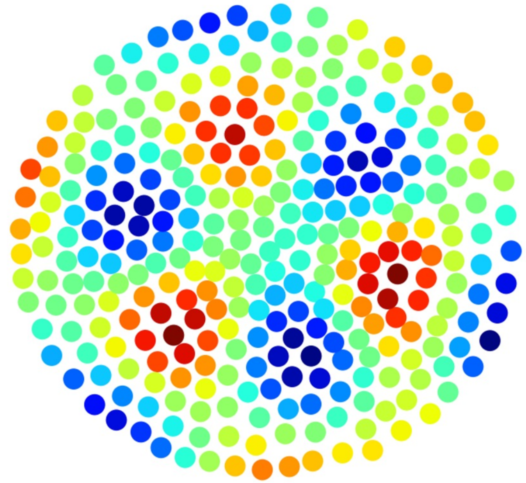
Poster: **#163 (last row)**
16:30 CAT – 18:30 CAT



Minimalistic Unsupervised Representation Learning with the Sparse Manifold Transform

Yubei Chen, Zeyu Yun, Yi Ma, Bruno Olshausen, Yann LeCun

Meta AI
NYU Center for Data Science
REDWOOD CENTER for Theoretical Neuroscience
BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

What is an unsupervised representation?

A general goal: transform raw data into a new space such that "similar" things are placed closer and meanwhile the new space is not collapsed.

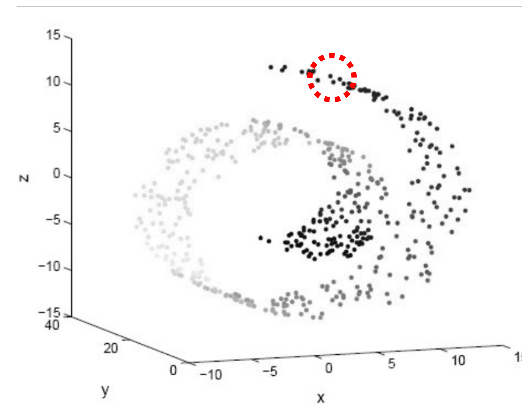
Where does “similarity” come from?

Three similarities:

Spatial co-occurrence

Temporal co-occurrence

Euclidean neighborhoods



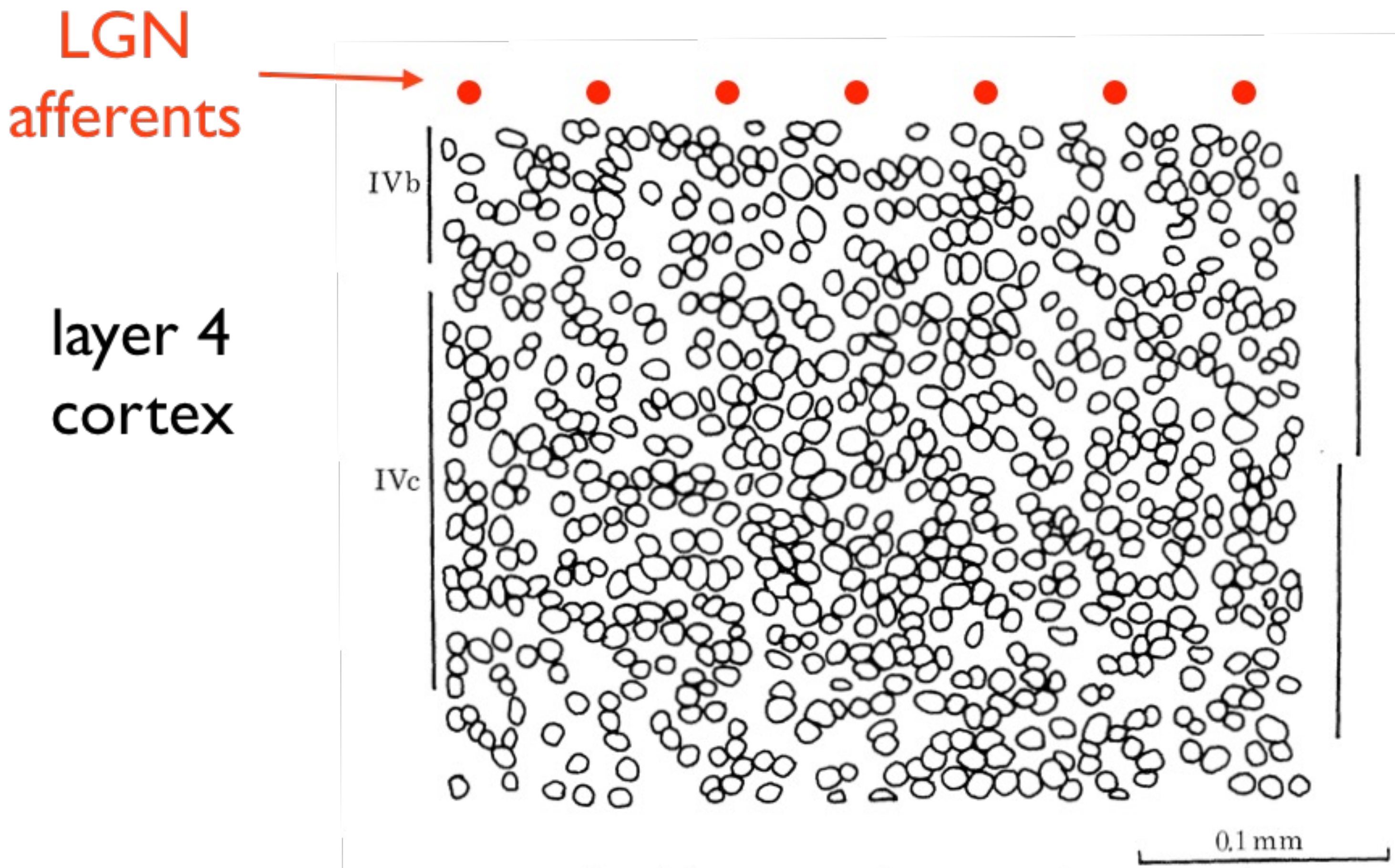
I have never met a **person** who is not interested in language

(Rumelhart, Hinton and Williams, 1986)
(Roweis and Lawrence, 2000)
(Tenenbaum, Silva and Langford, 2000)
(Wiskott and Sejnowski, 2002)
(Dumais, 2004)
(Mikolov et al., 2012)
⋮

How to establish a minimalistic unsupervised representation?

An unsupervised representation transform derived from neural and statistical principles

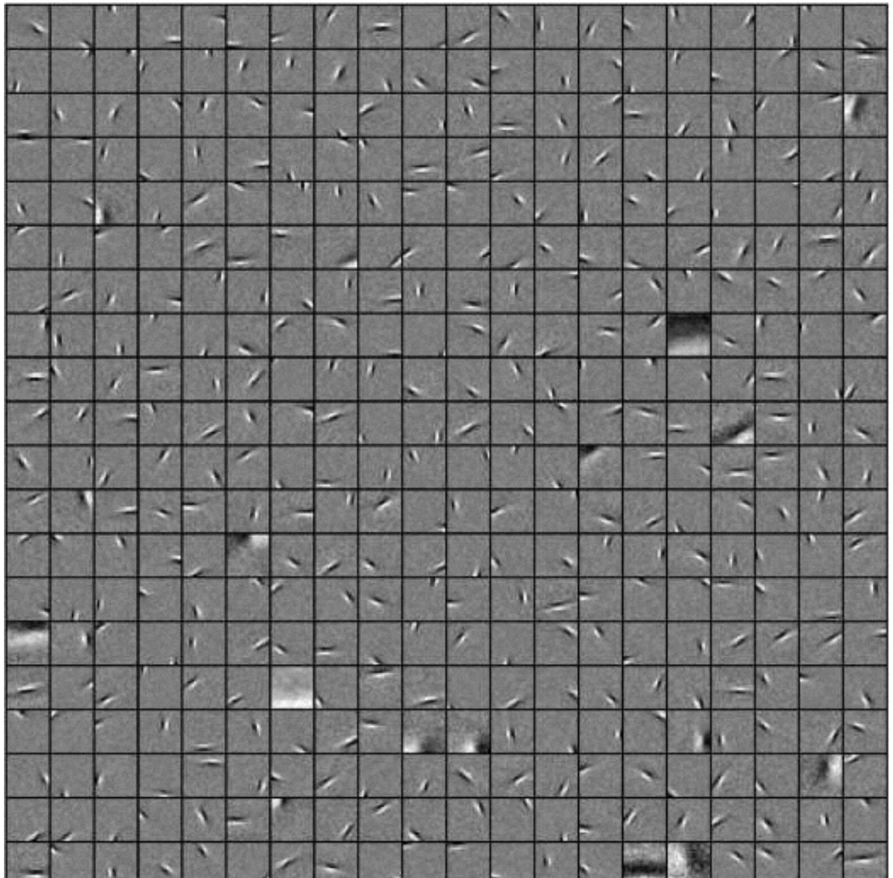
Neural principle: sparse coding



V1 is highly over-complete

(Barlow, 1981)

Neural principle: sparse coding



Learned features

Sparse coding image model

$$\vec{x} = \Phi \vec{\alpha} + \vec{n}$$

image

features

neural activities (sparse)

