

- **"Natural" privacy of an aggregated estimator:**

$$\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)$$

  - role of sketch size
    - sufficiently large for "task-level" information-preservation
    - *sufficiently small for "sample-level" information loss?*

- **Guaranteed differential privacy ?**
  - randomized sketching function ?
    - noise on training samples $\Psi(x_i) = \Phi(x_i + \xi_i)$
    - noise on random features $\Psi(x_i) = \Phi(x_i) + \xi_i$
    - partial random features $\Psi(x_i) = \mathtt{diag}(d_i) \cdot \Phi(x_i)$
    - combinations of the above … $\|\mathbf{d}_i\|_0 = \alpha m, \alpha < 1$

Private sketched learning ?

- **“Natural” privacy of an aggregated estimator:**

$$\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i)$$

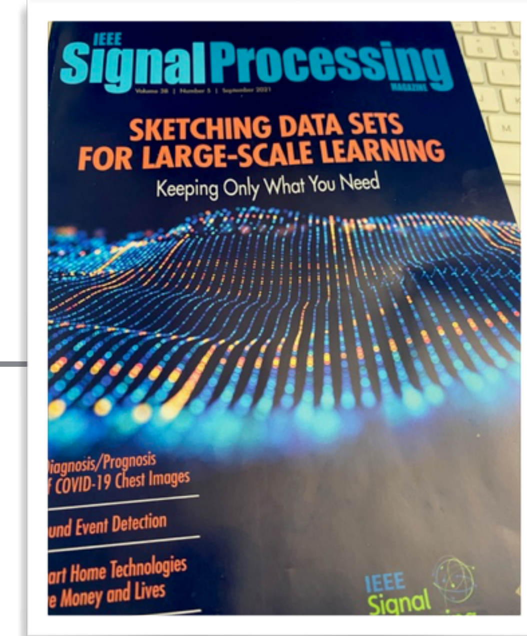
  - role of sketch size
    - sufficiently large for “task-level” information-preservation
    - *sufficiently small for “sample-level” information loss?*

- **Guaranteed differential privacy ?**
  - randomized sketching function ?
    - noise on training samples $\Psi(x_i) = \Phi(x_i + \xi_i)$
    - noise on random features ✓ $\Psi(x_i) = \Phi(x_i) + \xi_i$
    - partial random features ✓ $\Psi(x_i) = \mathtt{diag}(d_i) \cdot \Phi(x_i)$, $\|\mathbf{d}_i\|_0 = \alpha m, \alpha < 1$
    - combinations of the above …

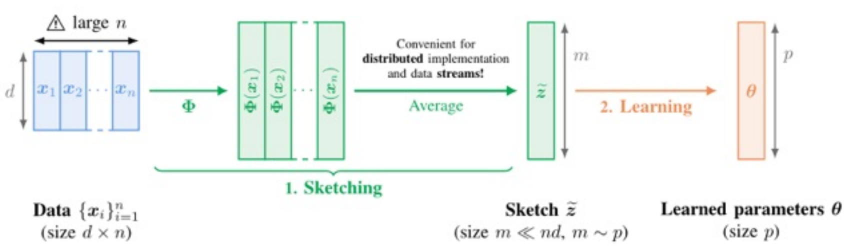
✓ *[Schellekens et al , Differentially Private Compressive k-Means,* ICASSP *2019]*
✓ *[Chatalic et al, Compressive Learning with Privacy Guarantees,* Information and Inference, *2021]*

- Principle of compressive learning
- Analogy with compressive sensing
- Inequalities between metrics on probability distribution
- Private sketching
- Take Home Messages

Summary



✓ Sketching framework



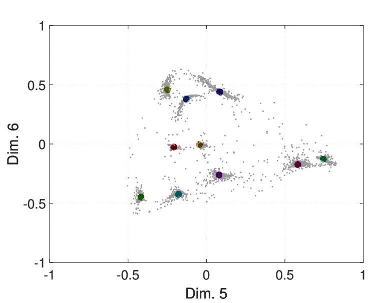
✓ Privacy guarantees
✓ Statistical guarantees

➡compressive PCA, k-means, GMM

➡***key links with kernels and optimal transport***

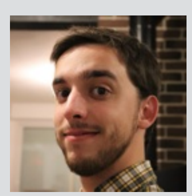
✓ Dimension reduction

✓ Empirical success



✤ **Open challenges:**

- **beyond *unsupervised*** sketched learning ?
  e.g.: classification, sparse matrix factorization



[with E. Lasalle, T. Vayer & P. Gonçalves
Compressive Recovery of Sparse Precision Matrices, preprint 2023]

- ***guaranteed* algorithms** to learn from a sketch ?
  e.g.: continuous OMP / sliding Frank-Wolfe

What's next ?

