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

Compressed Sensing 两层境界

高崇铭

Data Mining Lab, Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

Preface

- Feel free to interrupt me for asking any questions at any time!

- Viewer discretion is advised!

Two State of Compressed Sensing (CS)

数据挖掘实验室
Data Mining Lab

➢ Framework

1. Intuitive understanding
   - ✓ Basic idea in the Signal and Systems
   - ✓ Issues of CS

第一层境界：
掌握、理解CS

2. Math: Important theorems and proofs
   - ➢ Showed with explanations and proofs
   - ➢ Showed with intuitive explanations
   - ➢ Given directly without explanations

第二层境界：
CS背后的数学原理

3. Applications

   MRI: Magnetic Resonance Imaging
   Single pixel camera

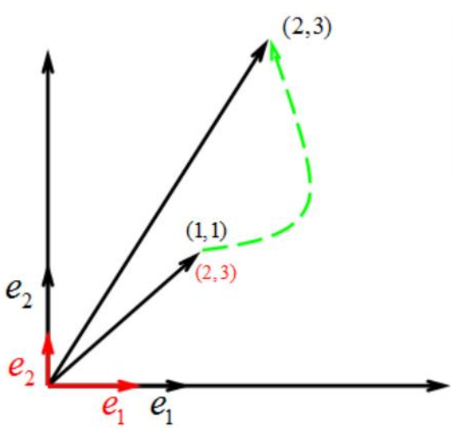
Prerequisite & Recommendations

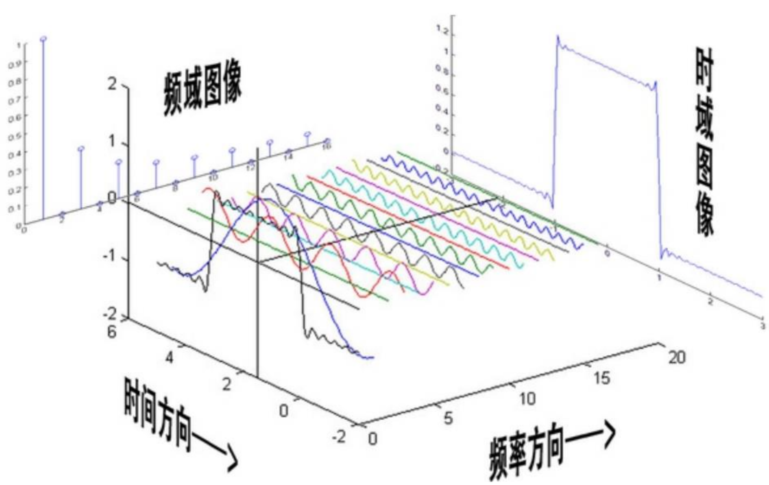
Prerequisite course:

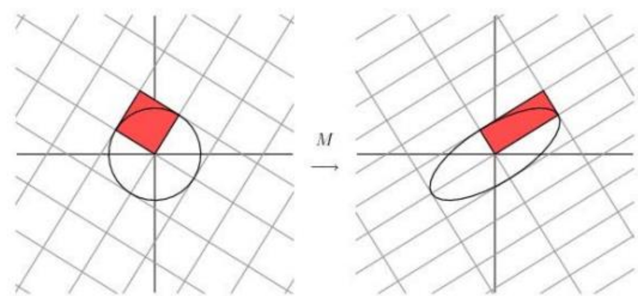
- Advanced Linear Algebra
- Signals and Systems or Digital Signal Processing

Recommended sources:

- ✓ 1.《Linear Algebra Done Right》(《线性代数应该这样学》)
- ✓ 2. 神奇的矩阵，黎文科，哈尔滨工程大学。

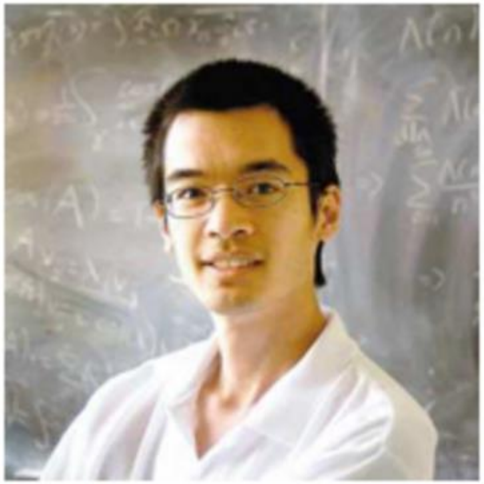






Story of CS

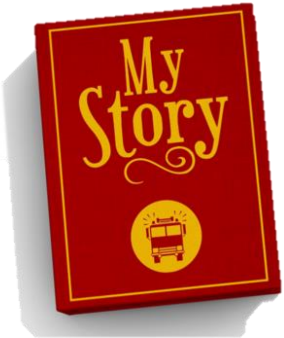
2004年…



陶哲轩

Emmanuel Candes

David Donoho



Story of CS

数据挖掘实验室
Data Mining Lab

- ✓ 压缩感知，又称压缩采样，压缩传感。它作为一个新的采样理论，它通过开发信号的稀疏特性，在远小于Nyquist采样率的条件下，用随机采样获取信号的离散样本，然后通过非线性重建算法完美的重建信号。

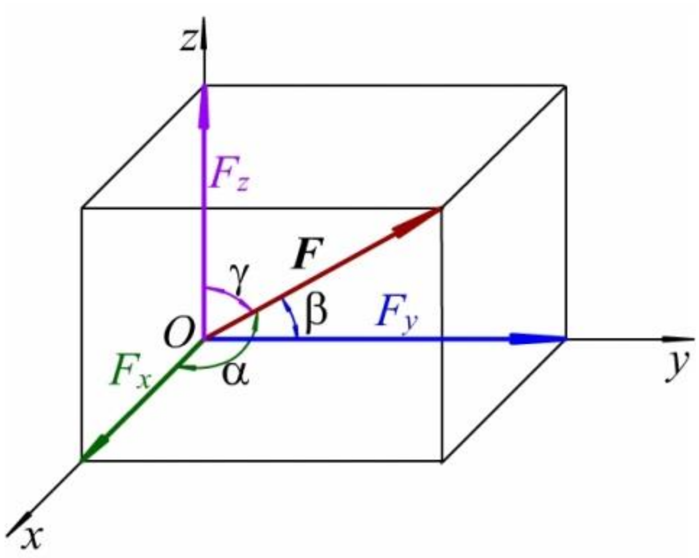
- ✓ 压缩感知理论一经提出，就引起学术界和工业界的广泛关注。它在信息论、图像处理、地球科学、光学、微波成像、模式识别、无线通信、大气、地质等领域受到高度关注，并被美国科技评论评为2007年度十大科技进展。

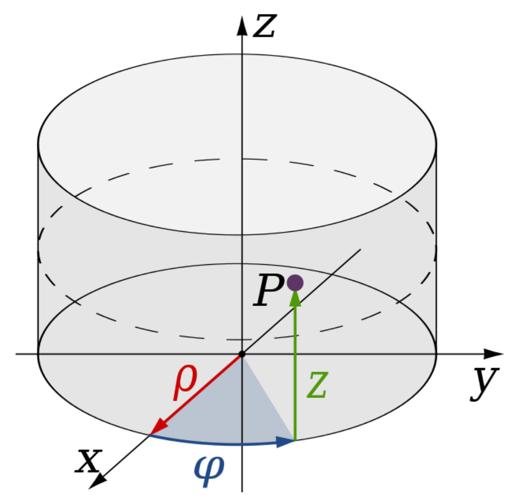
引自百度百科：压缩感知

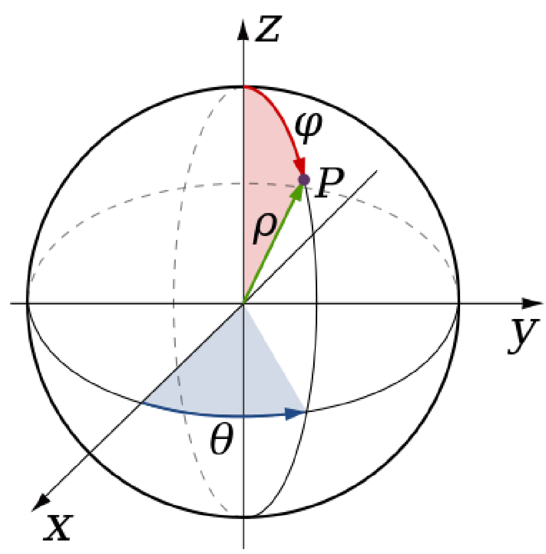
State 1: Intuitive understanding of CS

What's Sensing（感知）?

A kind of measurement！



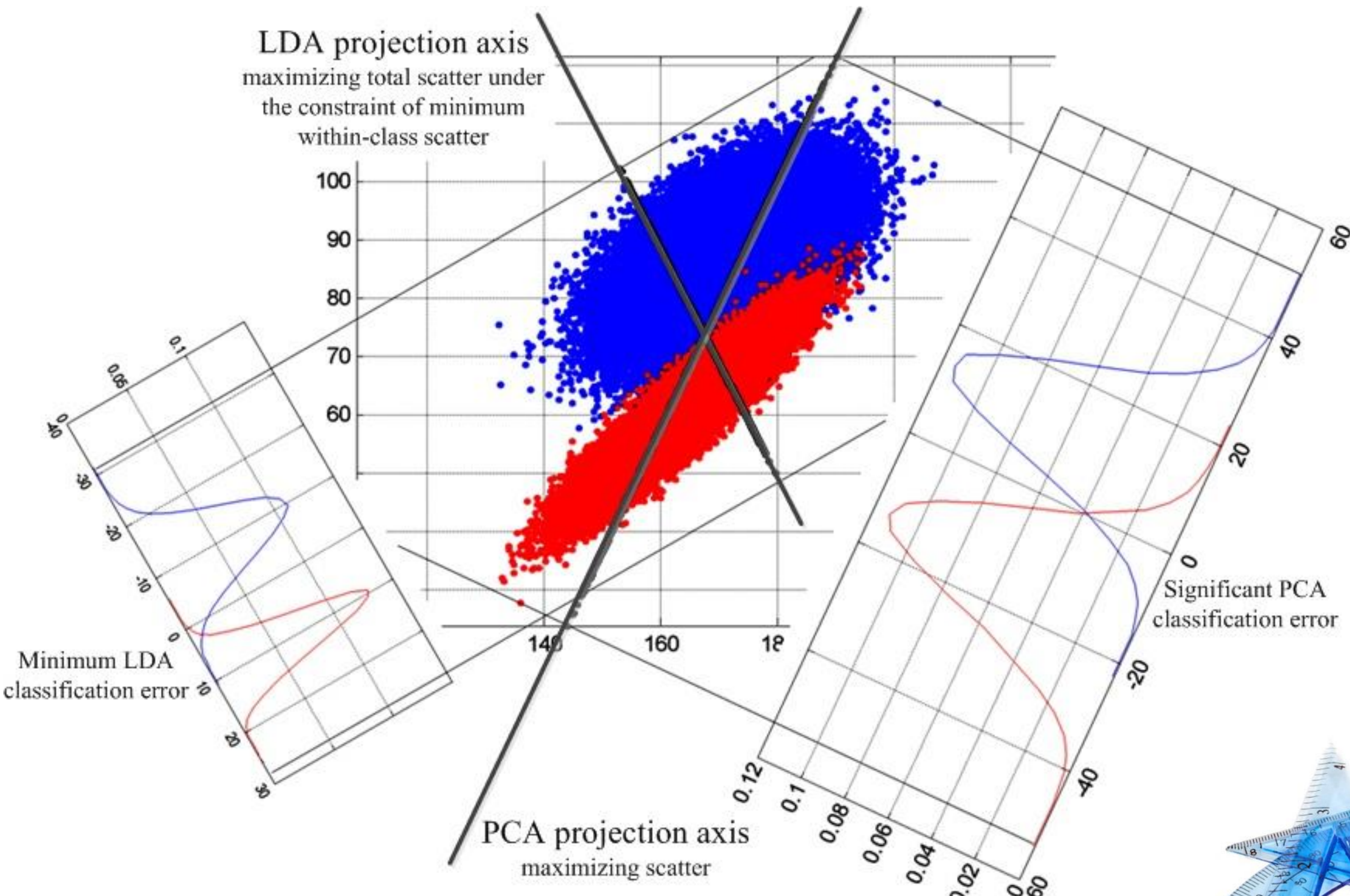




What's Sensing（感知）?

数据挖掘实验室
Data Mining Lab

A kind of measurement！



Sensing System

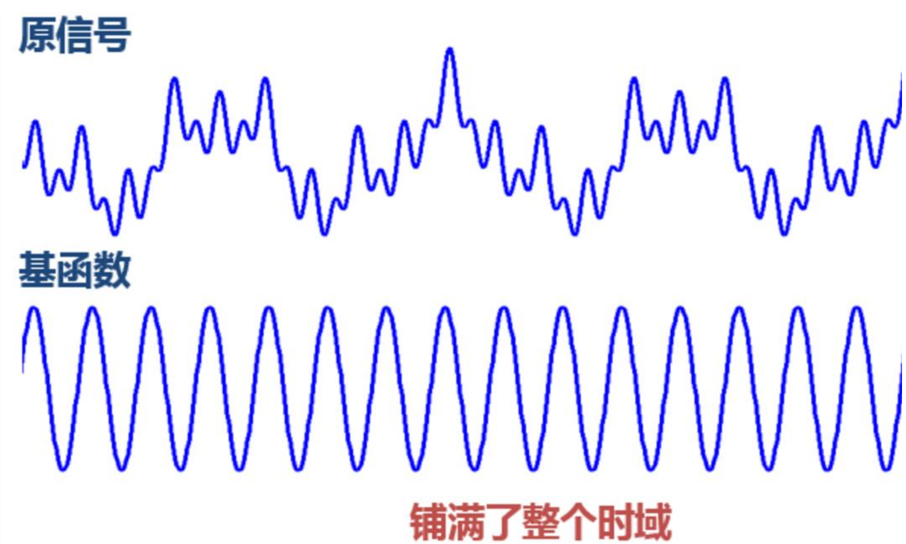
- Examples of sensing system

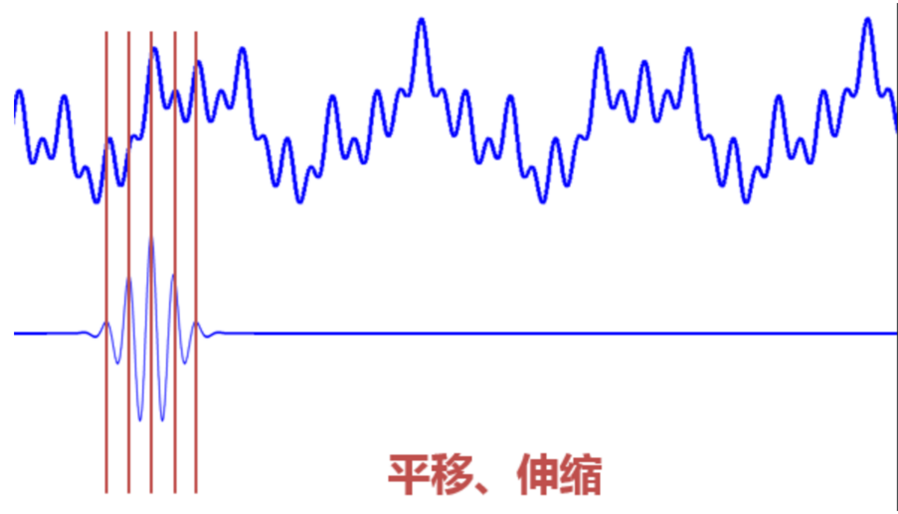
Fourier transform

$$F(w) = \int_{-\infty}^{\infty} f(t) * e^{-iwt} dt$$

Wavelet transform

$$WT(a,\tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) * \psi\left(\frac{t-\tau}{a}\right) dt$$