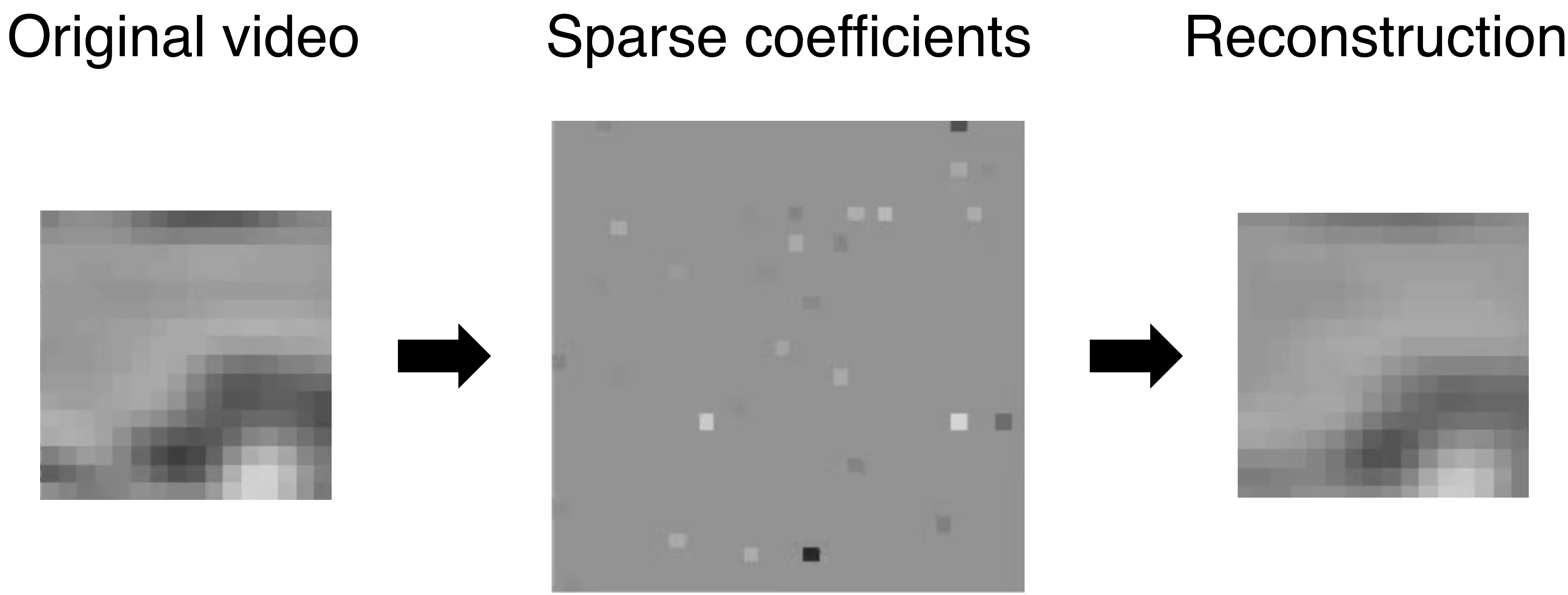
other stuff (noise)

Optimization:

$$\min_{\Phi,\ \vec{\alpha}_i,\ i \in 1,\ \cdots,\ N} \quad \frac{1}{N} \sum_{i=1}^{N} \|\vec{x}_i - \Phi\vec{\alpha}_i\|_2^2 + S(\vec{\alpha}_i)$$

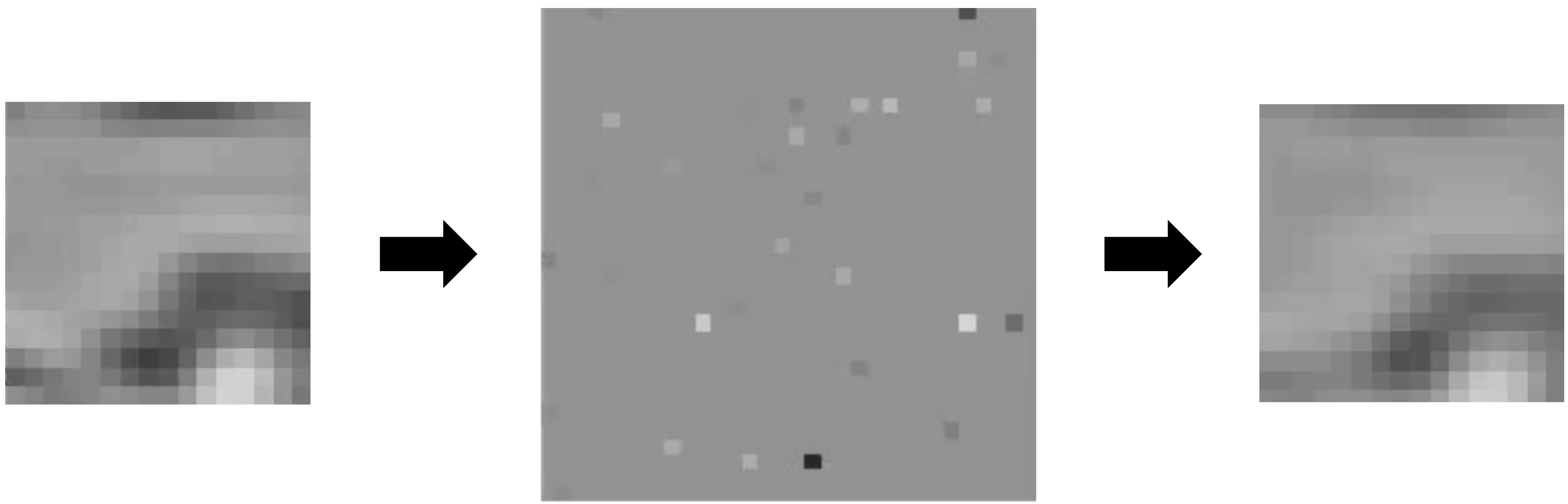
(Olshausen and Field, 1996)