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030 PROGRAMME DE RECHERCHE INTELLIGENCE ARTIFICIELLE

SHARP

Sharp Theoretical and Algorithmic Principles for frugal ML

Rémi Gribonval, LIP, Inria & ENS de Lyon, coordinator

https://project.inria.fr/sharp/

anr® agence nationale de la recherche

"Frugal learning" : an oxymoron ?

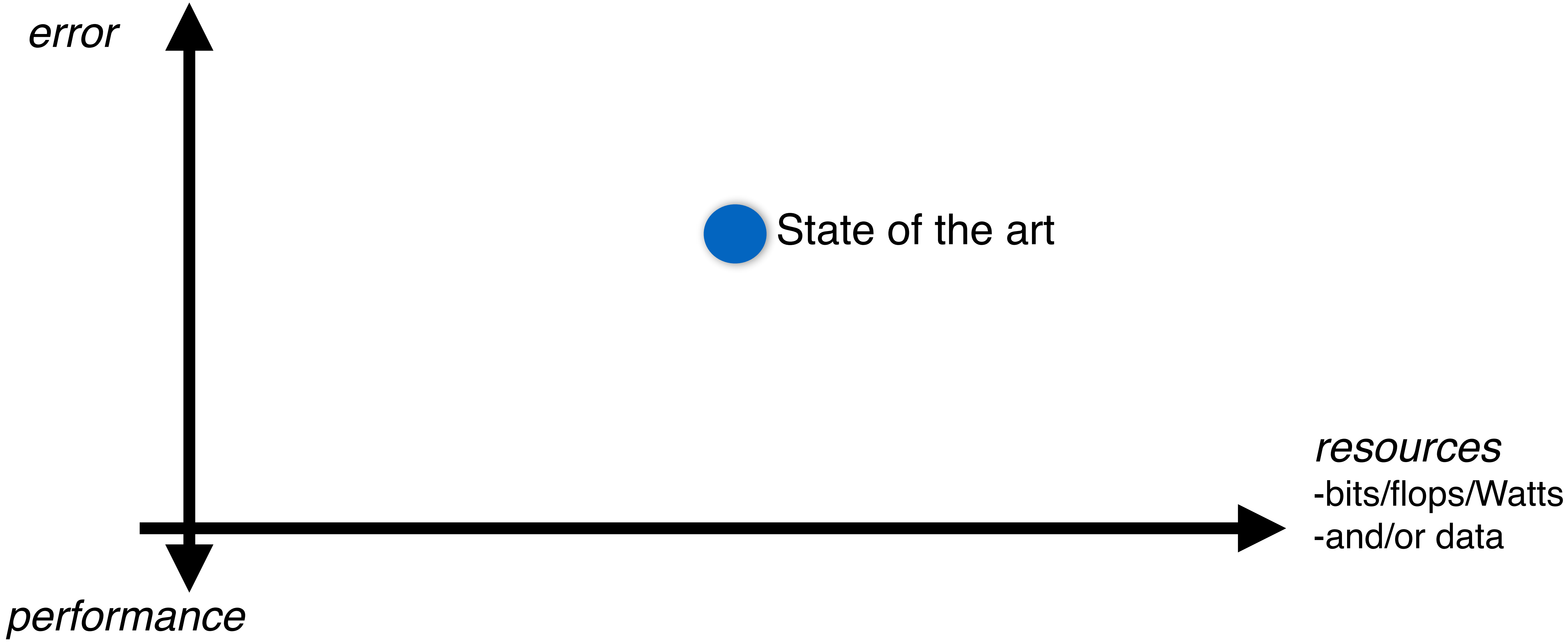
ML = evermore ?

