INFS4205/7205 Advanced Techniques for High Dimensional Data

When AI Meets High-Dimensional Data

Semester 1, 2021

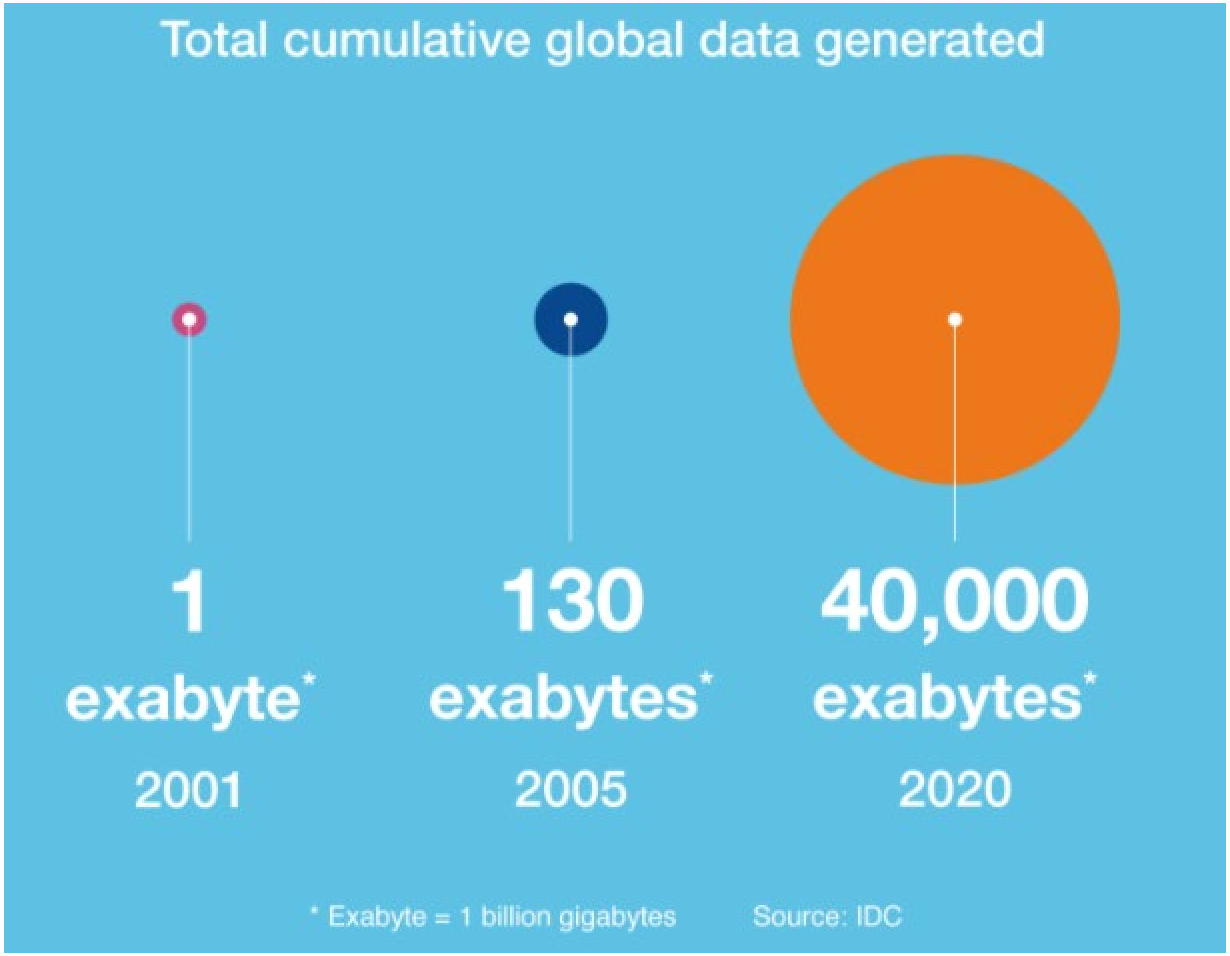
University of Queensland

+ Advanced Techniques for High Dimensional Data

- Course Introduction
- Introduction to Spatial Databases
- Spatial Data Organization
- Spatial Query Processing
- Managing Spatiotemporal Data
- Managing High-dimensional Data
- Introduction to Multimedia Database
- Route Planning in Road Network
- When AI Meets High-Dimensional Data
- Trends and Course Review

+ The Era of Big Data

- Social media
  - Facebook: over 300 million photos/day
  - YouTube: 300 hours of video/min
  - Douyin/TikTok: over 20 million short videos/day
- Commercial
  - Alibaba: 800 million orders for 15 million products, delivering 30 million packages/day
- Science
  - Gene sequencing: human genome with over 3 billion base pairs and 20,000 protein-encoding genes are sequenced in a few hours
- …



$1\text{EB} = 10^{30}$ GB

Besides data storage and retrieval, we need to extract information from data!

+ Classical Way of Data Collection

- Data acquisition process is expensive and time-consuming, so it is carefully designed and controlled
- Only necessary data are collected for a specific task to get only relevant information

Qualify customer expectations before the purchase

- What are you looking for on our site?
- What types of products are you interested in?
- What are your interests?
- For which occasion do you want to buy this type of product? (an anniversary, a farewell party, a wedding, etc.)
- What aroused your interest in our offer?
- What prompted you to visit our website?
- Did you find the products you were looking for easily?
- Did you use the dynamic search engine to do your research? If yes, what did you think? Did you use the dynamic search engine to do your research? If yes, what did you think of it?
- What do you think of our catalogue?

https://www.myfeelback.com/en/blog/questions-to-know-customers-needs-expectations

+ Classical Way of Data Collection

$$\text{Data} \approx \text{Information},$$

+ Modern Way of Data Collection

- Large amount of data are generated every day on Internet, smart phones, gene sequencing…
- We are entering a “data-rich” era

However

- Massive data are collected, without specific purpose in advance
- Data acquisition process is not under control

What is the relationship between data and information in the data-rich era?

+ Example I

- Task: tell the identity of a person

Size: 2500×2500

Size: 250×250

$$\textbf{Data} \gg \textbf{Information}$$

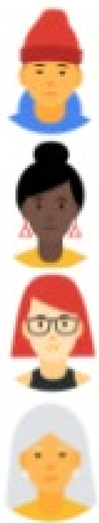
+ Example II

- Task: tell the identity of a person

https://economics.uq.edu.au/article/2018/03/uq-staff-team-deliver-culture-training

$$\text{Data} = \text{Information} + \text{Irrelevant Data}$$

+ Example III



	Harry Potter	The Triplets of Belleville	Shrek	The Dark Knight Rises	Memento
	✓	?	✓	✓	?
	?	✓	?	?	✓
	✓	✓	✓	?	?
	?	?	?	✓	✓

$$\text{Incomplete Data} \approx \text{Complete Information}$$

+ Objective in the Modern Era

- Recover specific information buries in highly redundant, irrelevant, seemingly incomplete high-dimensional data
- Such information is encoded as certain low-dimensional structure underlying the data

Use low-dimensional structures in high-dimensional space to extract information.

Sparsity

+ Example: MR Image

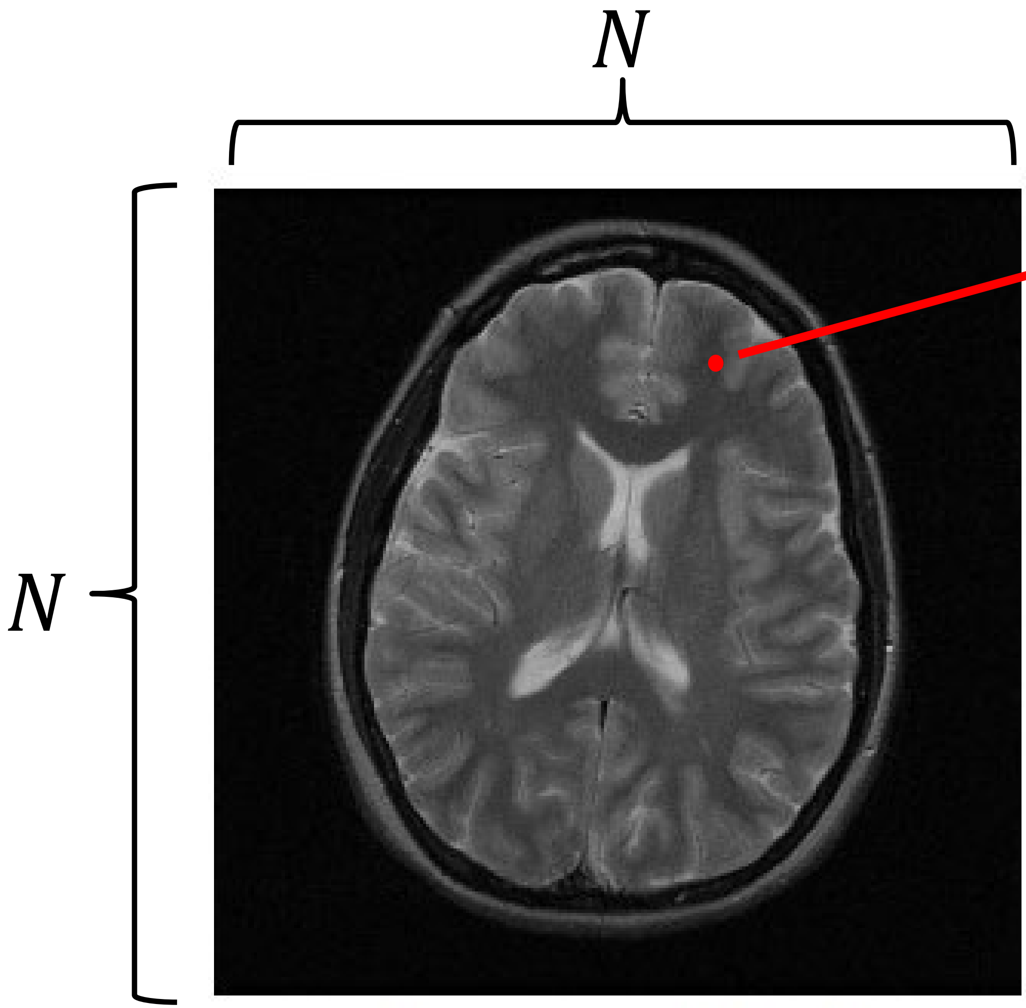


image $I(v)$

- Density of protons at a given spatial location inside the brain
- Patients is subjected to various magnetic fields causing the proton to oscillate at different frequencies
- A signal is collected per frequency, and used to estimate the proton density