Sparse encoding of a time-varying image



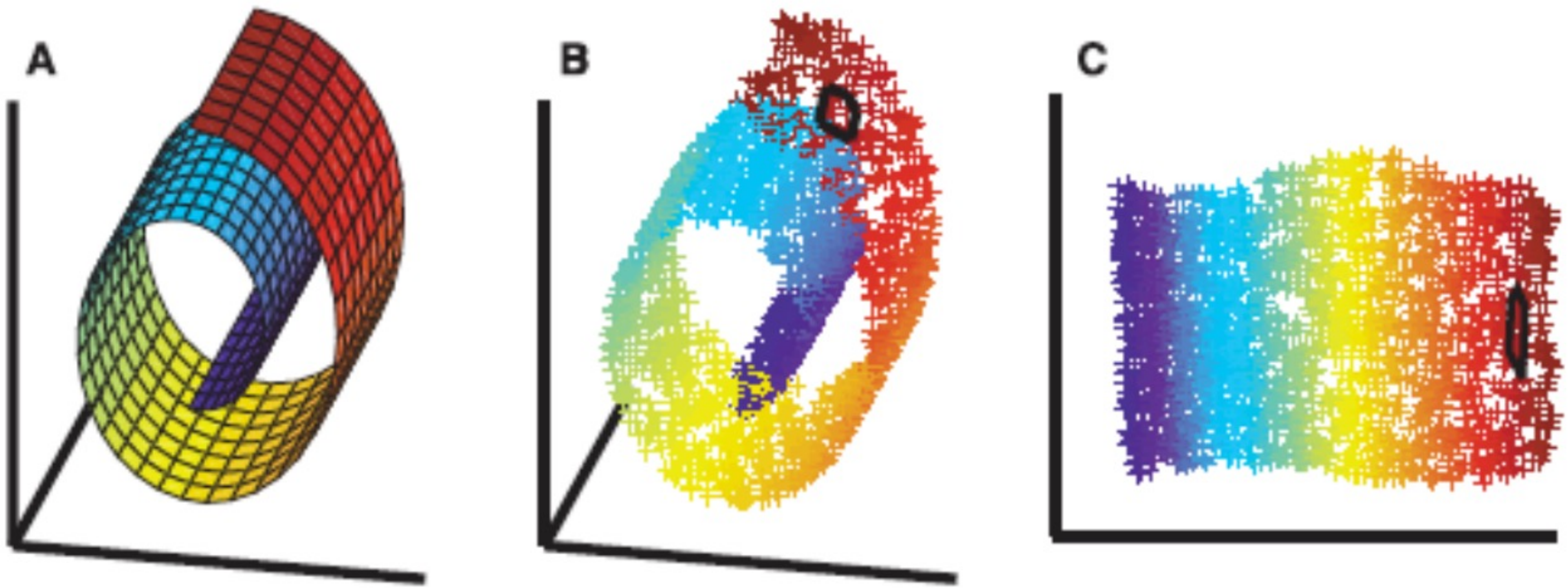
Sparse encoding of a time-varying image

Original video Sparse coefficients Reconstruction



But something is missing here: “similar” things are not placed closer

Statistical principle: manifold learning

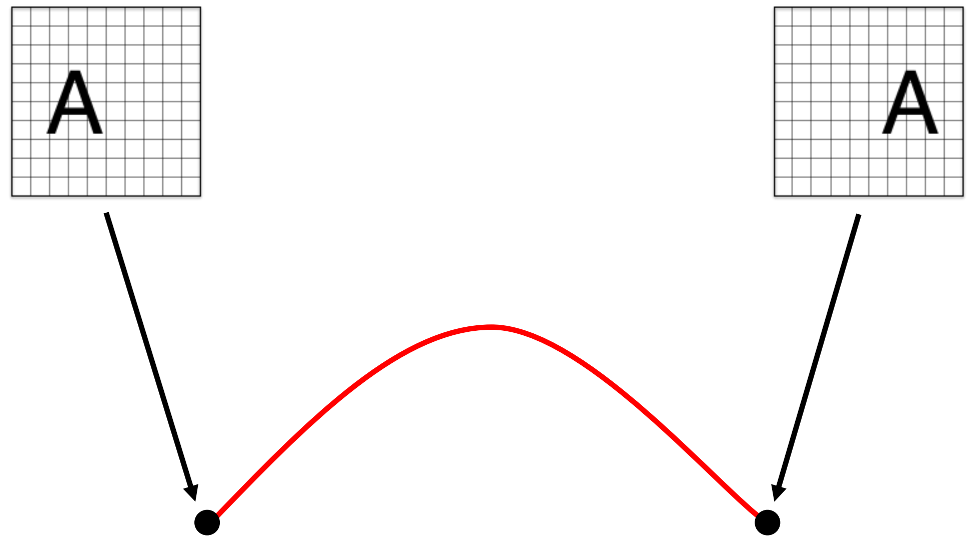


Manifold hypothesis

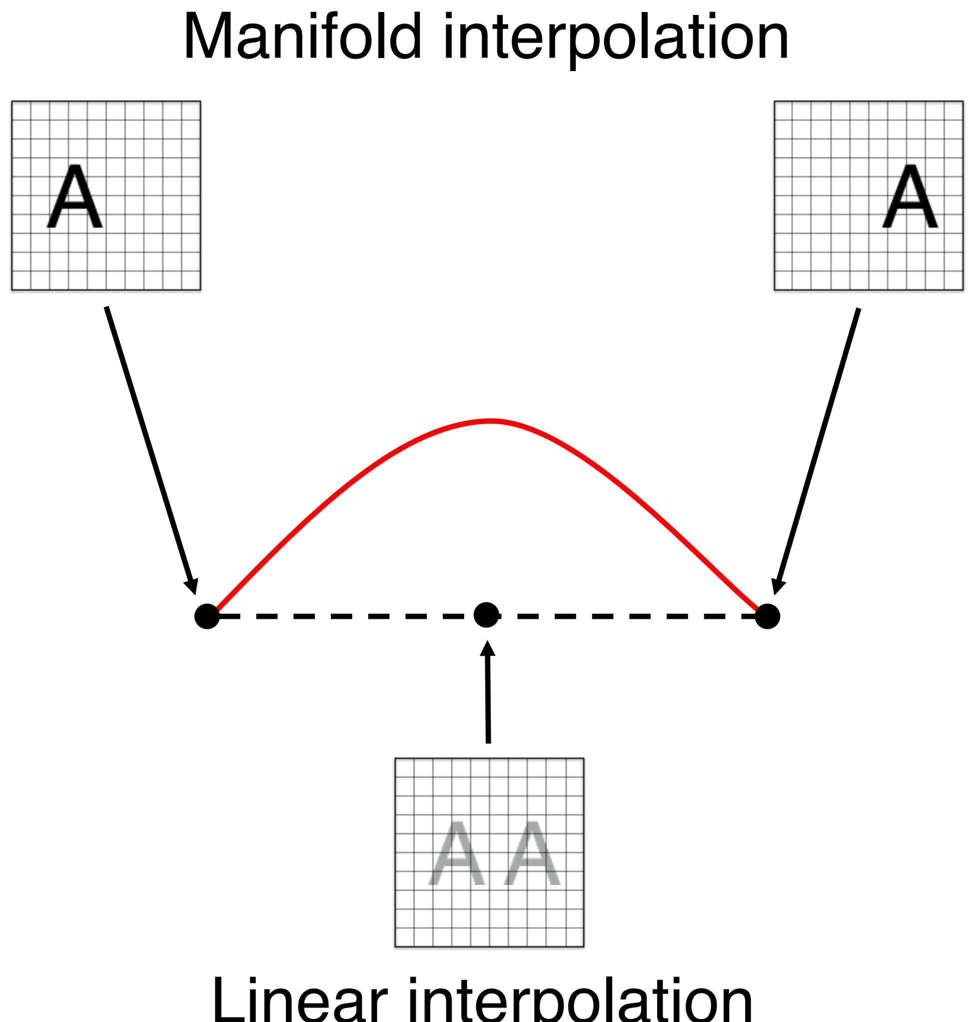
(Roweis and Saul, 2000)
(Tenenbaum, Silva and Langford, 2000)
(Saul and Roweis, 2002)

Natural signals can be highly nonlinear

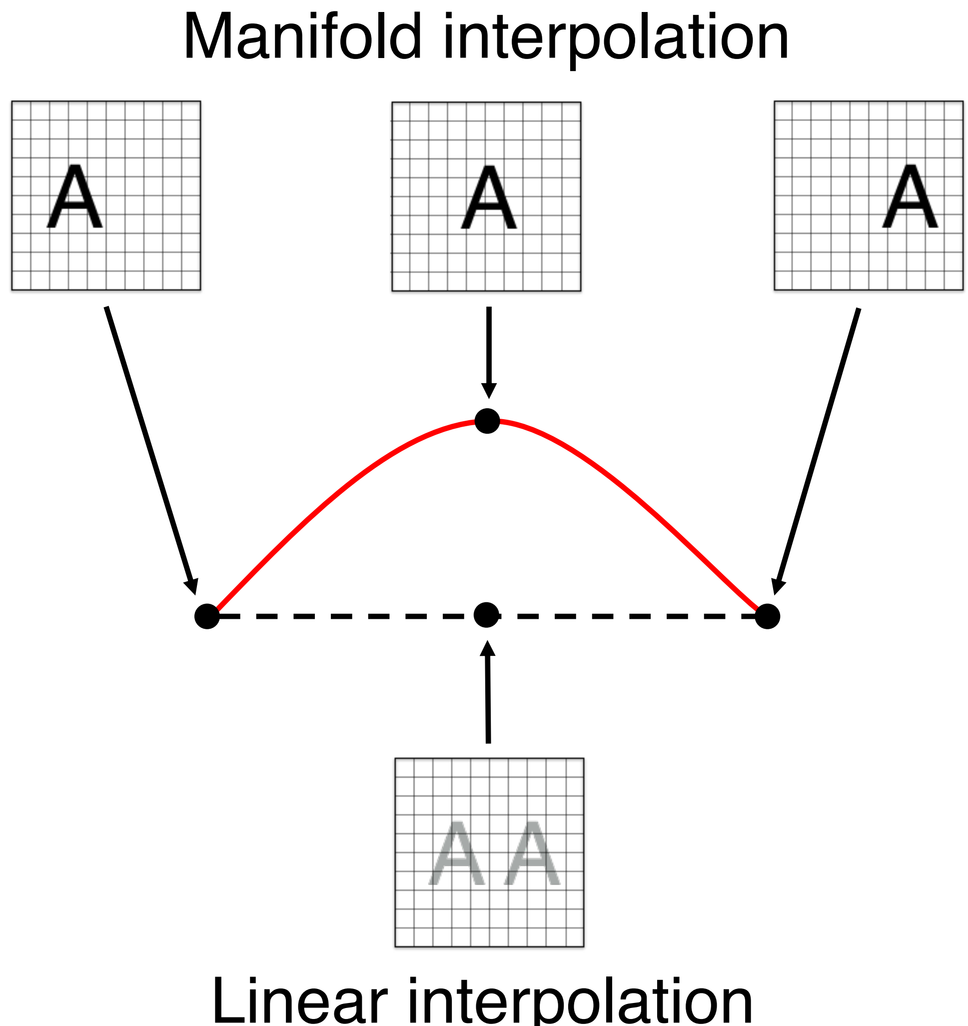
Manifold interpolation



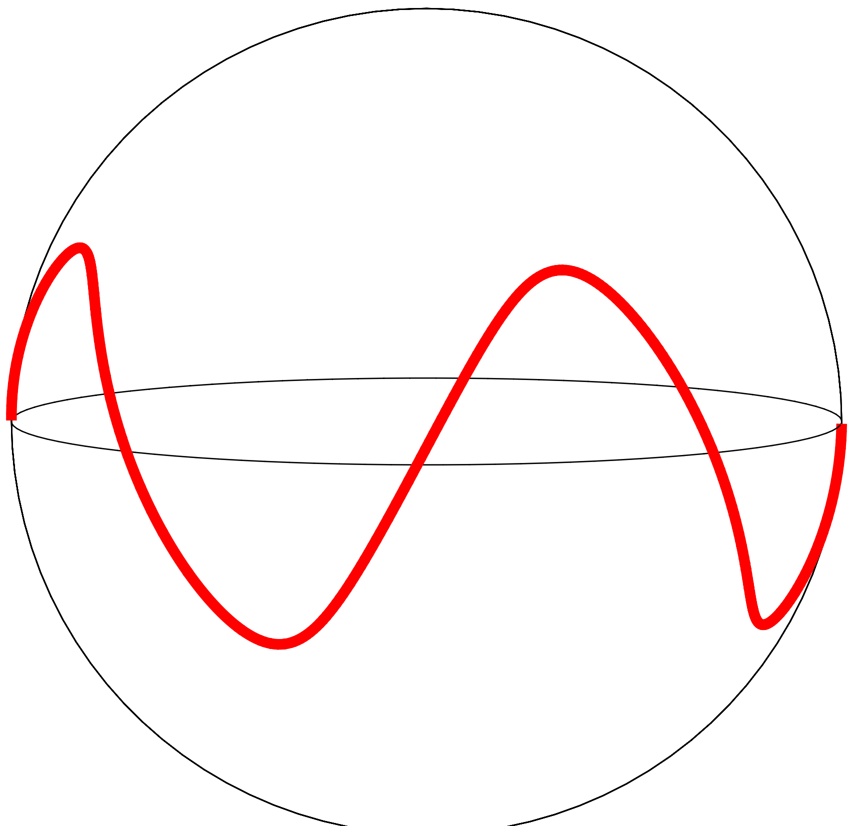
Natural signals can be highly nonlinear



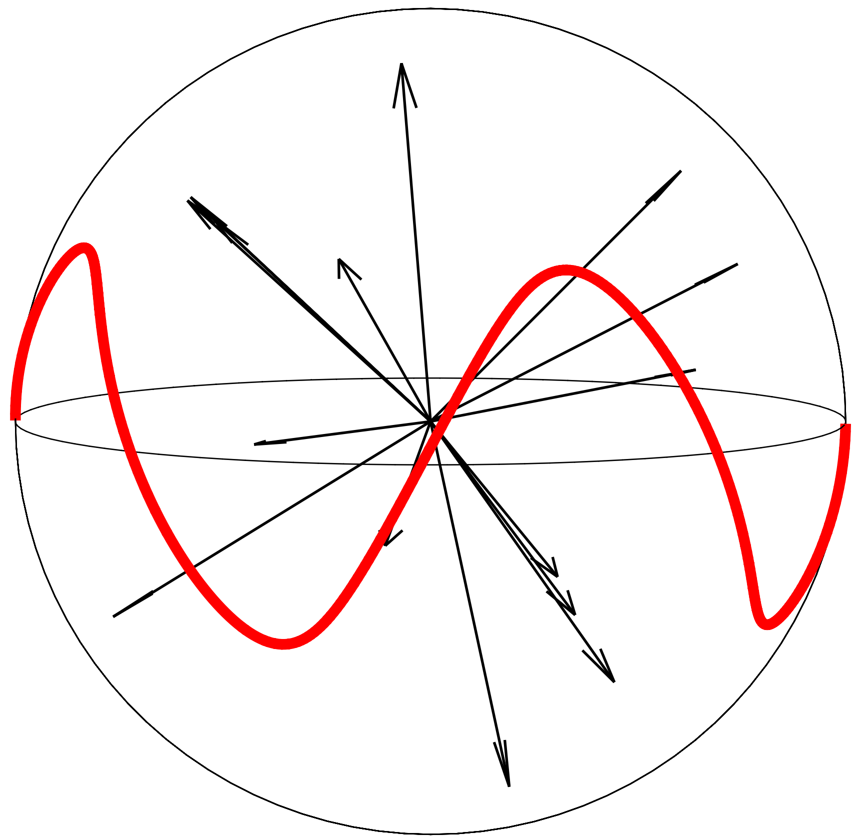
Natural signals can be highly nonlinear



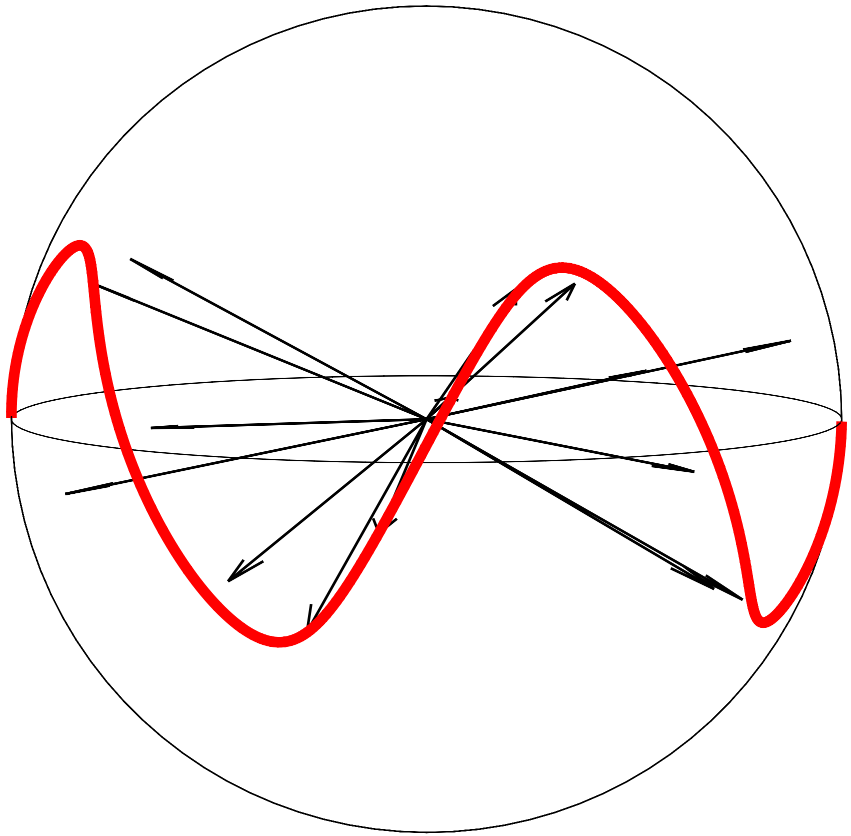
Dictionary elements learned by sparse coding form a locally linear approximation to the manifold