- Economic competition
- Maximizing performance
- Race towards giant models
- Unrestrained consumption

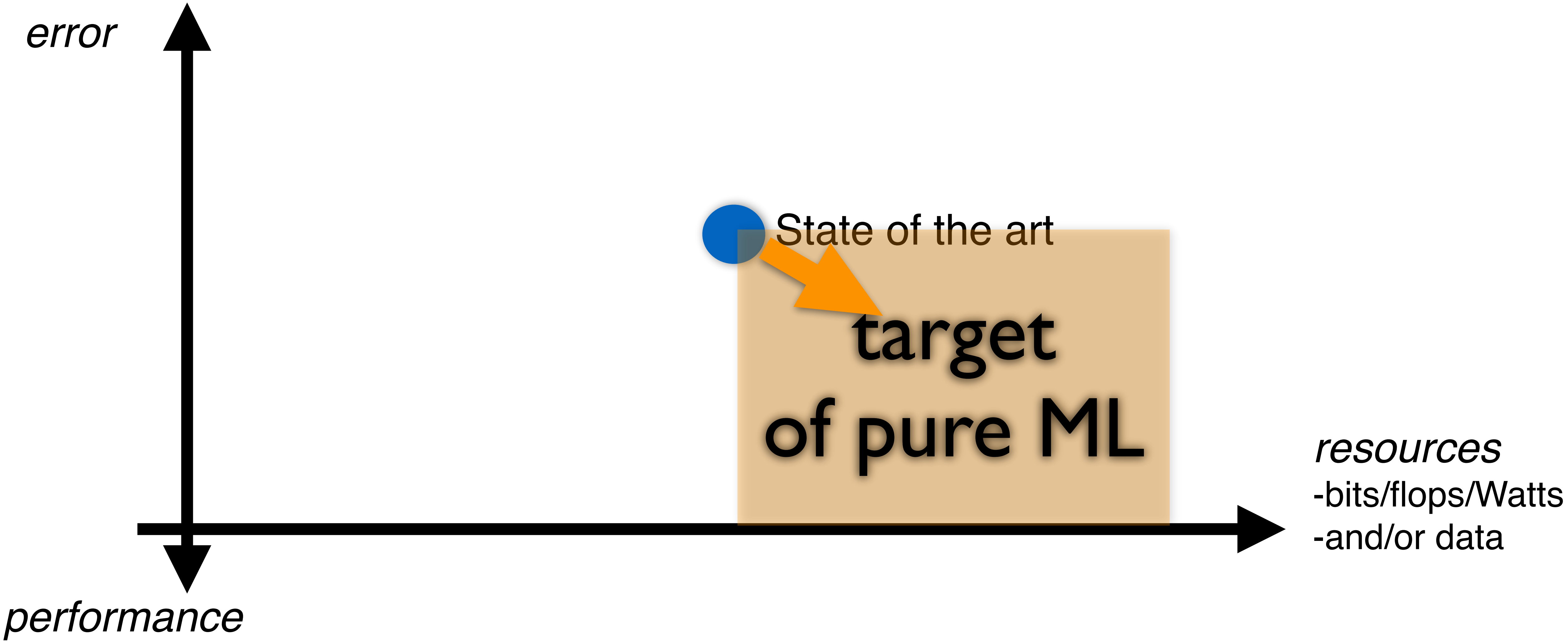
Frugality = sobriety ?

where to set the cursor ?

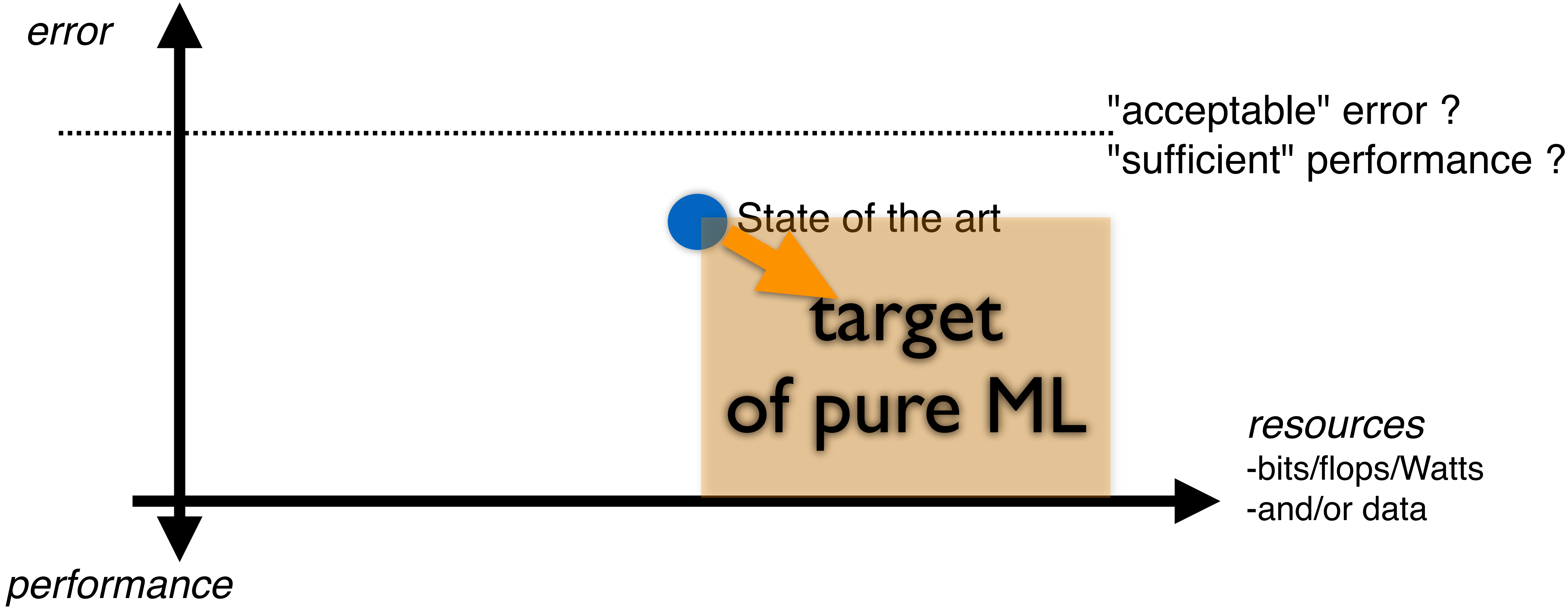
Frugality: what are we talking about ?



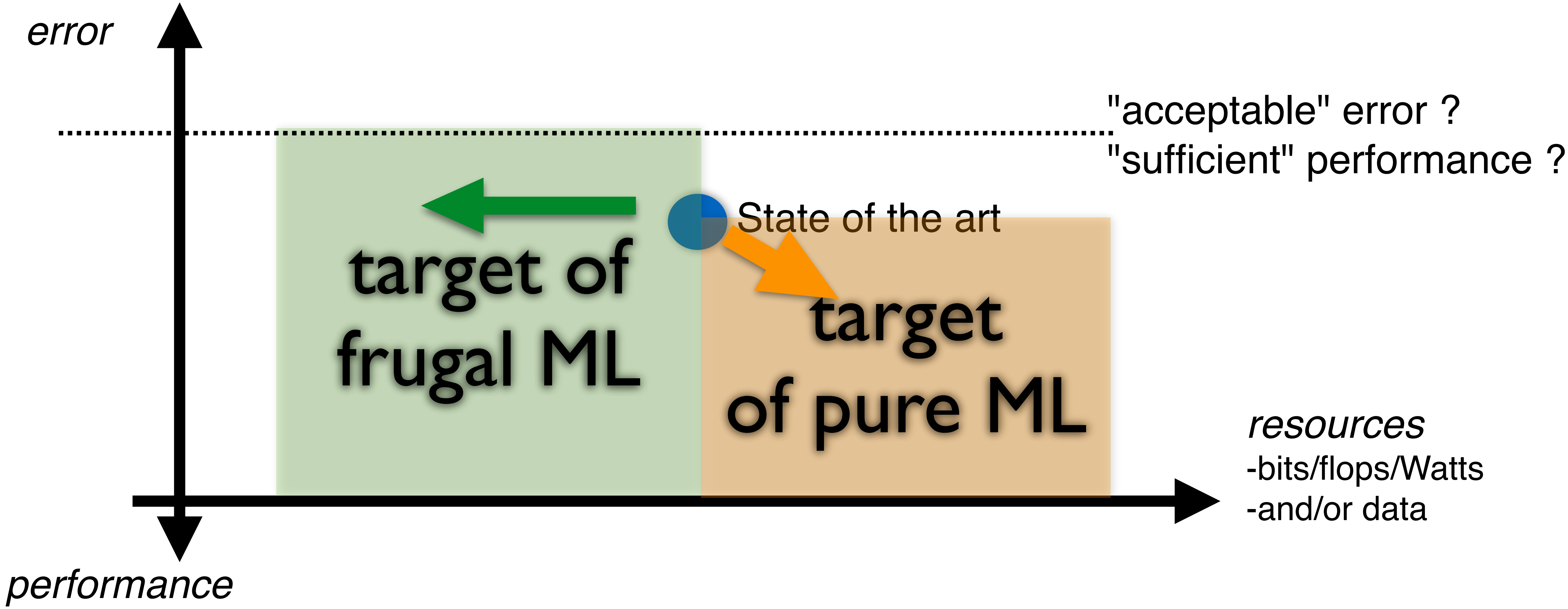
Frugality: what are we talking about ?



