Learning, Sparsity and Big Data

Is the cup half empty or half full
Data compression, Clustering, Regression

M. Magdon-Ismail
Rensselaer Polytechnic Institute

(Joint Work)

November 13, 2013.

The Promise of Big Data

Big Data: Lots of useful information.

The Promise of Big Data

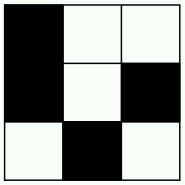
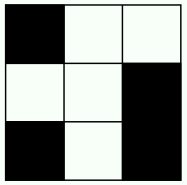
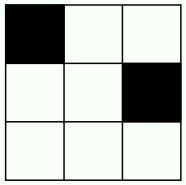
"Waiter! My glass is half empty."

Big Data: Lots of useful information.

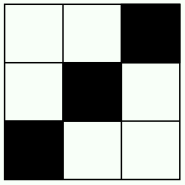
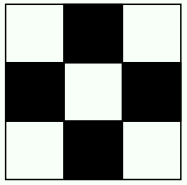
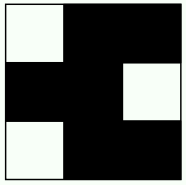
BIG Data: Lots of irrelevent misleading information.

Machine Learning: The art of sorting out the useful from the nonsense and making predictions.

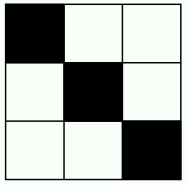
A Visual Prediction Task

NO

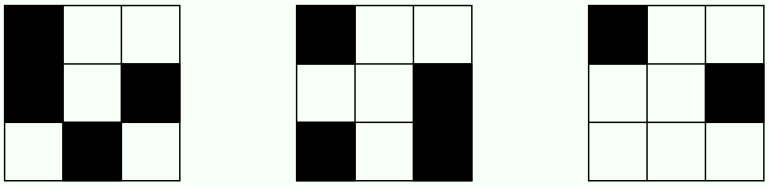




YES

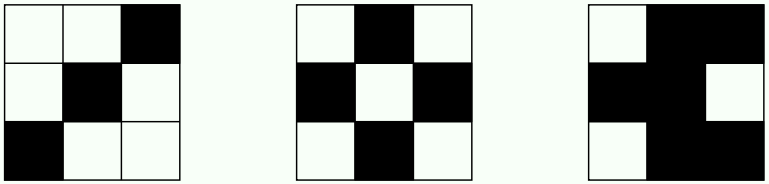


?

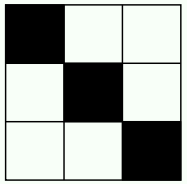
Out-of-Sample is What Counts



NO



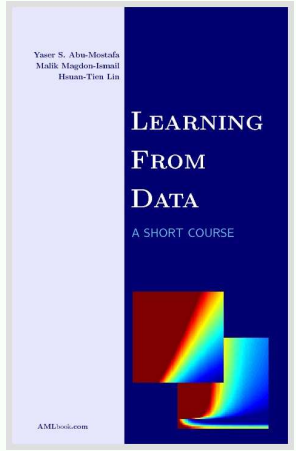
YES



?

- A **pattern** exists
- We **don't know it**
- We **have data** to learn it
- Tested on **new cases**

"Teaching Machine Learning to a Diverse Audience: the Foundation-based Approach," Lin, **M-I**, Abu-Mostafa, *ICML 2012*.



Data

Data Matrix

d dimensions ☹

n data points ☺

	name	age	debt	income	$\cdots$	hair	weight	sex
	$John$	$21yrs$	$-\$10K$	$\$65K$	$\cdots$	$black$	$175lbs$	M
	Joe	$74yrs$	$-\$100K$	$\$25K$	$\cdots$	$blonde$	$275lbs$	M
	$Jane$	$27yrs$	$-\$20K$	$\$85K$	$\cdots$	$blonde$	$135lbs$	F
	$\vdots$							
	Jen	$37yrs$	$-\$400K$	$\$105K$	$\cdots$	$brun$	$155lbs$	F

$$\mathrm{Z} \in \mathbb{R}^{n \times d}$$

Response Matrix

credit?	limit	risk
✓	$2K$	$high$
×	0	–
✓	$10K$	low
$\vdots$		
✓	$15K$	$high$

$$\mathrm{Y} \in \mathbb{R}^{n \times \omega}$$

Visual Data

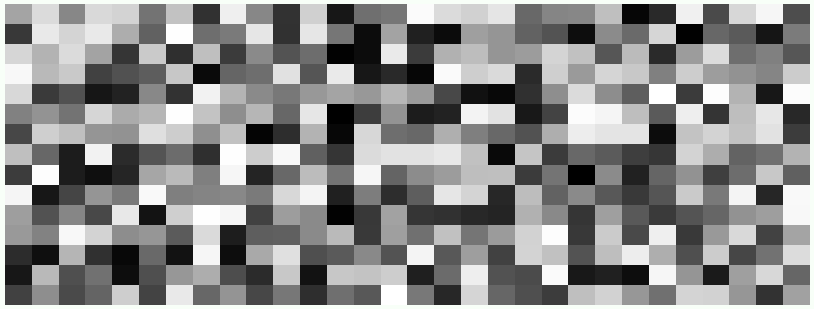
Data Matrix

d dimensions ☹

n data points ☺

	name	age	debt	income	$\cdots$	hair	weight	sex
	John	$21yrs$	$-\$10K$	$\$65K$	$\cdots$	*black*	$175lbs$	M
	Joe	$74yrs$	$-\$100K$	$\$25K$	$\cdots$	*blonde*	$275lbs$	M
	Jane	$27yrs$	$-\$20K$	$\$85K$	$\cdots$	*blonde*	$135lbs$	F
	$\vdots$							
	Jen	$37yrs$	$-\$400K$	$\$105K$	$\cdots$	*brun*	$155lbs$	F

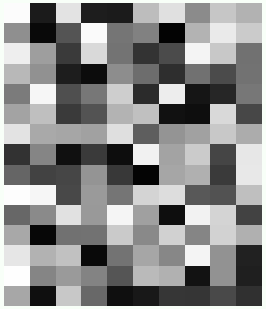
$$\mathrm{Z} \in \mathbb{R}^{n \times d}$$



Response Matrix

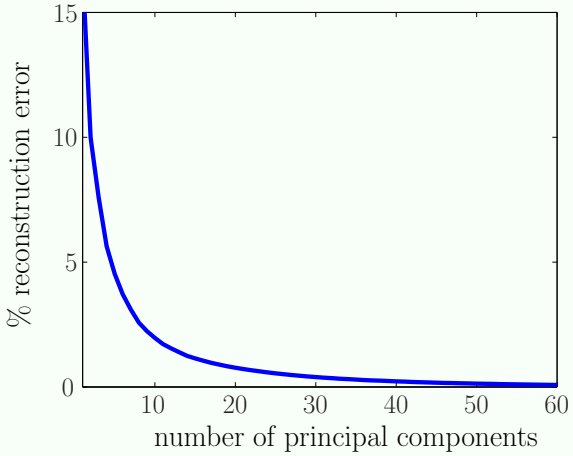
credit?	limit	risk
✓	$2K$	*high*
×	0	–
✓	$10K$	*low*
$\vdots$		
✓	$15K$	*high*

$$\mathrm{Y} \in \mathbb{R}^{n \times \omega}$$

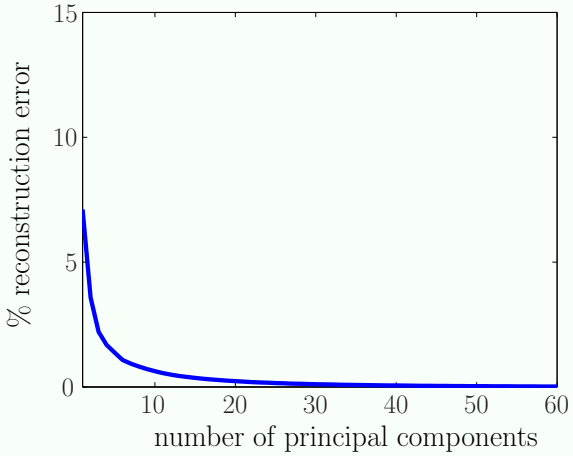


More Beautiful Data

$$Z \in \mathbb{R}^{231\times 174}$$



$$Y \in \mathbb{R}^{231\times 166}$$



Throwing Out Unnecessary Features is Good

'Sparse' solutions generalize to out-of-sample better – less *overfitting*.

Sparse solutions are easier to interpret – few important features.

Computations are more efficient.

Problem: How to find the few relevant features *quickly*.

Compact Data Representation

Principal Components Analysis (PCA): Reconstruct Z using **only a few** degrees of freedom.


Z


Z_{20}


Z_{40}


Z_{60}

Slow.

Compact but does not throw-away irrelevant features.

Fast PCA

$k = 20$ $k = 40$ $k = 60$



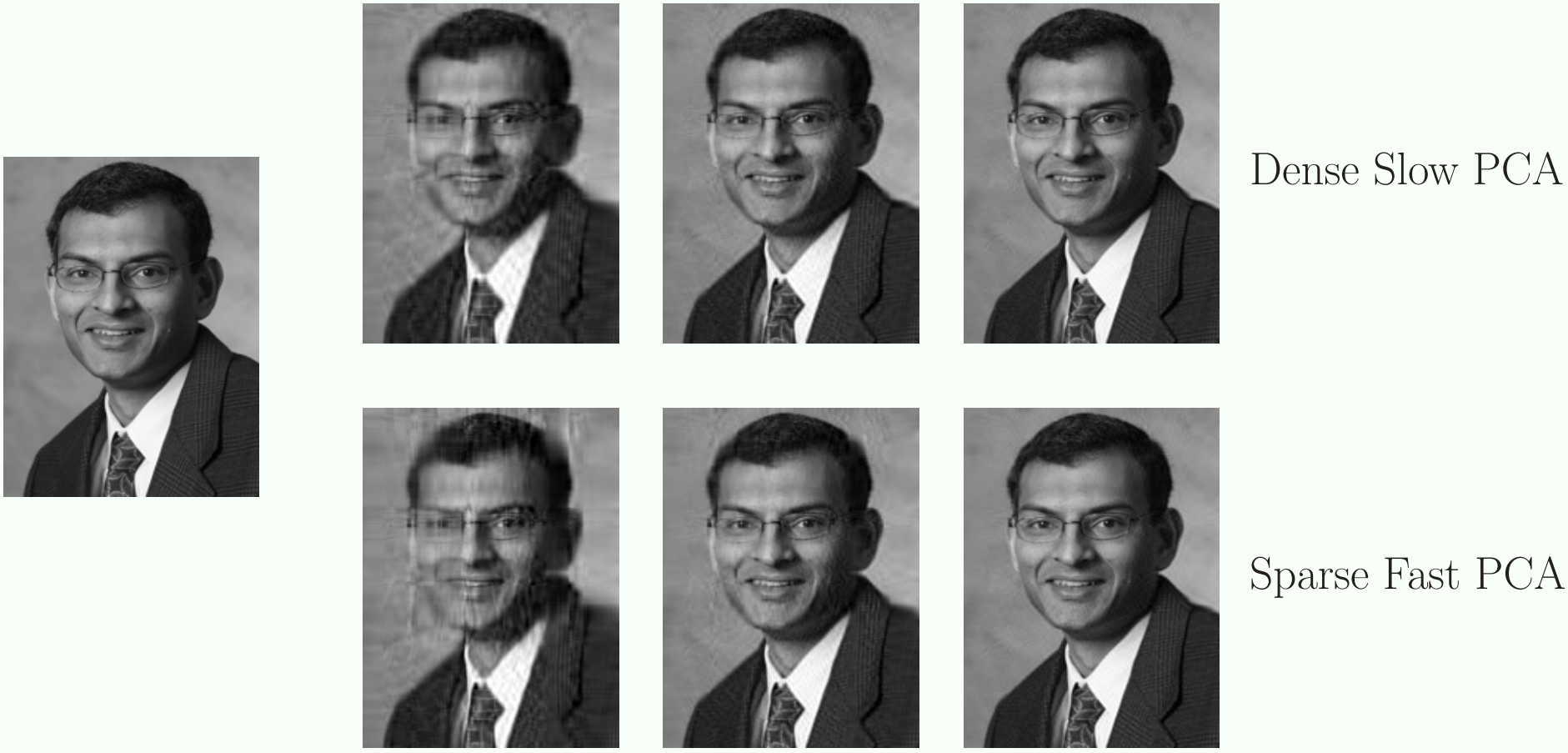
Exact, slow ☹



Approx, fast ☺

Theorem. We can *quickly* do **PCA** (principal components analysis), provably well.

Fast *Sparse* Compact Representations



Theorem. Can quickly find few 'relevant' features for reconstruction.

Clustering: K-Means

Full, slow

Fast, sparse

3 clusters

4 Clusters

Theorem. Can quickly find few important features giving comparable quality clusters.

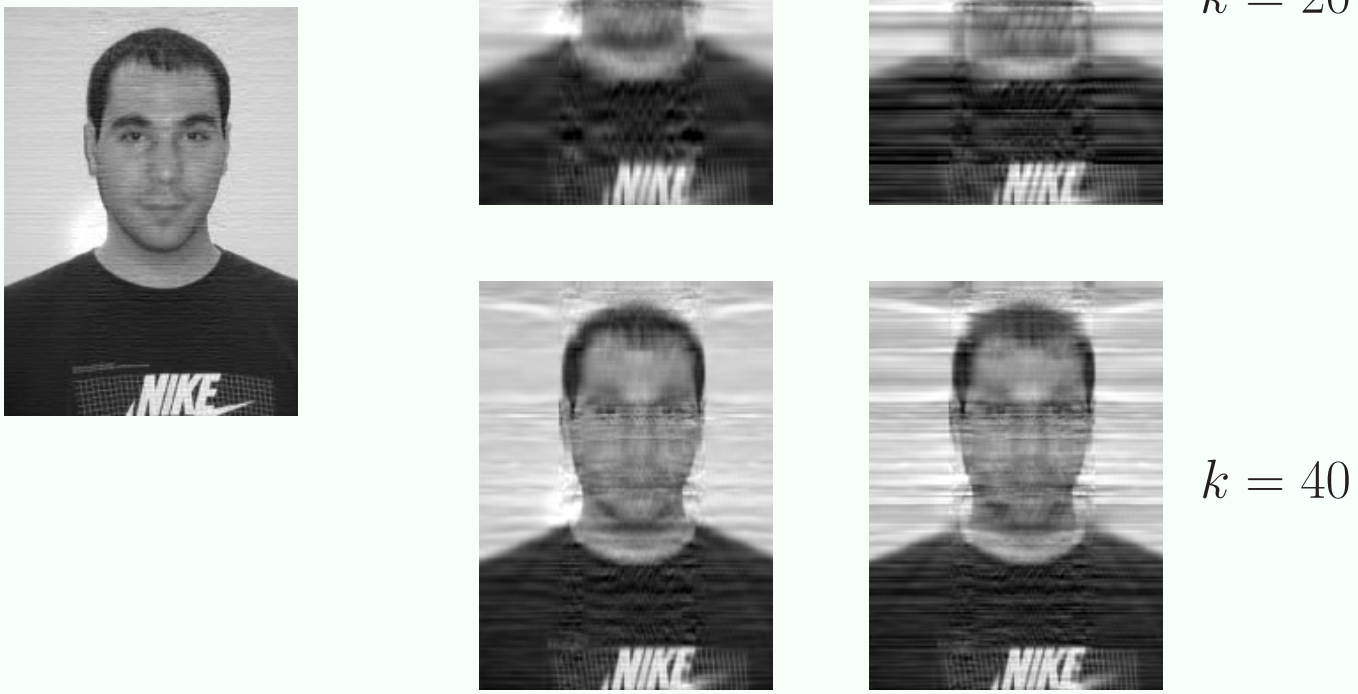
[BDM,2013]

Regression

$$Z \qquad \left[\quad\right] = \qquad Y$$

$$\hat{Y}$$

Fast Regression using Few Important Features



Theorem. Can find few important features which performs as well PCA.

[BDM,2013]

THANKS!

- **Data compression (PCA):**
  quick and reveals few important features

- **Unsupervised clustering:**
  quick and reveals few important features

- **Supervised Regression:**
  quick and reveals few important features

Few features: easy to interpret; better generalizers; faster computations.