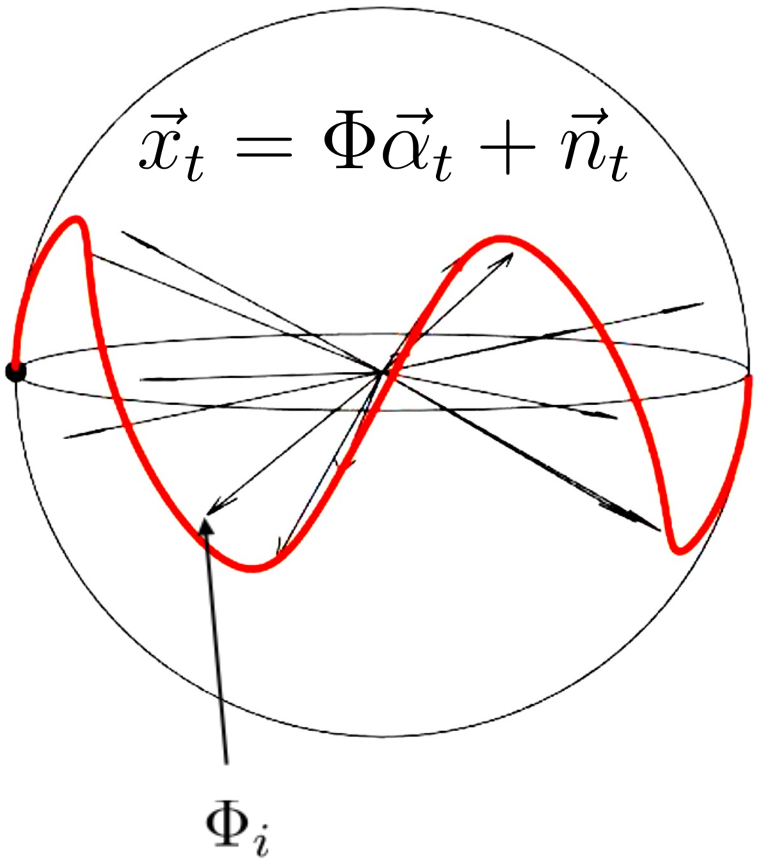
Dictionary elements learned by sparse coding form a locally linear approximation to the manifold

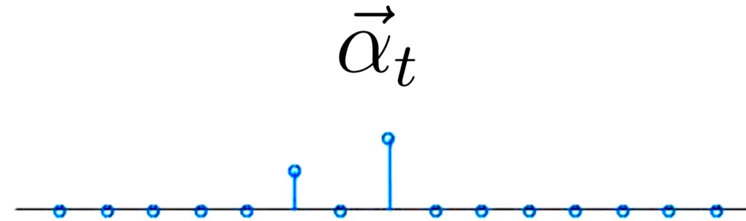


Dictionary elements learned by sparse coding form a locally linear approximation to the manifold

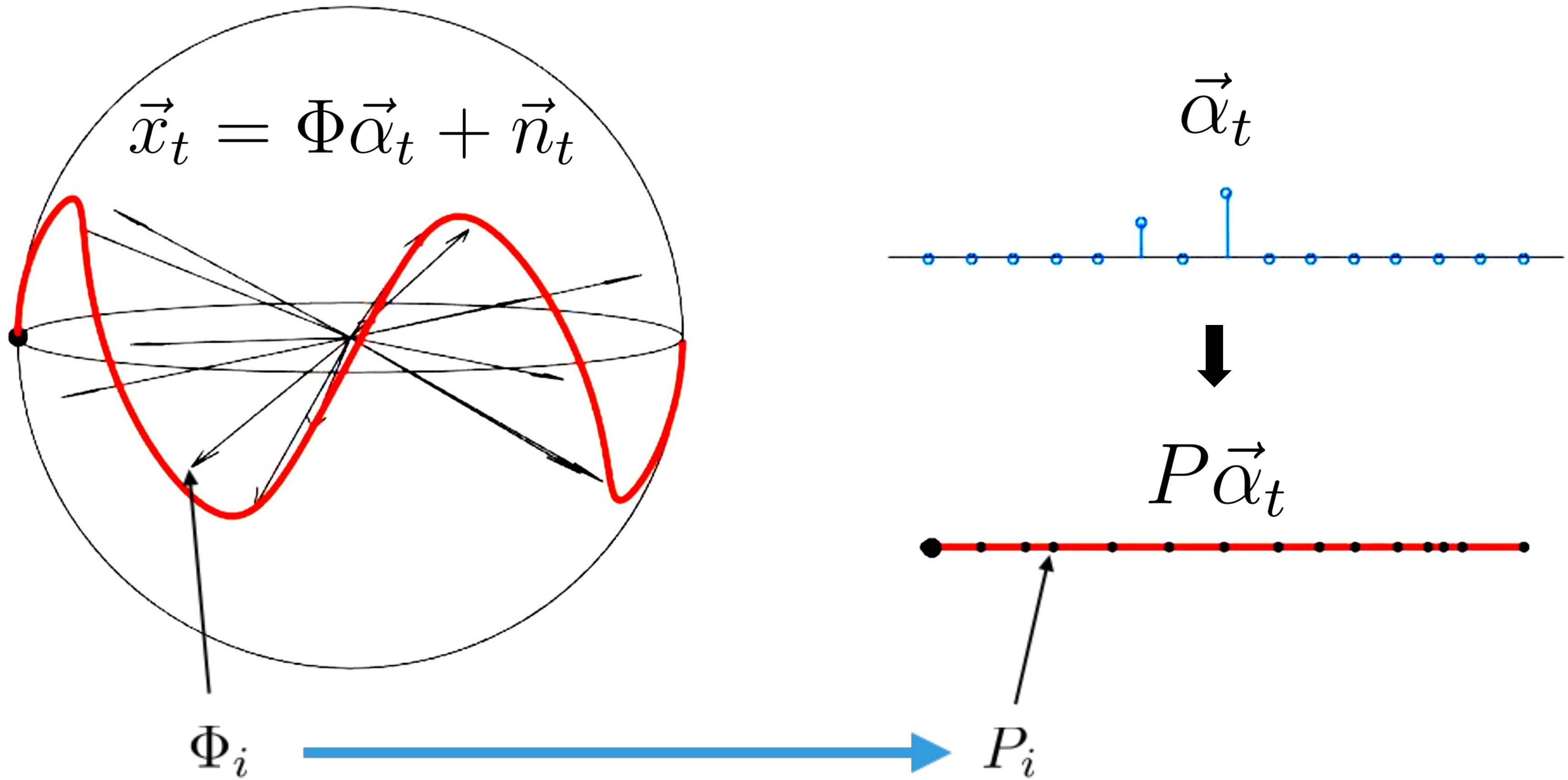


Dictionary elements learned by sparse coding form a locally linear approximation to the manifold

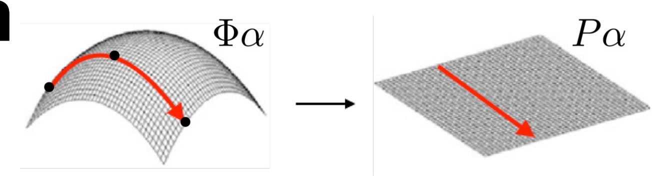




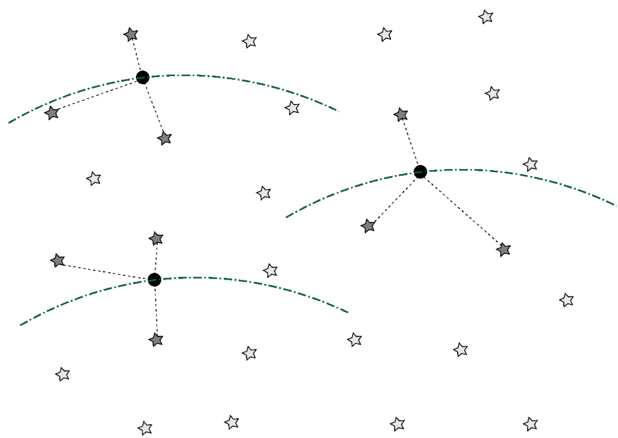
Dictionary elements learned by sparse coding form a locally linear approximation to the manifold

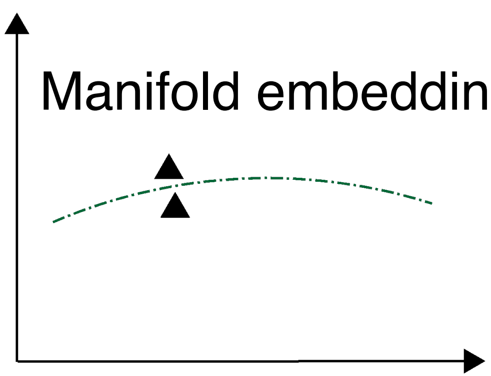


SMT formulation



Sparse coefficients





$$\vec{x}_t = \Phi\vec{\alpha}_t + \vec{n}_t$$

$$P\vec{\alpha}_t = \frac{1}{2}P(\vec{\alpha}_{t-1} + \vec{\alpha}_{t+1})$$

$$\ddot{\vec{\alpha}}_t = \vec{\alpha}_t - \frac{1}{2}(\vec{\alpha}_{t-1} + \vec{\alpha}_{t+1})$$