+ Example: MR Image

- Number of signals collected: m
- Number of unknows: N^2
- Ideally, we need to collect $m = N^2$ signals to get image I

However,

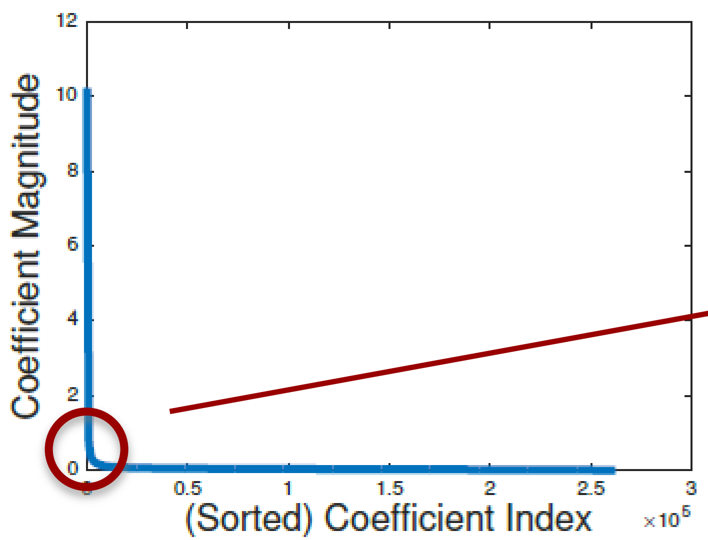
- Applying each frequency takes a long time
  - $m \ll N^2$
- Can we estimate the image I we need with $m \ll N^2$ signals?

+ Example: MR Image

- Wavelet transform of I

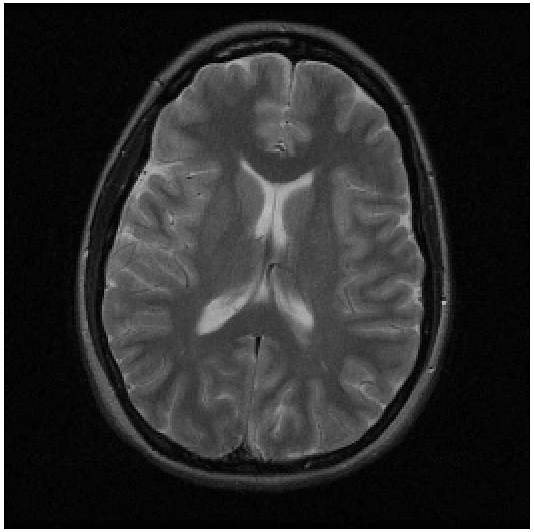
$$\underbrace{I}_{\text{image}} = \sum_{i=1}^{N^2} \underbrace{\psi_i}_{i\text{-th basis signal}} \times \underbrace{x_i}_{i\text{-th coefficient}}.$$

- Magnitude of the coefficients (sorted in descending order)

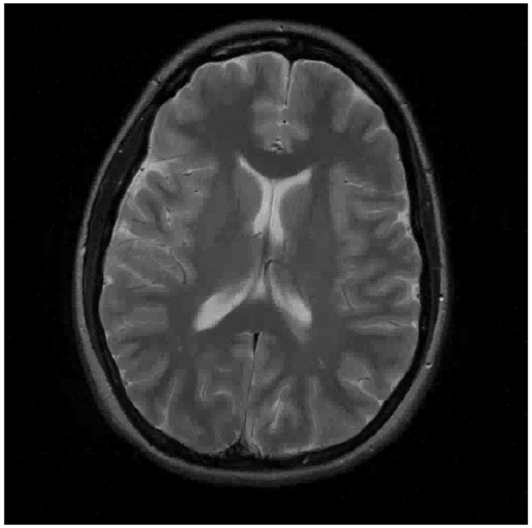


If we retain these small number of coefficients with large magnitudes, we can obtain an accurate approximation of I

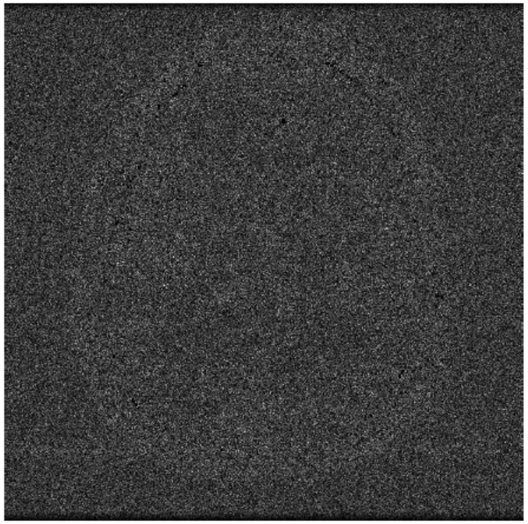
+ Example: MR Image



A: Original image



B: Approximated image using 7% of the coefficients



C = A – B

$$I \approx \sum_{i \in J} \psi_i x_i \quad = \quad \underbrace{\Psi}_{N^2 \times N^2 \text{ matrix}} \quad \underbrace{x}_{\text{sparse vector}},$$

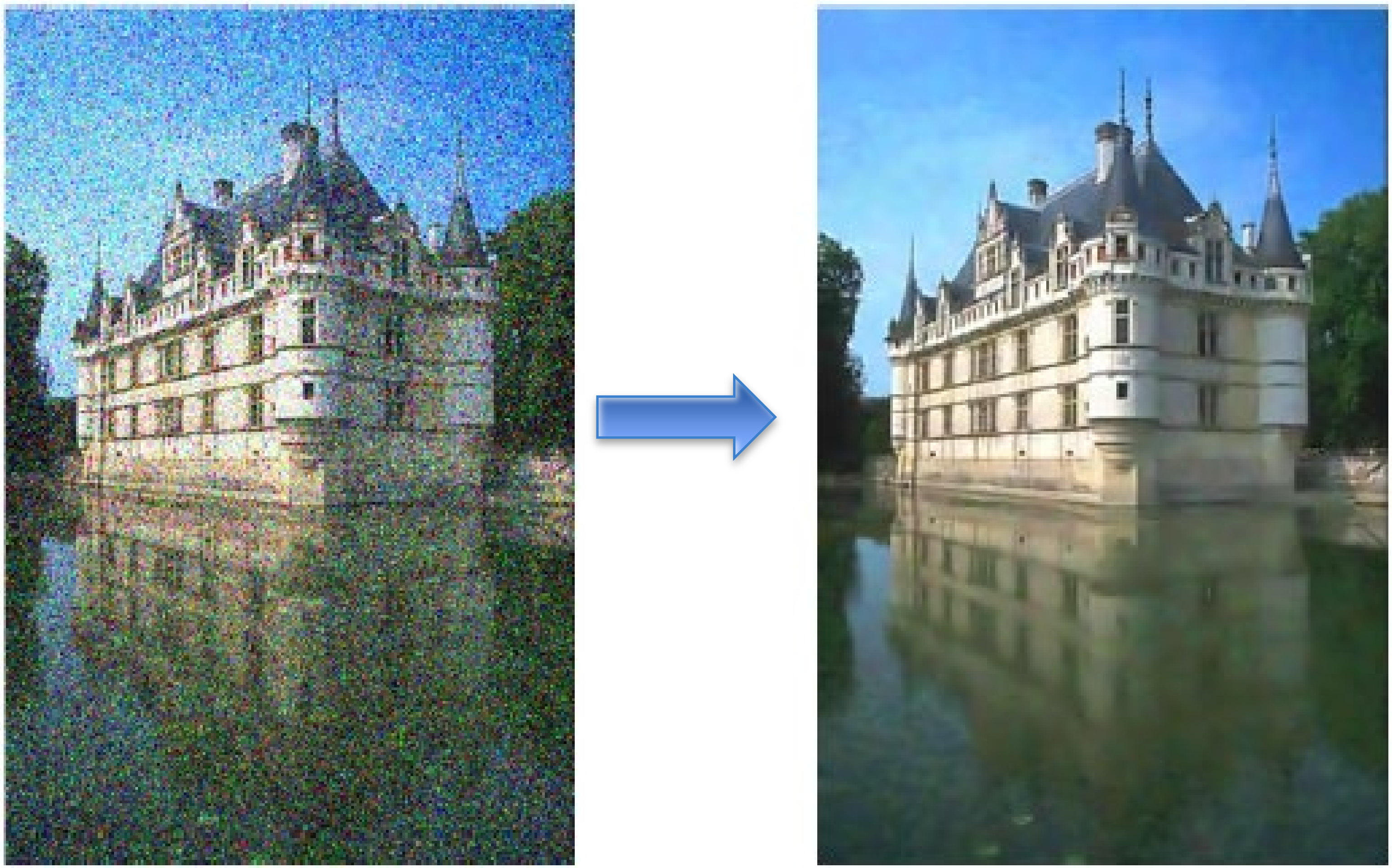
+ Example: Image Denoising

Patch dictionary $\boldsymbol{A}$

$$I_{\text{noisy}} = \underbrace{I_{\text{clean}}}_{\text{target image}} + \underbrace{\boldsymbol{z}}_{\text{sparse error}}.$$

