电子科技大学
University of Electronic Science and Technology of China

A Discussion of Learning Representation

The 2th Squad of Feature Engineering

Data Mining Lab,
Big Data Research Center, UESTC
Email：huangchen.uestc@gmail.com

Table of Contents

Draw an apple

Representation

➢ **Feature** is the general answer for *"What's this ? "*

Natural Language Processing

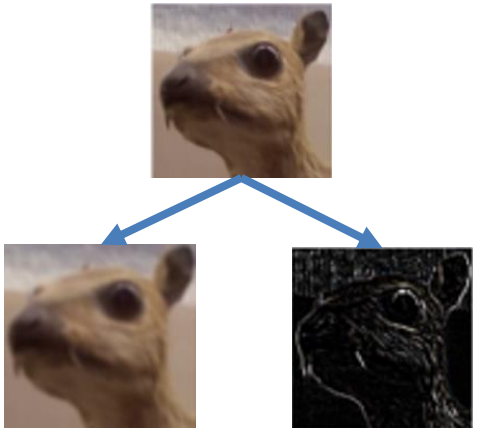
句子A：我/喜欢/看/电视，不/喜欢/看/电影。

句子B：我/不/喜欢/看/电视，也/不/喜欢/看/电影。

句子A：我 1，喜欢 2，看 2，电视 1，电影 1，不 1，也 0。

句子B：我 1，喜欢 2，看 2，电视 1，电影 1，不 2，也 1。

Computer vision



What about

- Trajectory data?
- Dynamic network?

So, what does representation means in data mining?

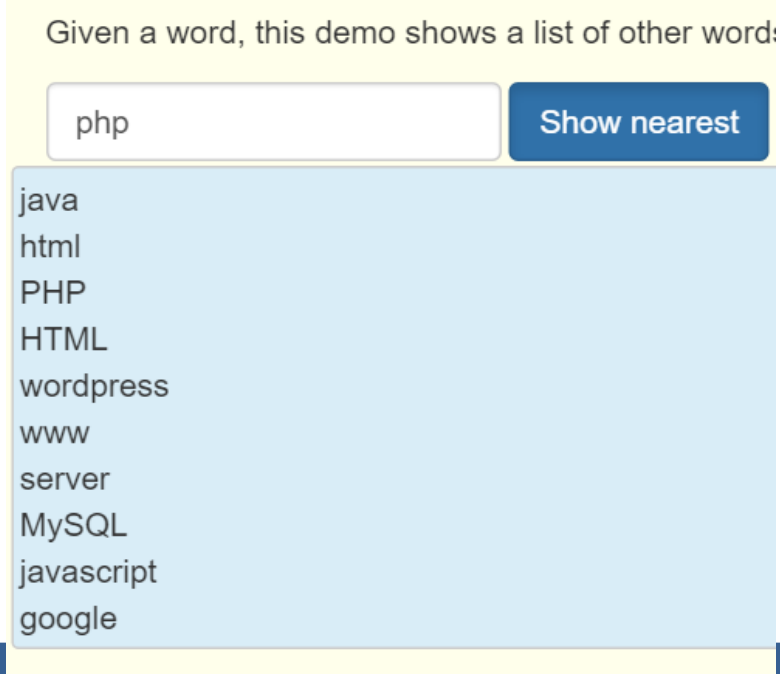
Feature → Numeric Abstraction (almost)

- **Feature** is the general answer for *"What's this ? "*
- **Some tricky cases:**
  - **Nominal attribute**

    [One-hot coding]: excellent\good\bad → 001\010\100
  - **Structured object**

    ***Sentences**, pics, sequential data, networks*

Word Embedding



Feature → Numeric Abstraction (almost)

- **Feature** is the general answer for *"What's this ? "*
- **Some tricky cases:**
  - **Nominal attribute**

    [One-hot coding]: excellent\good\bad → 001\010\100
  - **Structured object**

    Sentences, *pics, sequential data, networks*
  - **Discretize or not**
  - **Missing value**

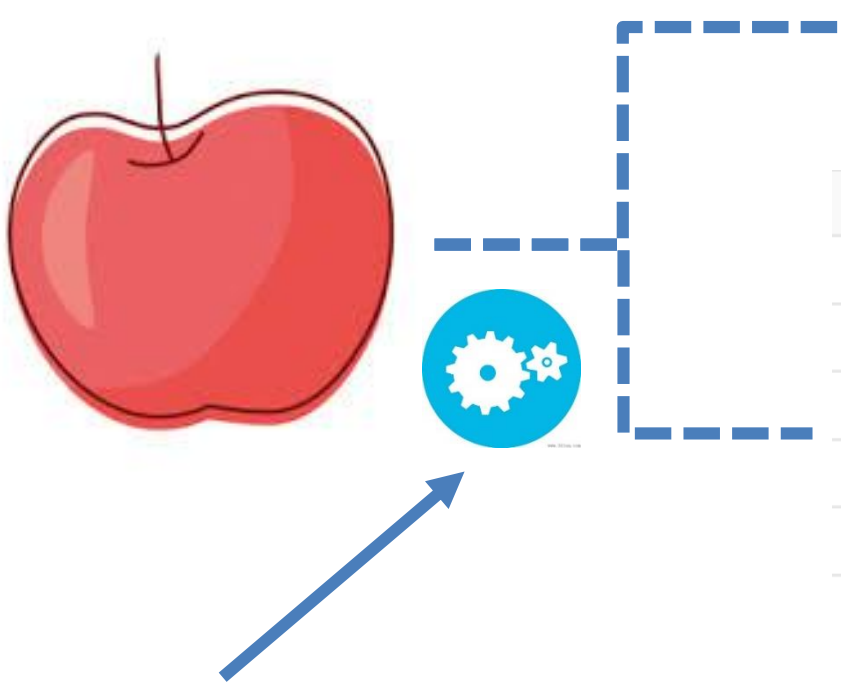
    Drop it, fill it or just leave it be?
  - **Ordered feature ?????**

3.1 排序特征

对原始特征中 1045 维 numeric 类型的特征从小到大进行排序，得到 1045 维排序特征。排序特征对异常数据都有较强的鲁棒性，使得模型更加稳定，降低过拟合的风险。

Feature (Cont.)

- **Feature** is the general answer for *"What's this ? "*
- **Label** is the general answer for *"What's your name ? "*



成分名称	含量	成分名称
可食部分	86%	水分（g）
能量（kJ）	200	蛋白质（g）
碳水化合物（g）	13.6	膳食纤维（g）
灰份（g）	0.2	维生素A（mg）
视黄醇（mg）	0	硫胺素（μg）
尼克酸（mg）	0.2	维生素C（mg）

- What data mining do ?
- What is model ?
  - General answer for *"How do they behave in **this** way?"*

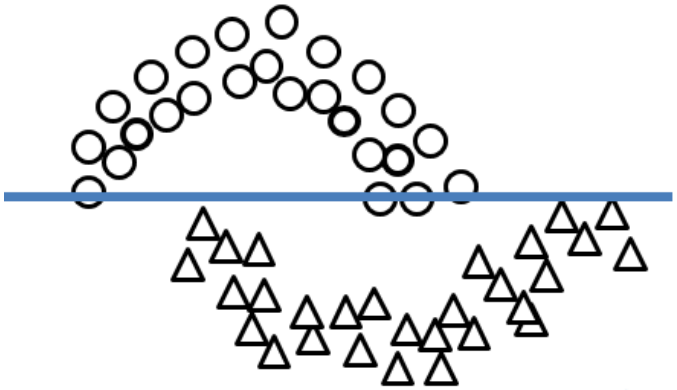
Feature V.S. Model

- Unsupervised learning is subjective
- Supervised learning is objective

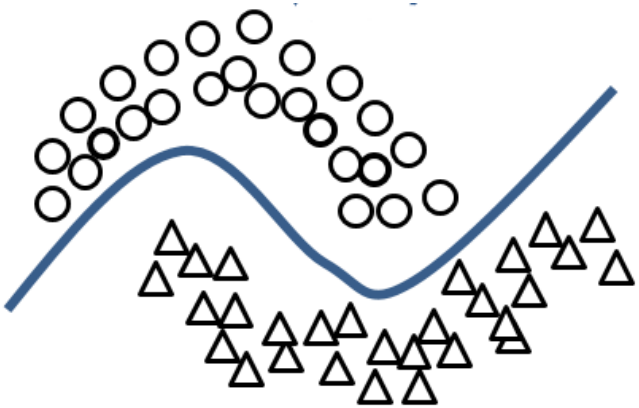
"We know it when we see it"

Feature —— Similarity? —— Model

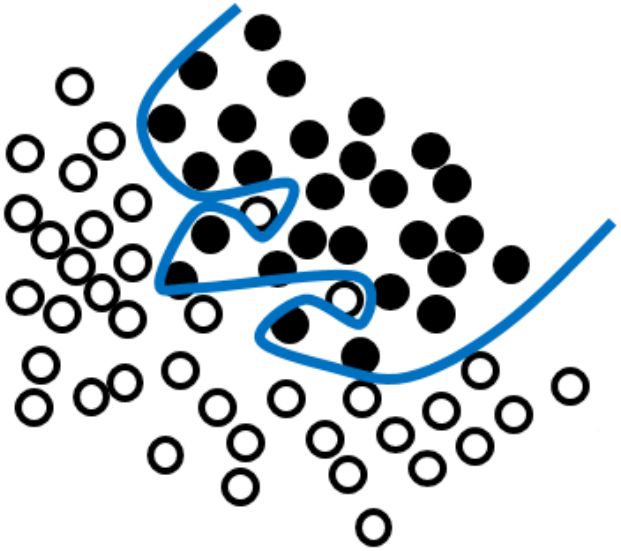
Feature V.S. Model



Good representation
Bad model



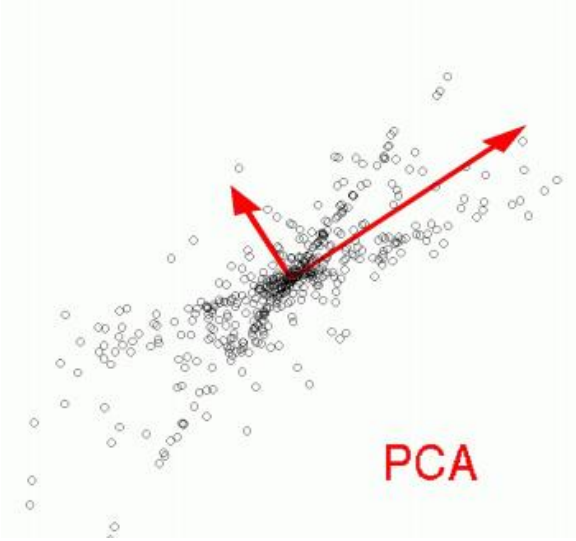
Good representation
Good model

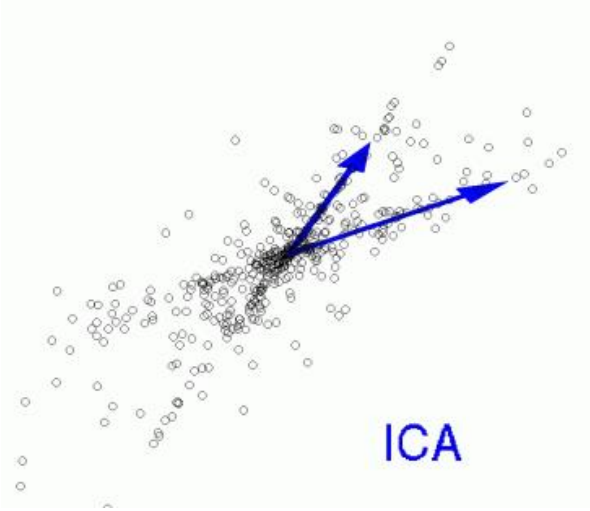


Bad representation
Good model

Learning Representation / Features

- **What's learning representation?**
- Don't panic ! You already know…





$$\boldsymbol{X_{new}} = \boldsymbol{f}(\boldsymbol{X}) \quad or \quad \boldsymbol{X_{new}} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{Y})$$

where $\boldsymbol{X}, \boldsymbol{Y}$ *is from one source or many*

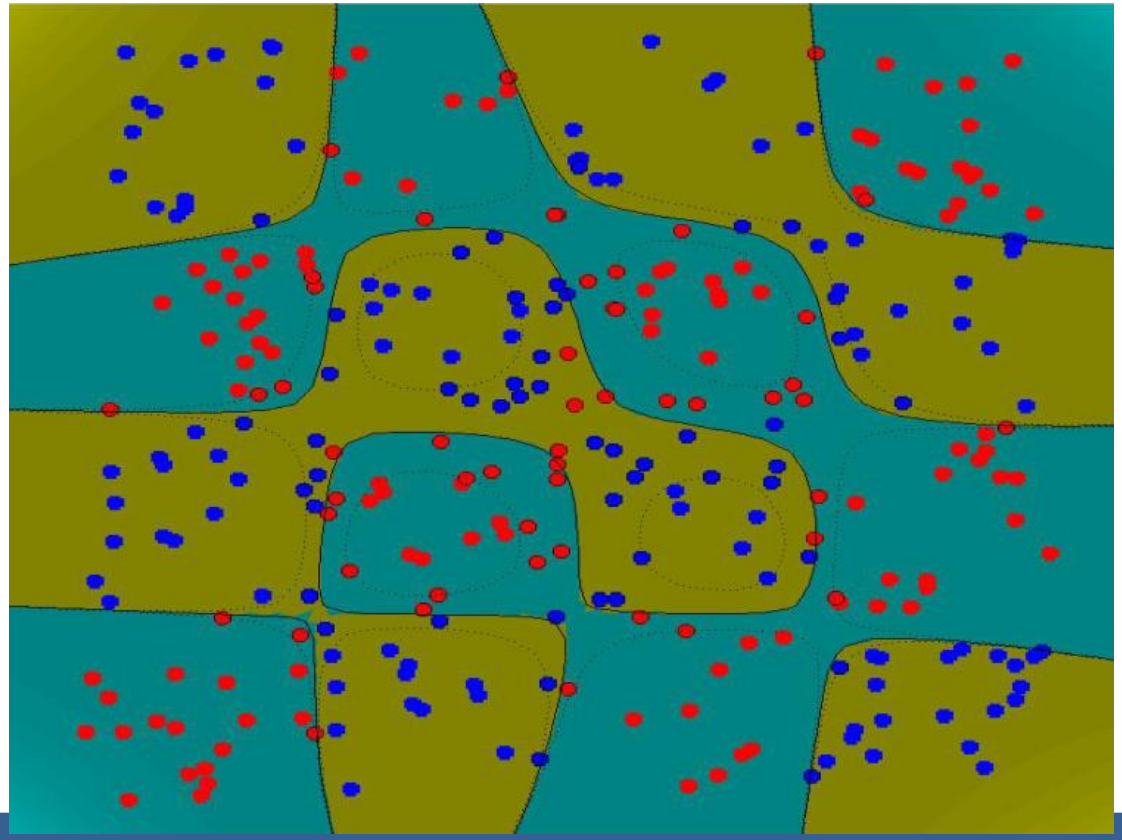
Learning Representation

➢Motivation

- Find a better representation (*for similarity measure or others*)

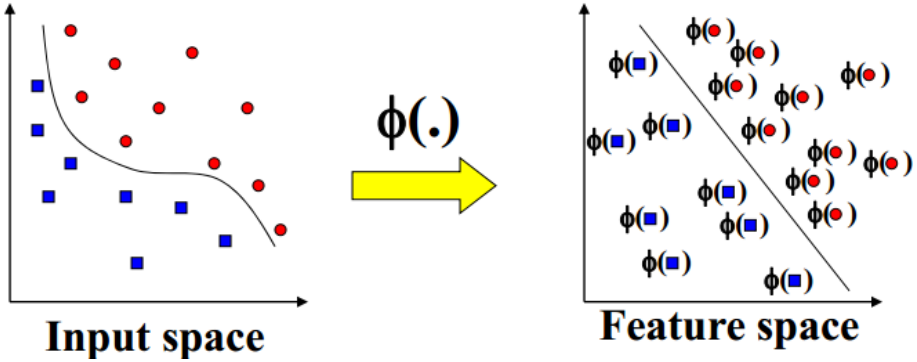
➢"better" how?

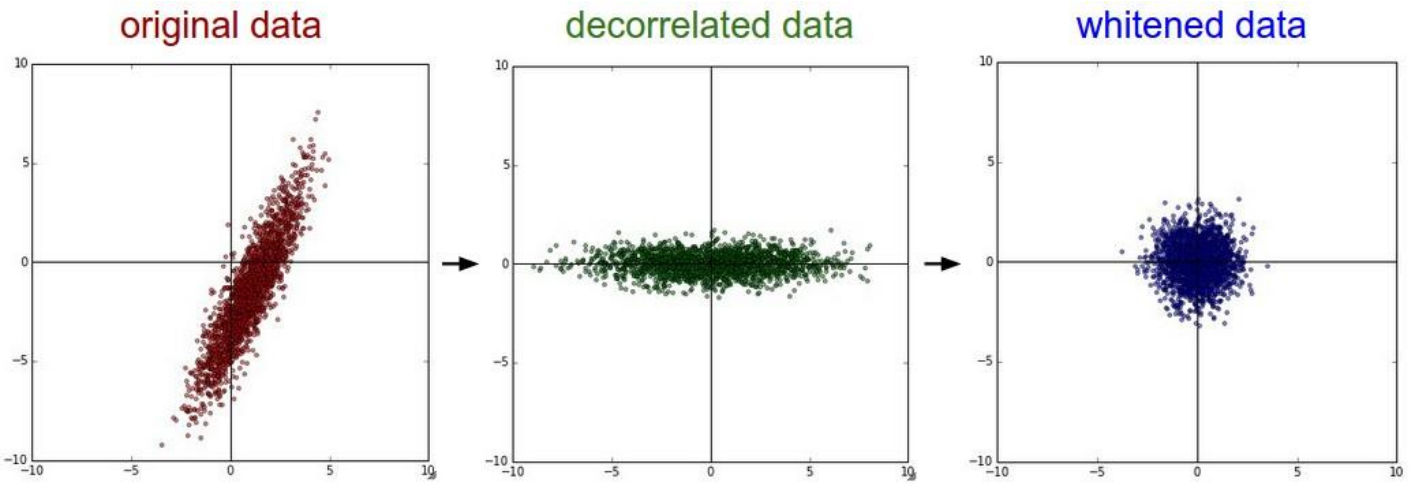
- What's wrong with my current representation?



What's your definition of "better" ?

- **What's your ideal form of representation?**
- **What do you want?**





What's your definition of "better" ?

- **Task-driven / Model-driven**
  - Classification, clustering, ranking
  - Whatever is best for your task
- **Interpretability**
  - Non-negative, …
- **Visualization / Storage**
- **Numerical calculation / Fast convergence / Tuning**
  - Normalization, Whiten, …

Any representation
fitting our **hypothesis**
of tasks/algorithms is good

Why we make hypothesis ?

➢ Additional prior

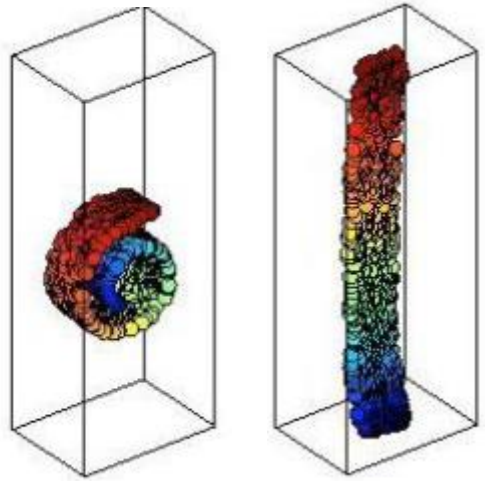
- Compositionality is useful to describe the world around us efficiently

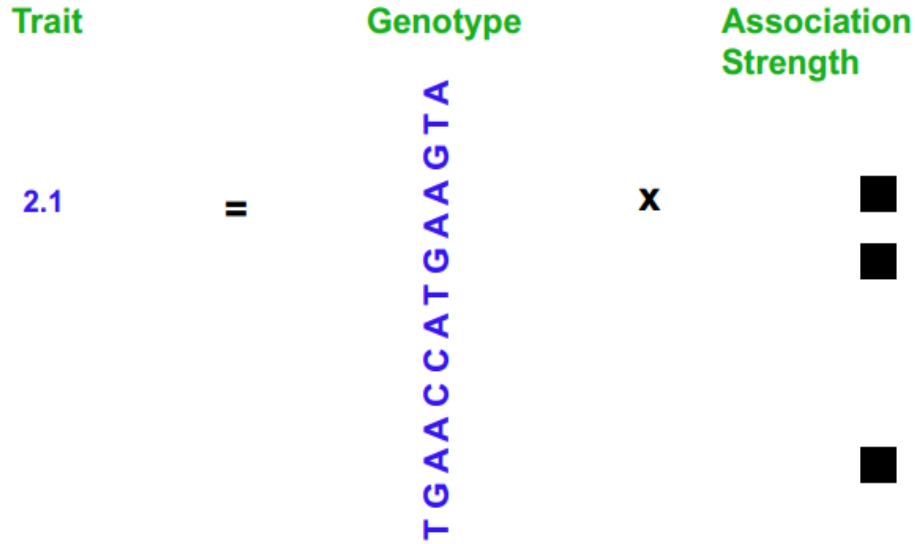
- To tell us the data generating distribution and recover a set of latent variables that describe a distribution over the observed data

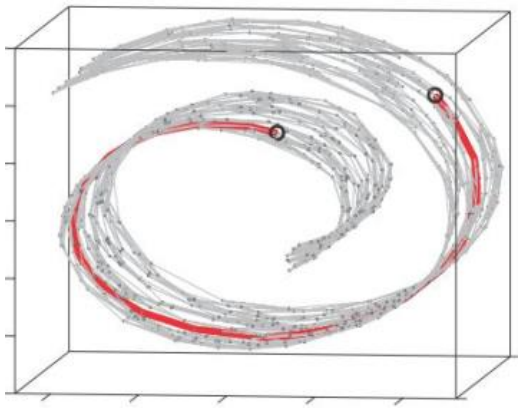
- *Help non-supervision be less subjective…

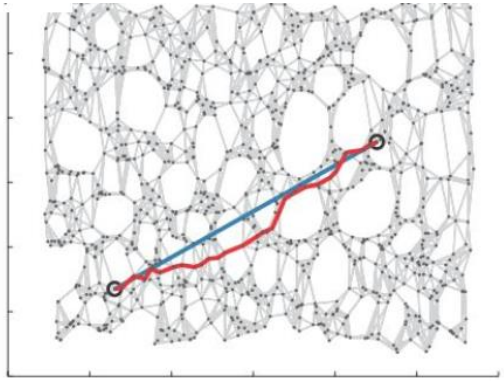
Hypothesis of "better"

- In your field, **what's the hypothesis you want?**



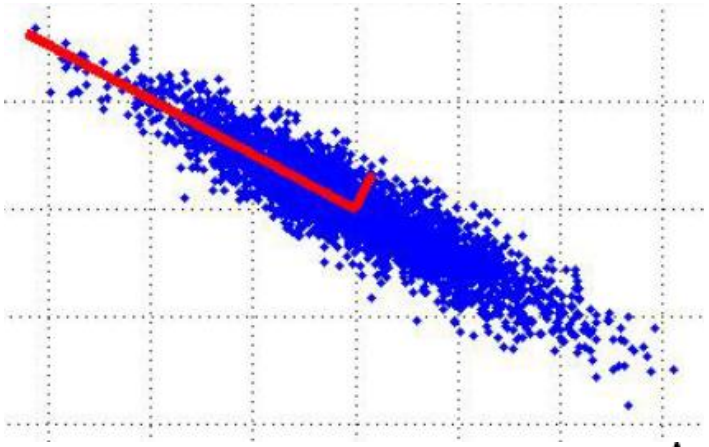




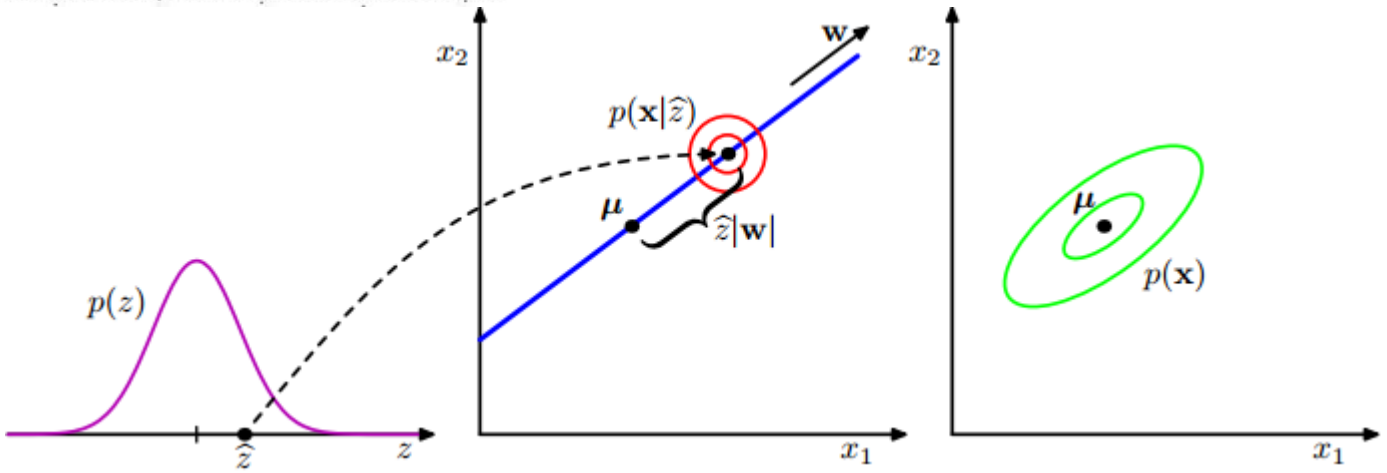


Hypothesis of "better"

- Example
  - What is the hypothesis of PCA ?



1. Maximum variance
2. Minimum projection error
3. Linear Gaussian with limited free parameters

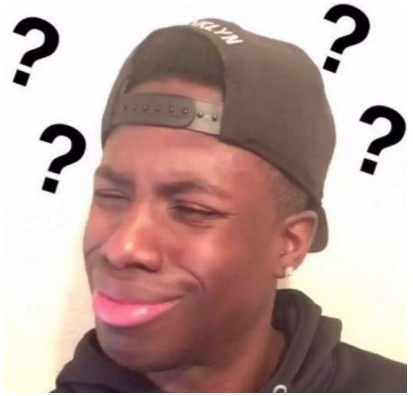


Hypothesis of "better"

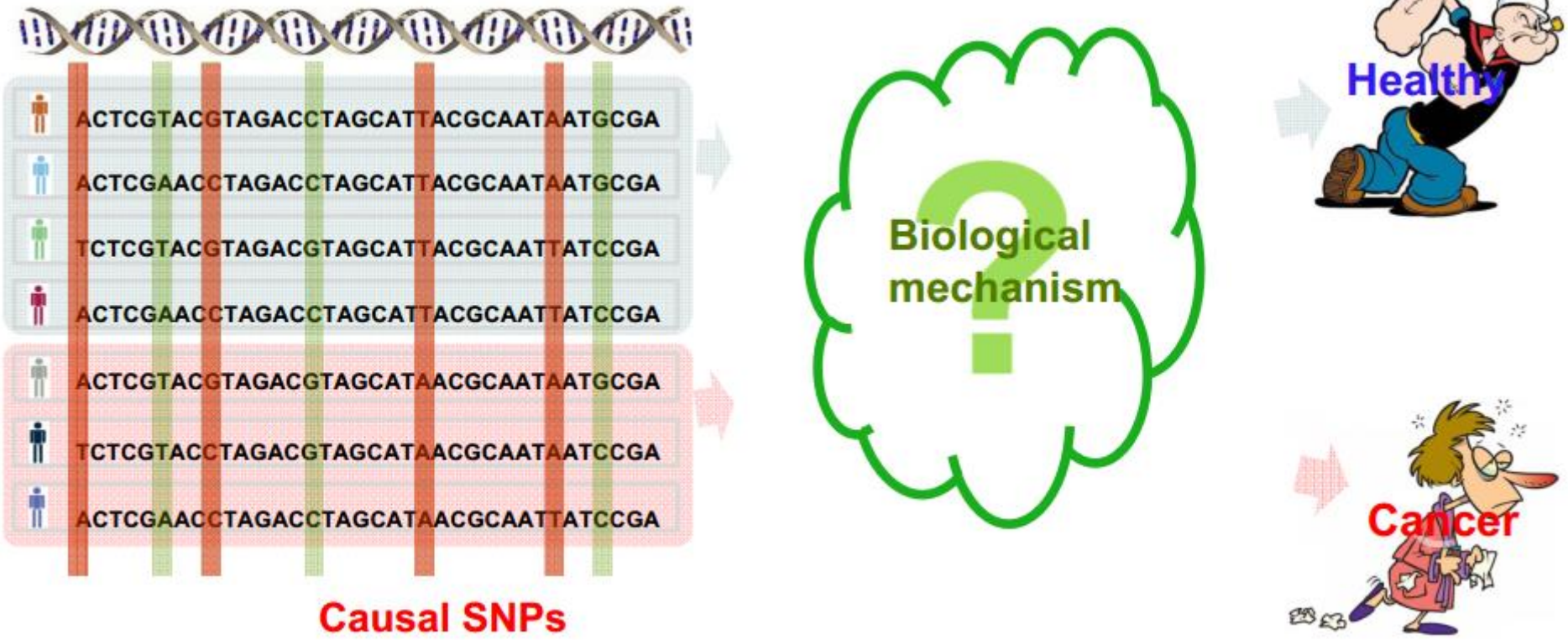
➢ **List of the hypothesis you want (maybe)**

- Sparse representation
- Low rank representation
- Common / Semantic space
- Dimension reduction / Kernel
- High-level / Deep features
- Temporal and spatial couple / coherence
- Structure-specified features
- Features decorrelation
- Discriminative features
- Separate manifolds/ clusters
- Normalization
- …

Mathematic language of hypothesis



Sparse representation



How to make our new presentation sparse?

Sparse representation

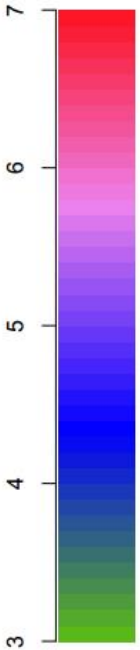
- L0 (*it'll do, if you could solve it…*)
- Generalized lasso

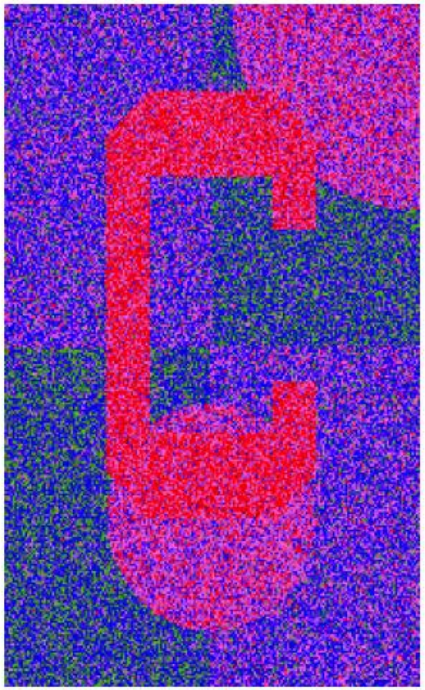
$$\min_{X_{new}} f(X_{new}) + \lambda \|DX_{new}\|_1$$

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}$$

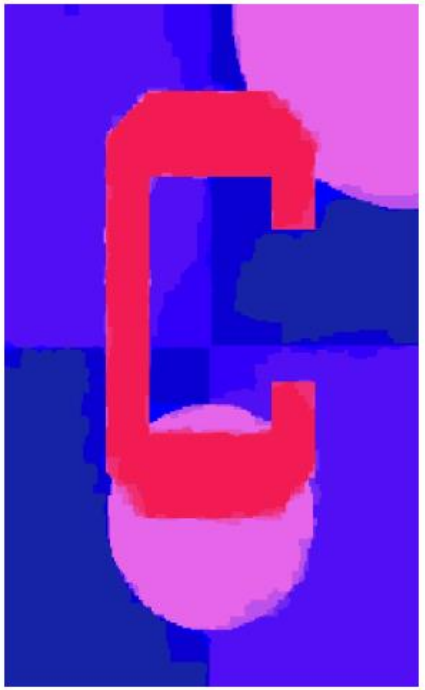
$$\|D\beta\|_1 = \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}|$$

Fused lasso example





Data (noisy image)



Solution (denoised image)

- Categorical variables are associated with separate manifolds, is it a **discrete sparse representation**?

$$z_n^{(t)} = \arg\max_k (x_n - \mu_k^{(t)})^T \Sigma_k^{-1(t)} (x_n - \mu_k^{(t)})$$

$$\downarrow$$

$$\underset{\mathcal{D},s}{\text{minimize}} \sum_i ||\mathcal{D}s^{(i)} - x^{(i)}||_2^2$$

$$\text{subject to } ||s^{(i)}||_0 \leq 1, \forall i$$

$$\text{and } ||\mathcal{D}^{(j)}||_2 = 1, \forall j$$

$$\downarrow$$

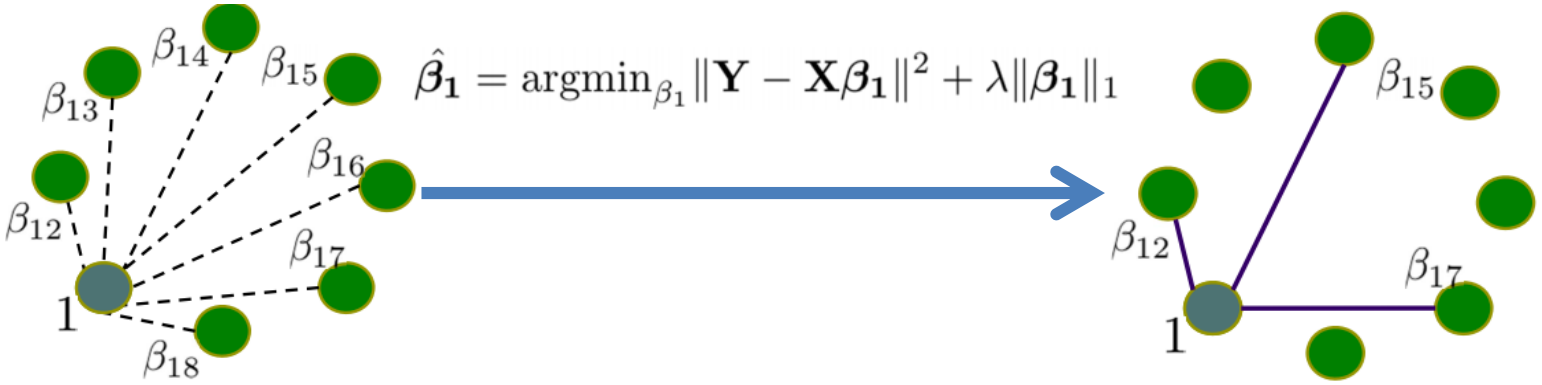
$$\underset{\mathcal{D},s}{\text{minimize}} \sum_i ||\mathcal{D}s^{(i)} - x^{(i)}||_2^2 + \lambda||s^{(i)}||_1$$

$$\text{subject to } ||D^{(j)}||_2 = 1, \forall j.$$

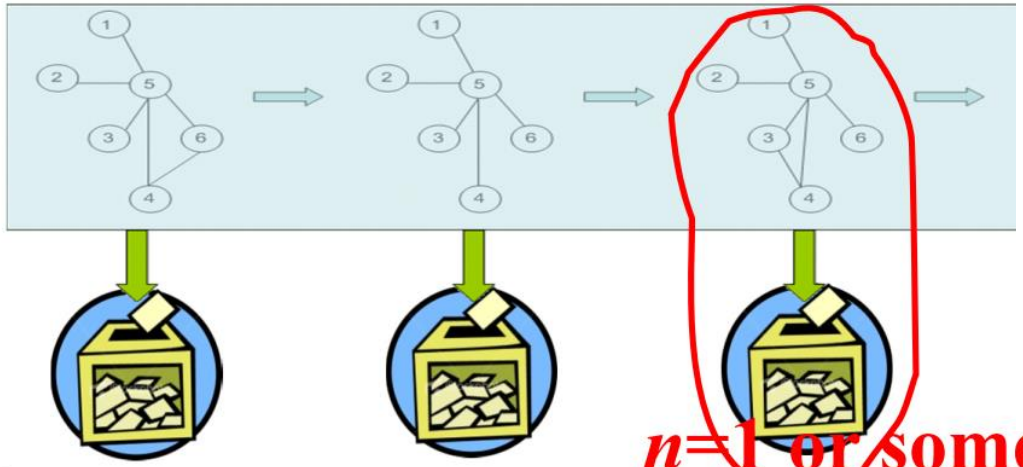
Sparsity in your field

- **Data stream ?**

- **Networks ?**



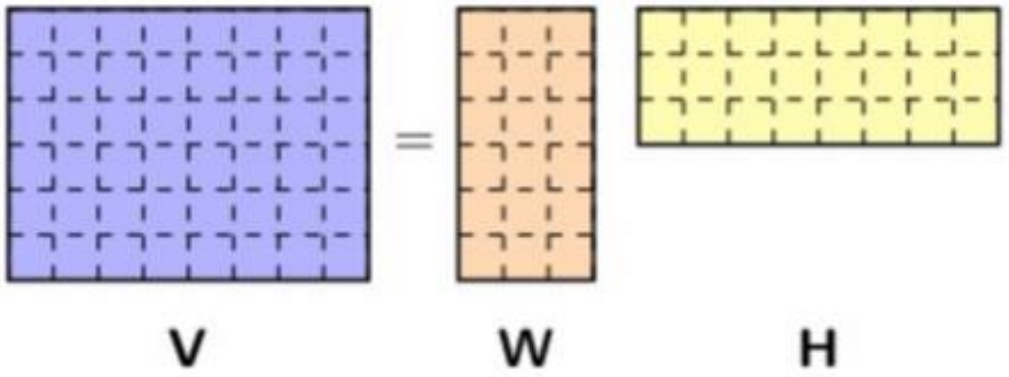
- **Temporal + Network?**



$$\hat{\theta}_i^1, \ldots, \hat{\theta}_i^T = \arg\min_{\theta_i^1,\ldots,\theta_i^T} \sum_{t=1}^{T} l_{avg}(\theta_i^t) + \lambda_1 \sum_{t=1}^{T} \| \theta_{-i}^t \|_1 + \lambda_2 \sum_{t=2}^{T} \| \theta_i^t - \theta_i^{t-1} \|_q^q,$$

Low Rank

- **Why low rank ????**
  - **2D – sparsity**, being sparse while coherent with other instances
  - Reducing correlation



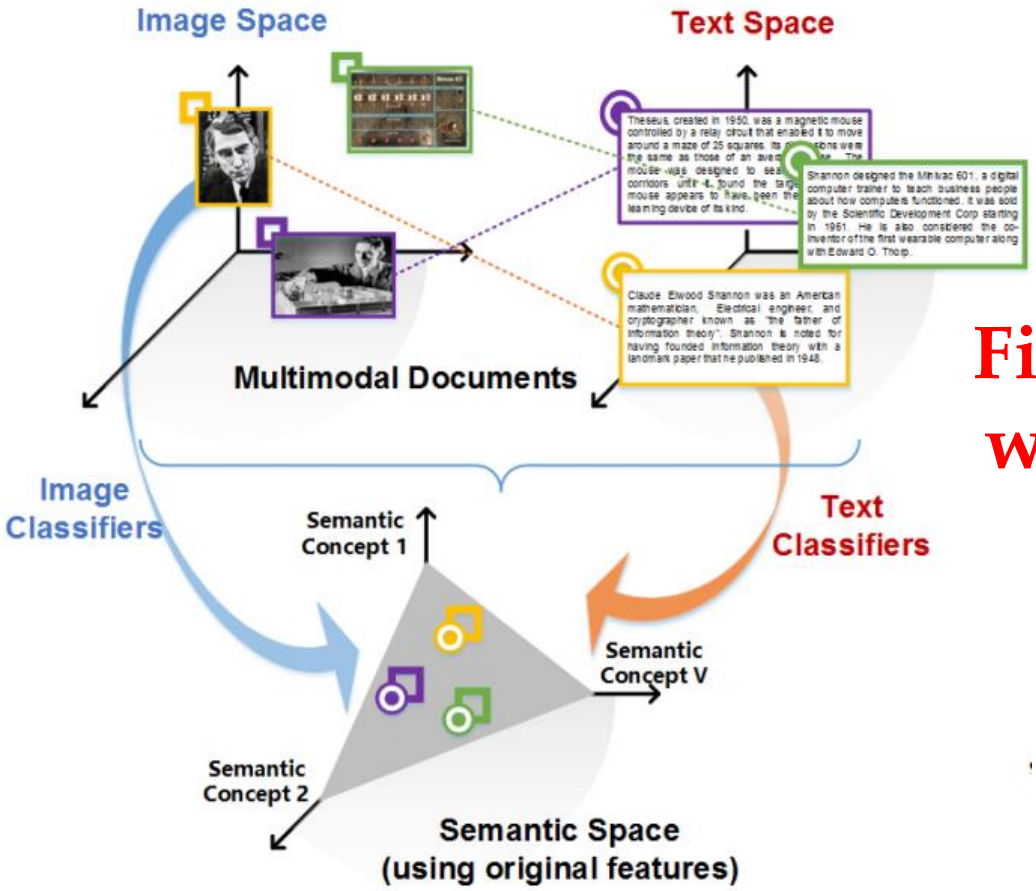
- **How to get there ???**
  - Nuclear norm

- **LR in your field??**

Image Rectification

Common representation

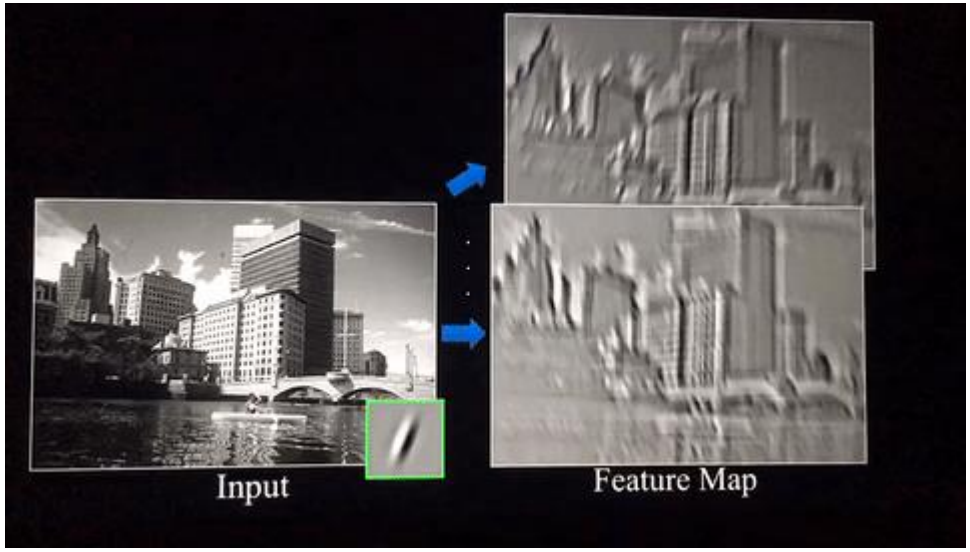
- Multi-view learning / multi-task learning/ transfer learning / domain adaption / semi-supervised learning / multi-source learning / … → **Integration**



Find a space that we could define a measure !

Deep representation

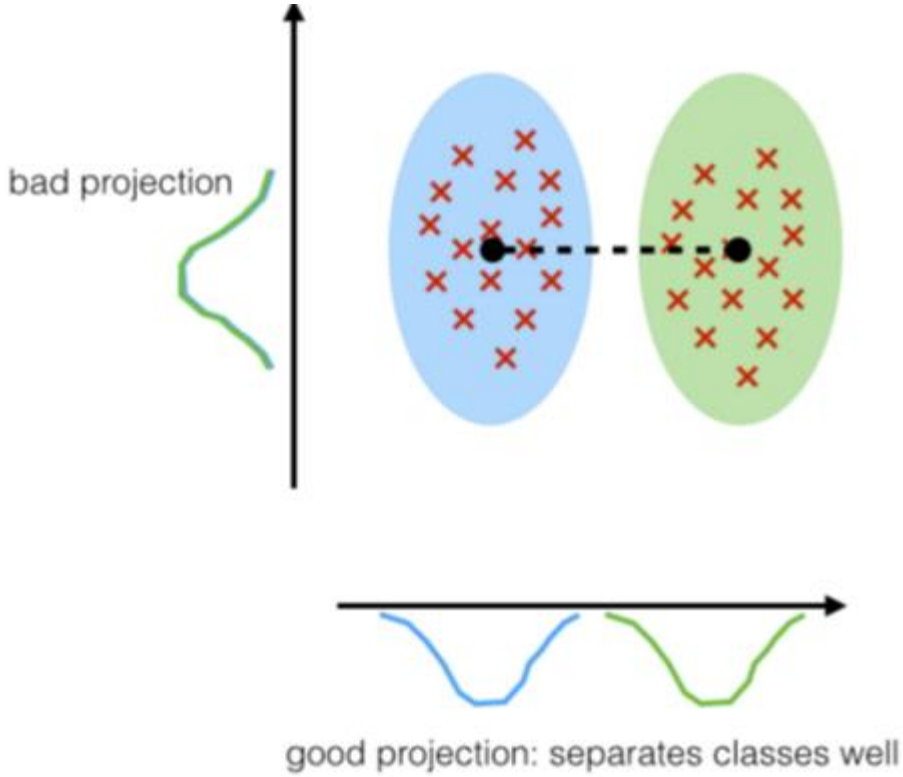
Operation	Filter	Convolved Image
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	



- Blank box
- Why it work

Discriminative Features

- We put constraints on labels !



LDA:
maximizing the component axes for class-separation

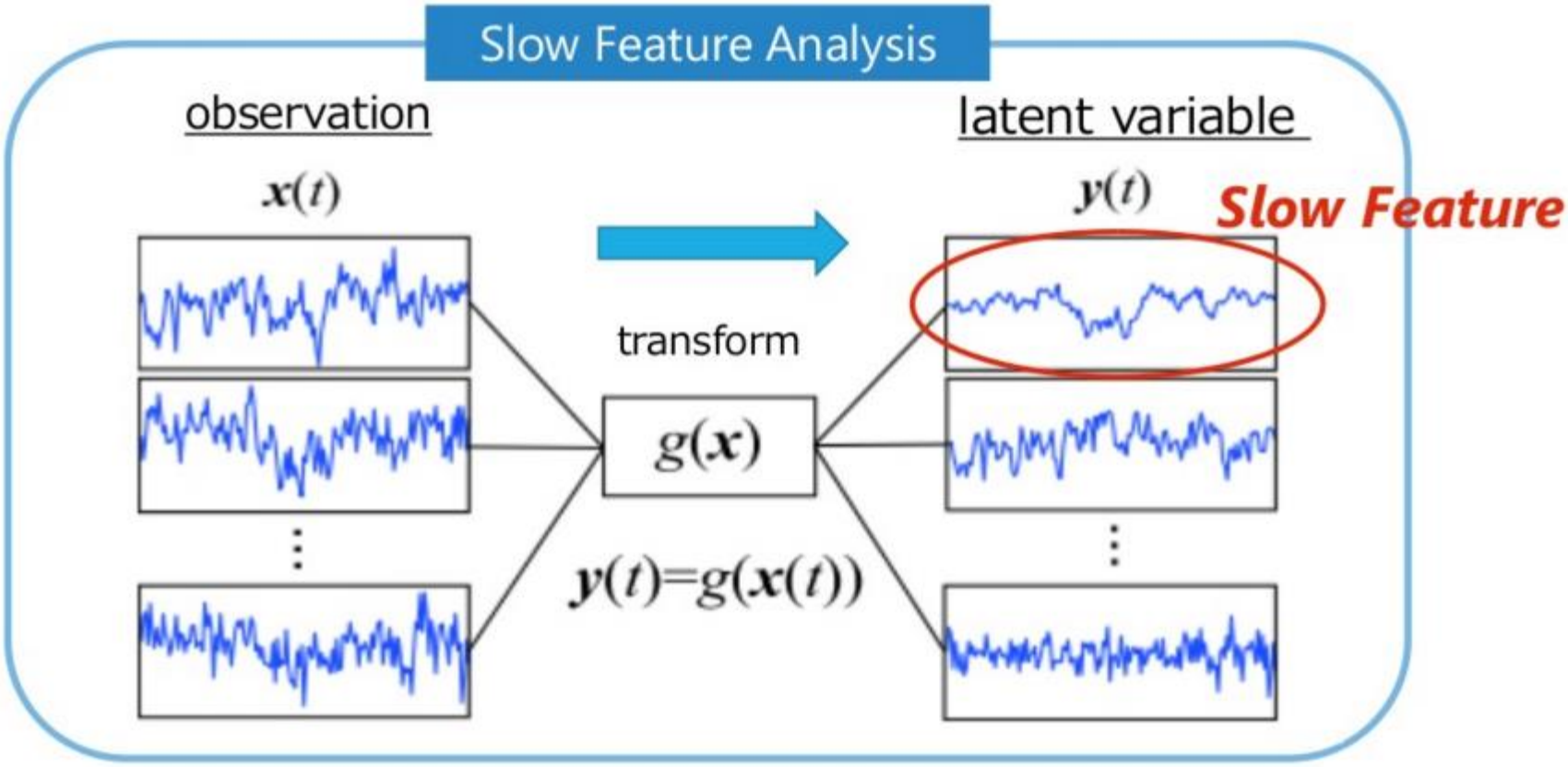
$$\mathcal{J} = tr((\mathbf{P}^T\mathbf{S}_t\mathbf{P})^{-1}(\mathbf{P}^T\mathbf{S}_b\mathbf{P}))$$

$$\eta = \sum_{i=1}^{l} c(y_i, f(\mathbf{x}_i)) + \gamma_1 \|f\|_{\mathcal{H}}^2 + \gamma_2 \|f\|_{\mathcal{D}}^2$$

Temporal and spatial coherence

- Slow-feature Analysis

Objective : Extract ***Slow Feature*** from ***Time series data*** .



Pre-processing 1 - rescale

➢ Normalization

- Why normalization ?

→ Weka experiment

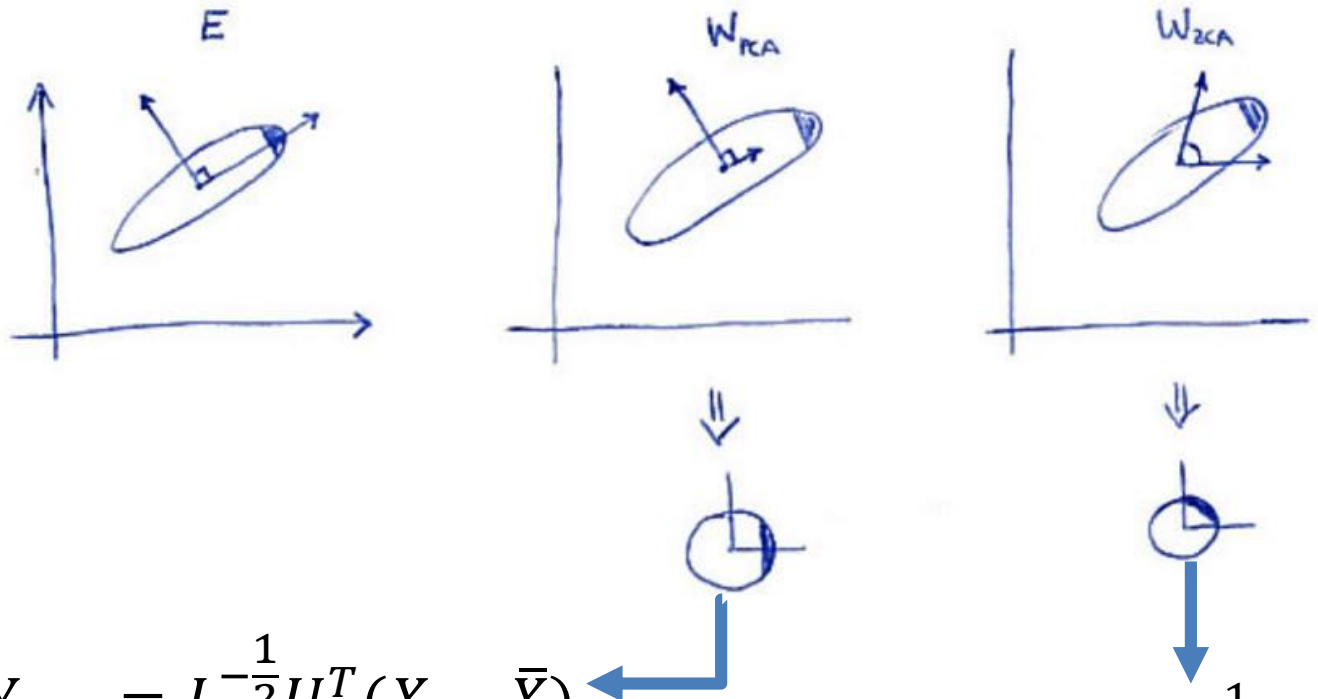
→ Training: 1000, Test: 500, 242 dimensions with SVM

	Original	Normalized
Time building model	13.29s	8.89s
Correctly classified	76(14.902%)	350(68.6275%)
Incorrectly classified	434(85.098%)	160(31.3725%)

Pre-processing 2 – feature decorrelation

➢ PCA / ZCA Whiten

- What is whiten ?



$$Y_{pca} = L^{-\frac{1}{2}}U^T(X - \bar{X})$$

$$Y_{zca} = UL^{-\frac{1}{2}}U^T(X - \bar{X})$$

- Why whiten ?
- Feature decorrelation

Ending

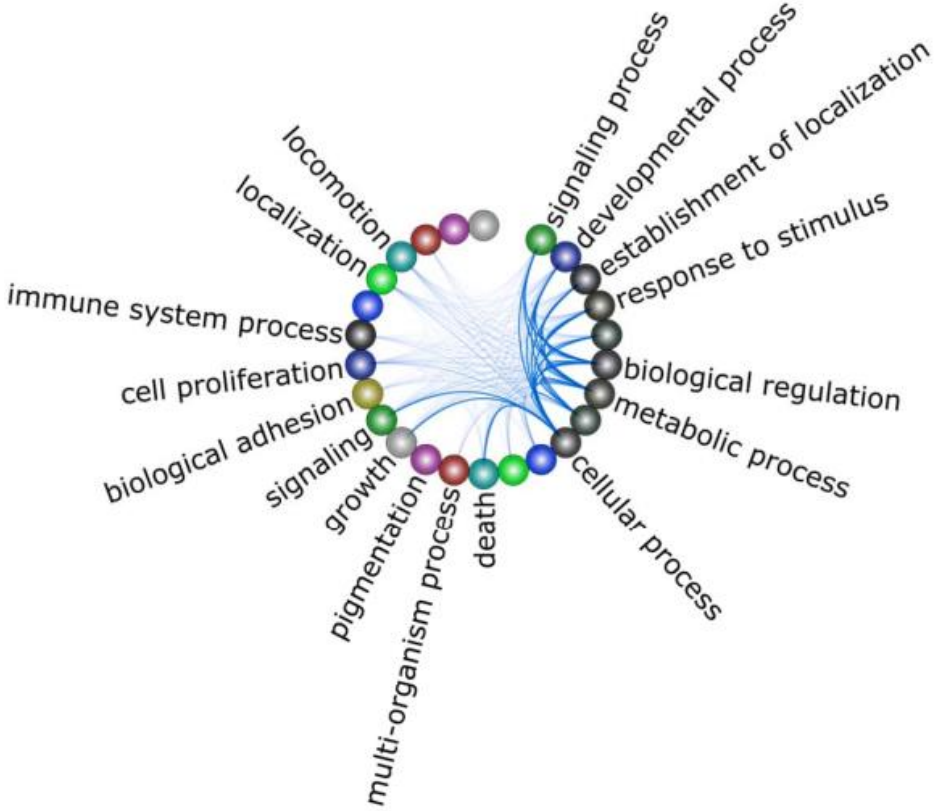
➢Representation

- What is feature
- Feature V.S. model

➢Learning representation

- What is good feature
- How to learn a good feature

1. By hypothesis
2. In action



Thanks