Optimization:

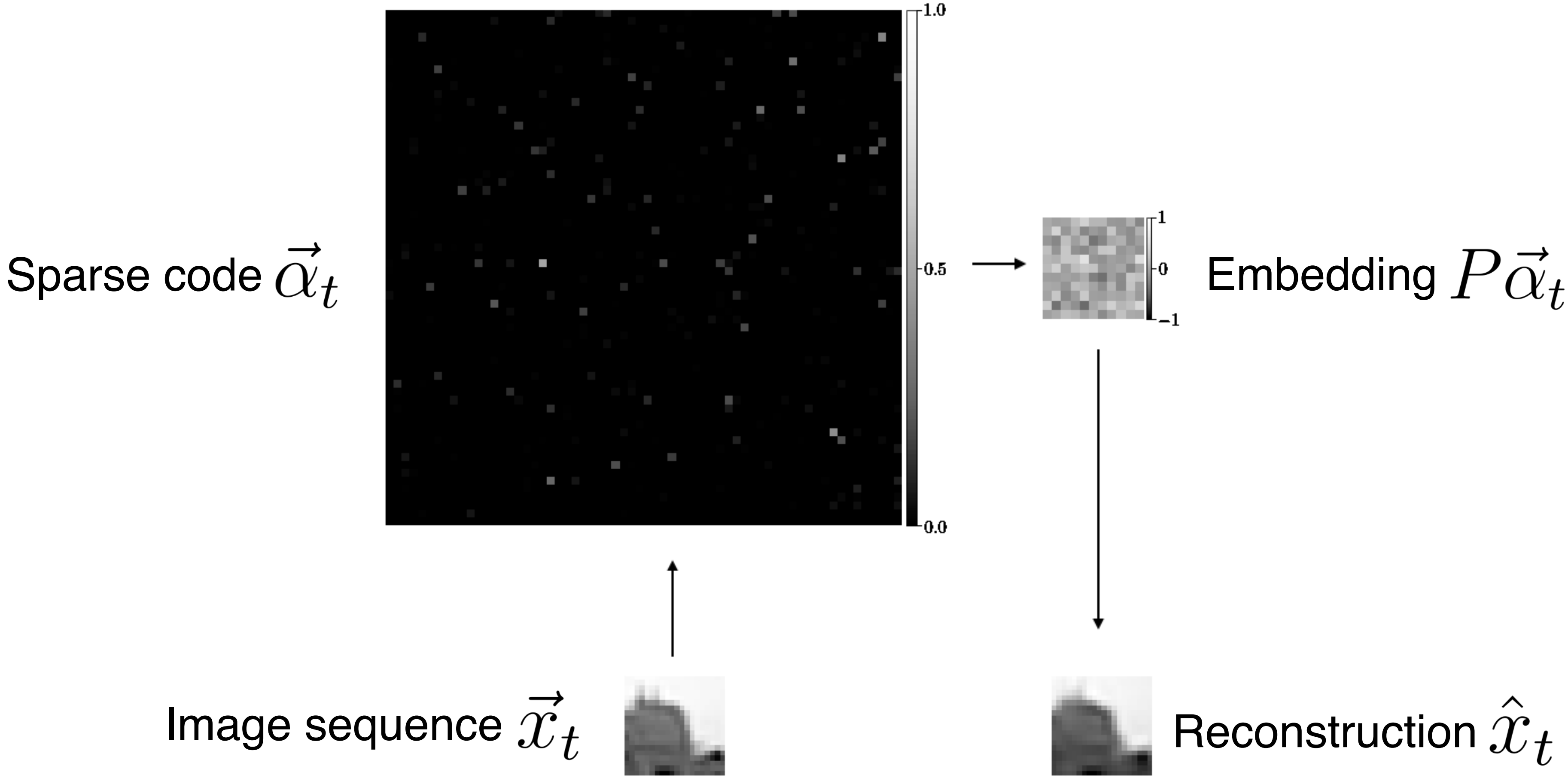
$$\min_{P} \quad tr\ P\sum_{t=1}^{T}(\ddot{\vec{\alpha}}_t\ddot{\vec{\alpha}}_t^{\,T})P^T$$

$$\text{s.t.} \quad PVP^T = I,\ (unit\ variance\ \&\ decorrelation)$$

$$V\ is\ covariance\ matrix\ of\ \vec{\alpha}$$

(**Chen** et al. NeurIPS'18)

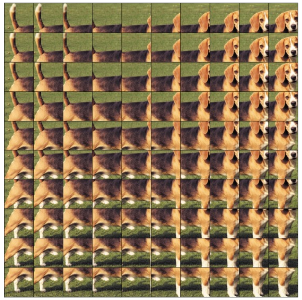
Encoding of a natural video sequence



(**Chen** et al. NeurIPS’18)

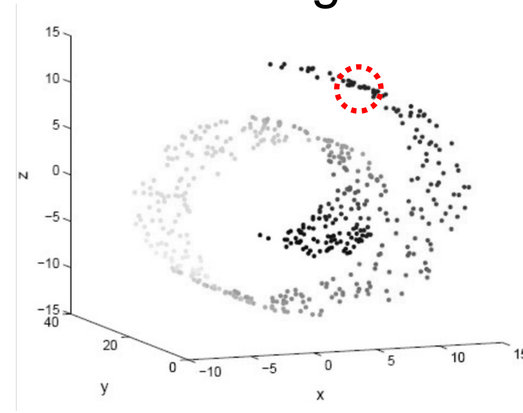
Let's generalize the formulation a little bit

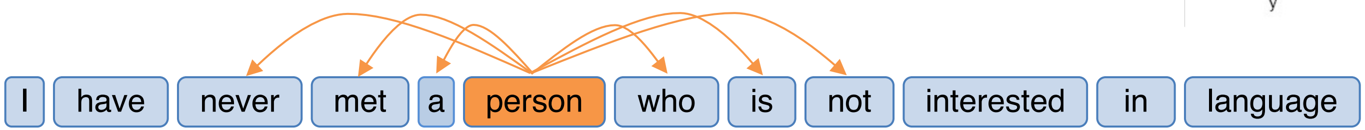
Spatial co-occurrence



Temporal co-occurrence

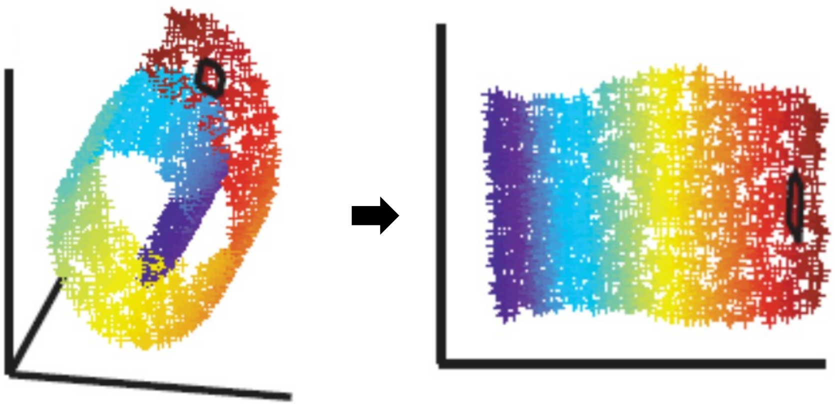
Euclidean neighborhoods



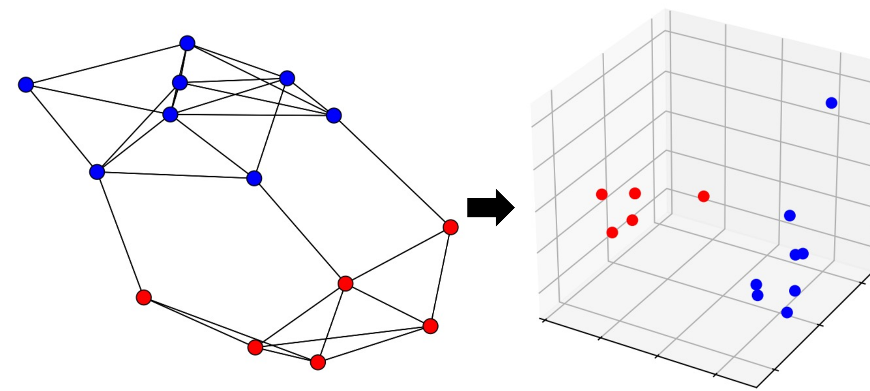


$P\vec{\alpha}_t = \frac{1}{2}P(\vec{\alpha}_{t-1} + \vec{\alpha}_{t+1})$	Temporal linearity (second-order derivative)
$P\vec{\alpha}_i - \sum_{j\in n(i)} w(i,j)P\vec{\alpha}_j = 0$	General linearity (second-order derivative)
$P\vec{\alpha}_i - P\vec{\alpha}_j = 0$	Similarity (first-order derivative)

Let's generalize the formulation a little bit



Manifold embedding



Graph embedding

$$P\vec{\alpha}_t = \frac{1}{2}P(\vec{\alpha}_{t-1} + \vec{\alpha}_{t+1})$$

Temporal linearity
(second-order derivative)