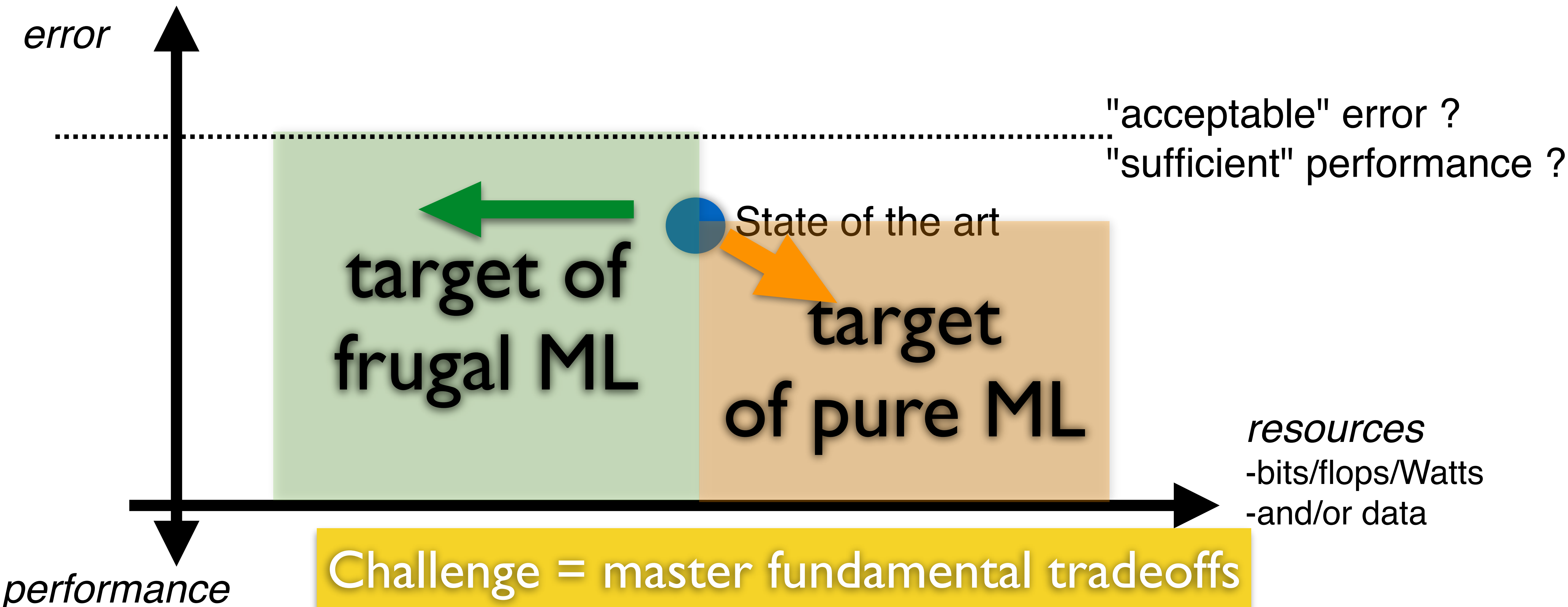
Frugality: what are we talking about ?



Frugality: what are we talking about ?



Frugality: what are we talking about ?



Challenge = master fundamental tradeoffs resources / performance (Pareto curve)

What's next ?

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Axis 1

Frugal
Architectures

Axis 2

Frugal
Principles

Axis 3

"Small" and "Raw"
Data

FRANCE 2030
PROGRAMME DE RECHERCHE
INTELLIGENCE ARTIFICIELLE

SHARP

Sharp Theoretical and Algorithmic Principles for frugal ML

Rémi Gribonval, LIP, Inria & ENS de Lyon, coordinator

https://project.inria.fr/sharp/

anr® agence nationale de la recherche