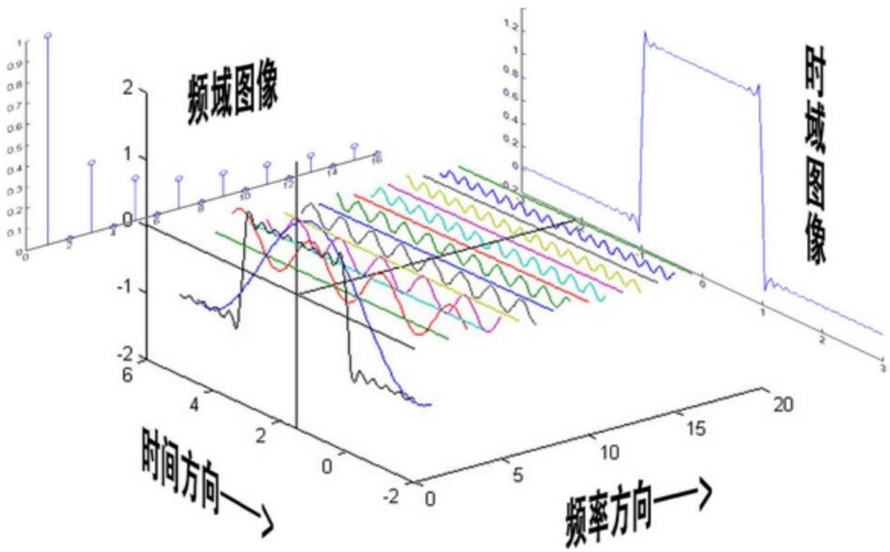




https://zhuanlan.zhihu.com/p/22450818

Sensing System

➢ Examples of sensing system



http://daily.zhihu.com/story/3935067
https://en.wikipedia.org/wiki/File:Fourier_series_and_transform.gif

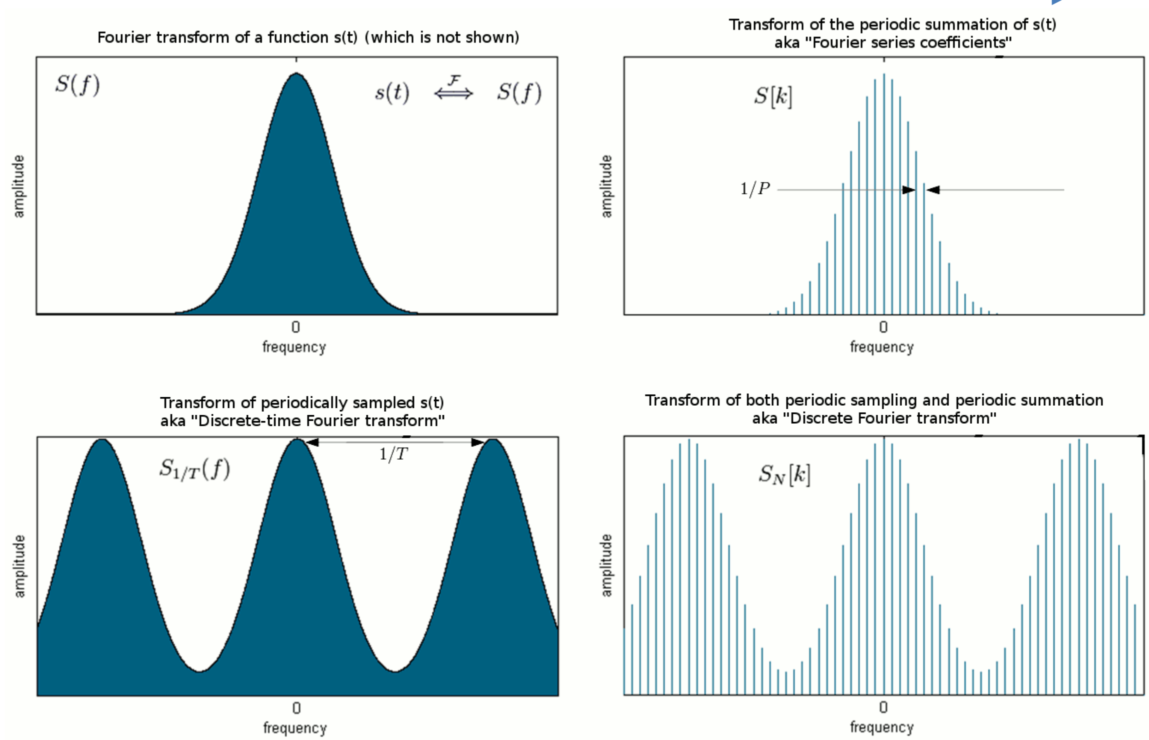
Sensing System

数据挖掘实验室
Data Mining Lab

➢DFT(Discrete Fourier Transformation)

频域离散化(数字存储)

时域离散化(采样)
频域周期化



$$X_k \stackrel{\text{def}}{=} \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi i k n/N}, \quad k \in \mathbb{Z} \quad (\text{integers})$$

https://en.wikipedia.org/wiki/Discrete_Fourier_transform

Sensing System

数据挖掘实验室
Data Mining Lab

- DFT(Discrete Fourier Transformation)

$$X_k \stackrel{\text{def}}{=} \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi i k n / N}, \quad k \in \mathbb{Z} \ \ (\text{integers})$$

$$X = Wx$$

$$W = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \omega^3 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \omega^6 & \cdots & \omega^{2(N-1)} \\ 1 & \omega^3 & \omega^6 & \omega^9 & \cdots & \omega^{3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \omega^{3(N-1)} & \cdots & \omega^{(N-1)(N-1)} \end{bmatrix}$$

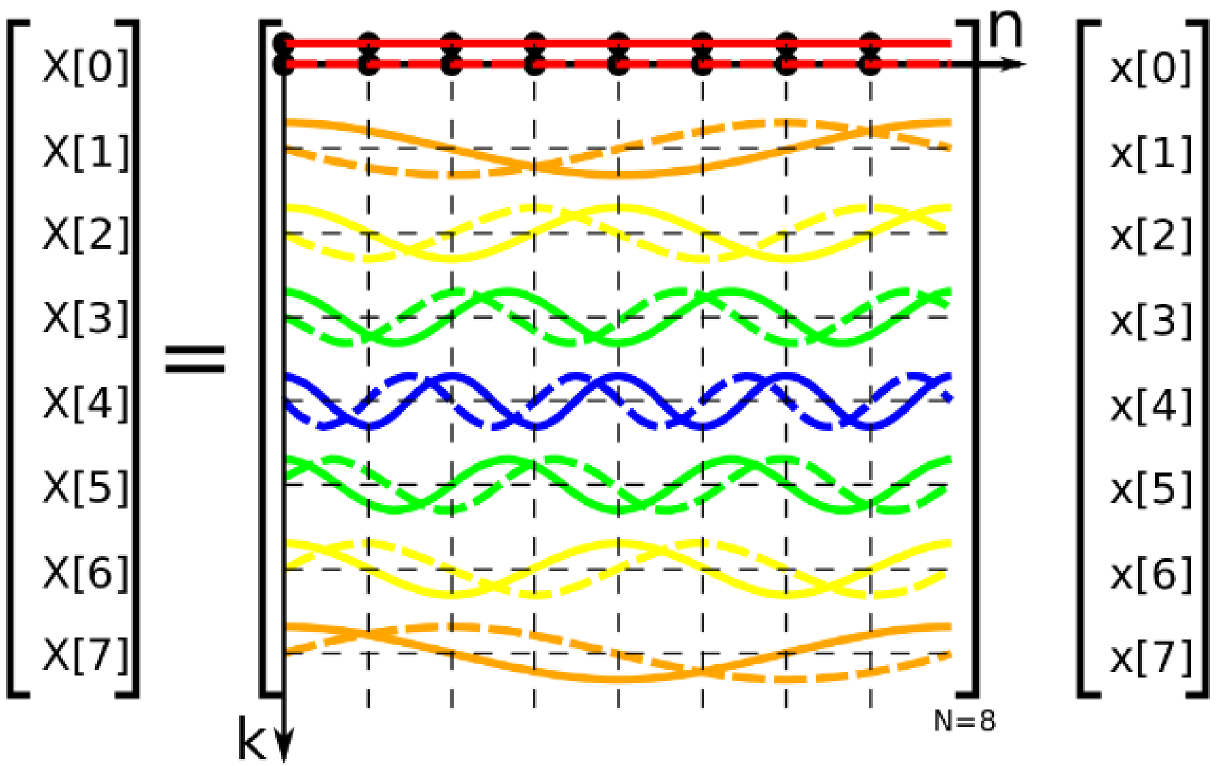
https://en.wikipedia.org/wiki/DFT_matrix

This is sensing of matrix form!

数据挖掘实验室
Data Mining Lab

- DFT(Discrete Fourier Transformation)

$$X_k \stackrel{\text{def}}{=} \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi i k n / N}, \quad k \in \mathbb{Z} \quad \text{(integers)}$$

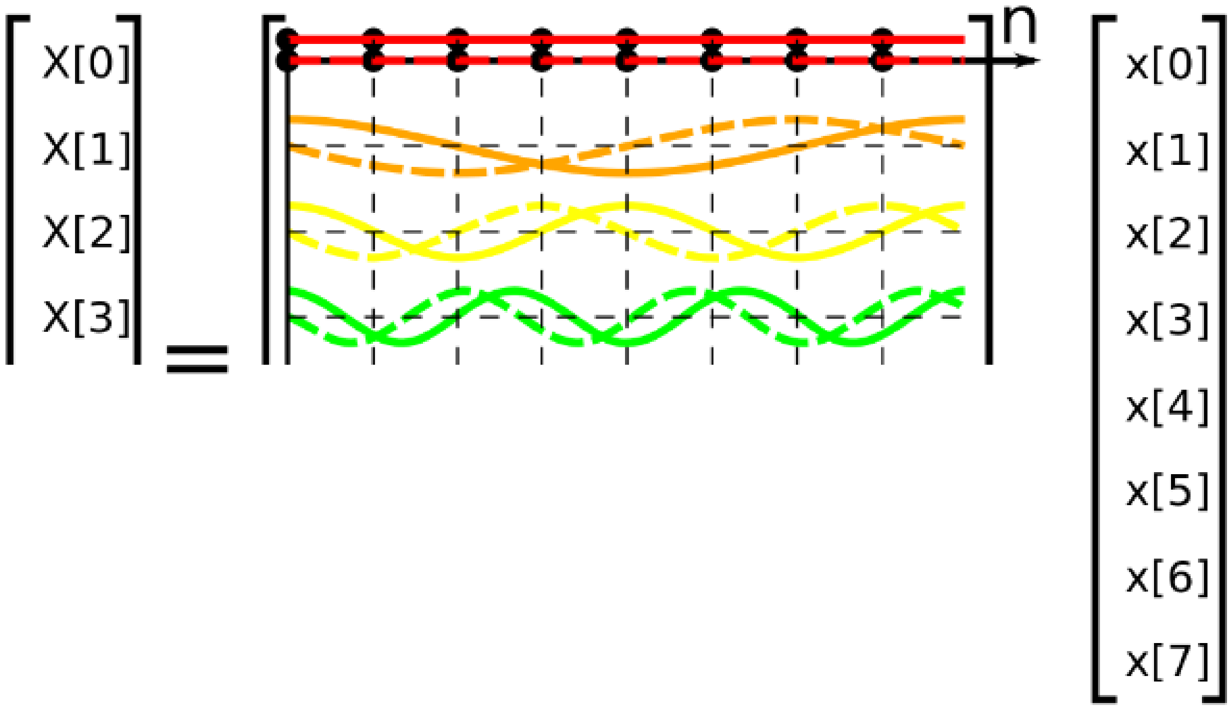


https://en.wikipedia.org/wiki/DFT_matrix

This is sensing of matrix form!

➤ DFT(Discrete Fourier Transformation)

$$X_k \stackrel{\text{def}}{=} \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi i k n / N}, \quad k \in \mathbb{Z} \quad (\text{integers})$$

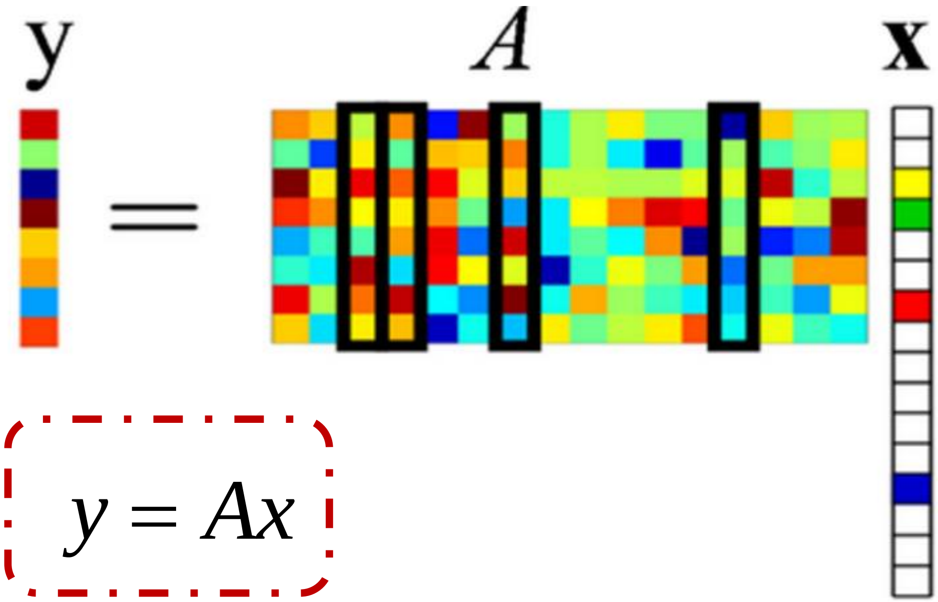


https://en.wikipedia.org/wiki/DFT_matrix

What's Compressed Sensing?

➢Idea of Compressed Sensing

$$X_k \stackrel{\text{def}}{=} \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi i k n / N}, \quad k \in \mathbb{Z} \quad \text{(integers)}$$



$$\boxed{y = Ax}$$

- ➢ Underestimated equation !
- ➢ The dimension of solution space is N-M
- ➢ But if **x** is sparse, under certain conditions the function has unique solution!

What's Compressed Sensing?

Essence：

Sparsity、Coherence → Compressed

Issues:

1. Sensing

考虑因素：如何构造感知矩阵

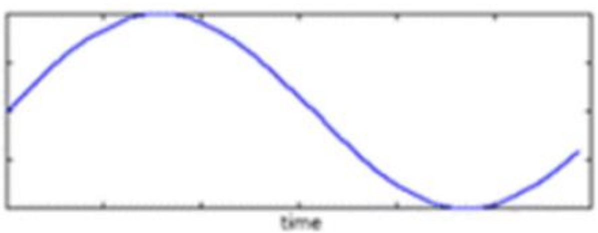
2. Recovery.

影响因素：信号稀疏度、矩阵的行列数目、矩阵的相关性(具体为列秩)

What's Compressed Sensing?

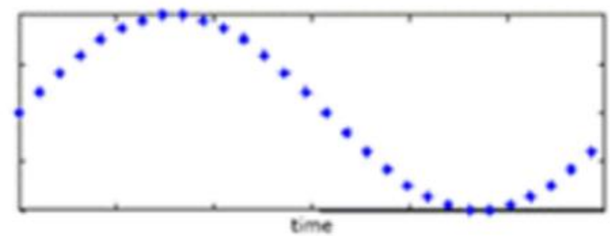
➢模拟信号采样

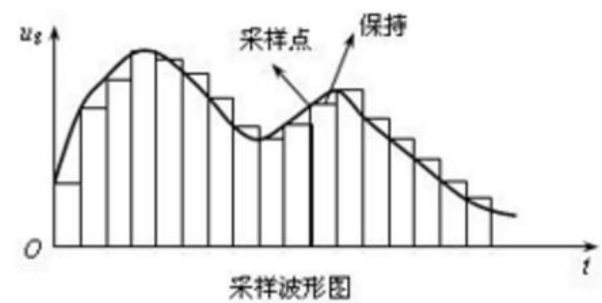
模拟信号



采样 →

数字信号





问题 用多大的采样频率？

What's Compressed Sensing?

➢奈奎斯特采样定律

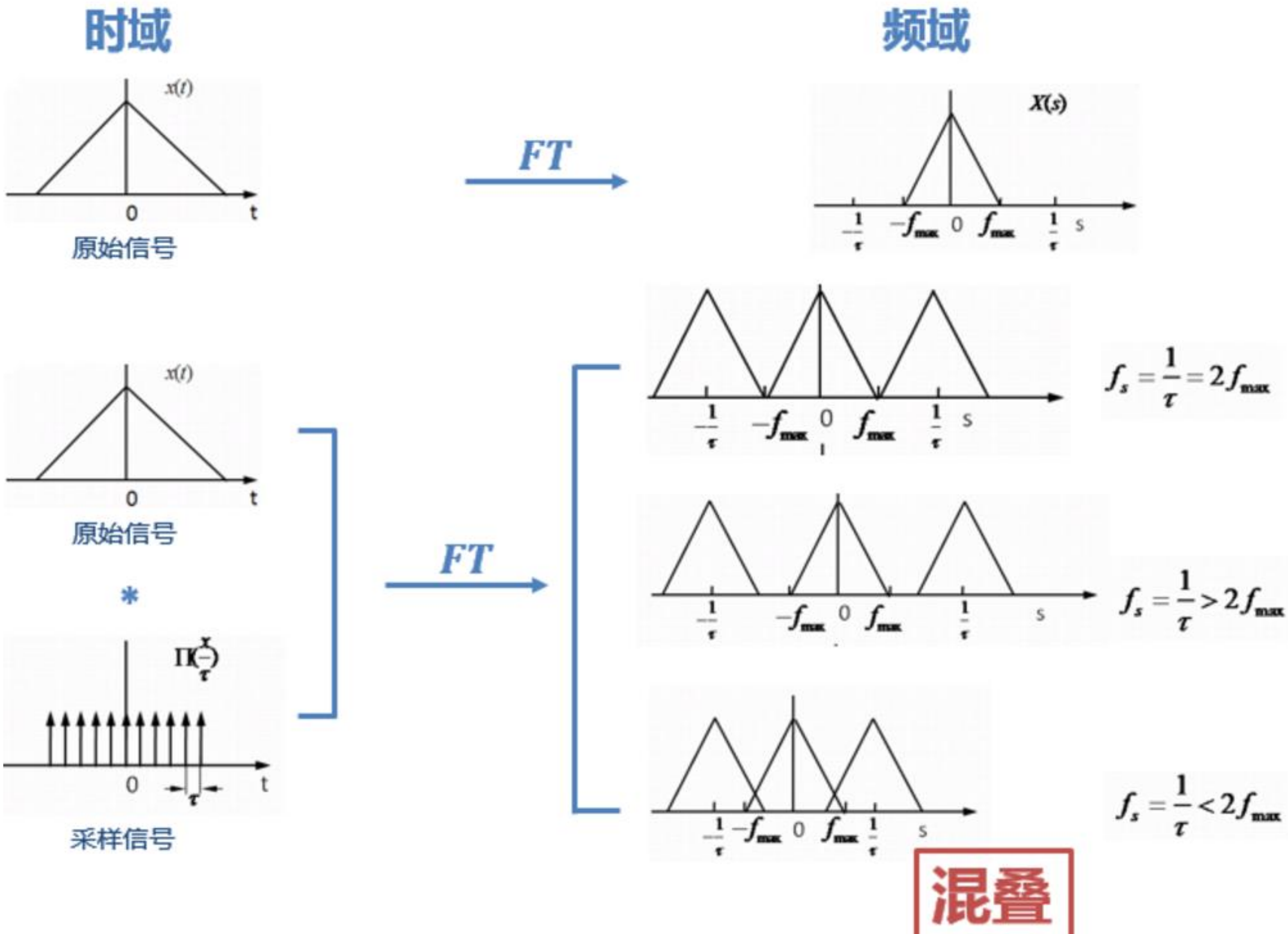
奈奎斯特，H.

1889-1976

要想让采样之后的数字信号完整保留原始信号中的信息，采样频率必须**大于信号中最高频率的2倍！**

What's Compressed Sensing?

➢奈奎斯特采样定律



What's Compressed Sensing?

2004年，几位大牛证明，如果信号是**稀疏的**，那么它可以由远**低于采样定理要求的采样点重建恢复**，并于**2007年**正式提出了“**压缩感知**”（Compressed Sensing）这个概念。

压缩感知

陶哲轩

Emmanuel Candes

David Donoho

What's Compressed Sensing?

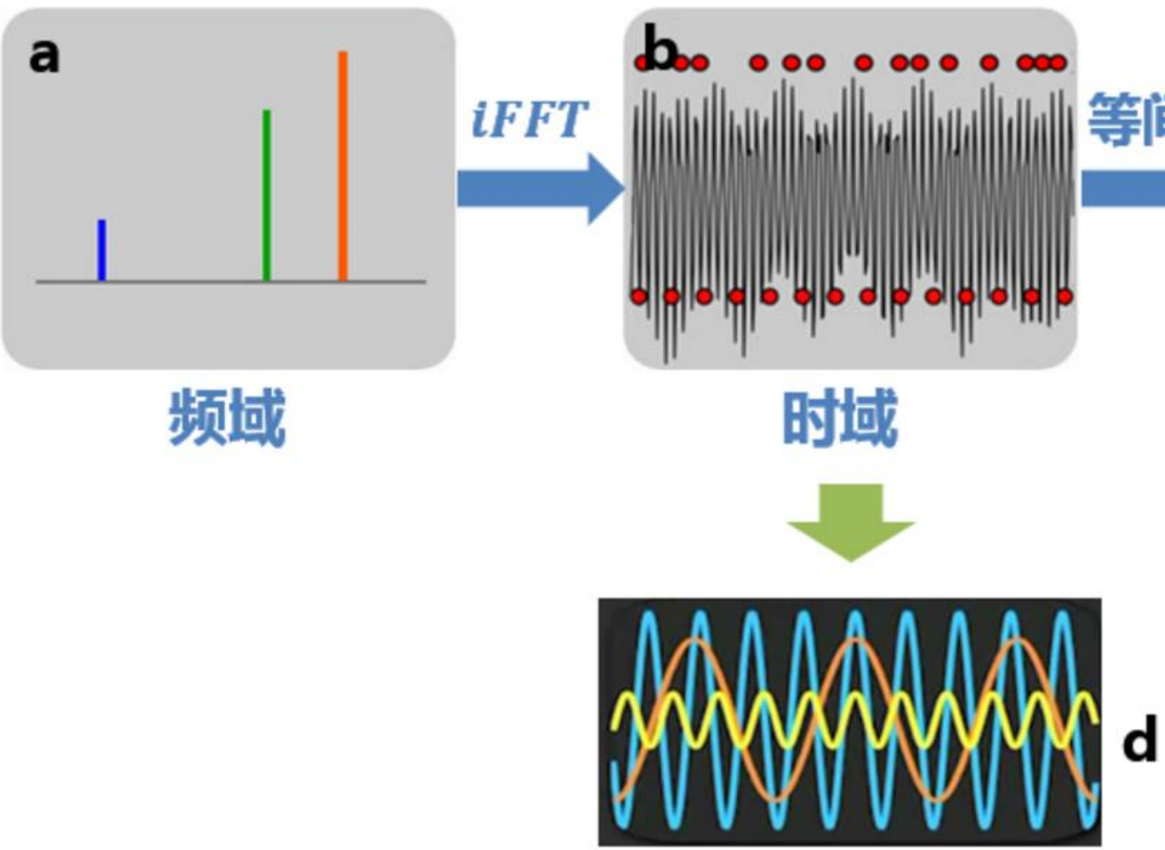
"采样频率须大于信号中最高频率的2倍" ——Nyquist

采样频率 $\xrightarrow{\text{意味着}}$ 等间距采样

- 等间距采样，频域将以1/τ为周期延拓，采样频率低必然引起混叠。
- 如果是不等间距采样呢？
- 如果是随机采样呢？

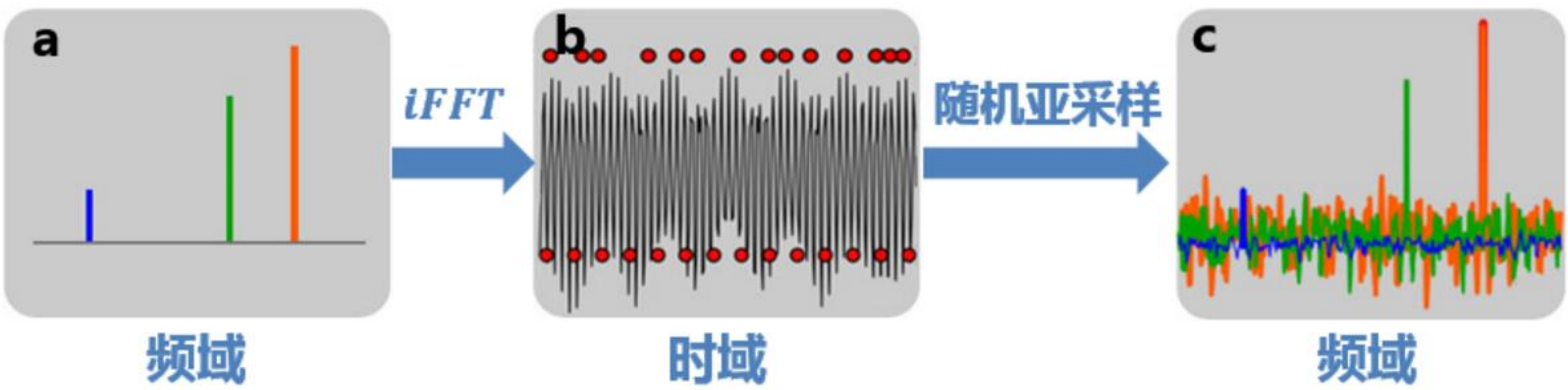
What's Compressed Sensing?

➢**传统等间距采样：**



What's Compressed Sensing?

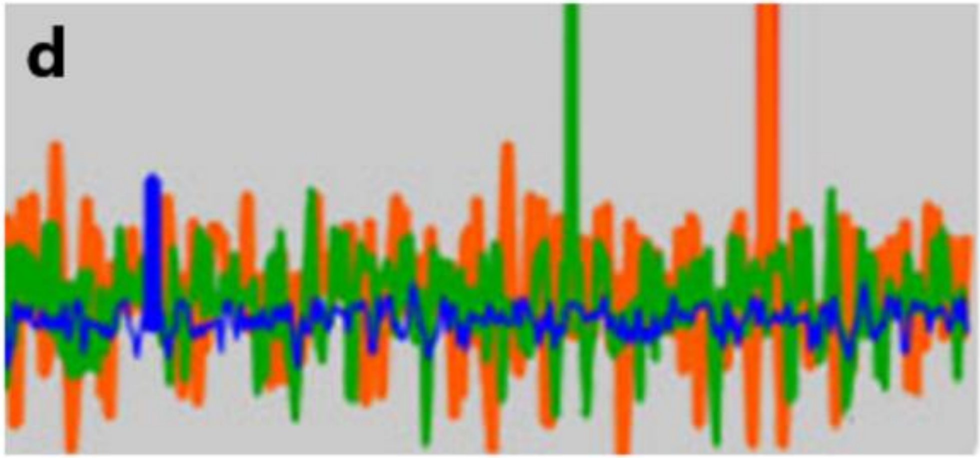
➢ **随机亚采样**



What's Compressed Sensing?

数据挖掘实验室
Data Mining Lab

➢随机亚采样



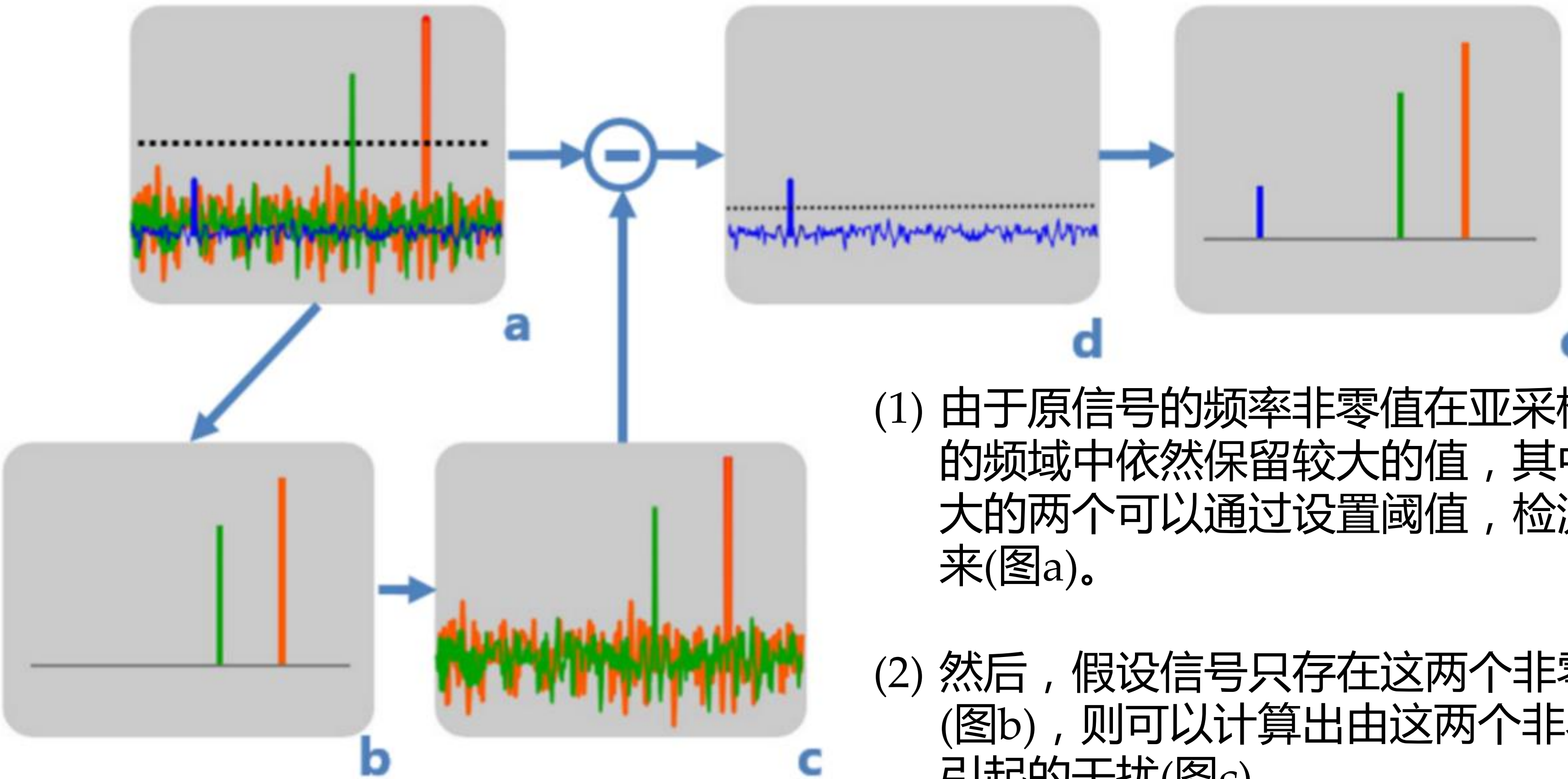
不相关的伪影，看起来像随机噪声

最大的几个峰值还依稀可见，只是一定程度上被干扰值覆盖。这些干扰值**看上去非常像随机噪声**，但**实际上是由于三个原始信号的非零值发生能量泄露导致的**

这可以理解成随机采样使得频谱**不再是整齐地搬移**，而是一小部分一小部分胡乱地搬移，**频率泄露均匀地分布在整个频域**，因而泄漏值都比较小，从而有了恢复的可能。

What's Compressed Sensing?

➢信号恢复：March Pursuit(匹配追踪思想)



(1) 由于原信号的频率非零值在亚采样后的频域中依然保留较大的值，其中较大的两个可以通过设置阈值，检测出来(图a)。

(2) 然后，假设信号只存在这两个非零值(图b)，则可以计算出由这两个非零值引起的干扰(图c)。

(3) 用a减去c，即可得到仅由蓝色非零值和由它导致的干扰值(图d），再设置阈值即可检测出它，得到最终复原频域（图e）

(4) 如果原信号频域中有更多的非零值，则可通过迭代将其一一解出

➢小结

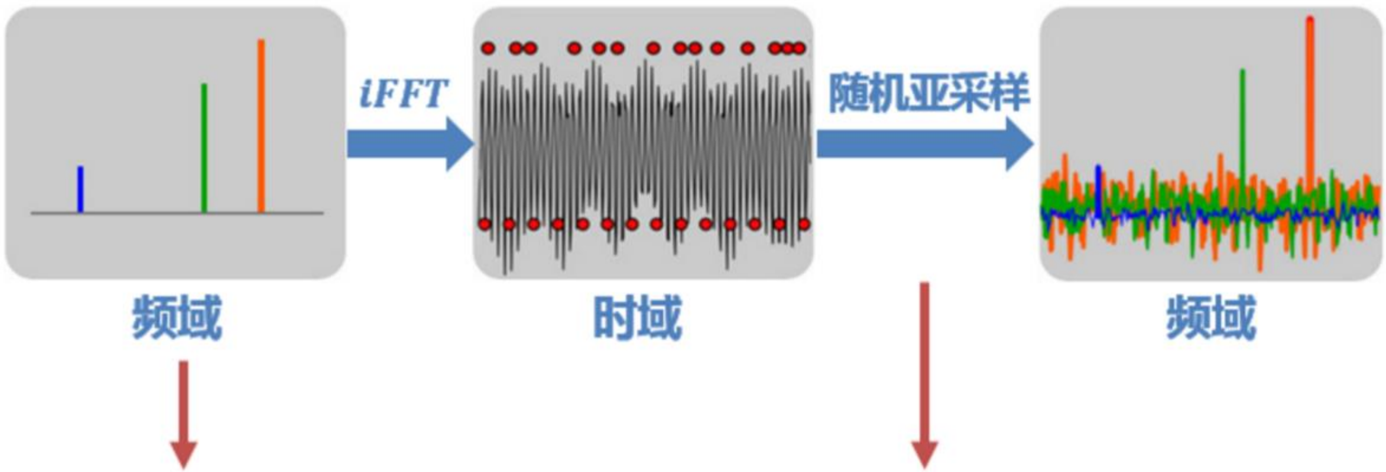
以上就是压缩感知理论的**核心思想**：

以比奈奎斯特采样频率要求的采样密度更稀疏的密度对信号进行随机亚采样，由于频谱是均匀泄露的，而不是整体延拓的，因此可以通过特别的追踪方法将原信号恢复。

What's Compressed Sensing?

➢Revisit：前提条件

刚才的例子中，满足了两个前提条件：



1. 信号在频域稀疏

2. 随机亚采样

1. 这个信号在频域只有3个非零值，所以可以较轻松地恢复出它们。

2. 采用了随机亚采样机制，因而使频率泄露均匀地分布在整个频域。

✓ 这两点对应了CS的两个前提条件——**稀疏性（sparsity）**、**不相关性（incoherence）**。

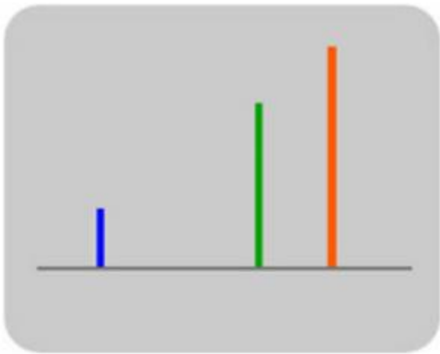
➢My Understandings of intrinsic nature of **Sparsity**

Objects or events have information much less than the representational ability of the coding system, which give rise to a compressed or sparse representation over specify basis.

如果物体或者事件本身包含的信息远小于编码系统能表示的最大信息范围，那么在这个编码系统的特定基下该物体可以表示为紧凑稀疏的形式。

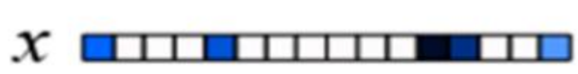
➢前提条件1：稀疏性

信号需要在某一个变换域具有稀疏性



信号在频域是稀疏的

如果信号在某个域中非零点远远小于信号总点数，则信号在该域中是稀疏的。

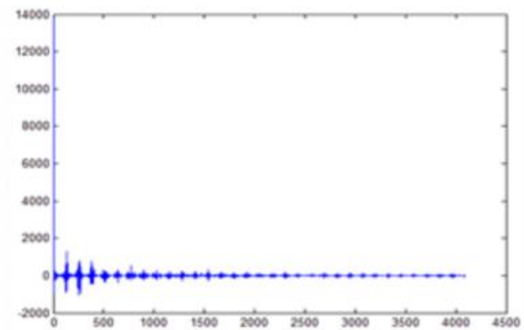


典型的一维稀疏信号，只有少量非零项

Sparsity

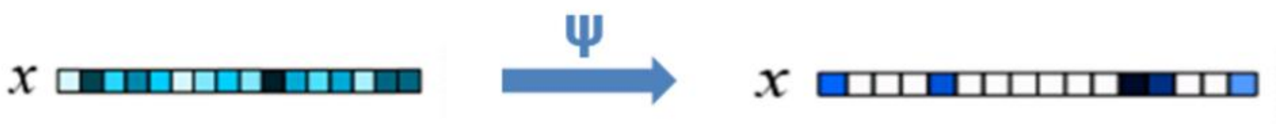
➢前提条件1：稀疏性

对于压缩感知：



信号只需要近似满足稀疏性，即为可压缩信号。

对于CS，只要它在某一个变换域满足近似稀疏特性即可，我们称之为稀疏域，重建将在稀疏域进行。



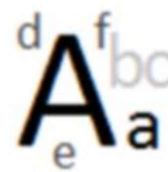
Ψ可以是频域、小波变换、离散余弦变换等

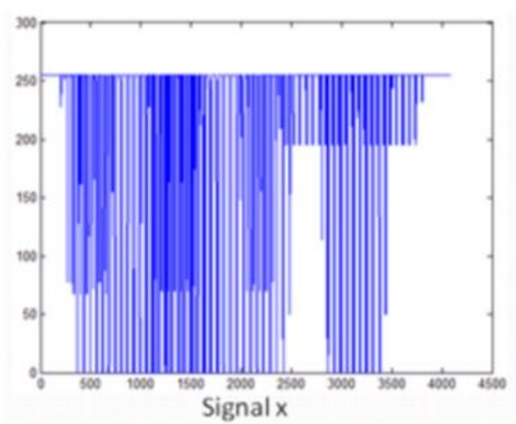
Sparsity

➢前提条件1：稀疏性

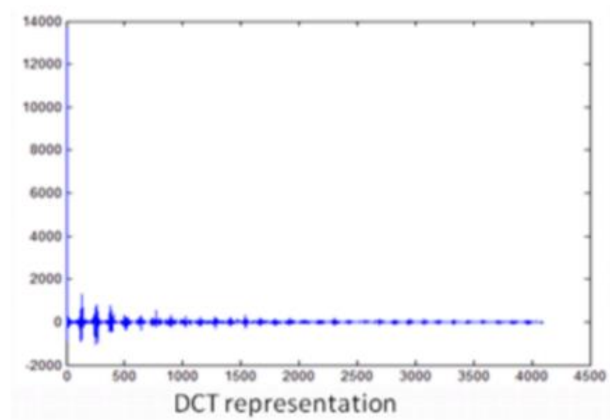
信号的稀疏性已经在图像压缩领域有了很好的应用。

如JPEG格式：





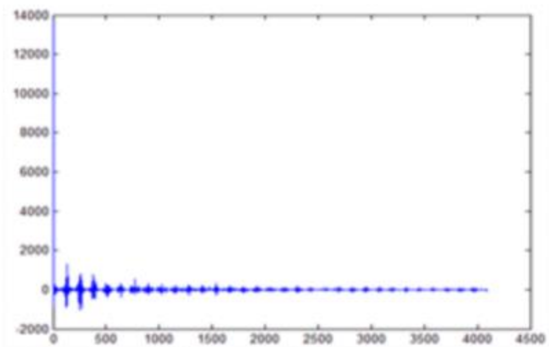
时域，不稀疏



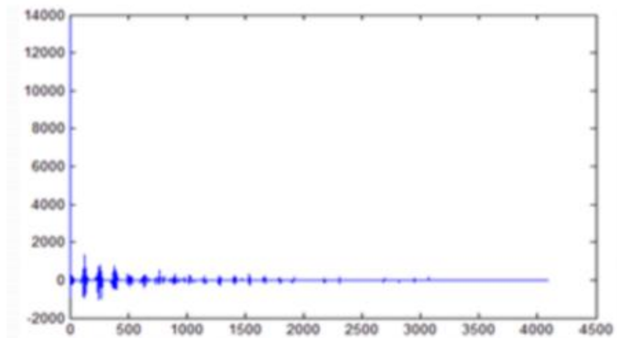
离散余弦变换域，稀疏

Sparsity

➢前提条件1：稀疏性

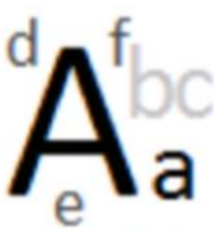


DCT representation



DCT representation if the threshold was set to 100

舍弃小于100的值



4096个点

282个点

仅占原图像数据大小的6.9%

Mechanism of Compressed Sensing

图像压缩

全采样 → 稀疏变换 大量小系数 少量大系数 → 保留大系数 丢弃小系数 → 压缩图像

压缩感知的原动力问题：

既然全采样了还要再丢弃，我们为什么不能直接少采样一些点？

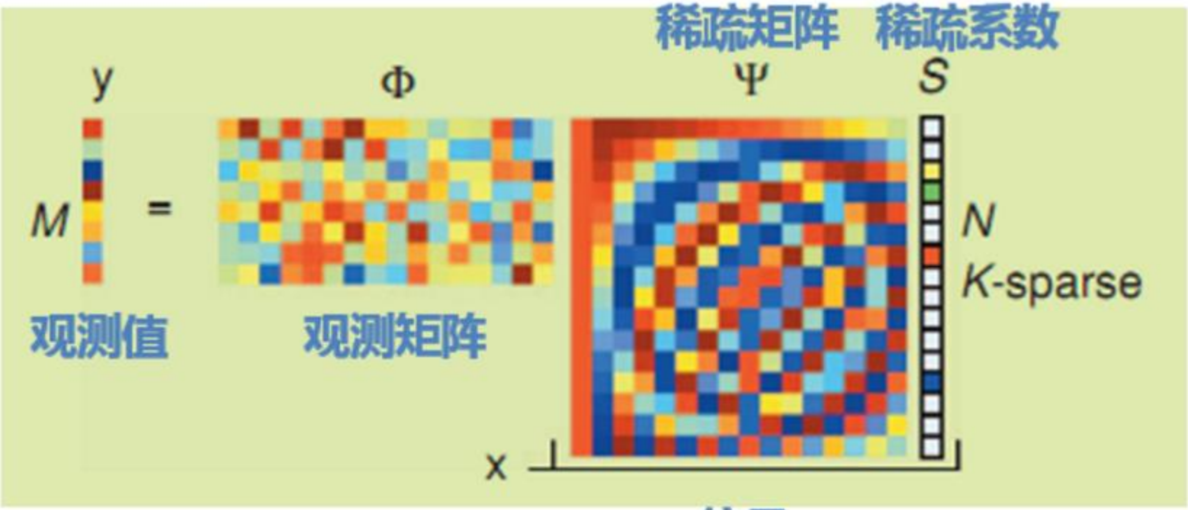
压缩感知

亚采样 → 在不相关的稀疏域导致伪影 → 消除伪影 → 复原信号

压缩感知直接在采样时就完成了压缩！

Mechanism of Compressed Sensing

数学表达



信号

$$y = \Phi\Psi s$$

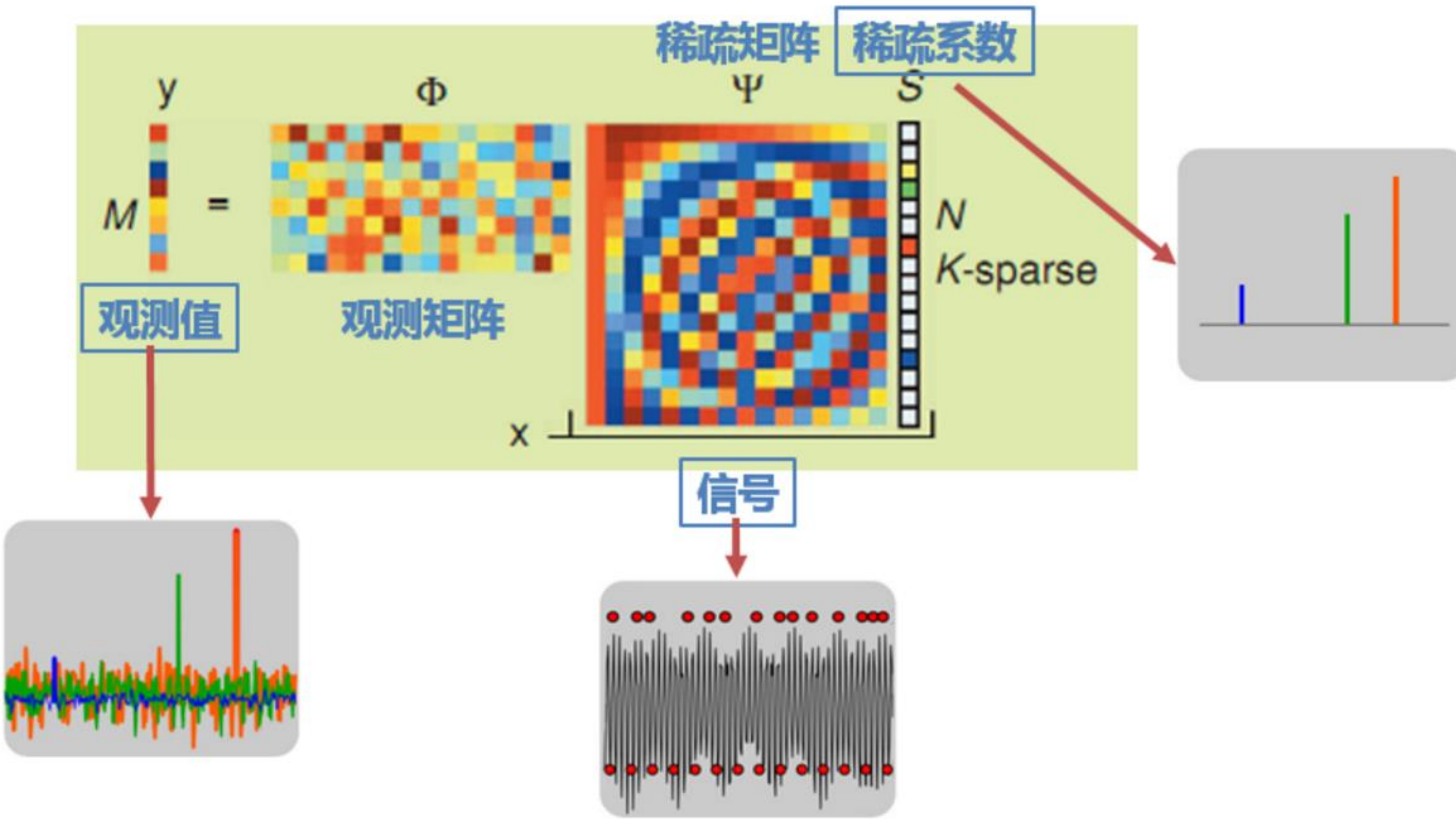
x ：信号
Φ ：观测矩阵
y ：观测值

- 观测矩阵Φ将高维信号x投影到低维空间。
- 对x在Ψ稀疏基上进行稀疏表示，x=Ψs，Ψ为稀疏基矩阵，s为稀疏系数。

Mechanism of Compressed Sensing

数据挖掘实验室
Data Mining Lab

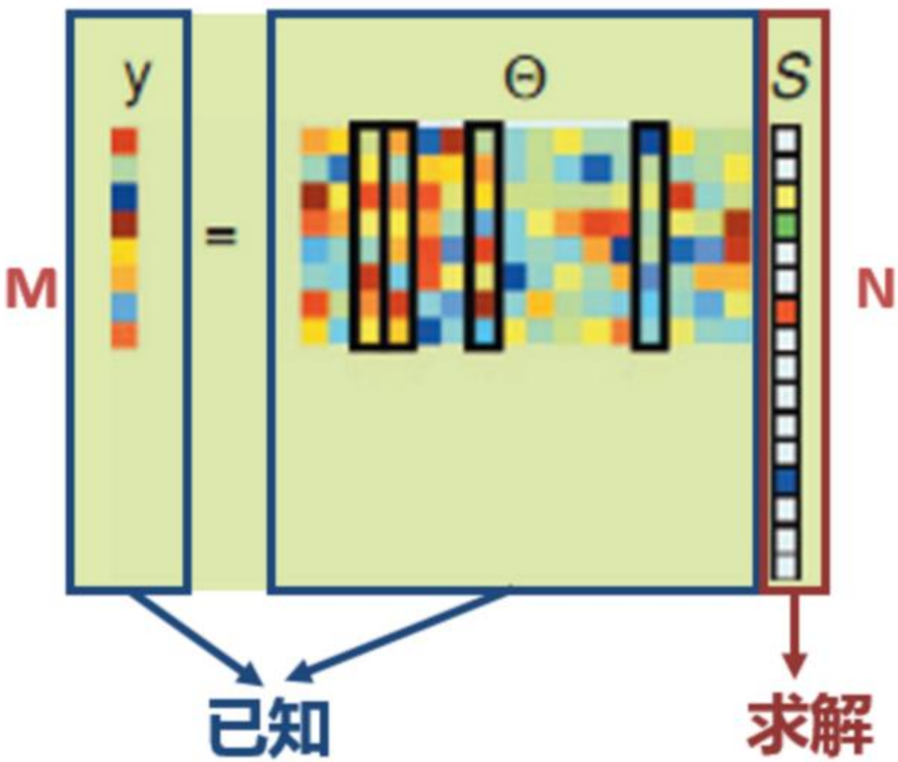
数学表达



数学表达

令$\Theta = \Phi\Psi$，Θ称为传感矩阵

$$y = \Theta s$$



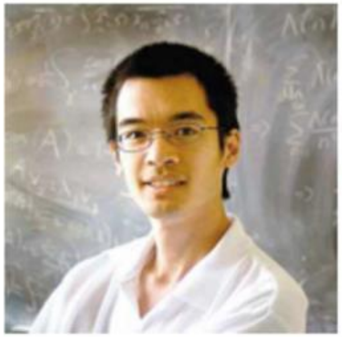
y、Θ ⟶ s

若M=N，则可轻松由y解出S

而M<N，可根据S的RIP特性重构

➢前提条件二：Coherence

之前已提到，采用随机亚采样才能实现信号的恢复。



陶哲轩和Candès于2005年给出了更为准确的要求：观测矩阵Φ应满足约束等距性条件(Restricted Isometry Property，简称RIP)：

即对于任意和常数，有:

$$(1-\delta_k)\|c\|_2^2 \le \|\phi c\|_2^2 \le (1+\delta_k)\|c\|_2^2$$

➢前提条件二：Coherence

Baraniuk证明：

RIP的等价条件是观测矩阵和稀疏表示基不相关（incoherent）

$$y = \Phi\Psi s$$

$$\Phi \xleftrightarrow{\text{不相关}} \Psi$$

相关性的定义：

$$\mu(\Phi, \Psi) = \sqrt{n} \cdot \max_{1 \leq k, j \leq n} |\langle \varphi_k, \psi_j \rangle|.$$

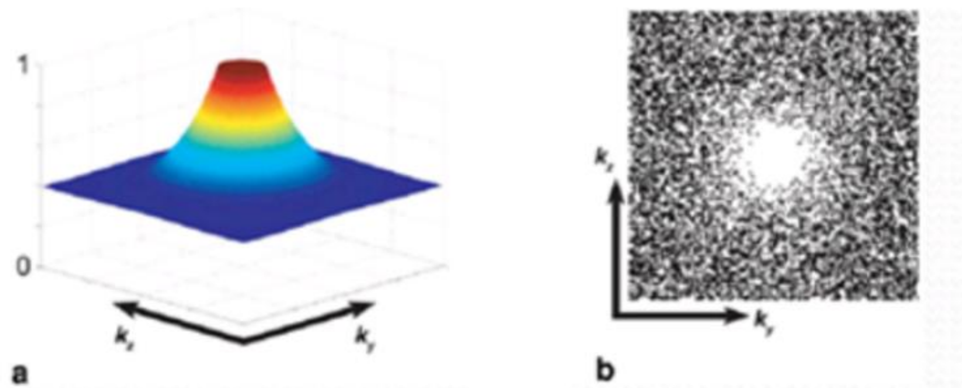
μ的范围： $\mu(\Phi, \Psi) \in [1, \sqrt{n}]$

μ越小，Φ和Ψ越不相关

➢Measure矩阵的构造

陶哲轩和Candès证明：

独立同分布的高斯随机测量矩阵可以成为普适的压缩感知测量矩阵。



- 对于二维信号，往往就采用如右上图所示的采样矩阵对图像进行亚采样。
- 对于一维信号，采用前文提到的随机不等间距的亚采样即可。

➢总结

如果一个信号在某个变换域是稀疏的，那么就可以用一个与变换基不相关的观测矩阵将变换所得高维信号投影到一个低维空间上，然后通过求解一个优化问题就可以从这些少量的投影中以高概率重构出原信号。

Example of CS

- 12 balls problem

Question: You are give 12 identical looking balls. One of them is fake (could be heavier or lighter) than the rest of the 11 (all the others weight exactly the same). You a provided with a simple mechanical balance and you are restricted to only 3 uses. Find the fake ball.

http://freemind.pluskid.org/machine-learning/a-compressed-sense-of-compressive-sensing-ii/

Example of CS

➢12 balls problem：solution

$$y = wx = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & -1 & 1 & -1 & 0 & 1 & -1 & 1 \\ 0 & 1 & 0 & -1 & -1 & 0 & -1 & 0 & 1 & 1 & 1 & -1 \\ 0 & 0 & 1 & 0 & -1 & -1 & 0 & 1 & -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} m \\ m \\ m \\ m \\ m \\ m \\ M \\ m \\ m \\ m \\ m \\ m \end{pmatrix}$$

$$Wx = (M - m)W_{:k}$$

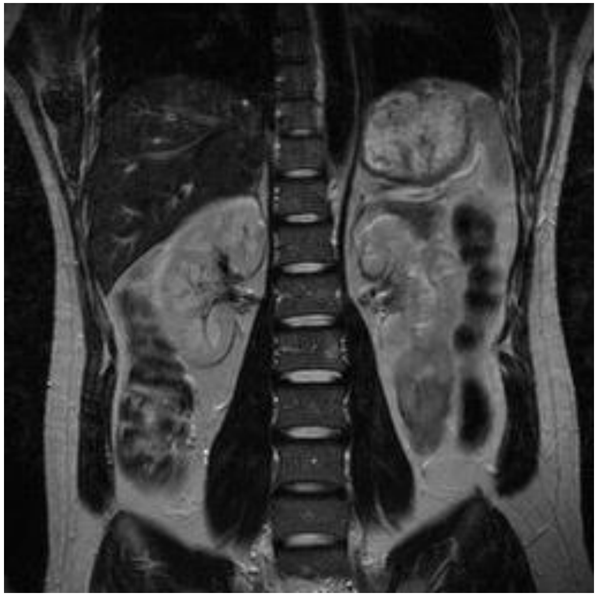
$$\mathrm{sign}(Wx) = \mathrm{sign}\left((M - m)W_{:k}\right) = \mathrm{sign}(M - m)W_{:k}$$

我们的观察值实际上将会是矩阵的第k列或者该列的相反数，具体取决于异类球的质量是比正常球大还是小。所以说，如果我们能将矩阵W设计为任意两列以及它们的相反数都互不相等的话（当然还必须满足每一行的正负1个数相等），就可以根据这个测量矩阵找出异类球的位置 了。

Real application of CS

数据挖掘实验室
Data Mining Lab

➢MRI: Magnetic Resonance Imaging



- 断层扫描（CT）技术和核磁共振（MRI）技术中，**仪器所采集到的都不是直接的图像像素，而是图像经历过全局傅立叶变换后的数据。**
- 也就是说，每一个单独的数据都在某种程度上包含了全图像的信息。在这种情况下，去掉一部分采集到的数据并不会导致一部分图像信息永久的丢失（它们仍旧被包含在其它数据里）。

- 我们可以在采集数据的时候只简单采集一部分数据（**压缩感知**），然后把复杂的部分交给数据还原的这一端来做，正好匹配了我们期望的格局。

- 在医学图像领域里，这个方案特别有好处，因为采集数据的过程往往是对病人带来很大麻烦甚至身体伤害的过程。以 X 光断层扫描为例，压缩感知意味着我们可以用比经典方法少得多的辐射剂量来进行数据采集，这在医学上的意义是不言而喻的。

State 2: Mathematics of CS

➢Bases & Frames (1. Bases)

A set $\{\phi_i\}_{i=1}^n$ is called a basis for $\mathbb{R}^n$ if the vectors in the set span $\mathbb{R}^n$ and are linearly independent.[2] This implies that each vector in the space has a unique representation as a linear combination of these basis vectors. Specifically, for any $x \in \mathbb{R}^n$, there exist (unique) coefficients $\{c_i\}_{i=1}^n$ such that

$$x = \sum_{i=1}^{n} c_i \phi_i.$$

Note that if we let Φ denote the $n \times n$ matrix with columns given by ϕ_i and let c denote the length-n vector with entries c_i, then we can represent this relation more compactly as

$$x = \Phi c.$$

An important special case of a basis is an orthonormal basis, defined as a set of vectors $\{\phi_i\}_{i=1}^n$ satisfying

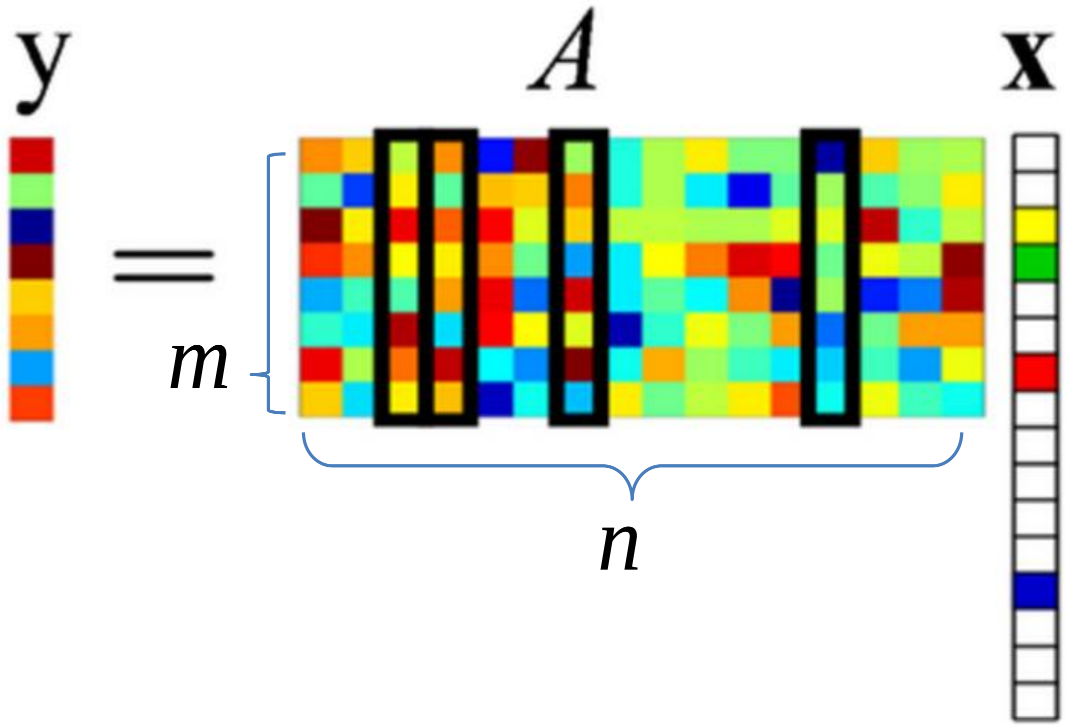
$$\langle \phi_i, \phi_j \rangle = \begin{cases} 1, & i = j; \\ 0, & i \neq j. \end{cases}$$

An orthonormal basis has the advantage that the coefficients c can be easily calculated as

$$c_i = \langle x, \phi_i \rangle,$$

Some Basic Concepts

- Bases & Frames (2. Frames)



由于A矩阵中列数大于行数，故一定线性相关！

➢Bases & Frames (2. Frames)

It is often useful to generalize the concept of a basis to allow for sets of possibly linearly dependent vectors, resulting in what is known as a *frame* [48, 55, 65, 163, 164, 182]. More formally, a frame is a set of vectors $\{\phi_i\}_{i=1}^n$ in $\mathbb{R}^d$, $d < n$ corresponding to a matrix $\Phi \in \mathbb{R}^{d\times n}$, such that for all vectors $x \in \mathbb{R}^d$,

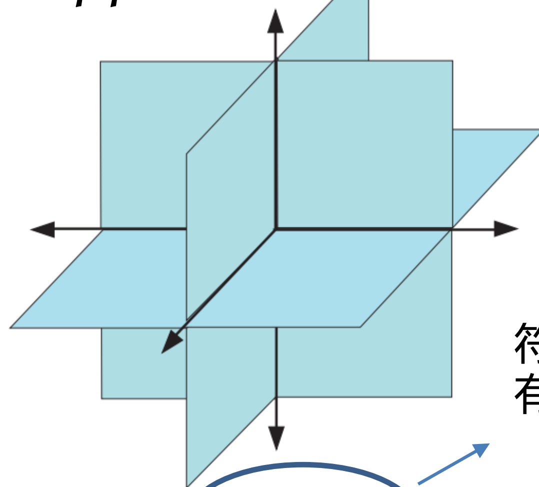
$$A\|x\|_2^2 \le \left\|\Phi^T x\right\|_2^2 \le B\|x\|_2^2$$

with $0 < A \le B < \infty$. Note that the condition $A > 0$ implies that the rows of Φ must be linearly independent. When A is chosen as the largest possible value and B as the smallest for these inequalities to hold, then we call them the *(optimal) frame bounds*. If A and B can be chosen as $A = B$, then the frame is called *A-tight*, and if $A = B = 1$, then Φ is a *Parseval frame*. A frame is called *equal-norm*, if there exists some $\lambda > 0$ such that $\|\phi_i\|_2 = \lambda$ for all $i = 1, \ldots, n$, and it is *unit-norm* if $\lambda = 1$. Note also that while the concept of a frame is very general and can be defined in infinite-dimensional spaces, in the case where Φ is a $d \times n$ matrix A and B simply correspond to the smallest and largest eigenvalues of $\Phi\Phi^T$, respectively.

Some Basic Concepts

数据挖掘实验室
Data Mining Lab

- *Sparsity and nonlinear approximation*



符号说明：二维子空间，其坐标最多有2个非0值

Figure 1.5 Union of subspaces defined by $\Sigma_2 \subset \mathbb{R}^3$, i.e., the set of all 2-sparse signals in $\mathbb{R}^3$.

Geometry of sparse signals

Sparsity is a highly nonlinear model, since the choice of which dictionary elements are used can change from signal to signal [77]. This can be seen by observing that given a pair of k-sparse signals, a linear combination of the two signals will in general no longer be k sparse, since their supports may not coincide. That is, for any $x, z \in \Sigma_k$, we do not necessarily have that $x + z \in \Sigma_k$ (although we do have that $x + z \in \Sigma_{2k}$). This is illustrated in Fig. 1.5, which shows Σ_2 embedded in $\mathbb{R}^3$, i.e., the set of all 2-sparse signals in $\mathbb{R}^3$.

➤ *Sparsity and nonlinear approximation*

are well approximated by the first few principal components [139]. In fact, we can quantify the compressibility by calculating the error incurred by approximating a signal x by some $\widehat{x} \in \Sigma_k$:

$$\sigma_k(x)_p = \min_{\widehat{x} \in \Sigma_k} \|x - \widehat{x}\|_p \, . \tag{1.2}$$

If $x \in \Sigma_k$, then clearly $\sigma_k(x)_p = 0$ for any p. Moreover, one can easily show that the thresholding strategy described above (keeping only the k largest coefficients) results in the optimal approximation as measured by (1.2) for all ℓ_p norms [77].

其中$\hat{x}$是经过解方程：$y = Ax$得到的解$\hat{x}$，这个解与原x的相似程度，**自然与A矩阵的结构有关！**

Important Variables of Matrix

➢三个描述矩阵A的incoherent的量

1、**Spark**
2、**NSP (null space property)**
3、**RIP (restricted isometry property)**

Transform

$\mu(A)$：**coherence of a matrix A**

Important Variables and Guarantees

数据挖掘实验室
Data Mining Lab

➢Null space(零空间) property and Spark

A natural place to begin is by considering the null space of A, denoted

$$\mathcal{N}(A) = \{z : Az = 0\}.$$

If we wish to be able to recover *all* sparse signals x from the measurements Ax, then it is immediately clear that for any pair of distinct vectors $x, x' \in \Sigma_k$, we must have $Ax \neq Ax'$, since otherwise it would be impossible to distinguish x from x' based solely on the measurements y. More formally, by observing that if $Ax = Ax'$ then $A(x - x') = 0$ with $x - x' \in \Sigma_{2k}$, we see that A uniquely represents all $x \in \Sigma_k$ if and only if $\mathcal{N}(A)$ contains no vectors in Σ_{2k}. While there are many equivalent ways of characterizing this property, one of the most common is known as the *spark* [86].

Definition 1.1. *The spark of a given matrix A is the smallest number of columns of A that are linearly dependent.*

Important Variables and Guarantees

数据挖掘实验室
Data Mining Lab

➢Spark guarantee

Theorem 1.1 (Corollary 1 of [86]). *For any vector $y \in \mathbb{R}^m$, there exists at most one signal $x \in \Sigma_k$ such that $y = Ax$ if and only if* $\operatorname{spark}(A) > 2k$.

Proof. We first assume that, for any $y \in \mathbb{R}^m$, there exists at most one signal $x \in \Sigma_k$ such that $y = Ax$. Now suppose for the sake of a contradiction that $\operatorname{spark}(A) \leq 2k$. This means that there exists some set of at most $2k$ columns that are linearly independent, which in turn implies that there exists an $h \in \mathcal{N}(A)$ such that $h \in \Sigma_{2k}$. In this case, since $h \in \Sigma_{2k}$ we can write $h = x - x'$, where $x, x' \in \Sigma_k$. Thus, since $h \in \mathcal{N}(A)$ we have that $A(x - x') = 0$ and hence $Ax = Ax'$. But this contradicts our assumption that there exists at most one signal $x \in \Sigma_k$ such that $y = Ax$. Therefore, we must have that $\operatorname{spark}(A) > 2k$.

Now suppose that $\operatorname{spark}(A) > 2k$. Assume that for some y there exist $x, x' \in \Sigma_k$ such that $y = Ax = Ax'$. We therefore have that $A(x - x') = 0$. Letting $h = x - x'$, we can write this as $Ah = 0$. Since $\operatorname{spark}(A) > 2k$, all sets of up to $2k$ columns of A are linearly independent, and therefore $h = 0$. This in turn implies $x = x'$, proving the theorem. □

It is easy to see that $\operatorname{spark}(A) \in [2, m + 1]$. Therefore, Theorem 1.1 yields the requirement $m \geq 2k$. 其中m为矩阵的行数，即列秩的最大值

Important Variables and Guarantees

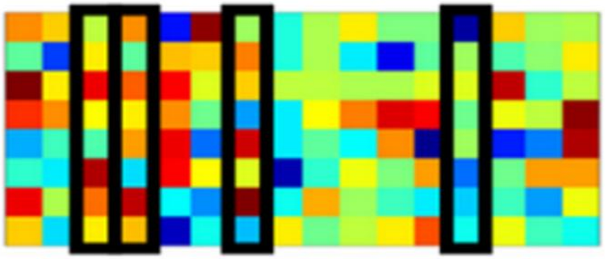
➢NSP(Null space property) 定义

tors that are sparse. In order to state the formal definition we define the following notation that will prove to be useful throughout much of this book. Suppose that $\Lambda \subset \{1, 2, \ldots, n\}$ is a subset of indices and let $\Lambda^c = \{1, 2, \ldots, n\} \backslash \Lambda$. By x_Λ we typically mean the length n vector obtained by setting the entries of x indexed by Λ^c to zero. Similarly, by A_Λ we typically mean the $m \times n$ matrix obtained by setting the columns of A indexed by Λ^c to zero.[6]

Definition 1.2. *A matrix A satisfies the* null space property *(NSP) of order k if there exists a constant $C > 0$ such that,*

$$\|h_\Lambda\|_2 \le C \frac{\|h_{\Lambda^c}\|_1}{\sqrt{k}} \tag{1.5}$$

holds for all $h \in \mathcal{N}(A)$ and for all Λ such that $|\Lambda| \le k$.



Important Variables and Guarantees

数据挖掘实验室

Data Mining Lab

➢NSP(Null space property) guarantee

不证明

To fully illustrate the implications of the NSP in the context of sparse recovery, we now briefly discuss how we will measure the performance of sparse recovery algorithms when dealing with general non-sparse x. Towards this end, let $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ represent our specific recovery method. We will focus primarily on guarantees of the form

$$\|\Delta(Ax) - x\|_2 \leq C\frac{\sigma_k(x)_1}{\sqrt{k}} \tag{1.6}$$

Theorem 1.2 (Theorem 3.2 of [57]). *Let $A : \mathbb{R}^n \to \mathbb{R}^m$ denote a sensing matrix and $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ denote an arbitrary recovery algorithm. If the pair (A, Δ) satisfies (1.6) then A satisfies the NSP of order $2k$.*

Proof. Suppose $h \in \mathcal{N}(A)$ and let Λ be the indices corresponding to the $2k$ largest entries of h. We next split Λ into Λ_0 and Λ_1, where $|\Lambda_0| = |\Lambda_1| = k$. Set $x = h_{\Lambda_1} + h_{\Lambda^c}$ and $x' = -h_{\Lambda_0}$, so that $h = x - x'$. Since by construction $x' \in \Sigma_k$, we can apply (1.6) to obtain $x' = \Delta(Ax')$. Moreover, since $h \in \mathcal{N}(A)$, we have

$$Ah = A(x - x') = 0$$

so that $Ax' = Ax$. Thus, $x' = \Delta(Ax)$. Finally, we have that

$$\|h_\Lambda\|_2 \leq \|h\|_2 = \|x - x'\|_2 = \|x - \Delta(Ax)\|_2 \leq C\frac{\sigma_k(x)_1}{\sqrt{k}} = \sqrt{2}C\frac{\|h_{\Lambda^c}\|_1}{\sqrt{2k}},$$

Important Variables and Guarantees

数据挖掘实验室
Data Mining Lab

➢RIP(Restricted isometry property) guarantee

最出名

Candès and Tao introduced the following isometry condition on matrices A and established its important role in CS.

Definition 1.3. *A matrix A satisfies the* restricted isometry property *(RIP) of order k if there exists a $\delta_k \in (0,1)$ such that*

$$\boxed{(1-\delta_k)\,\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta_k)\,\|x\|_2^2\,,} \tag{1.7}$$

holds for all $x \in \Sigma_k$.

If a matrix A satisfies the RIP of order $2k$, then we can interpret (1.7) as saying that A approximately preserves the distance between any pair of k-sparse vectors. This will clearly have fundamental implications concerning robustness to noise. Moreover, the potential applications of such *stable embeddings* range

➢RIP等价定义形式

最出名

It is important to note that while in our definition of the RIP we assume bounds that are symmetric about 1, this is merely for notational convenience. In practice, one could instead consider arbitrary bounds

$$\boxed{\alpha \|x\|_2^2 \leq \|Ax\|_2^2 \leq \beta \|x\|_2^2}$$

where $0 < \alpha \leq \beta < \infty$. Given any such bounds, one can always scale A so that it satisfies the symmetric bound about 1 in (1.7). Specifically, multiplying A by $\sqrt{2/(\beta+\alpha)}$ will result in an $\widetilde{A}$ that satisfies (1.7) with constant $\delta_k = (\beta - \alpha)/(\beta+\alpha)$. While we will not explicitly show this, one can check that all of the theorems in this chapter based on the assumption that A satisfies the RIP actually hold as long as there exists some scaling of A that satisfies the RIP. Thus, since we can always scale A to satisfy (1.7), we lose nothing by restricting our attention to this simpler bound.

Important Variables and Guarantees

数据挖掘实验室
Data Mining Lab

➢RIP(Restricted isometry property) guarantee

最出名

Candès and Tao introduced the following isometry condition on matrices A and established its important role in CS.

Definition 1.3. *A matrix A satisfies the* restricted isometry property *(RIP) of order k if there exists a $\delta_k \in (0,1)$ such that*

$$(1-\delta_k)\,\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta_k)\,\|x\|_2^2\,, \tag{1.7}$$

holds for all $x \in \Sigma_k$.

If a matrix A satisfies the RIP of order $2k$, then we can interpret (1.7) as saying that A approximately preserves the distance between any pair of k-sparse vectors. This will clearly have fundamental implications concerning robustness to noise. Moreover, the potential applications of such *stable embeddings* range

Important Variables and Guarantees

数据挖掘实验室 Data Mining Lab

➢定义 *pair* (*A*, Δ) 的C-stable来证明RIP下界

Definition 1.4. *Let* $A : \mathbb{R}^n \to \mathbb{R}^m$ *denote a sensing matrix and* $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ *denote a recovery algorithm. We say that the pair* (A, Δ) *is* C-stable *if for any* $x \in \Sigma_k$ *and any* $e \in \mathbb{R}^m$ *we have that*

$$\boxed{\|\Delta(Ax + e) - x\|_2 \le C \|e\|_2 .}$$

C-stable 说明了在y=Ax的输出端（观测端）产生的扰动（噪音）e对原始信号x的很小，上界为扰动的二范数乘一个常数！！

Important Variables and Guarantees

数据挖掘实验室
Data Mining Lab

➤RIP上界：不是很必要

One might respond to this result by arguing that since the upper bound is not necessary, we can avoid redesigning A simply by rescaling A so that as long as A satisfies the RIP with $\delta_{2k} < 1$, the rescaled version αA will satisfy (1.8) for any constant C. In settings where the size of the noise is independent of our choice of A, this is a valid point — by scaling A we are essentially adjusting the gain on the "signal" part of our measurements, and if increasing this gain does not impact the noise, then we can achieve arbitrarily high signal-to-noise ratios, so that eventually the noise is negligible compared to the signal.

However, in practice we will typically not be able to rescale A to be arbitrarily large. Moreover, in many practical settings the noise is not independent of A. For example, consider the case where the noise vector e represents quantization noise produced by a finite dynamic range quantizer with B bits. Suppose the measurements lie in the interval $[-T, T]$, and we have adjusted the quantizer to capture this range. If we rescale A by α, then the measurements now lie between $[-\alpha T, \alpha T]$, and we must scale the dynamic range of our quantizer by α. In this case the resulting quantization error is simply αe, and we have achieved *no reduction* in the reconstruction error.

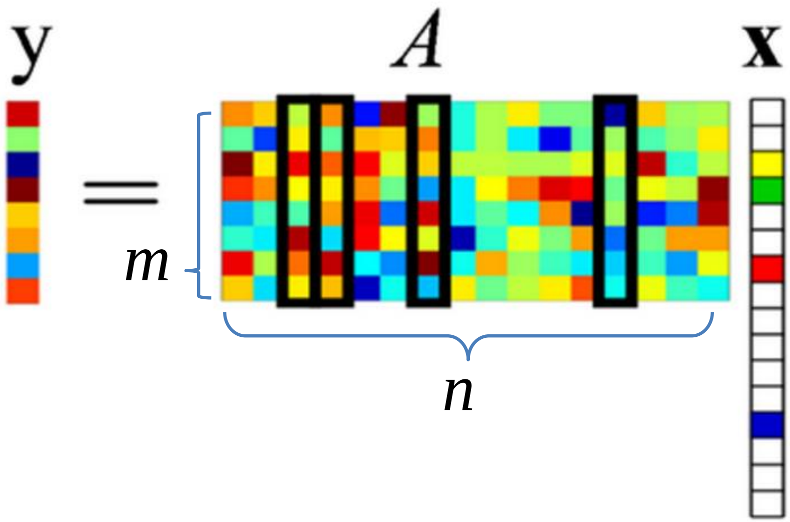
Important Variables and Guarantees

➢由RIP可以得到的重要结论

Theorem 1.4 (Theorem 3.5 of [67]). *Let A be an $m \times n$ matrix that satisfies the RIP of order $2k$ with constant $\delta \in (0, \frac{1}{2}]$. Then*

$$\boxed{m \geq Ck \log\left(\frac{n}{k}\right)}$$

where $C = 1/2 \log(\sqrt{24} + 1) \approx 0.28$.



不同的推导方式可以得出不同的下界！尤其是C的值。

Important Variables and Guarantees

数据挖掘实验室

Data Mining Lab

LESS IS MORE

➢衡量矩阵列间相关性的直接指标

不证明

Definition 1.5. *The coherence of a matrix A, $\mu(A)$, is the largest absolute inner product between any two columns a_i, a_j of A:*

$$\boxed{\mu(A) = \max_{1 \le i < j \le n} \frac{|\langle a_i, a_j \rangle|}{\|a_i\|_2 \|a_j\|_2}.}$$

Lemma 1.4. *For any matrix A,*

$$\operatorname{spark}(A) \ge 1 + \frac{1}{\mu(A)}.$$

Theorem 1.7 (Theorem 12 of [86]). *If*

$$k < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right),$$

then for each measurement vector $y \in \mathbb{R}^m$ there exists at most one signal $x \in \Sigma_k$ such that $y = Ax$.

Lemma 1.5. *If A has unit-norm columns and coherence $\mu = \mu(A)$, then A satisfies the RIP of order k with $\delta = (k-1)\mu$ for all $k < 1/\mu$.*

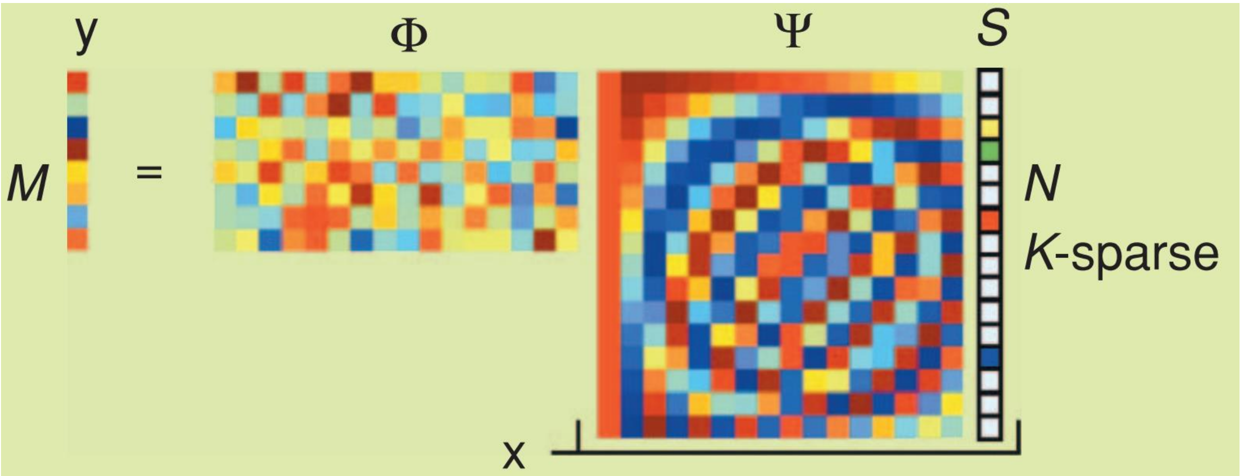
找到了μ(*A*)与Spark、NSP、RIP的关系

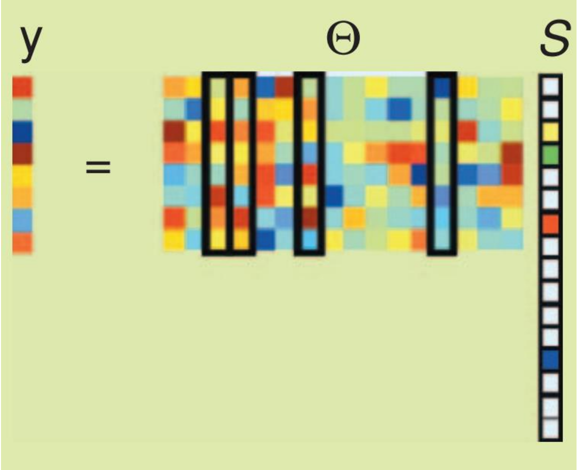
Issues of CS

数据挖掘实验室
Data Mining Lab

➢Matrix Construction

不证明





思路：随机构造矩阵，可以以很高概率恢复！

于是我们构造正态分布。正态分布的线性变换还是正态分布！

工程中的硬件不能做到随机构造，于是有各种构造方法

Issues of CS

数据挖掘实验室
Data Mining Lab

➢信号恢复(L0,L1 minimization)

不证明

Given measurements y and the knowledge that our original signal x is sparse or compressible, it is natural to attempt to recover x by solving an optimization problem of the form

$$\widehat{x} = \arg\min_{z} \|z\|_0 \quad \text{subject to} \quad z \in \mathcal{B}(y), \tag{1.10}$$

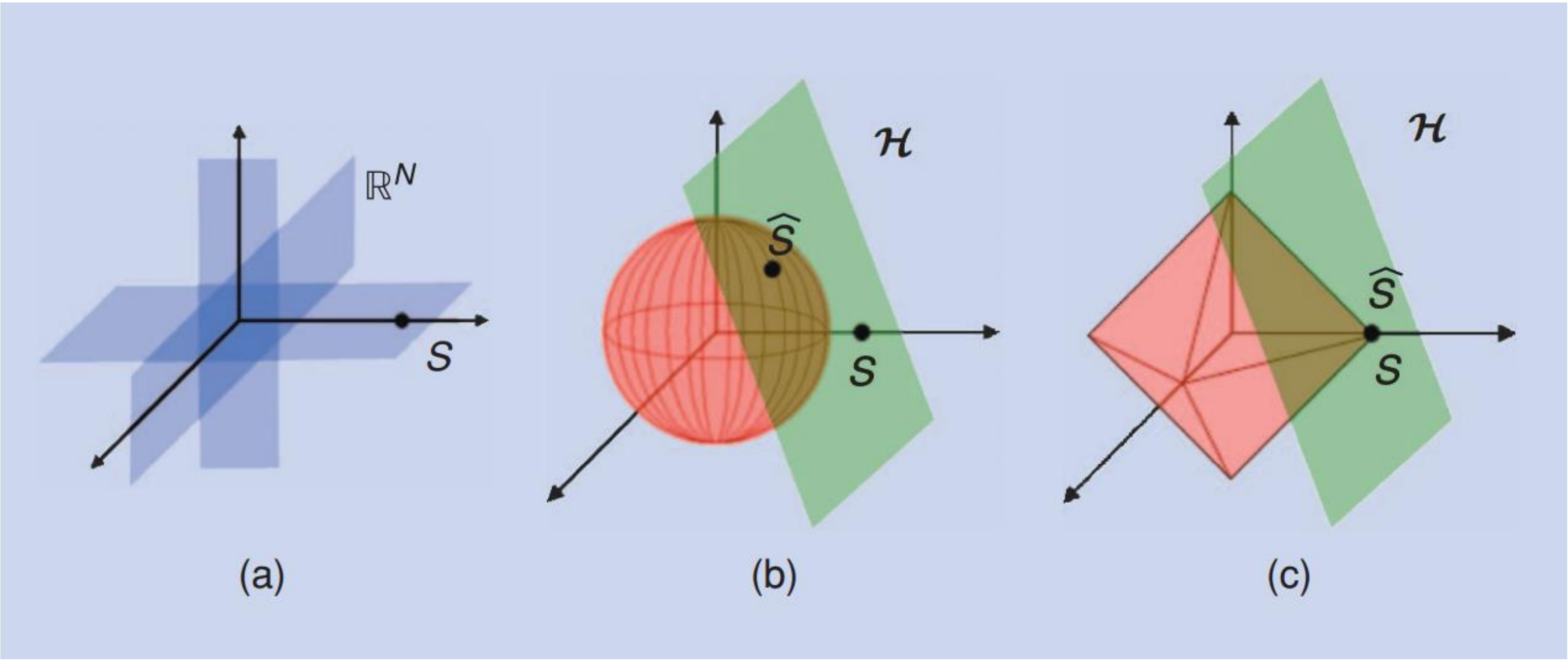
$$\Updownarrow$$

One avenue for translating this problem into something more tractable is to replace $\|\cdot\|_0$ with its convex approximation $\|\cdot\|_1$. Specifically, we consider

$$\widehat{x} = \arg\min_{z} \|z\|_1 \quad \text{subject to} \quad z \in \mathcal{B}(y). \tag{1.12}$$

Provided that $\mathcal{B}(y)$ is convex, (1.12) is computationally feasible. In fact, when $\mathcal{B}(y) = \{z : Az = y\}$, the resulting problem can be posed as a linear program [53].

➢L1 最小化几何解释



Issues of CS

数据挖掘实验室
Data Mining Lab

➢最出名的方法：MP

不证明

$$\min \|x\|_0, \quad s.t.\ Ax = b$$

并已知 $k < 1/2\mu$。OMP 算法的步骤如下：

1. 初始化：$x^0 = 0, r^0 = Ax^0 - b = b, S = \emptyset$
2. For $\ell = 1, \ldots, k$
    1. 令 j 为 $\arg\max |\langle A_j, r^{\ell-1}\rangle| / \|A_j\|_2^2$
    2. $S \leftarrow S \cup j$
    3. $r^\ell \leftarrow \mathrm{Proj}_{U^\perp}(b)$，其中 $U = \mathrm{span}(A_S)$
    4. 如果 $r^\ell = 0$ 则退出循环
3. 求解 $A_S x_S = b$ 得出 x 的非零系数。

首先注意到每一次迭代都会有一个新的下标 j 进入集合 S，因为如果 j 在之前的某一步中被加入到 S 中了的话，那么 residual r 和 A_j 的内积会保持为零，所以不会选到该下标，亦即 j 不会进入集合两次。所以经过 k 次迭代之后将会得到 k 个下标。接下来我们要说明这 k 个下标刚好对应了 x 的 k 个非零下标。令 $T = \{j : x_j \neq 0\}$，则我们要证明的是 $S = T$。

➢Equivalence of Several L1 Sparsity Problem

不证明

有这么几个 ℓ_1-norm 相关的稀疏优化问题在某种意义下是等价的。首先是如下问题：

$$x^a(\lambda) = \arg\min_x \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1$$

这个形式通常来自于 regularized linear regression 等问题，其中使用 ℓ_1-norm 作为 regularizer 会偏向于得到稀疏的解，我们之前也曾简要地讨论过这个问题。同样的形式还可以解释为使用 Laplace prior 的 MAP 参数估计所得到的目标函数。

另一个问题是主要来自于 Compressive Sensing 的如下形式：

$$x^b(\epsilon) = \arg\min_x \|x\|_1 \quad s.t. \|Ax - y\|^2 \le \epsilon$$

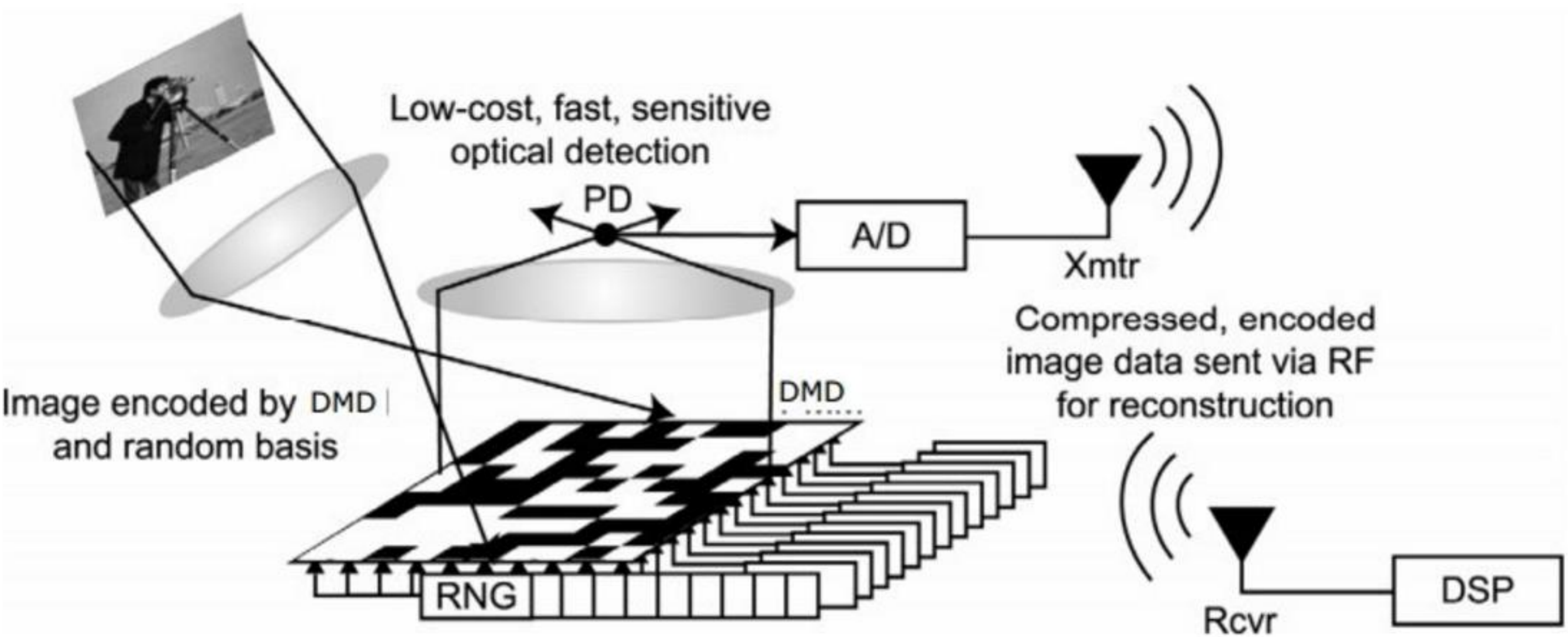
第三个是把目标函数和 constraint 的位置反过来一下的形式，这实际上是 LASSO 的原始形式：

$$x^c(t) = \arg\min_x \frac{1}{2}\|Ax - y\|^2 \quad s.t. \|x\|_1 \le t$$

http://freemind.pluskid.org/machine-learning/a-compressed-sense-of-compressive-sensing-iii/

Applications

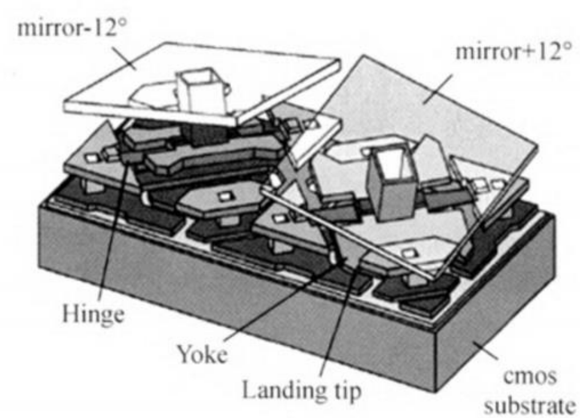
➢ Single-pixel cameras



http://blog.csdn.net/jbb0523/article/details/41288573

Applications

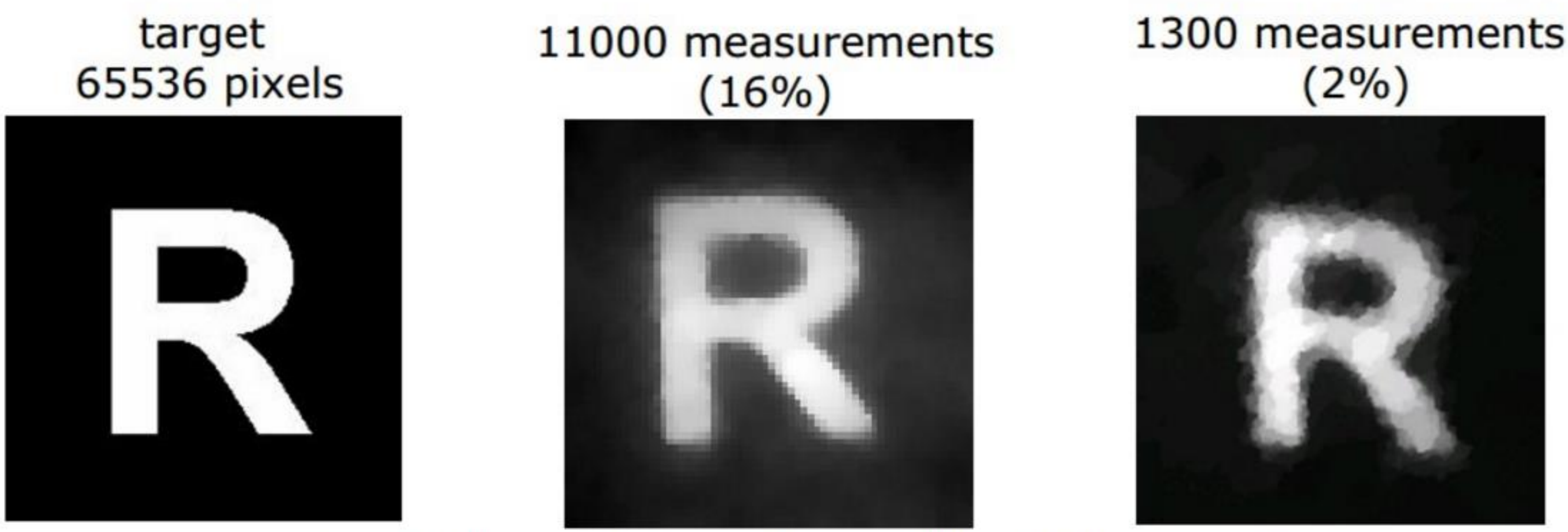
➢ Single-pixel cameras



http://blog.csdn.net/jbb0523/article/details/41288573

Applications

- Single-pixel cameras



test image(65536 pixels) and CS construction using 11000 and 1300 measurements

https://terrytao.wordpress.com/2007/04/13/compressed-sensing-and-single-pixel-cameras/

Conclusions

➢Compressed sensing 实际上是对信号采集的颠覆性的理论，打破了乃奎斯特采样（也称香农采样）。

➢我们主要关注的问题：

- **Matrix Construction**
- **Signals Recovery**

➢Compressed sensing 在数学推导上非常美，但发展十年来，由于其计算量大、限制条件难以满足，在实际应用中的作用十分局限。

References:

数据挖掘实验室
Data Mining Lab

1. Davenport, Mark A., et al. "Introduction to compressed sensing." *Preprint* 93.1 (2011): 2.
2. https://zhuanlan.zhihu.com/p/22445302
3. http://blog.csdn.net/jbb0523/article/details/41288573
4. http://freemind.pluskid.org/machine-learning/a-compressed-sense-of-compressive-sensing-i/
5. http://blog.csdn.net/abcjennifer/article/details/7721834

Epilogue

当初法拉第最初发现电磁感应的重要原理时，做了一个报告。

有个妇女做完等他做完报告之后，问法拉第：“这个又有什么用处呢！”

法拉第回答说："刚生下来的婴儿有什么用呢？"

想想看，现在我们周围不是全部都是电磁相关的东西吗？

CS sources:

数据挖掘实验室
Data Mining Lab

1. http://dsp.rice.edu/cs （英文论文最全资料库）
2. http://blog.csdn.net/abcjennifer/article/details/7724360 （中文论文最全资料库）
3. http://blog.csdn.net/abcjennifer/article/details/7721834
4. https://sites.google.com/site/igorcarron2/cs

Thanks

Chongming Gao
Yingcai Experimental School
gchongming@126.com

Linear transformation and its matrix

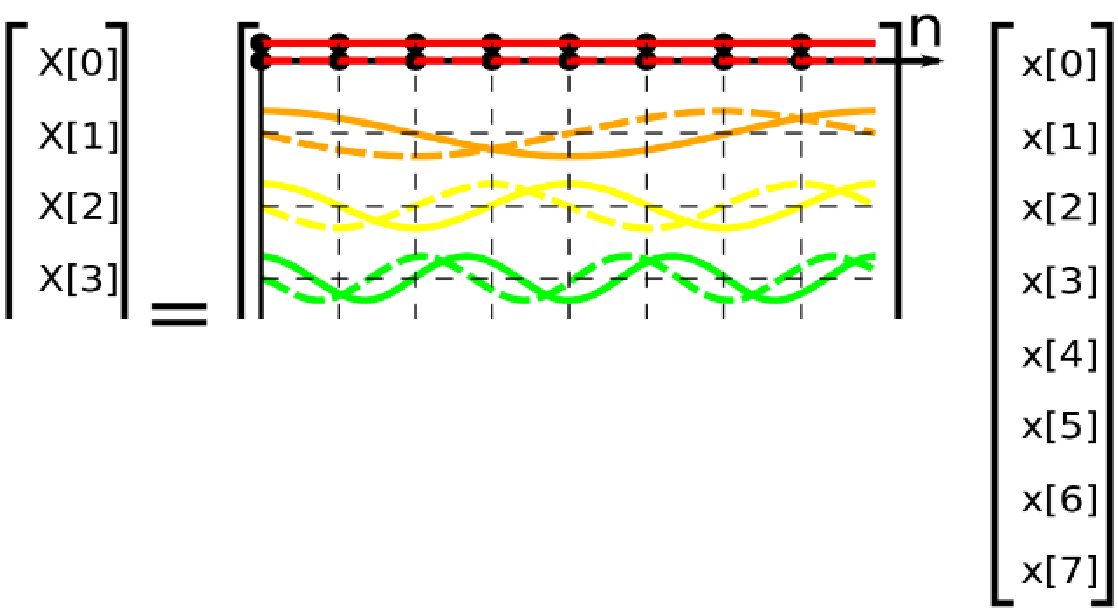
A specific linear transformation

$$T \in L(V, W)$$

$$Tv_k = a_{1,k}\omega_1 + \cdots + a_{1,m}\omega_m$$

The matrix of T on the basis V and W

$$\begin{array}{c c} & \begin{array}{ccccc} v_1 & \cdots & v_k & \cdots & v_n \end{array} \\ \begin{array}{c} \omega_1 \\ \vdots \\ \omega_m \end{array} & \left[\begin{array}{ccccc} & & a_{1,k} & & \\ & & \vdots & & \\ & & a_{m,k} & & \end{array}\right] \end{array}$$



Linear transformation and its matrix

A specific linear transformation

$$T \in L(V, W)$$

$$Tv_k = a_{1,k}\omega_1 + \cdots + a_{1,m}\omega_m$$

The matrix of T on the basis V and W

$$\begin{array}{cc} & \begin{array}{ccccc} v_1 & \cdots & v_k & \cdots & v_n \end{array} \\ \begin{array}{c} \omega_1 \\ \vdots \\ \omega_m \end{array} & \begin{bmatrix} & & a_{1,k} & & \\ & & \vdots & & \\ & & a_{m,k} & & \end{bmatrix} \end{array}$$

Example:

$$T(x, y) = (x + 3y, 2x + 5y, 7x + 9y)$$

$$\begin{bmatrix} 1 & 3 \\ 2 & 5 \\ 7 & 9 \end{bmatrix}$$

$$T(1,0) = (1,2,7), T(0,1) = (3,5,9)$$