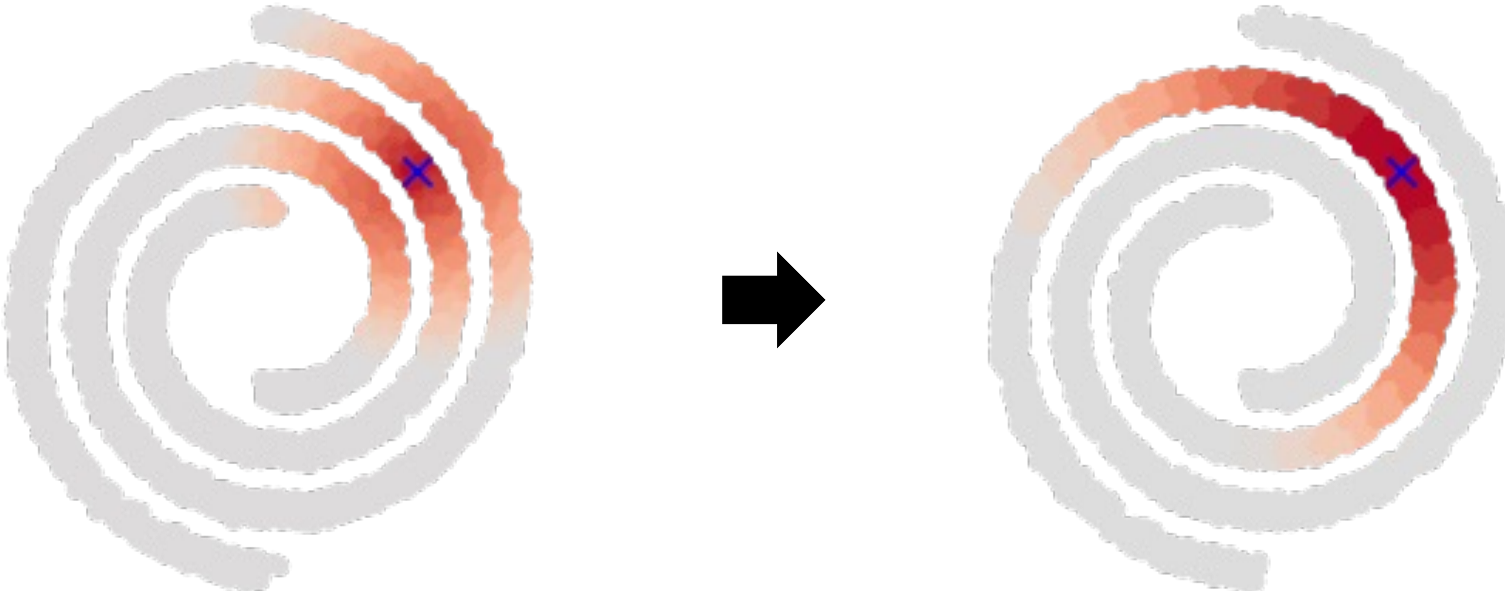
$$P\vec{\alpha}_i - \sum_{j \in n(i)} w(i,j)P\vec{\alpha}_j = 0$$

General linearity
(second-order derivative)

$$P\vec{\alpha}_i - P\vec{\alpha}_j = 0$$

Similarity
(first-order derivative)

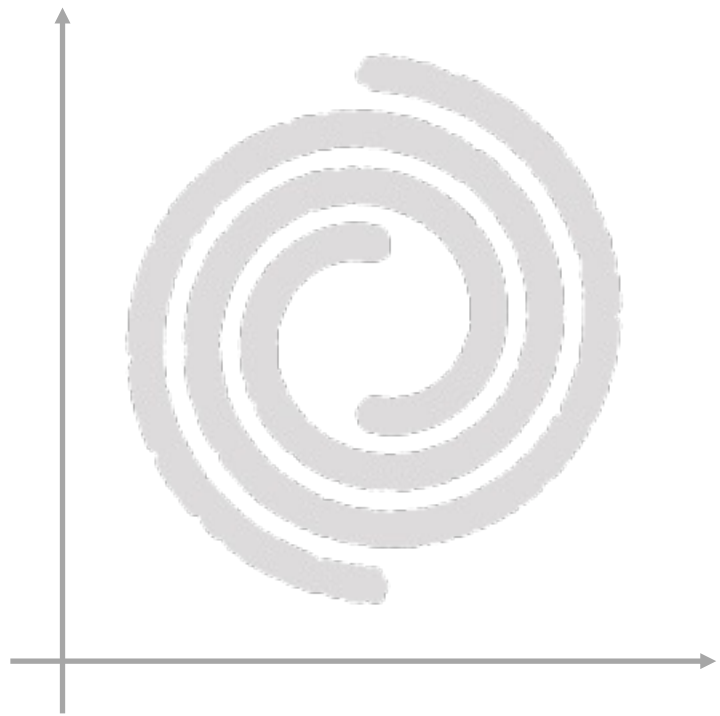
The similarity is better reflected in the representation space



Sensory signal neighborhood   Representation neighborhood

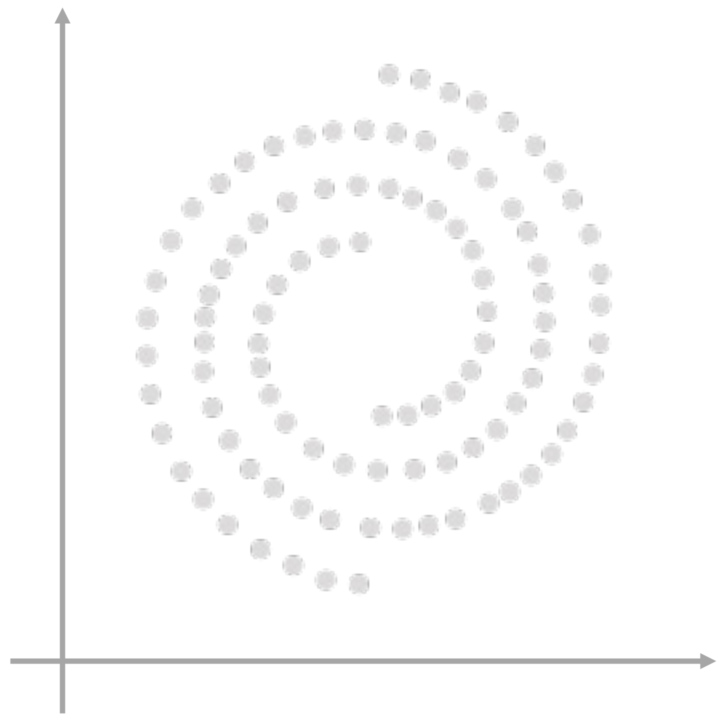
Our goal: transform raw data into a new space such that similar things are placed closer and meanwhile the new space is not collapsed.

An illustration of the sparse manifold transform steps



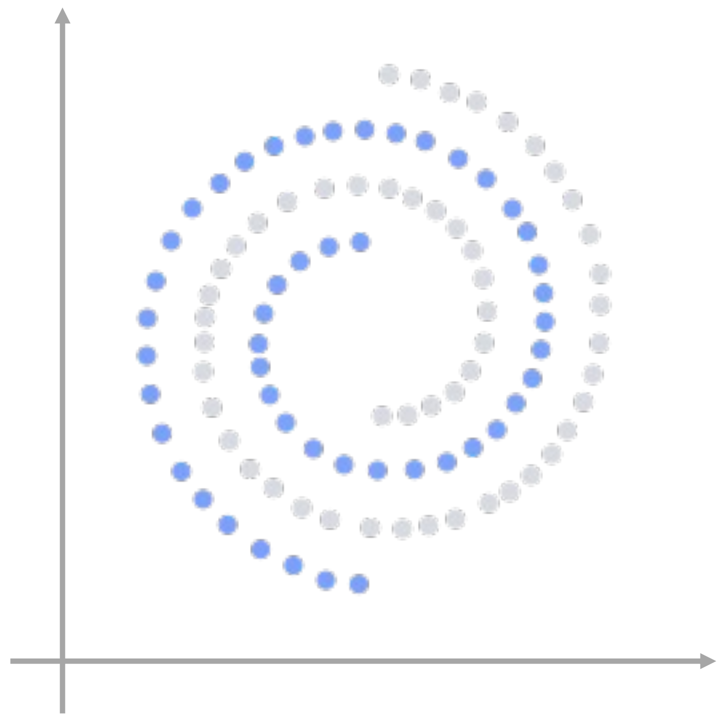
An illustration of the sparse manifold transform steps

Sparse coding tiles the data manifold and provides a support in the data space.



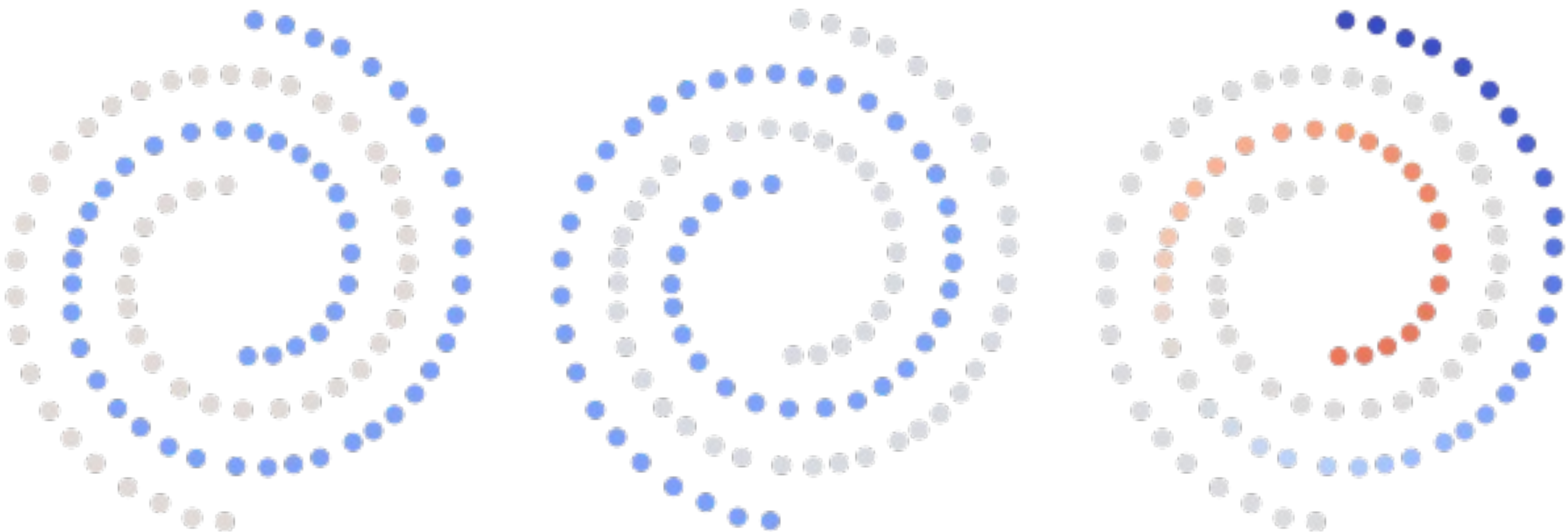
An illustration of the sparse manifold transform steps

Spectral embedding establishes similarity on the support.