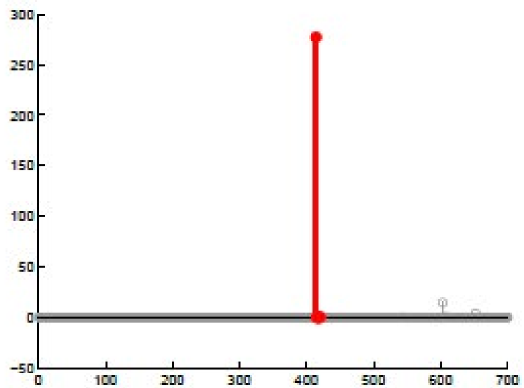
$$\boldsymbol{y}_{i\text{clean}} \approx \underbrace{\boldsymbol{A}}_{\text{patch dictionary}} \times \underbrace{\boldsymbol{x}_i}_{\text{sparse coefficient vector}}.$$

+ Example: Image Denoising



+ Example: Face Recognition

=



Sparse coefficients

×

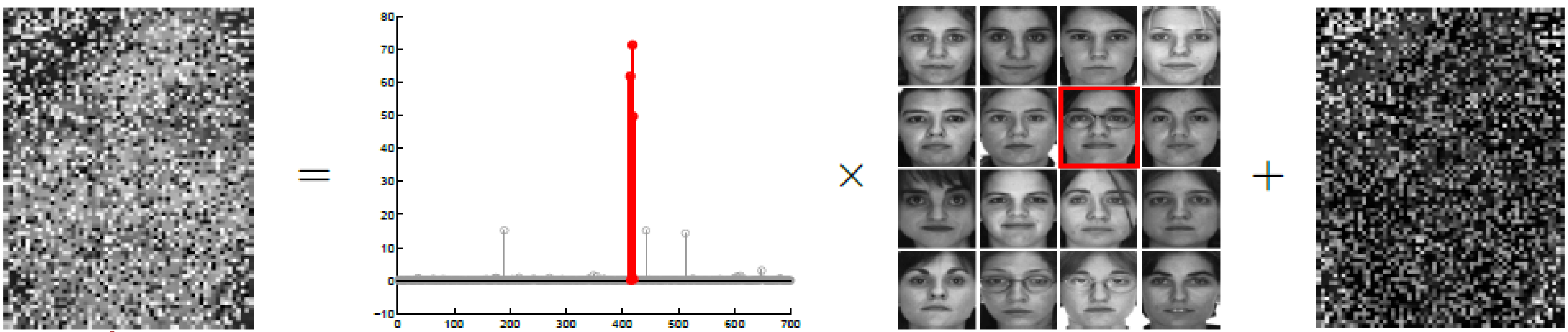
Training faces

+

Sparse noise

A new image taken under some new lighting condition, or occluded

+ Example: Face Recognition



50% pixels arbitrarily corrupted

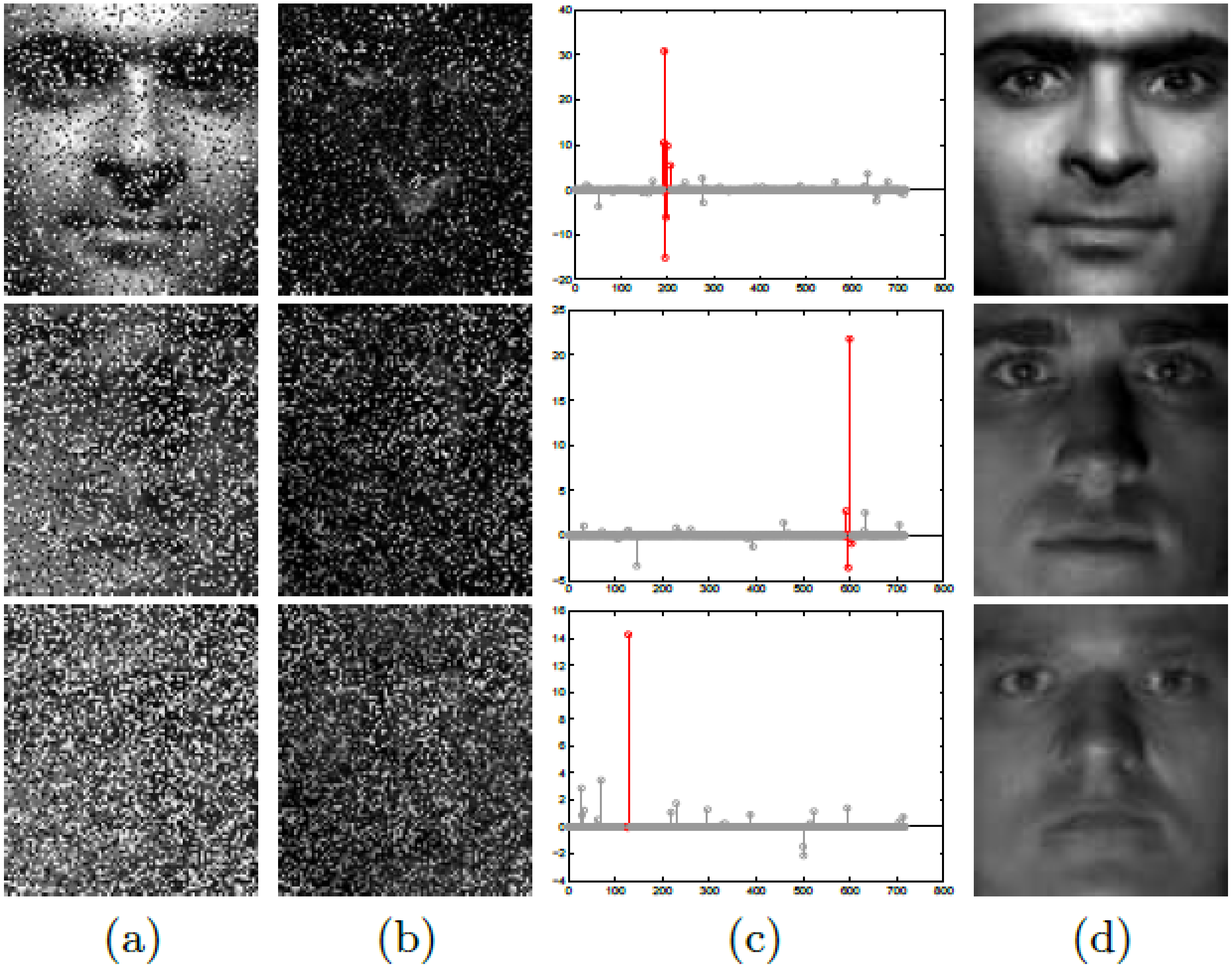
+ Example: Face Recognition

Pixels corrupted

30%

50%

70%



+ L0 norm

- The L0 norm for a vector: the number of nonzero entries in it

$$\|\boldsymbol{x}\|_0 = \#\{i \mid \boldsymbol{x}(i) \neq 0\}$$

- Mathematically, $\boldsymbol{x}$ is sparse whenever $||\boldsymbol{x}||_0$ is small

+ Get a Sparse Solution

- We observe vector $\boldsymbol{y}$
- We have known matrix $\mathbf{A}$
- We know there is a relationship $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}_{\mathbf{0}}$ and $\boldsymbol{x}_{\mathbf{0}}$ is sparse
- We want to calculate the unknown $\boldsymbol{x}_{\mathbf{0}}$, what should we do?

$$\begin{array}{ll} \min & \|\boldsymbol{x}\|_0 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array}$$

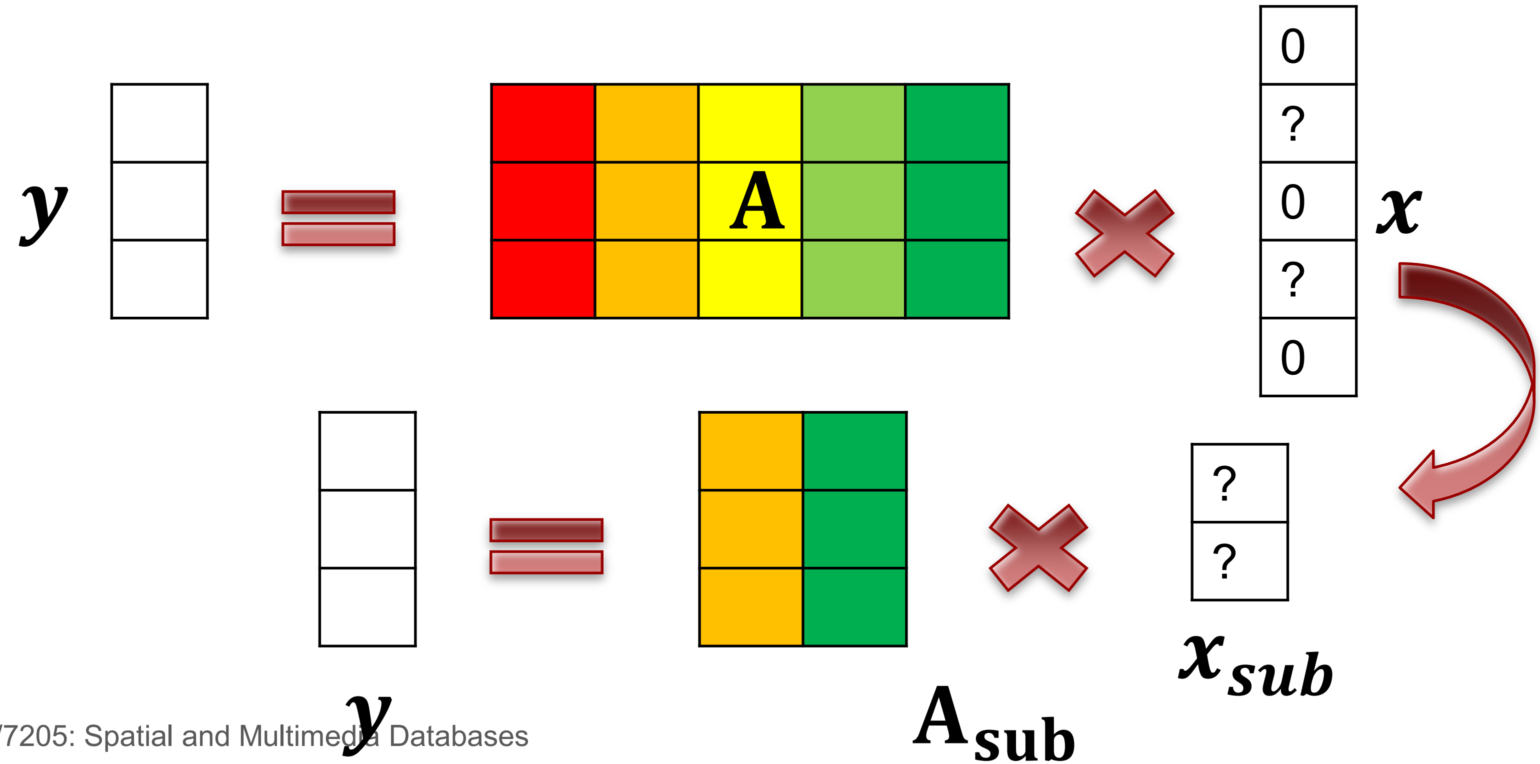
+ Questions about the Optimization

$$\begin{array}{ll} \min & \|\boldsymbol{x}\|_0 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array}$$

- Q1: How can we solve the optimization?
- Q2: If we solve the optimization (i.e., get some solution to the optimization), is it really the $\boldsymbol{x_0}$ we want?
- Q3: Can we solve the optimization efficiently?

+ Q1: A naïve solver-exhaustive search

- For every possible sparsity, try to find a solution
- Example: $\boldsymbol{x}$ is of length 5, and $\boldsymbol{y}$ of length 3
  - All possible sparsity: any subsets of {1,2,3,4,5} to denote nonzero entries' positions
  - The sparsity {2, 4} is $x_1 = 0, x_2 \neq 0, x_3 = 0, x_4 \neq 0, x_5 = 0$

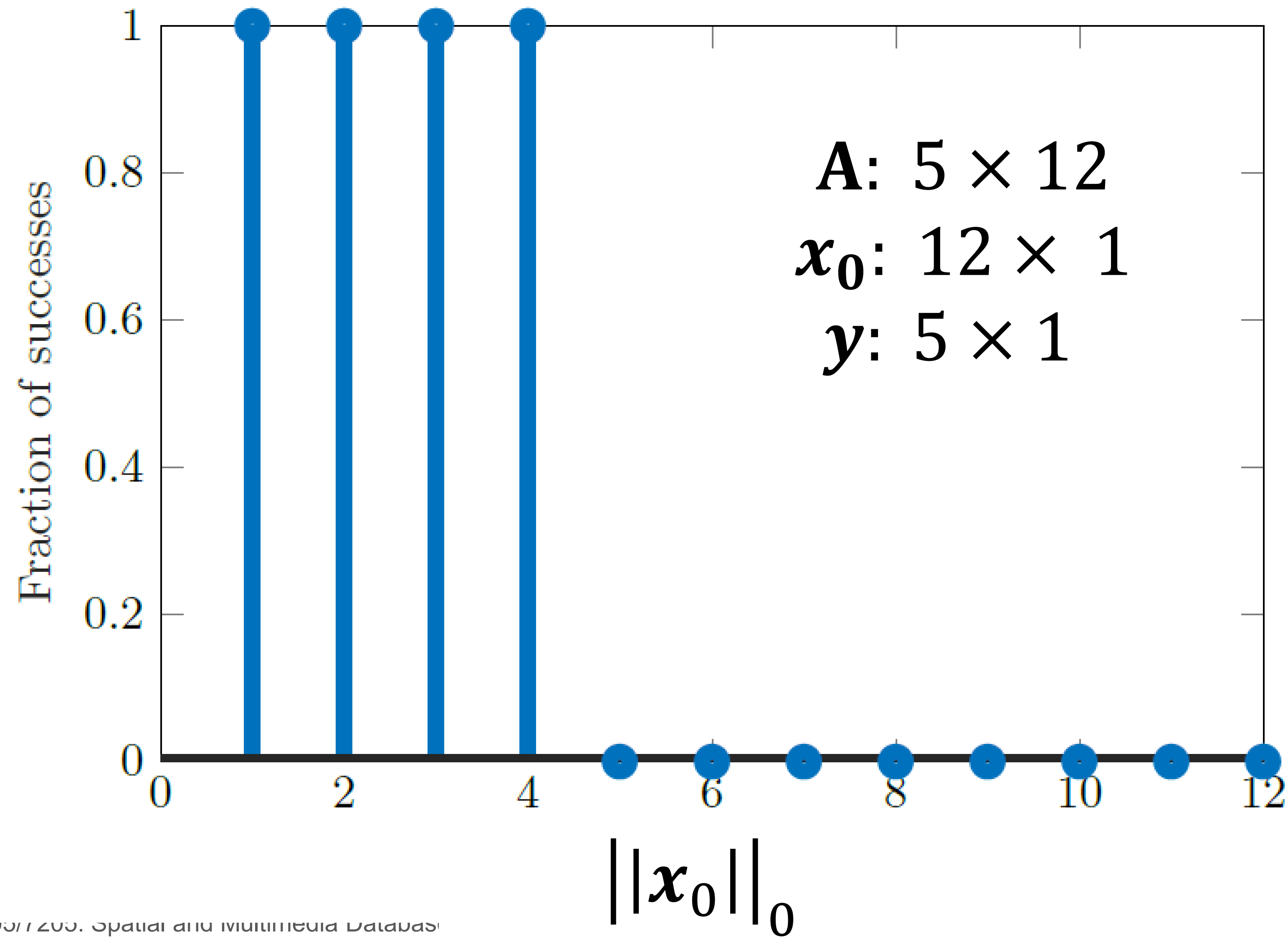


$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} A_{12} & A_{14} \\ A_{22} & A_{24} \\ A_{13} & A_{34} \end{bmatrix}}_{\mathbf{A_{sub}}} \times \underbrace{\begin{bmatrix} x_2? \\ x_4? \end{bmatrix}}_{\boldsymbol{x_{sub}}}$$

$$\begin{cases} y_1 = A_{12}x_2 + A_{14}x_4 \\ y_2 = A_{22}x_2 + A_{24}x_4 \\ y_3 = A_{32}x_2 + A_{34}x_4 \end{cases}$$

- Return the solution if it has one

+ Q2: Is the solution what we want?



+ Theoretical result

Definition of Kruskal Rank *The* Kruskal rank *of a matrix* $\boldsymbol{A}$, *written as* $\operatorname{krank}(\boldsymbol{A})$, *is the largest number r such that every subset of r columns of $\boldsymbol{A}$ is linearly independent.*

Theorem *Suppose that* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$,

$$\|\boldsymbol{x}_o\|_0 \;\le\; \tfrac{1}{2}\operatorname{krank}(\boldsymbol{A}).$$

Then $\boldsymbol{x}_o$ is the unique optimal solution to the ℓ^0 minimization problem

$$\begin{array}{ll} \min & \|\boldsymbol{x}\|_0 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array}$$

For a "generic" **A**, *the Kruskal rank is quite large*

+ Q3: Can we solve the optimization efficiently?

- The exhaustive search approach
  - Worse-case: all non-empty subsets of $\{1, 2, 3, \ldots, n\}$ if $\boldsymbol{x}$ is of length n, can be as large as $2^n - 1$
  - It is estimated that on a standard laptop, if $m = 50, n = 200, ||\boldsymbol{x}_0||_0 = 10$, require 140 centuries for solving

- It can be proved that the problem is NP-hard

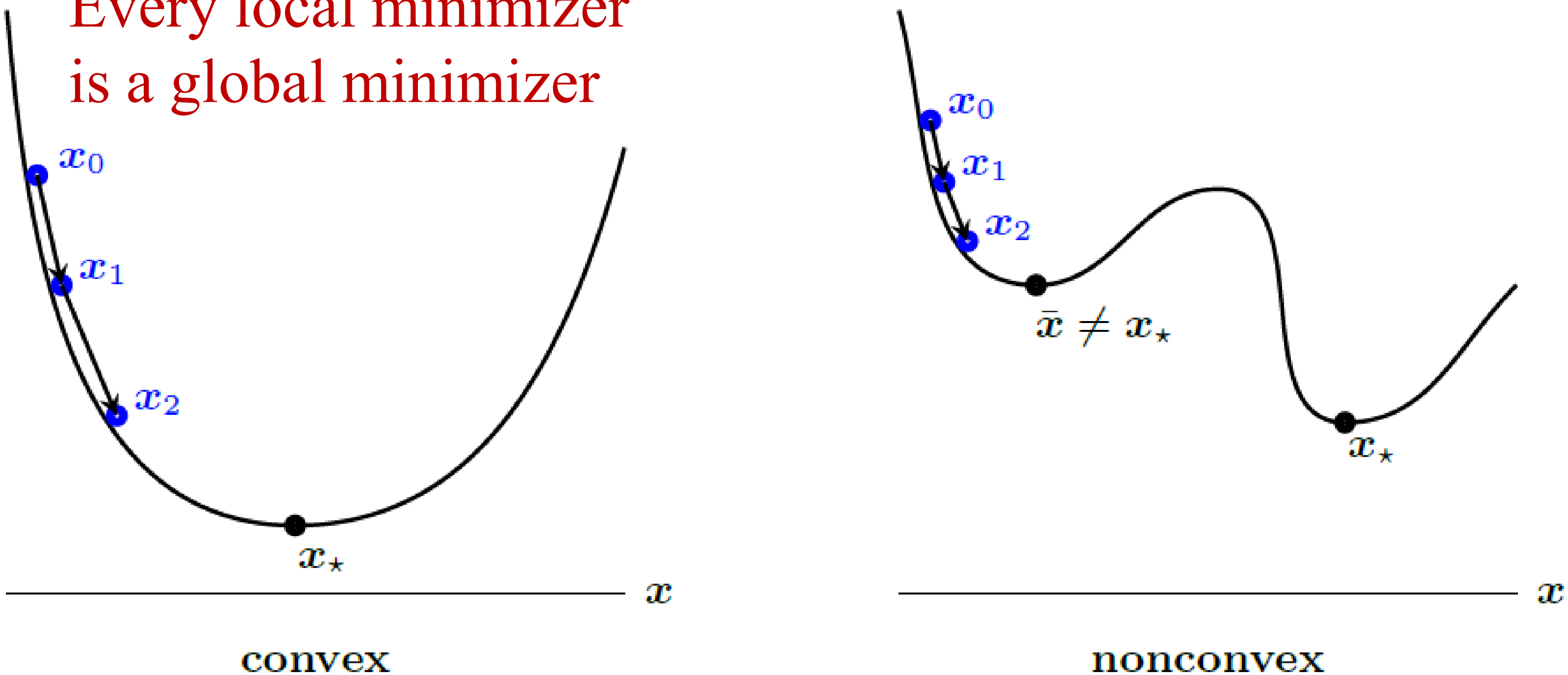
$$\begin{array}{ll} \min & \|\boldsymbol{x}\|_0 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array}$$

Win one million dollars if you can have a formal proof that P=NP or P≠ NP, refer to http://www.claymath.org/sites/default/files/millennium_prize_rules.pdf

Even if we cannot solve the problem efficiently in the worse-case, there may be a subclass of instances of interest which can be solved efficiently!

+ Convex function and optimization

Every local minimizer is a global minimizer

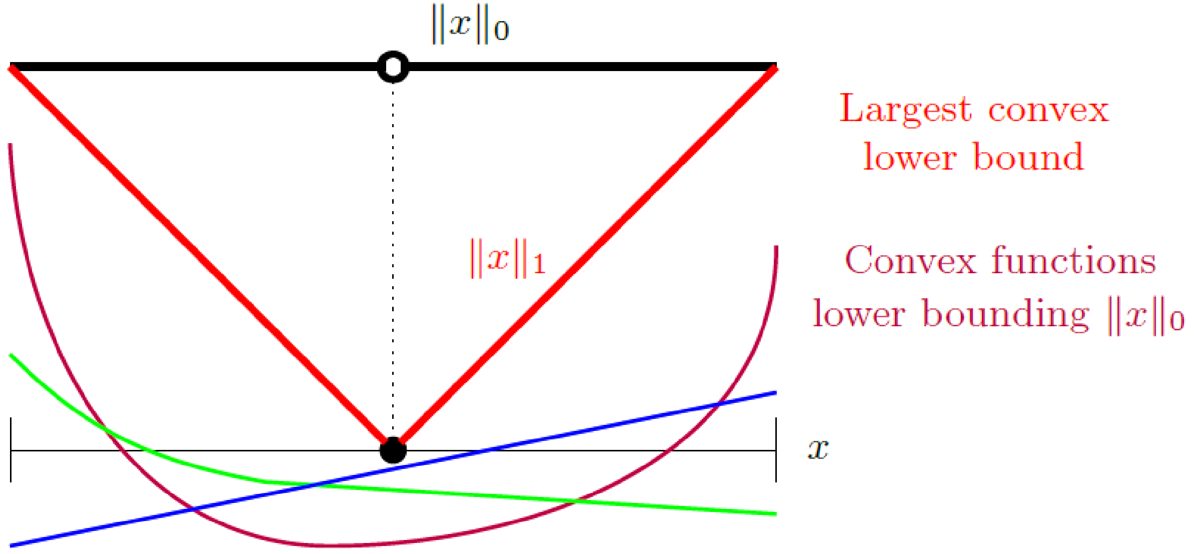


Iterative methods for finding the minimizer of a convex function:
starting from x_0 and around x_0, find a new point making the objective function decreases

$$x_1 = x_0 - t\nabla f(x_0)$$

+ The convex surrogate for L0 norm

- L0 norm $$\|\boldsymbol{x}\|_0 = \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{x}(i)\neq 0}.$$
- L1 norm $$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |\boldsymbol{x}(i)|$$
- L1 norm is the largest convex lower bound of L0 norm



+ The relaxed optimization

- Using L1 norm, the optimization is relaxed to

$$\begin{array}{ll} \min & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \end{array}$$

- It can be optimized efficiently, refer to Sec. 2.3.3. (and the reference there) of the reference for details

When will optimizing L1 norm gives us a solution we want?

+ When will optimizing L1 norm work?

- $\boldsymbol{Ax}$ can preserve the norm of the sparse vector $\boldsymbol{x}$, formally as "Restricted Isometry Property (RIP)"

$$\forall\, \boldsymbol{x} \;\; k\text{-}sparse, \quad (1-\delta)\,\|\boldsymbol{x}\|_2^2 \;\leq\; \|\boldsymbol{Ax}\|_2^2 \;\leq\; (1+\delta)\,\|\boldsymbol{x}\|_2^2$$

- The inequality definite hold if δ is quite large
- $\delta_k(\boldsymbol{A})$ is the smallest number from $[0,1)$ the above inequality holds

THEOREM (ℓ^1 Recovery under RIP). *Suppose that* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, *with* $k = \|\boldsymbol{x}_o\|_0$. *If* $\delta_{2k}(\boldsymbol{A}) < \sqrt{2} - 1$, *then* $\boldsymbol{x}_o$ *is the unique optimal solution to*

If $\delta_{2k}(\boldsymbol{A})$ <1, then optimizing L0 norm works

$$\begin{array}{ll} \min & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \boldsymbol{Ax} = \boldsymbol{y}. \end{array}$$

+ When will $\boldsymbol{A}$ be RIP?

THEOREM (RIP of Gaussian Matrices). *There exists a numerical constant $C > 0$ such that if $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a random matrix with entries independent $\mathcal{N}\left(0, \frac{1}{m}\right)$ random variables, with high probability, $\delta_k(\boldsymbol{A}) < \delta$, provided*

$$m \geq Ck \log(n/k)/\delta^2.$$

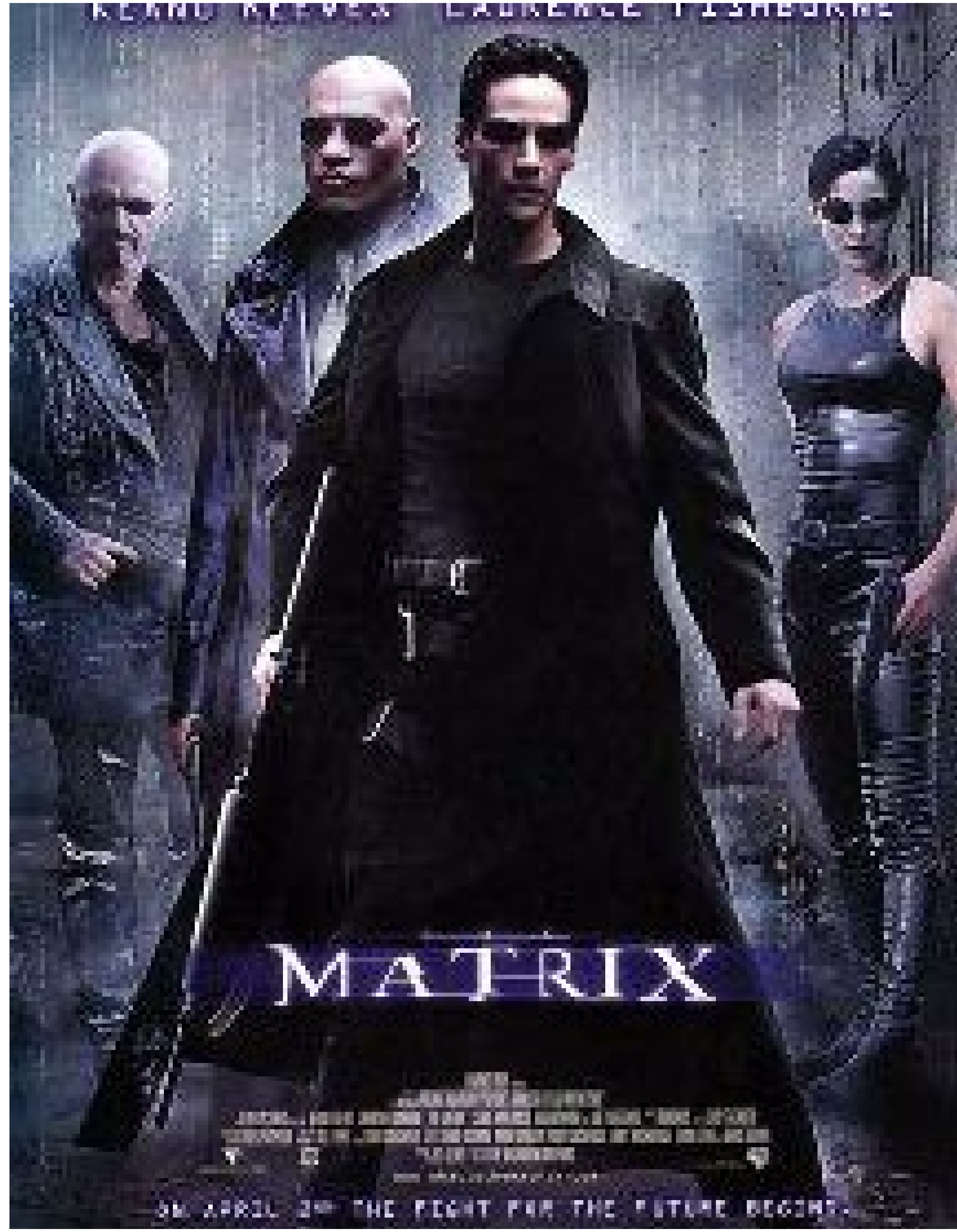
It gives some insights that

- If $\boldsymbol{x}_0$ has k non-zero elements, then $\boldsymbol{y}$ of length $O(k)$ can lead to recovery of the $\boldsymbol{x}_0$

Low rankness

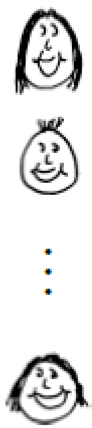
“The matrix is everywhere. It is all around us.”

- Morpheus from movie *The Matrix*

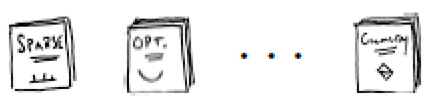


+ Example: Online Recommendation Systems

Users



$$\begin{bmatrix} 5 & 3 & \ldots & ? \\ ? & 2 & \ldots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 5 & ? & \ldots & ? \end{bmatrix} = \mathcal{P}_\Omega \left(\begin{bmatrix} 5 & 3 & \ldots & 5 \\ 4 & 2 & \ldots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 5 & 5 & \ldots & 3 \end{bmatrix} \right)$$



Items

Observed (Incomplete) Ratings $\boldsymbol{Y}$

Complete Ratings $\boldsymbol{X}$

$$\Omega \doteq \left\{ (i,j) \mid \text{user } i \text{ has rated product } j \right\}$$

$$\mathcal{P}_\Omega[\boldsymbol{X}](i,j) = \begin{cases} X_{ij} & (i,j) \in \Omega, \\ 0 & \text{else.} \end{cases}$$

$$\underbrace{\boldsymbol{Y}}_{\text{Observed ratings}} = \mathcal{P}_\Omega \left[\underbrace{\boldsymbol{X}}_{\text{Complete ratings}} \right]$$

+ Example: Online Recommendation Systems

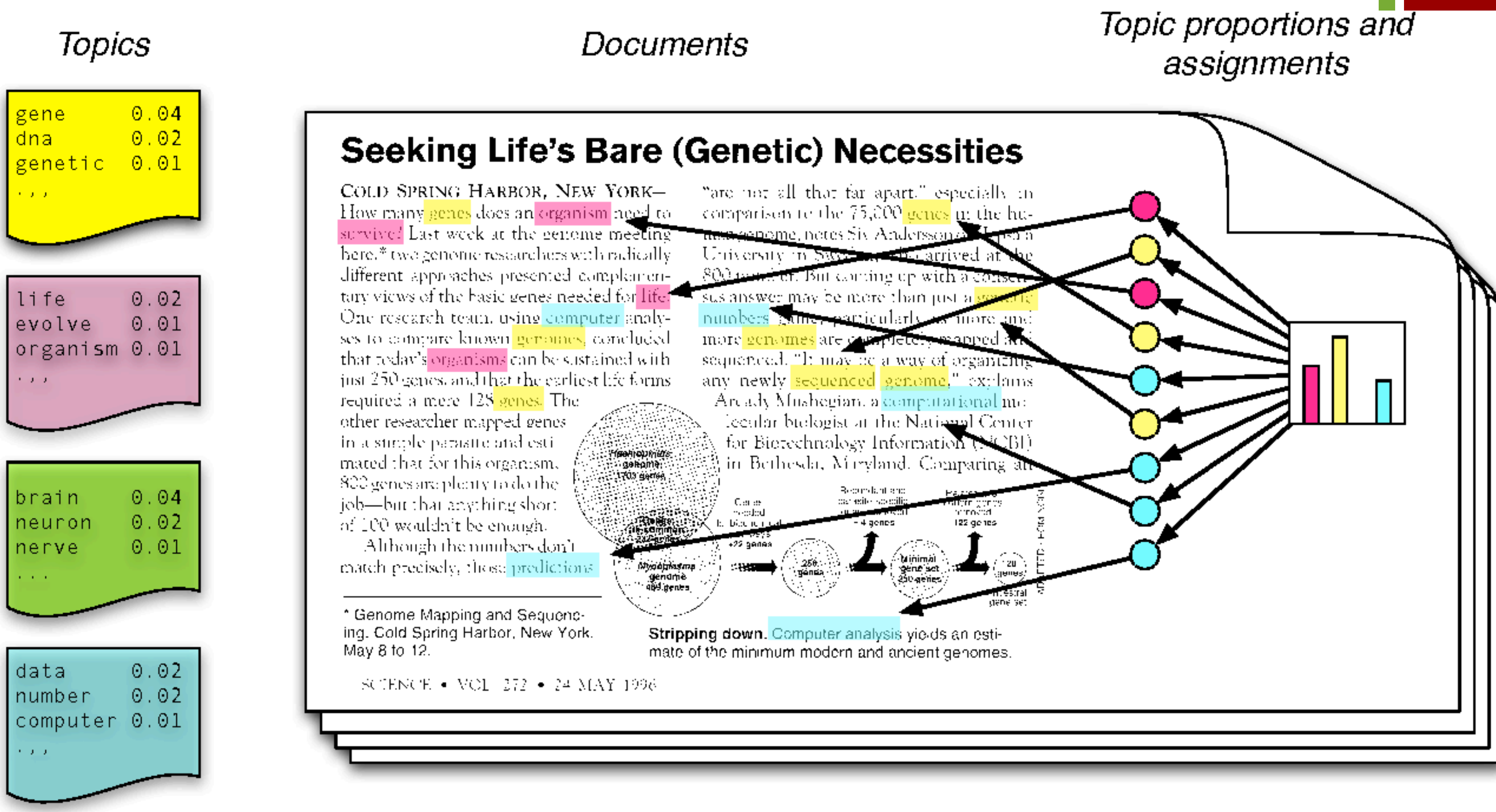
- Popular assumption: ratings of users or products are correlated
  - The rating matrix is determined by a few factors
  - The rating matrix is low-rank

+ Example: Topic Model

	Word 1	Word 2	…	Word n_1-1	Word n_1
Document 1					
Document 2					
Document 3					
…					
Document n_2					

- **Y**: Y_{ij} is the fraction of occurrences of word i in document j
- Assume there are r topics, $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_r$, each is a probability distribution on $\{1, 2, \ldots, n_1\}$

+ Example: Topic Model

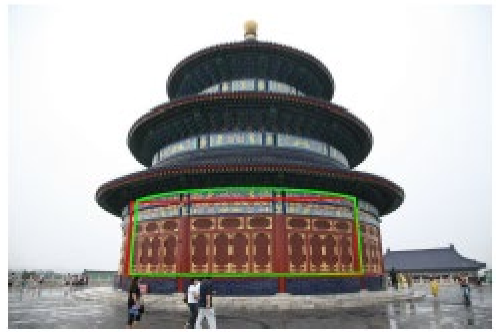


+ Example: Topic Model

	Word 1	Word 2	…	Word n_1-1	Word n_1
Document 1					
Document 2					
Document 3					
…					
Document n_2					

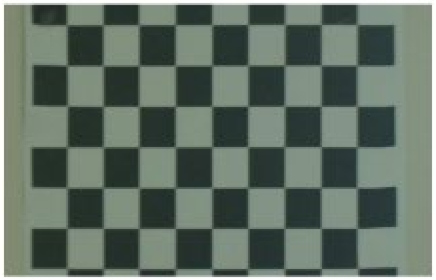
- **Y**: Y_{ij} is the fraction of occurrences of word i in document j
- Assume there are r topics, $\boldsymbol{t}_1, \dots, \boldsymbol{t}_r$, each is a probability distribution on $\{1, 2, \dots, n_1\}, r \ll n_1, n_2$
- **Y** is determined by a few topics
  - **Y** is low rank

+ Example: Texture Images



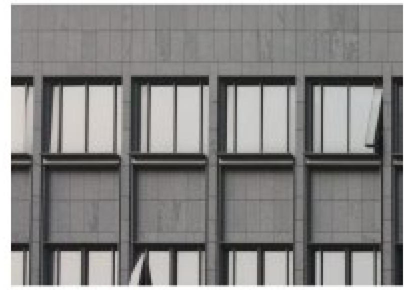
- The image **I** is composed of regular patterns
- **I** is determined by a few little patches
    - **I** is low-rank

+ Example: Texture Images



(a) a calibration rig

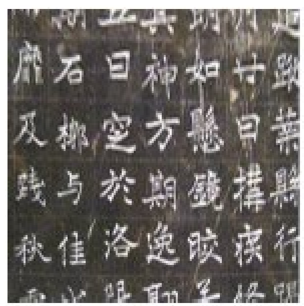
(b) a carpet



(c) windows

(d) a door

(e) a license plate



(f) characters

(g) a car

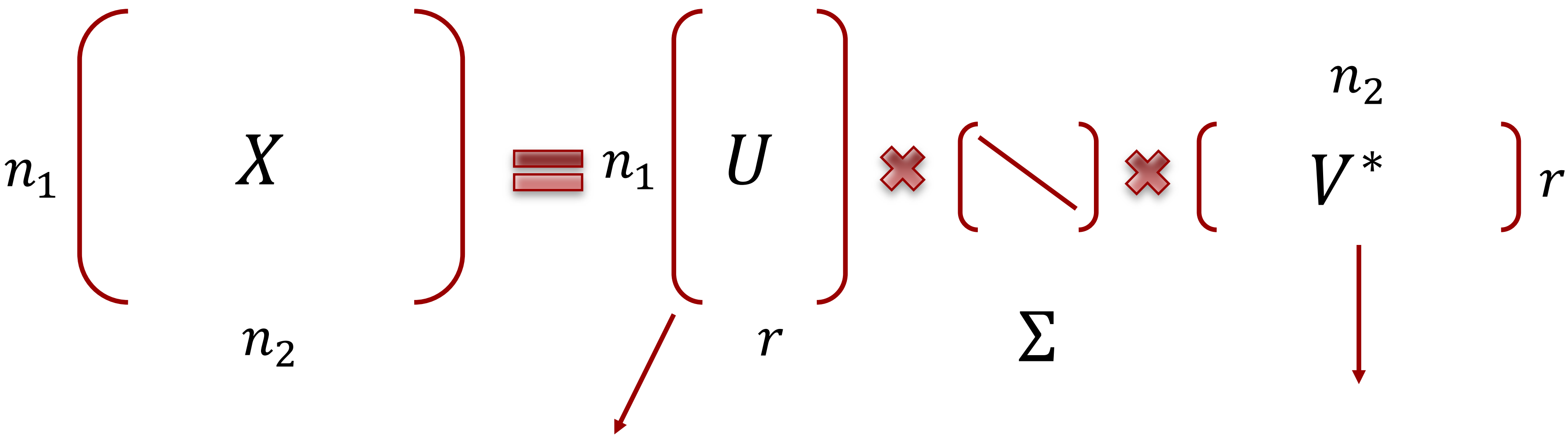
(h) a face

These images viewed as matrices are all low-rank or approximately low-rank

+ Representing a Low-Rank Matrix via SVD

$$X = U\Sigma V^* = \sum_{i=1}^{r} \sigma_i u_i v_i^*.$$

$$\sigma_1 \geq \sigma_2 \geq \cdots$$



Each column has L2 norm 1 and columns are orthogonal to each other ($UU^* = I$)

Each row has length 1 and rows are orthogonal to each other ($V^*V = I$)

+ Rank and L0 norm

$$\begin{array}{c} \\ n_1 \begin{pmatrix} & & \\ & X & \\ & & \end{pmatrix} \\ n_2 \end{array} = \begin{array}{c} \\ n_1 \begin{pmatrix} \\ U \\ \\ \end{pmatrix} \\ r \end{array} \times \begin{pmatrix} \diagdown \end{pmatrix} \times \begin{array}{c} n_2 \\ \begin{pmatrix} & V^* & \end{pmatrix} r \\ \end{array}$$

$$\text{diag}(\Sigma) = \boldsymbol{\sigma}$$

$$\text{rank}(\boldsymbol{X}) = \|\boldsymbol{\sigma}(\boldsymbol{X})\|_0$$

+ Best Low-rank Approximation

- If we know a matrix $\mathbf{Y}$ of size $n_1 \times n_2$, whose rank is larger than r
- We want to find rank-r matrix $\mathbf{X}$, which can best approximate $\mathbf{Y}$

$$\begin{array}{ll} \min & \|\boldsymbol{X} - \boldsymbol{Y}\|_F\,, \\ \text{subject to} & \operatorname{rank}(\boldsymbol{X}) \le r. \end{array}$$

- Full singular value decomposition of $\mathbf{Y}$

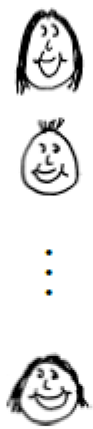
$$\boldsymbol{Y} = \sum_{i=1}^{\min(n_1, n_2)} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*$$

- The optimal solution is

$$\hat{\boldsymbol{X}} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*;$$

+ Recovering a Matrix

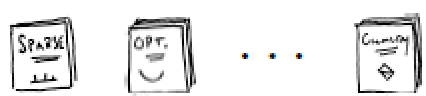
- Reconsider the recommendation problem



Users

$$\begin{bmatrix} 5 & 3 & \cdots & ? \\ ? & 2 & \cdots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 5 & ? & \cdots & ? \end{bmatrix} = \mathcal{P}_{\Omega}\left(\begin{bmatrix} 5 & 3 & \cdots & 5 \\ 4 & 2 & \cdots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 5 & 5 & \cdots & 3 \end{bmatrix} \right)$$

Complete Ratings $\boldsymbol{X}$



Items

Observed (Incomplete) Ratings $\boldsymbol{Y}$

- Without assumption, we can have arbitrary ways to fill the matrix at those ?

+ Recover a Low-rank Matrix

- With a low-rank assumption

$$\underbrace{\begin{array}{cc} & n_2 \\ n_1 & \begin{pmatrix} \quad X \quad \end{pmatrix} \end{array}}_{n_1 n_2 \text{ entries}} = \underbrace{n_1 \begin{pmatrix} U \end{pmatrix} \times \underset{\Sigma}{\begin{pmatrix} \ddots \end{pmatrix}} \times \begin{array}{cc} n_2 & \\ \begin{pmatrix} \quad V^* \quad \end{pmatrix} & r \end{array}}_{n_1 r + n_2 r + r = O(\max(n_1, n_2)) \text{ entries}}$$

$n_1 r + n_2 r + r = O(\max(n_1, n_2))$ entries
if $r \ll n_1, n_2$

Intuitively, even if X is quite large, observing only a small part of it can still enable recovery of the whole matrix, due to that the matrix is determined by a few factors!

+ The Optimization of Rank

- A matrix recovery problem:
  - The target matrix (unknown) X
  - The observed matrix Y

$$\begin{array}{ll} \min & \operatorname{rank}(\boldsymbol{X}), \\ \text{subject to} & \mathcal{P}_{\Omega}[\boldsymbol{X}] = \boldsymbol{Y}. \end{array}$$

- More generally,

$$\begin{array}{ll} \min & \operatorname{rank}(\boldsymbol{X}), \\ \text{subject to} & \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \end{array}$$

$\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a linear map.

+ The Optimization of Rank

- L0 norm of the singular values $\operatorname{diag}(\Sigma) = \boldsymbol{\sigma}$

$$\begin{array}{ll} \min & \operatorname{rank}(\boldsymbol{X}), \\ \text{subject to} & \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \end{array} \Rightarrow \begin{array}{ll} \min & \|\boldsymbol{\sigma}(\boldsymbol{X})\|_0 \\ \text{subject to} & \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \end{array}$$

- If X is a diagonal matrix, it is the same as the following NP-hard problem

$$\begin{array}{ll} \min & \|\boldsymbol{x}\|_0 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array}$$

- In the worse case the matrix recovery problem is at least NP-hard

+ Convex Relaxation of Rank Minimization

- L1 norm of the singular values $\mathrm{diag}(\Sigma) = \boldsymbol{\sigma}$

$$\|\boldsymbol{\sigma}(\boldsymbol{X})\|_1 = \sum_i \sigma_i(\boldsymbol{X}).$$

- The sum of singular values is called **nuclear norm** or **trace norm** of a matrix

$$\|\boldsymbol{X}\|_* = \sum_i \sigma_i(\boldsymbol{X}).$$

- Optimization with nuclear norm

$$\begin{array}{ll} \min & \|\boldsymbol{X}\|_*, \\ \text{subject to} & \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \end{array}$$

- Q1: How can we solve the optimization efficiently?
- Q2: If we solve the optimization (i.e., get some solution to the optimization), is it really the X we want?

+ Q1. How can we solve the optimization

- A property of nuclear norm

$$\|X\|_* = \min_{U,V} \tfrac{1}{2}(\|U\|_F^2 + \|V\|_F^2),\ s.t.\ X = UV^*$$

- Then the problem can be equivalently written as

$$\begin{array}{ll} \min & \|X\|_*, \\ \text{subject to} & \mathcal{A}[X] = y. \end{array} \Rightarrow \begin{array}{l} \min\limits_{U,V} \frac{1}{2} ||U||_F^2 + ||V||_F^2 \\ \text{subject to } A[UV^*] = y \end{array}$$

- This is called "matrix factorization"

- It is a non-convex optimization, but it is proved that under some conditions, the global optimum can be achieved

Prateek Jain, Praneeth Netrapalli, Sujay Sanghavi. Low-rank matrix completion using alternating minimization. STOC 2013: 665-674

+ Q2: Is the solution what we want?

- Rank-Restricted Isometry Property

$$(1-\delta)\|\boldsymbol{X}\|_F^2 \;\leq\; \|\mathcal{A}[\boldsymbol{X}]\|_2^2 \;\leq\; (1+\delta)\|\boldsymbol{X}\|_F^2$$

- Theory

THEOREM (Nuclear Norm Minimization). *Suppose that* $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_o]$ *with* $\operatorname{rank}(\boldsymbol{X}_o) \leq r$, *and that* $\delta_{4r}(\mathcal{A}) \leq \sqrt{2}-1$. *Then* $\boldsymbol{X}_o$ *is the unique optimal solution to the nuclear norm minimization problem*

$$\begin{array}{ll} \min & \|\boldsymbol{X}\|_* \\ \text{subject to} & \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \end{array}$$

+ Recovery from noisy observation

- If the observation is corrupted by as small additive noise

$$\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_o] + \boldsymbol{z}, \quad \|\boldsymbol{z}\|_2 \leq \varepsilon.$$

- The optimization becomes

$$\begin{array}{ll} \min & \|\boldsymbol{X}\|_* \\ \text{subject to} & \|\mathcal{A}[\boldsymbol{X}] - \boldsymbol{y}\|_2 \leq \varepsilon. \end{array}$$

The Netflix prize winner

- Winner: **BellKor's Pragmatic Chaos**

The Pragmatic Theory solution to the Netflix Grand Prize

https://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf

The BellKor Solution to the Netflix Grand Prize

Yehuda Koren
August 2009

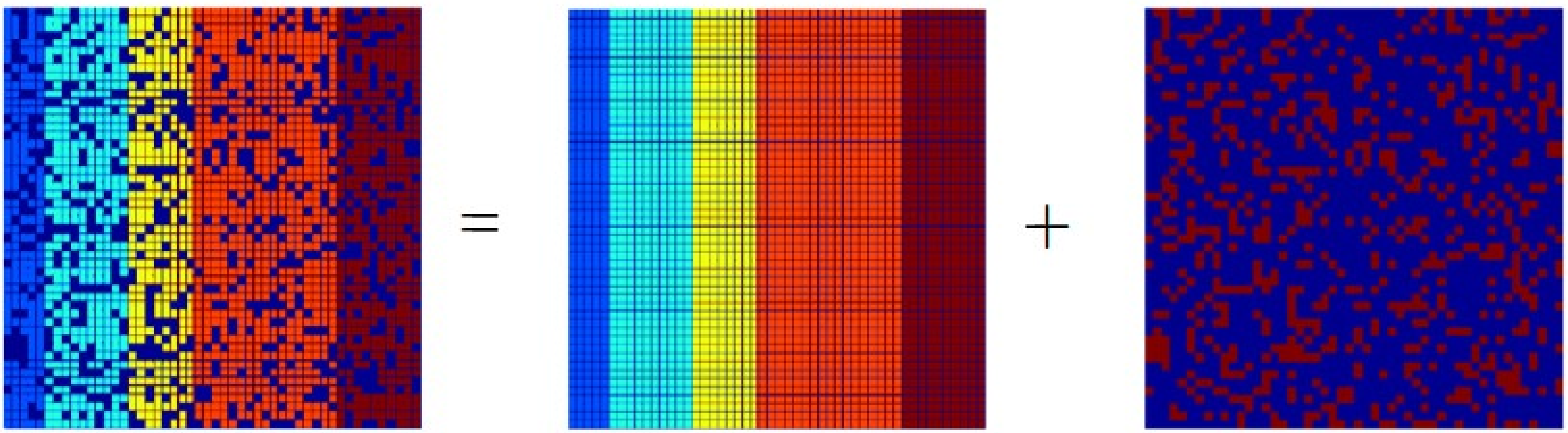
https://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf

Low rank and sparse

+ Problem Formulation

- The observed matrix is the sum of a low-rank and a sparse matrix

$$\boldsymbol{Y} = \boldsymbol{L}_o + \boldsymbol{S}_o$$



- Optimization problem

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \\ \text{subject to} & \boldsymbol{L} + \boldsymbol{S} = \boldsymbol{Y}\,. \end{array}$$

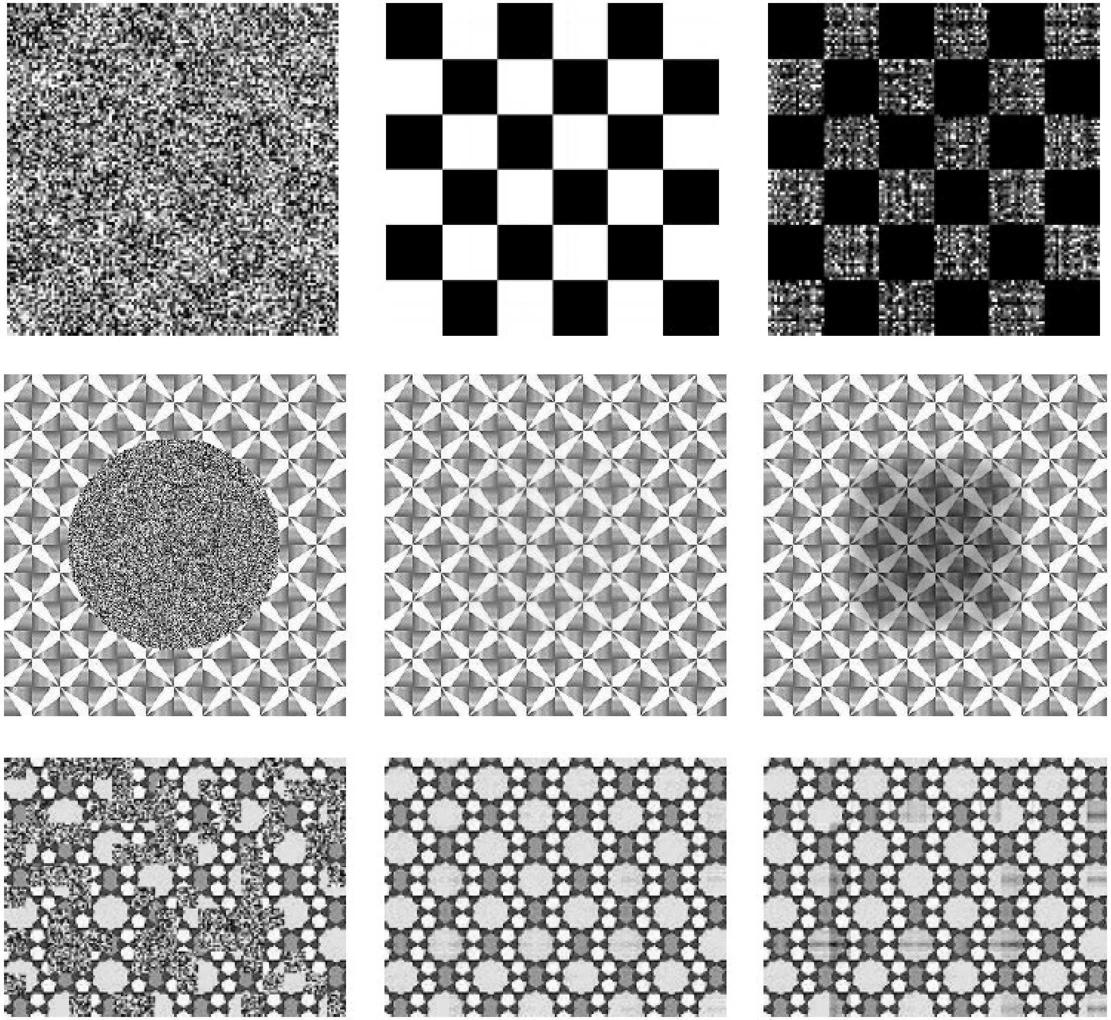
+ Background modelling from video

(a) Original frames (b) Low-rank $\hat{L}$ (c) Sparse $\hat{S}$

+ Removing shadows, specularities and saturations

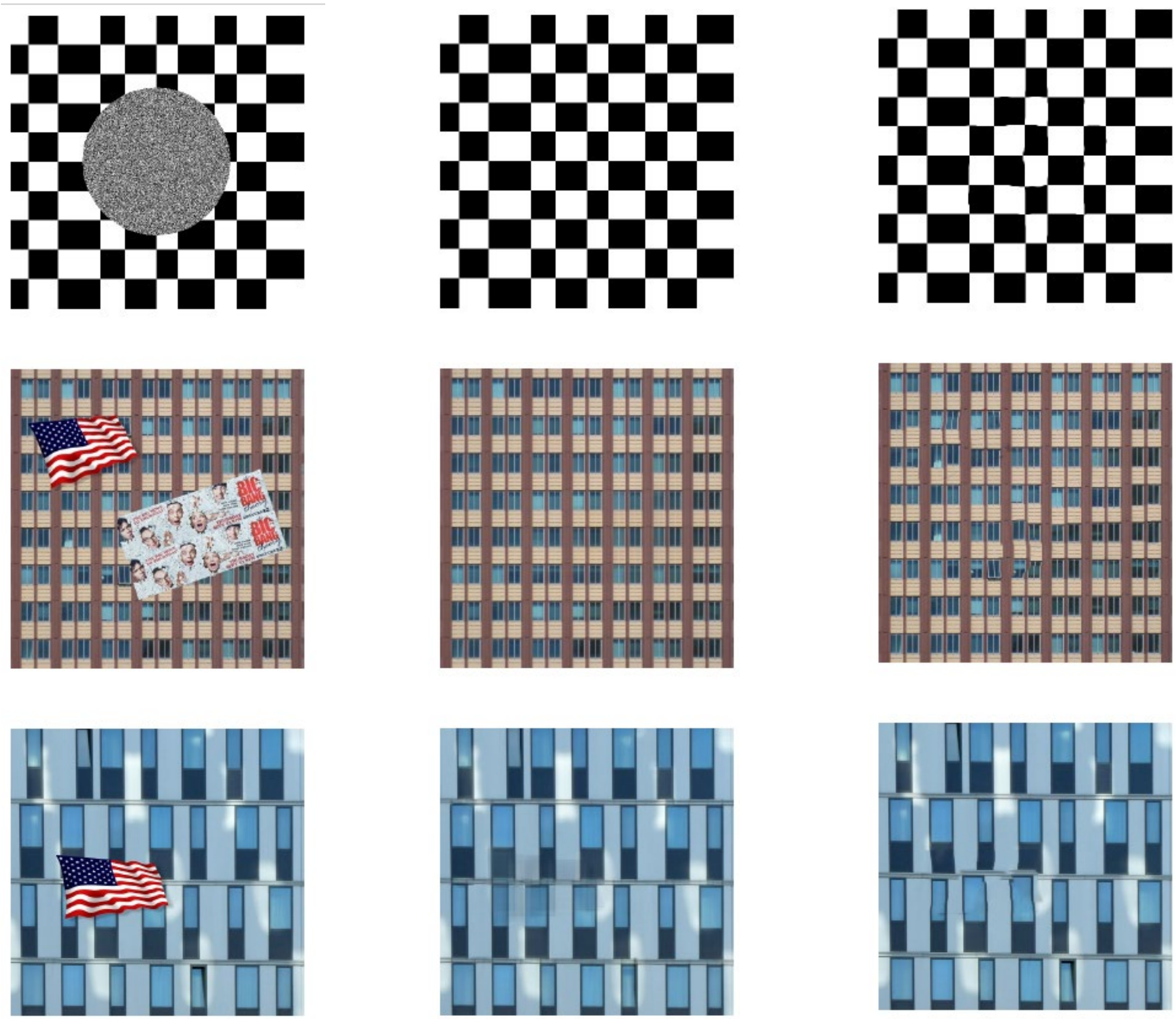
(a) $\boldsymbol{Y}$ (b) $\hat{\boldsymbol{L}}$ (c) $\hat{\boldsymbol{S}}$

+ Structured text recovery



(a) Input image (b) Sparse + Low-rank (c) Low-rank only

+ Structured text recovery




+ Further watching

- Neural network as a white box
  - https://www.youtube.com/watch?v=z2bQXO2mYPo

+ Reference

- John Wright and Yi Ma. High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications. Cambridge University Press, 2021.