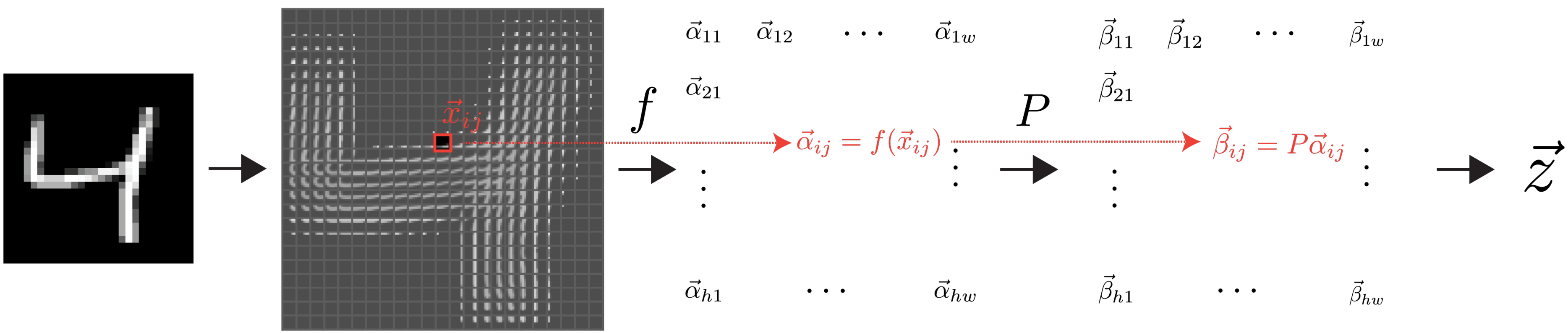
An illustration of the sparse manifold transform steps

Spectral embedding establishes similarity on the support. Each embedding dimension is a low-frequency function on the support.



2nd embedding dim 1st embedding dim 3rd embedding dim

SMT representation for natural images



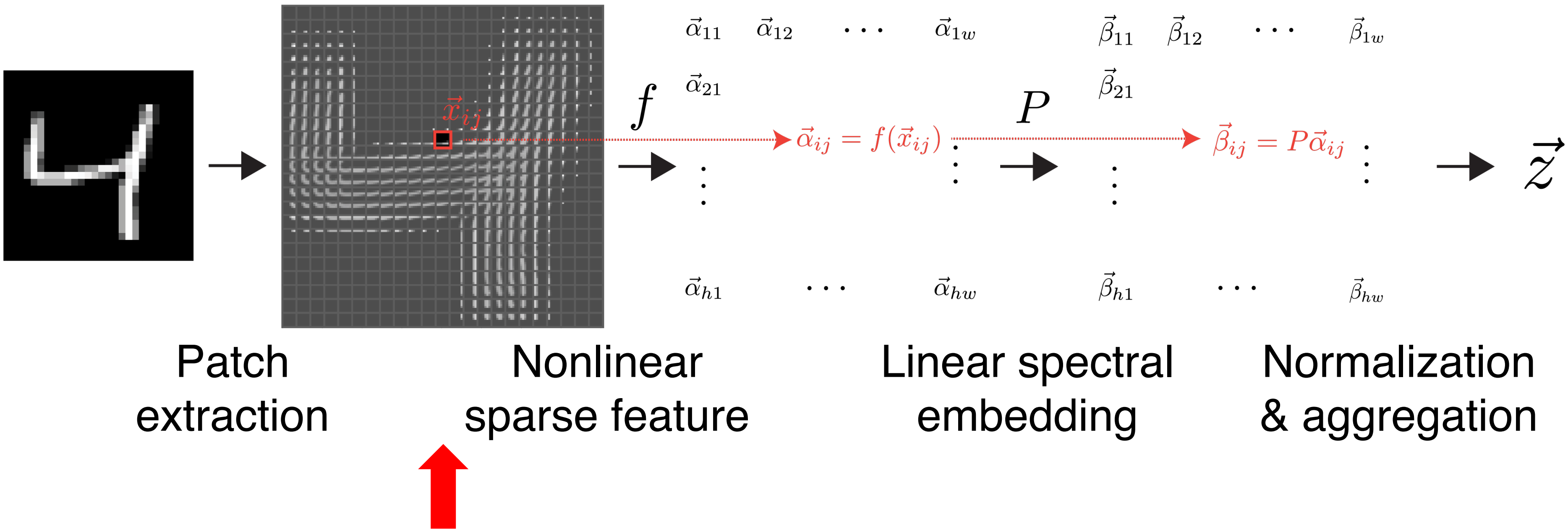
Patch extraction

Nonlinear sparse feature

Linear spectral embedding

Normalization & aggregation

SMT representation for natural images



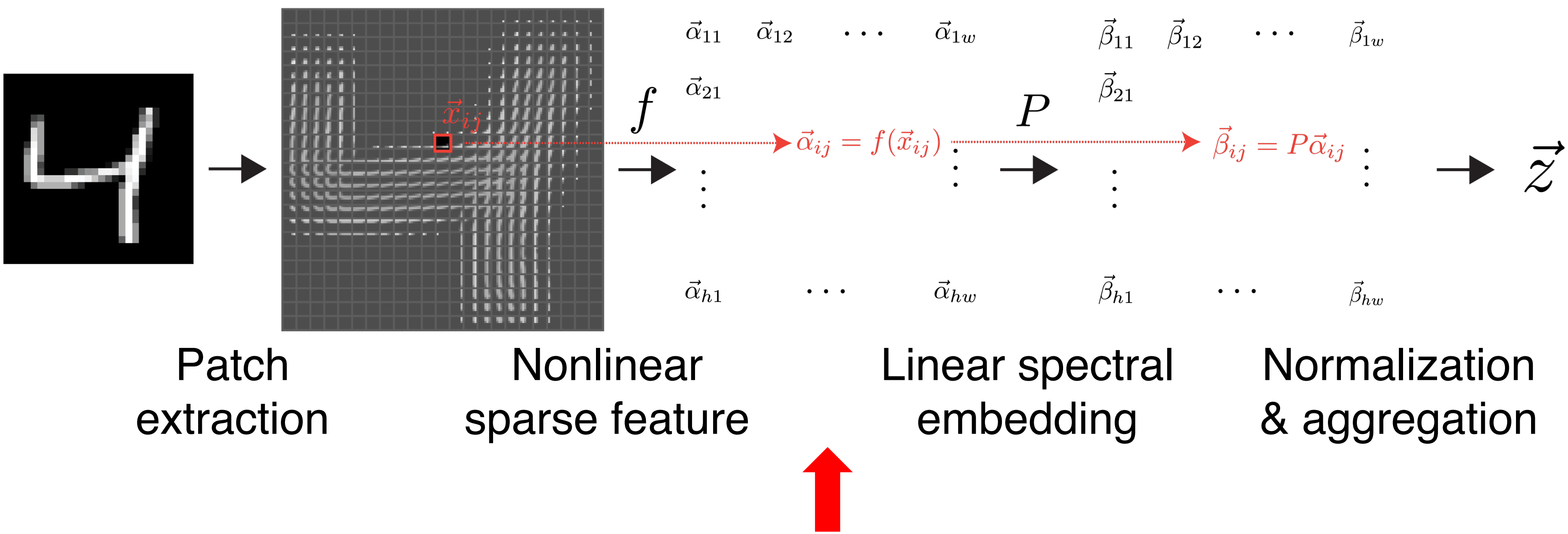
Patch extraction

Nonlinear sparse feature

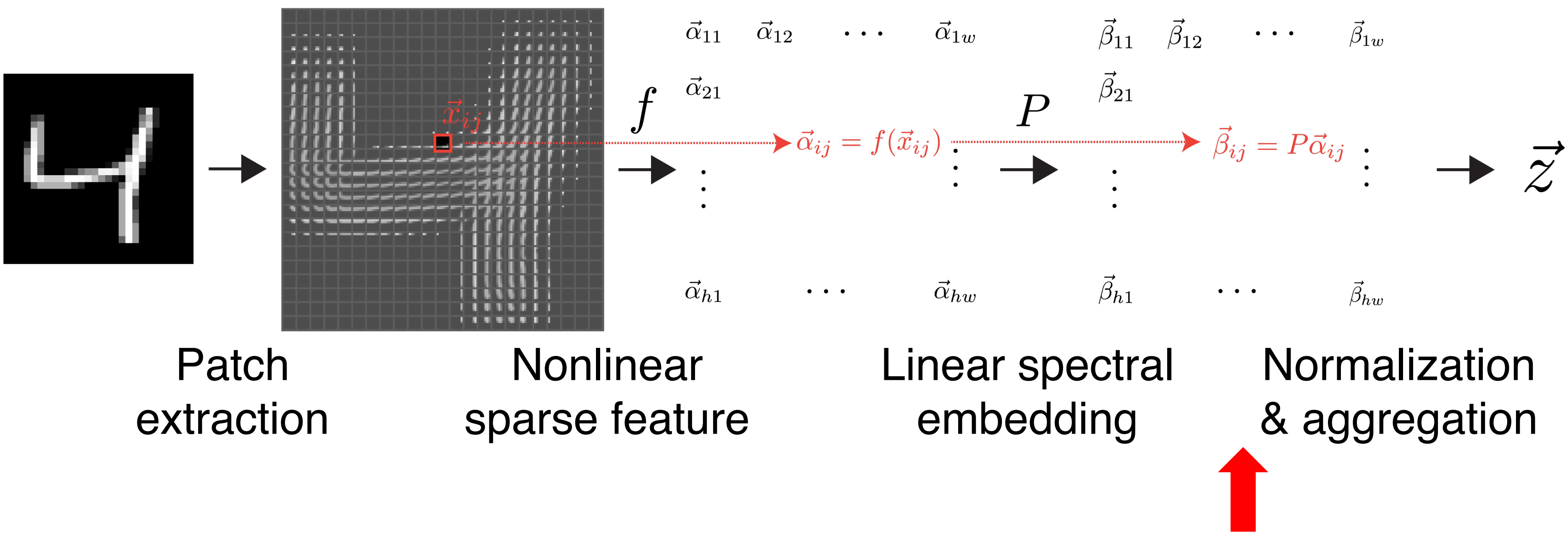
Linear spectral embedding

Normalization & aggregation

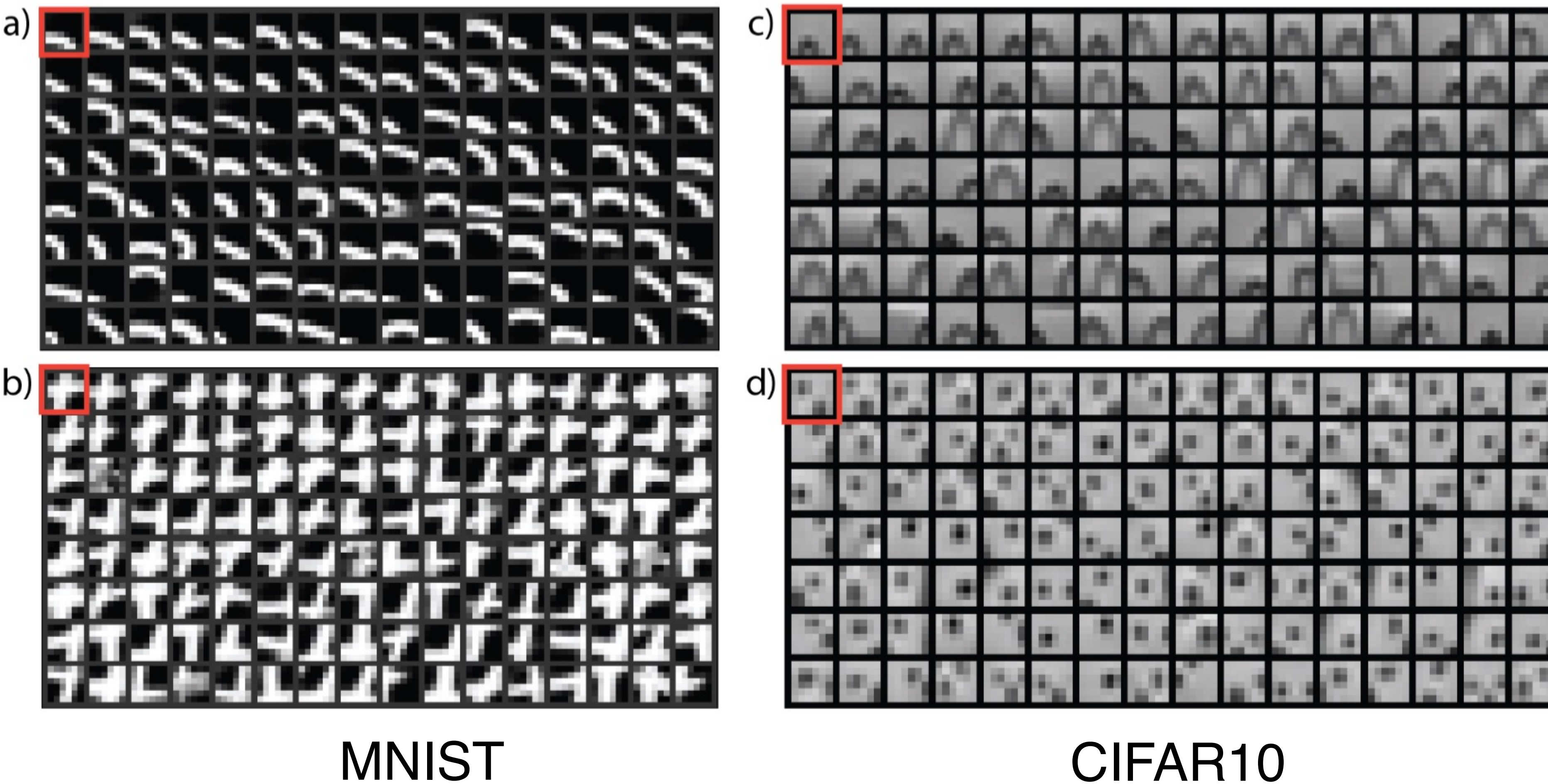
SMT representation for natural images



SMT representation for natural images



SMT is a local distance manipulation



SMT representation for natural images

MNIST:

Co-Occurrence Context Range	SMT-VQ (16384)	SMT-VQ (65536)	SMT-VQ (GloVe, 100K)
Whole Image	99.0%	98.9%	98.8%
3 Pixels	99.2%	99.3%	99.0%

CIFAR10:

Color Augmentation	SMT-VQ (100K)	SMT-GQ (8192)	SMT-GQ (65536)	SimCLR (ResNet18)	VICReg (ResNet18)
Original Image	—	79.2%	**81.1%**	68.3%	70.2%
Grayscale Image Only	78.4%	77.5%	78.9%	80.6%	**81.3%**
Original + Grayscale	—	81.4%	83.2%	**85.7%**	83.7%
Full (ColorJitter etc.)	—	—	—	90.1%	**91.1%**

CIFAR100:

Color Augmentation	SMT-VQ (100K)	SMT-GQ (8192)	SMT-GQ (65536)	SimCLR (ResNet18)	VICReg (ResNet18)
Original Image	—	50.8%	**53.2%**	32.4%	32.6%
Grayscale Image	46.6%	45.8%	**48.9%**	43.0%	43.9%
Original + Grayscale	—	53.7%	**57.0%**	48.9%	46.0%
Full (ColorJitter etc.)	—	—	—	63.7%	**65.4%**

Deep SSL:

18 layers

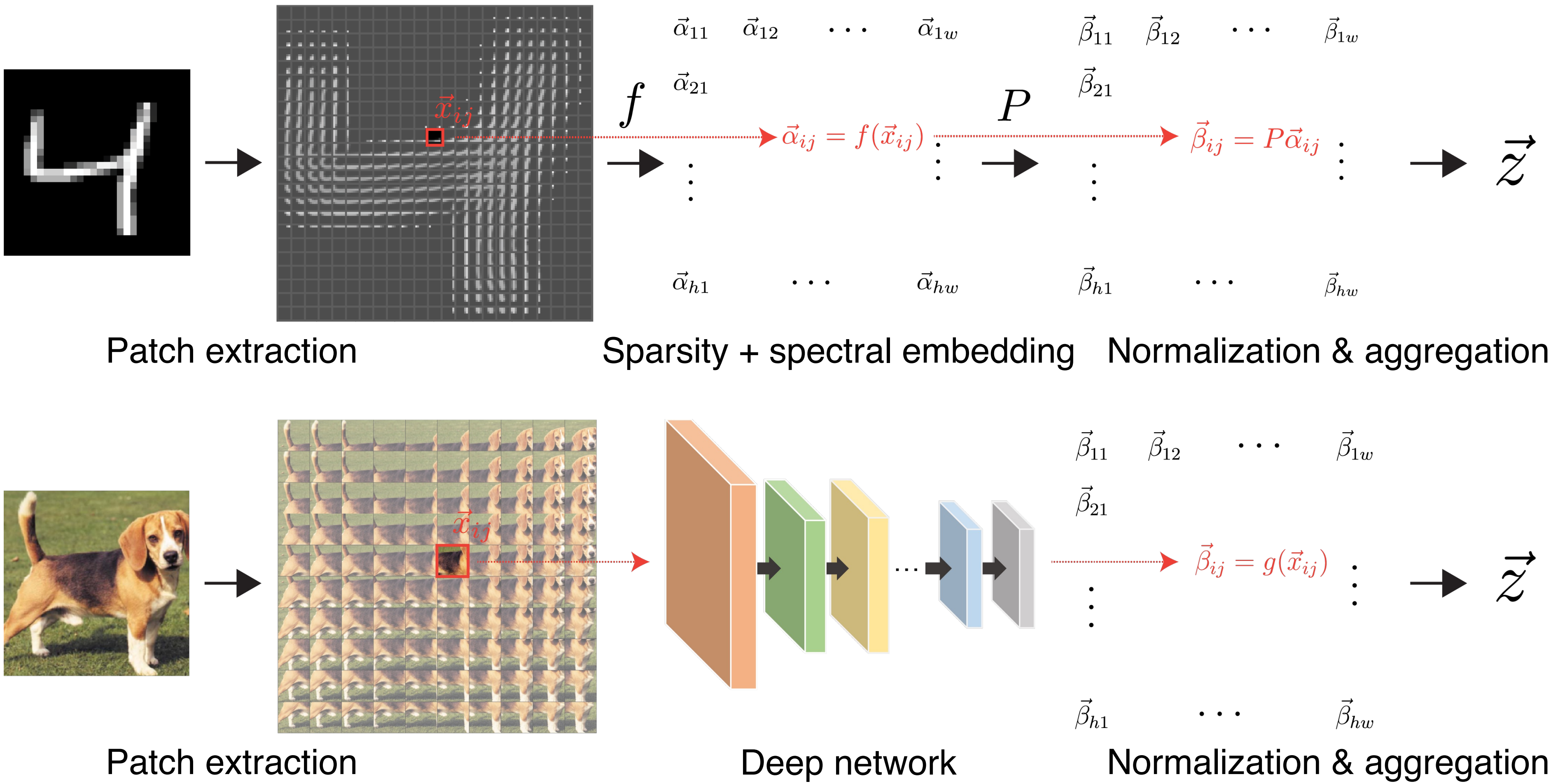
1000 training epochs

SMT:

2 layers

1 training epoch

The convergence



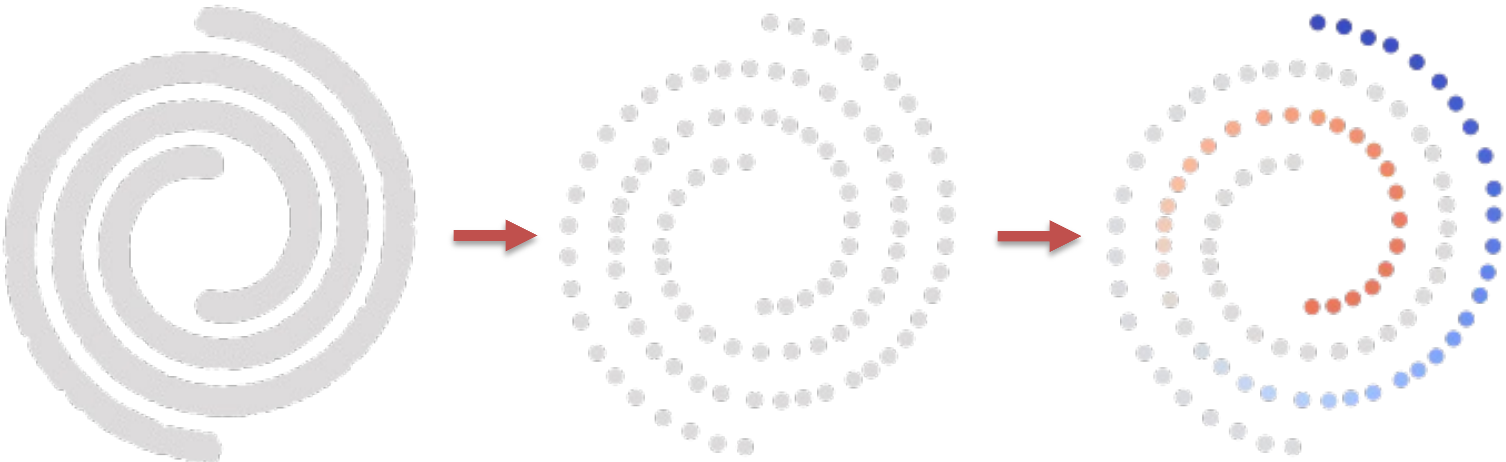
(Bardes et al., ICLR'22)
(**Chen** et al., arXiv'22)
(Tong et al., arXiv'23)

Main points

Poster: **#163 (last row)**
16:30 CAT – 18:30 CAT

Unsupervised representation from neural and statistical principles

- Sparse coding tiles the data space and provides the support
- Spectral embedding establishes similarity and linearity on the support



Data distribution | Sparse feature | Spectral embedding

Meta AI | NYU Center for Data Science | REDWOOD CENTER for Theoretical Neuroscience | BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH