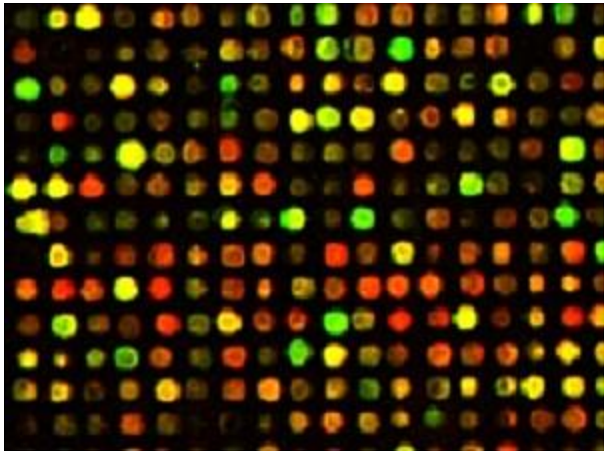
The Fundamentals of Compressive Sensing

Mark A. Davenport

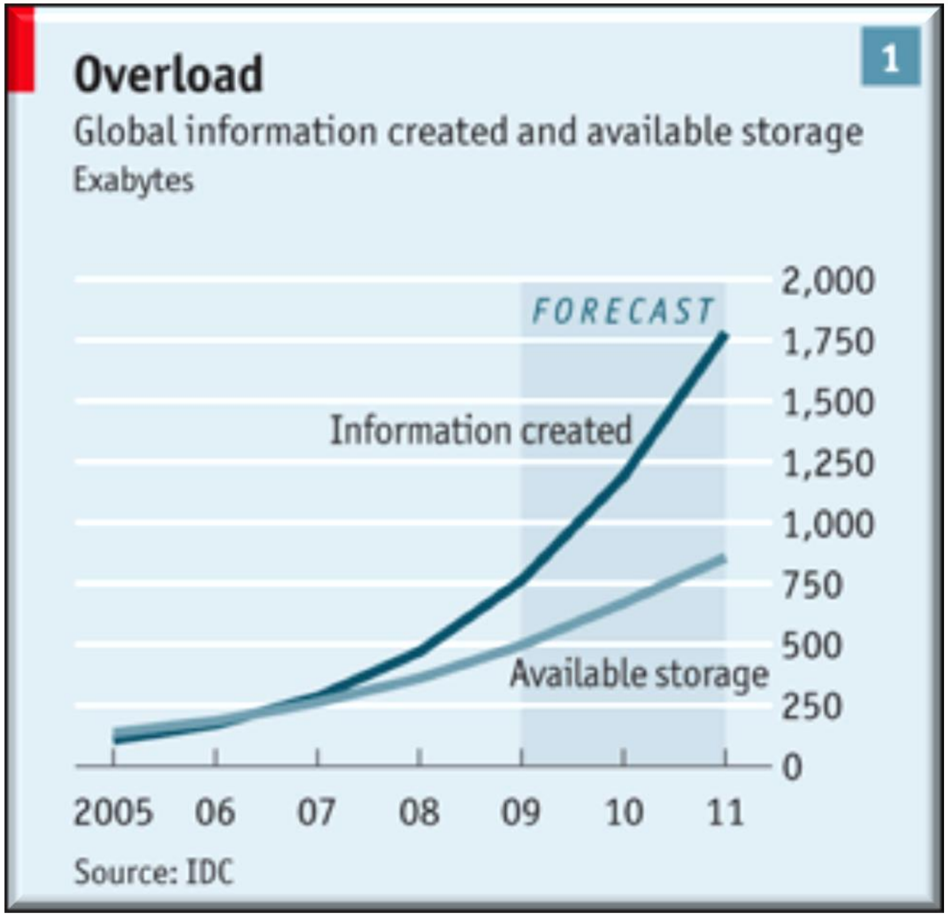
Georgia Institute of Technology
School of Electrical and Computer Engineering

Sensor explosion

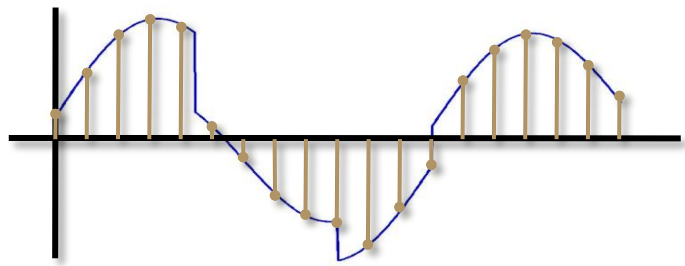


Data deluge


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

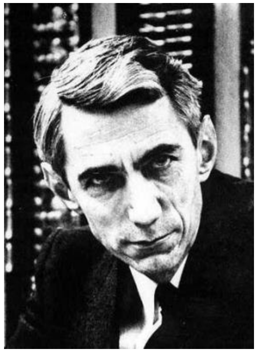

Overload
1
Global information created and available storage
Exabytes
FORECAST
Information created
Available storage
2,000
1,750
1,500
1,250
1,000
750
500
250
0
2005
06
07
08
09
10
11
Source: IDC

Digital revolution



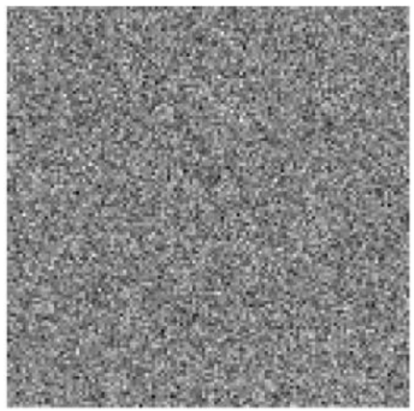
"If we sample a signal at twice its highest frequency, then we can recover it exactly."

Whittaker-Nyquist-Kotelnikov-Shannon



Dimensionality reduction

Data with high-frequency content is often not intrinsically high-dimensional



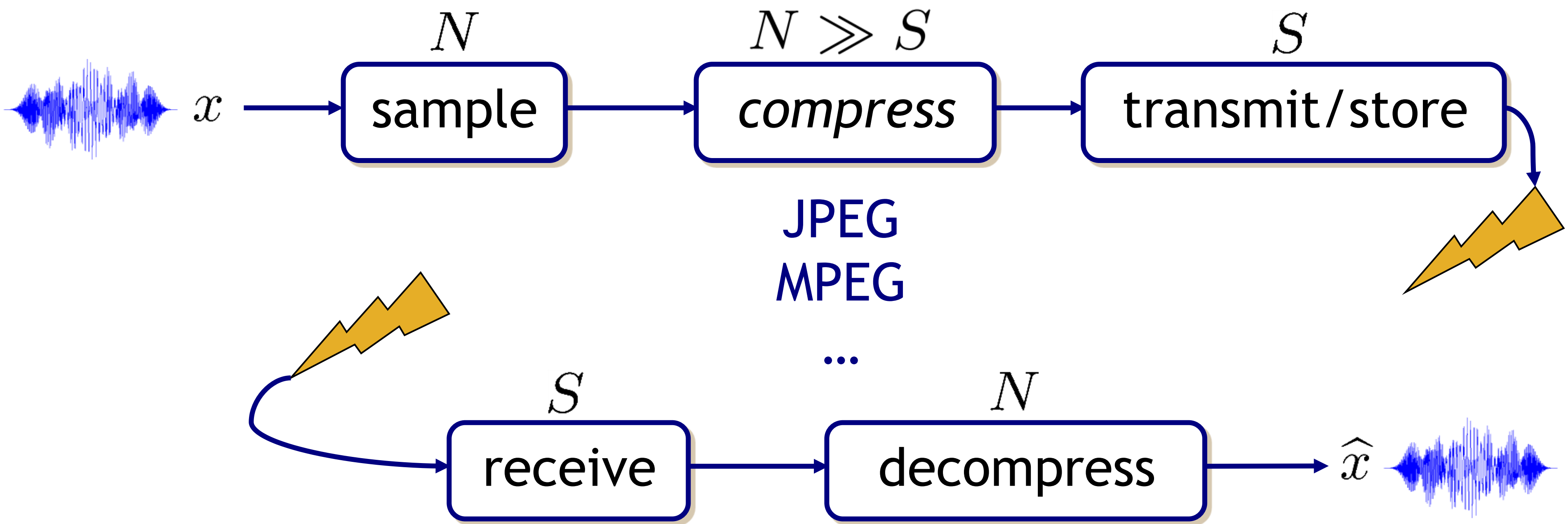
Signals often obey *low-dimensional models*

- sparsity
- manifolds
- low-rank matrices

The "intrinsic dimension" S can be much less than the "ambient dimension" N

Sample-then-compress paradigm

- Standard paradigm for digital data acquisition
  - *sample* data (ADC, digital camera, …)
  - *compress* data (signal-dependent, nonlinear)

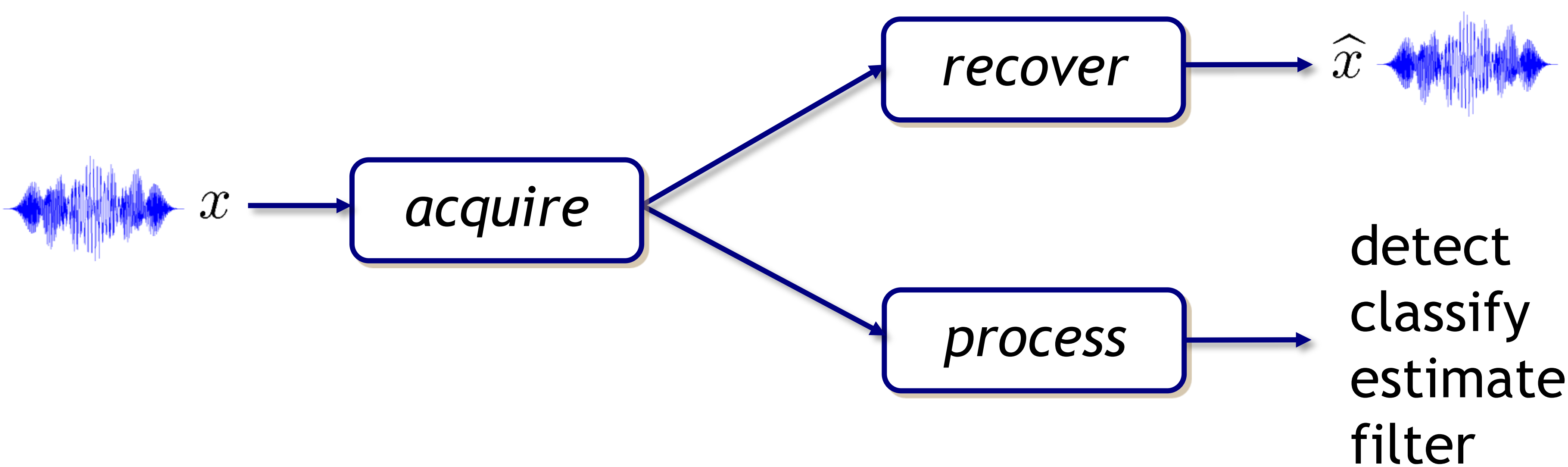


- Sample-and-compress paradigm is *wasteful*
  - samples cost $$$ and/or time

Exploiting low-dimensional structure

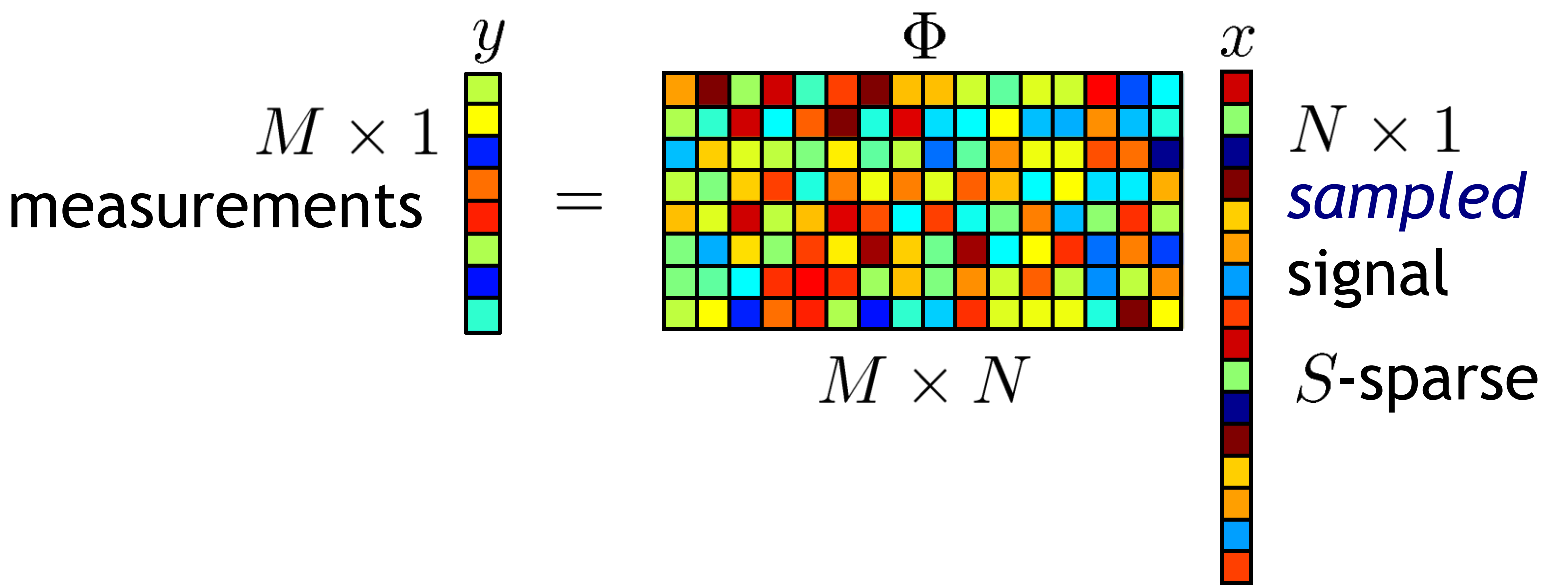
> How can we exploit low-dimensional structure in the design of signal processing algorithms?

We would like to operate at the *intrinsic dimension* at all stages of the information-processing pipeline



Compressive sensing

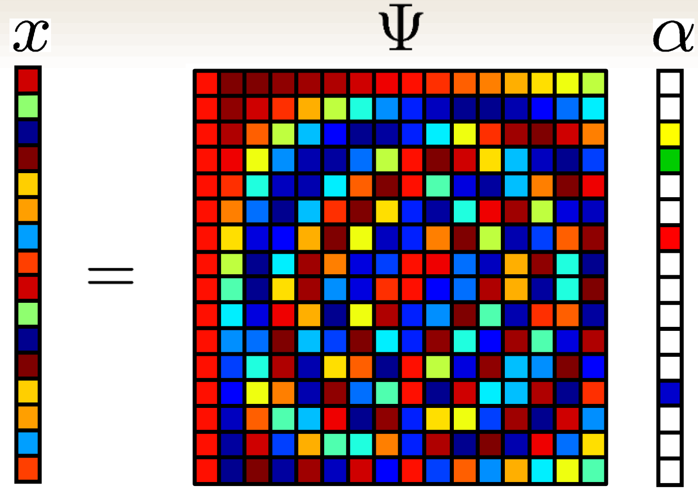
Replace samples with general *linear measurements*



[Donoho; Candès, Romberg, Tao - 2004]

Sparsity

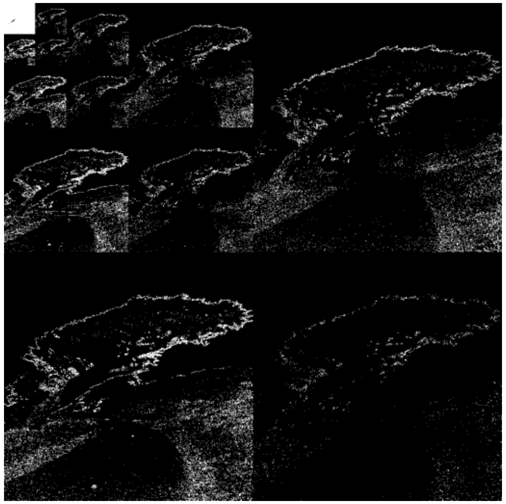
$$x = \sum_{j=1}^{N} \alpha_j \psi_j$$
$$= \Psi \alpha$$



S nonzero entries

$$\|\alpha\|_0 = S$$

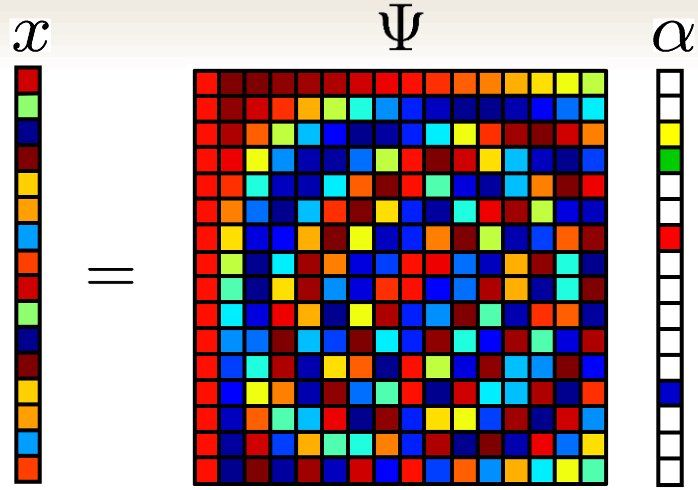
N pixels



$S \ll N$ large wavelet coefficients

Sparsity

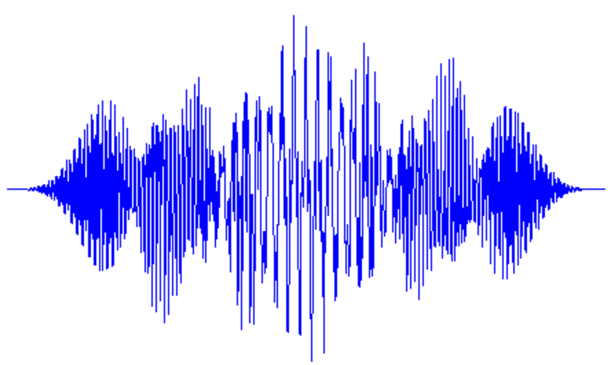
$$x = \sum_{j=1}^{N} \alpha_j \psi_j$$
$$= \Psi \alpha$$

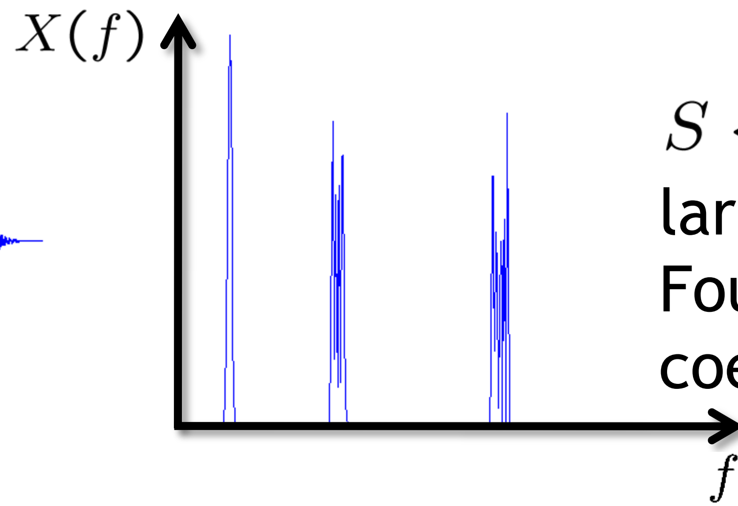


S nonzero entries

$\|\alpha\|_0 = S$

N samples

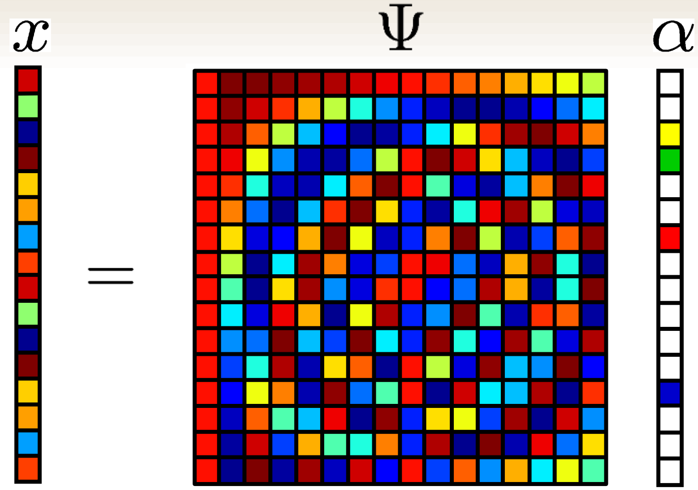




$S \ll N$ large Fourier coefficients

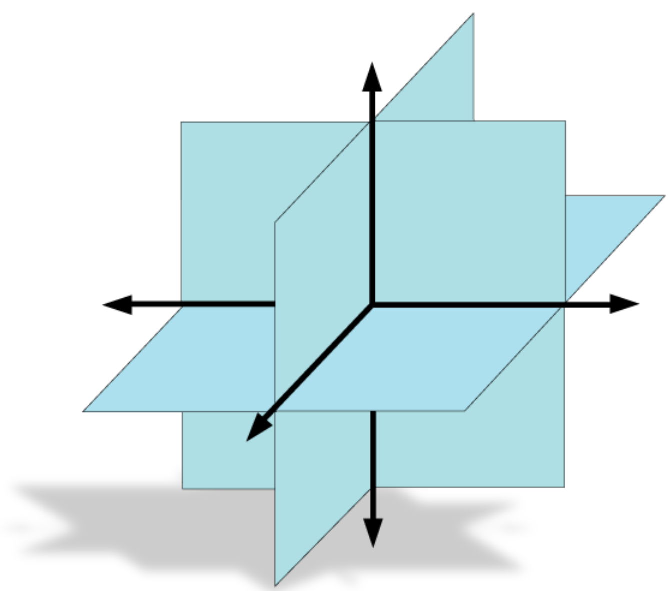
Sparsity

$$x = \sum_{j=1}^{N} \alpha_j \psi_j$$
$$= \Psi \alpha$$



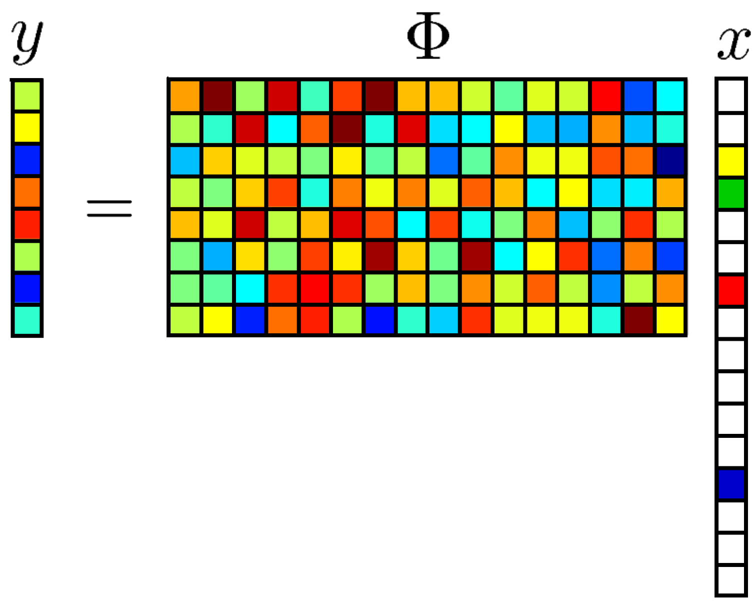
S nonzero entries

$$\|\alpha\|_0 = S$$



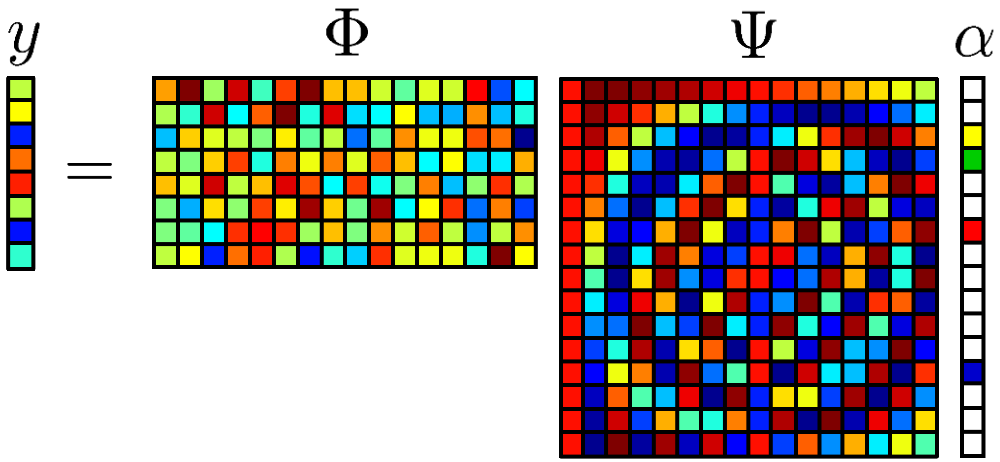
Core theoretical challenges

- How should we design the matrix Φ so that M is as small as possible?



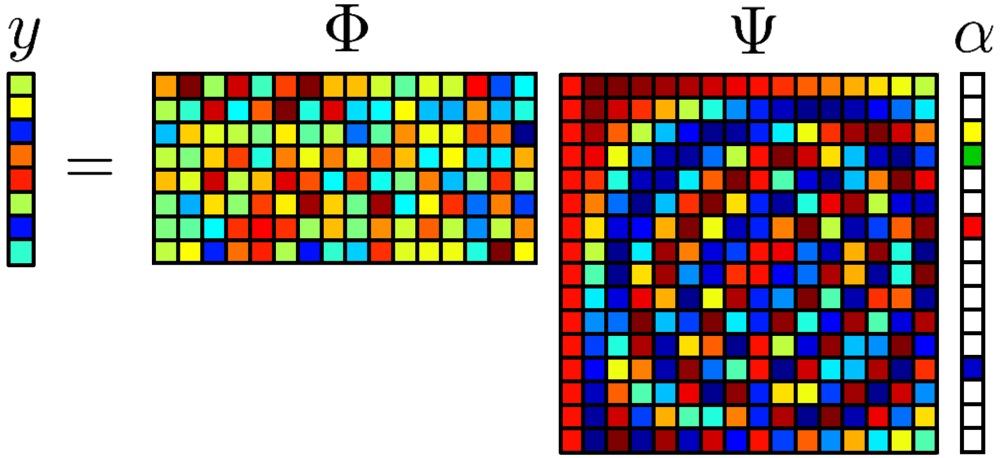
Core theoretical challenges

- How should we design the matrix Φ so that M is as small as possible?



Core theoretical challenges

- How should we design the matrix Φ so that M is as small as possible?



- How can we recover x from the measurements y ?

Outline

- Sensing matrices and real-world compressive sensors
  - (structured) randomness
  - tomography, cameras, ADCs, …

- Sparse signal recovery
  - convex optimization
  - greedy algorithms

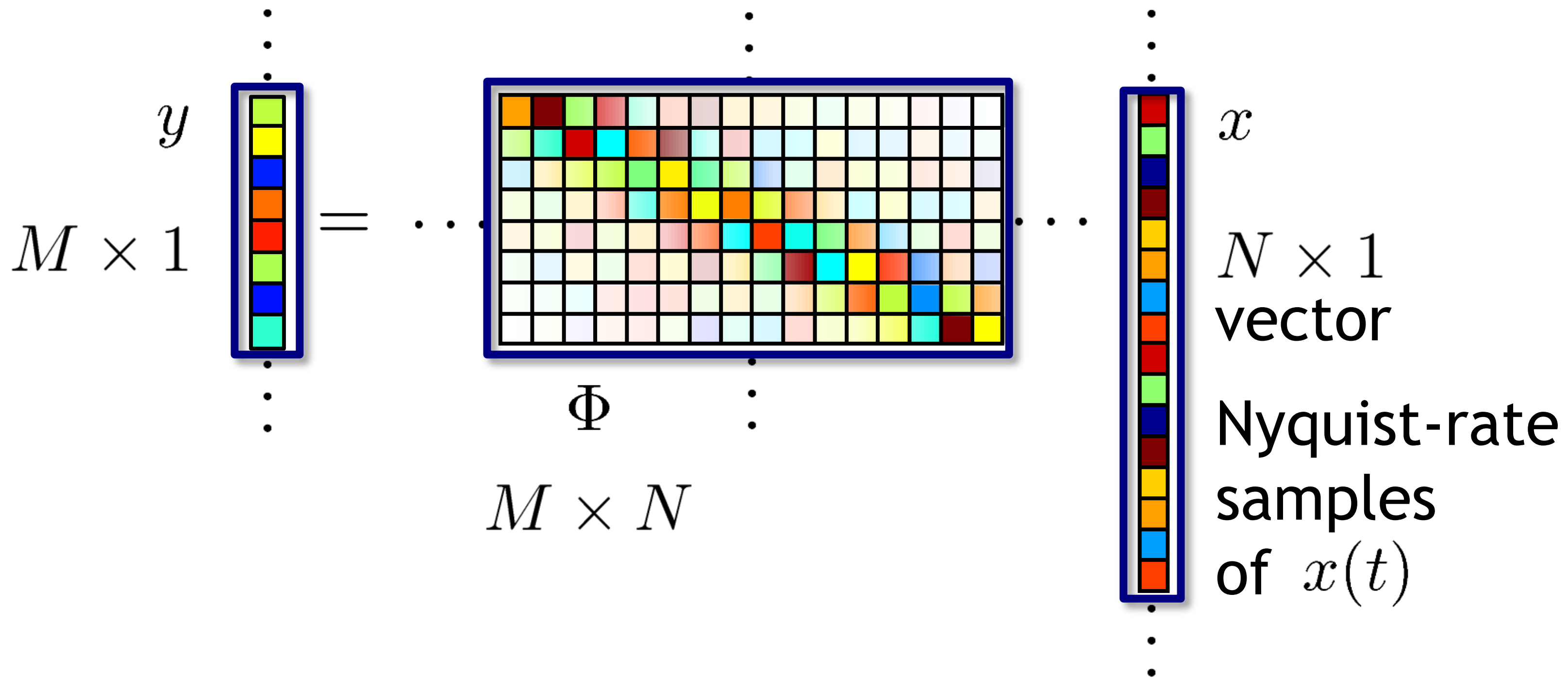
- Beyond sparsity
  - parametric models, manifolds, low-rank matrices, …

Sensing Matrix Design

Analog sensing *is* matrix multiplication

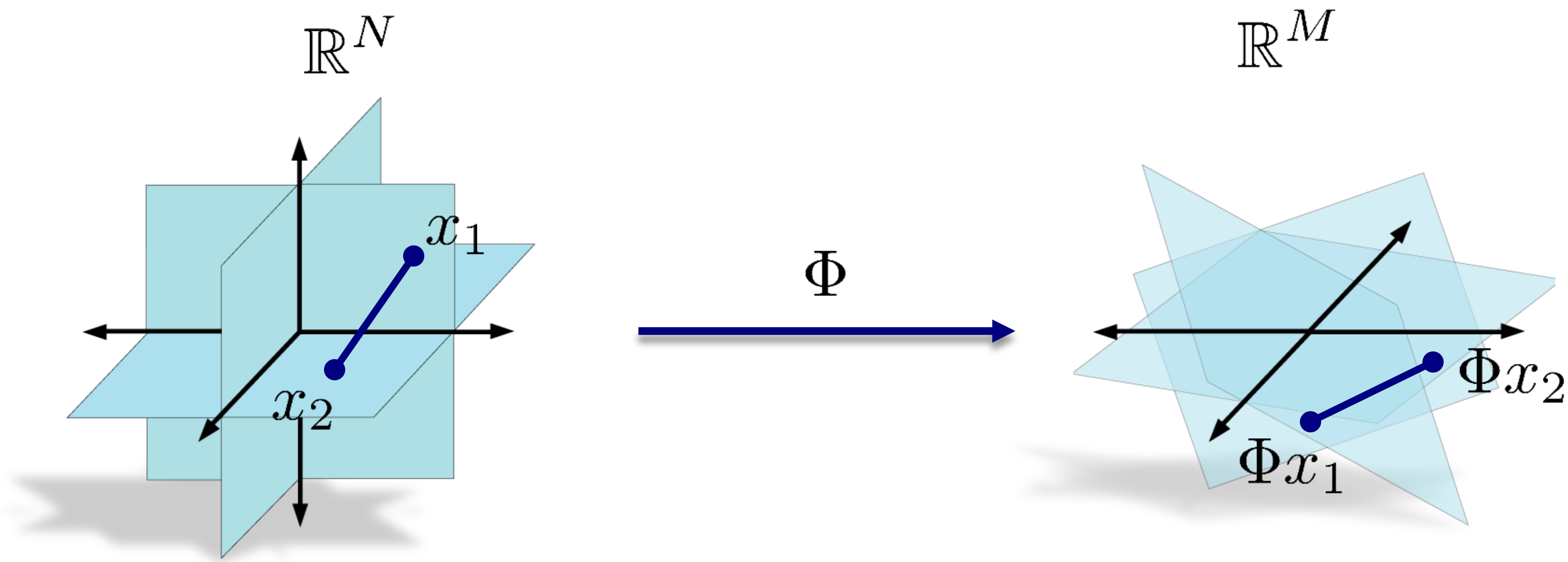
If $x(t)$ is bandlimited,

$$y[m] = \langle \phi_m(t), x(t) \rangle = \sum_{n=-\infty}^{\infty} x[n] \langle \phi_m(t), \mathrm{sinc}(t/T_s - n) \rangle$$



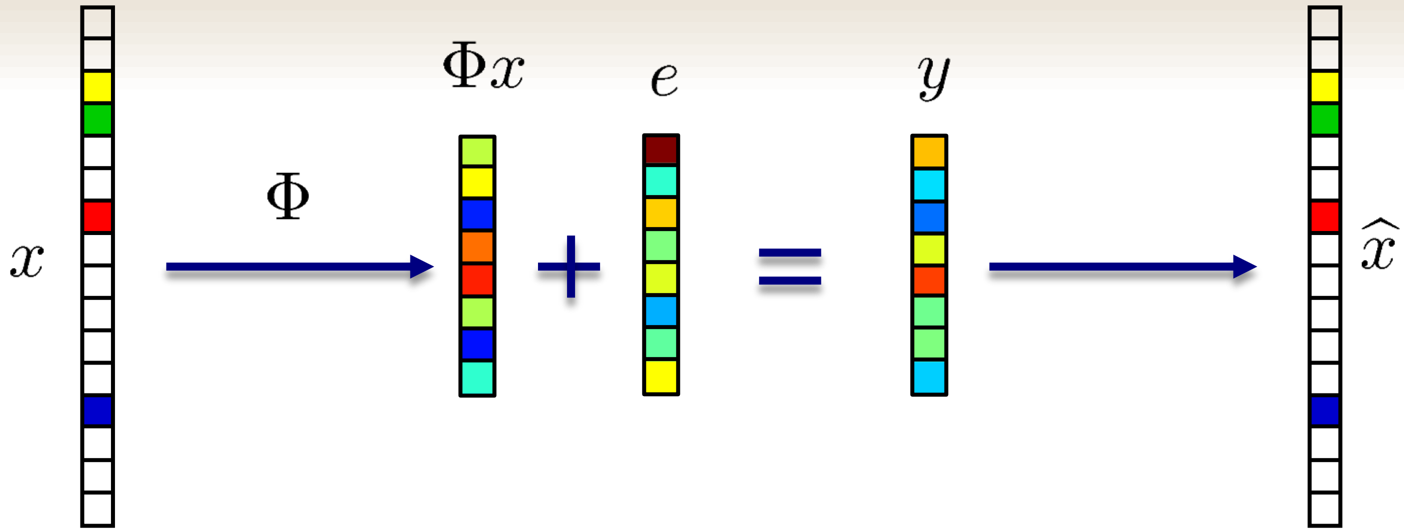
Restricted Isometry Property (RIP)

$$1 - \delta \leq \frac{\|\Phi x_1 - \Phi x_2\|_2^2}{\|x_1 - x_2\|_2^2} \leq 1 + \delta \qquad \|x_1\|_0, \|x_2\|_0 \leq S$$



$$1 - \delta \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq 1 + \delta \qquad \|x\|_0 \leq 2S$$

RIP and stability



If we want to guarantee that

$$\|x - \widehat{x}\|_2 \leq C \|e\|_2$$

then we must have

$$\boxed{\frac{1}{C} \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \qquad \|x\|_0 \leq 2S}$$

Sub-Gaussian distributions

- As a first example of a matrix Φ which satisfies the RIP, we will consider *random* constructions

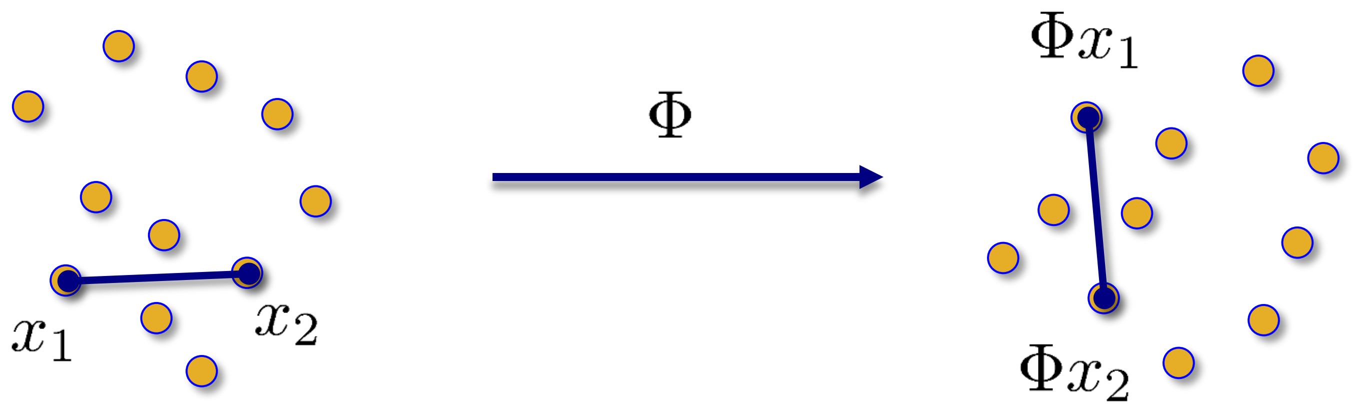
- Sub-Gaussian random variable: $\mathbb{E}\left(e^{Xt}\right) \leq e^{c^2t^2/2}$
  - Gaussian
  - Bernoulli/Rademacher (± 1)
  - any bounded distribution

- For any x, if the entries of Φ are sub-Gaussian, then there exists a δ such that with high probability

$$(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta)\|x\|_2^2$$

Johnson-Lindenstrauss Lemma

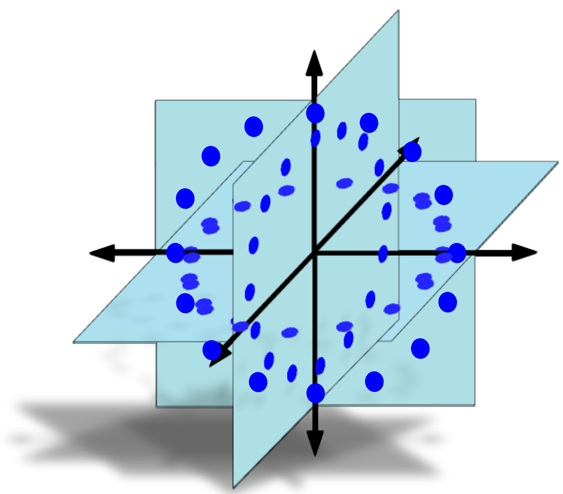
- Stable projection of a discrete set of P points



- Pick Φ at *random* using a sub-Gaussian distribution
- For any fixed x, $\|\Phi x\|_2$ concentrates around $\|x\|_2$ with (exponentially) high probability
- We preserve the length of all $O(P^2)$ difference vectors simultaneously if $M = O(\log P^2) = O(\log P)$.

JL Lemma meets RIP

$$1 - \delta \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq 1 + \delta \qquad \|x\|_0 \leq 2S$$

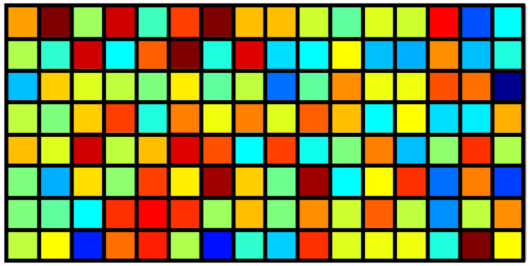


$$P = O\left((N/S)^S\right) \quad \Rightarrow \quad M = O(S \log(N/S))$$

[Baraniuk, Davenport, DeVore, Wakin –2008]

RIP matrix: Option 1

- Choose a *random matrix*
  - fill out the entries of Φ with i.i.d. samples from a sub-Gaussian distribution
  - project onto a “random subspace”

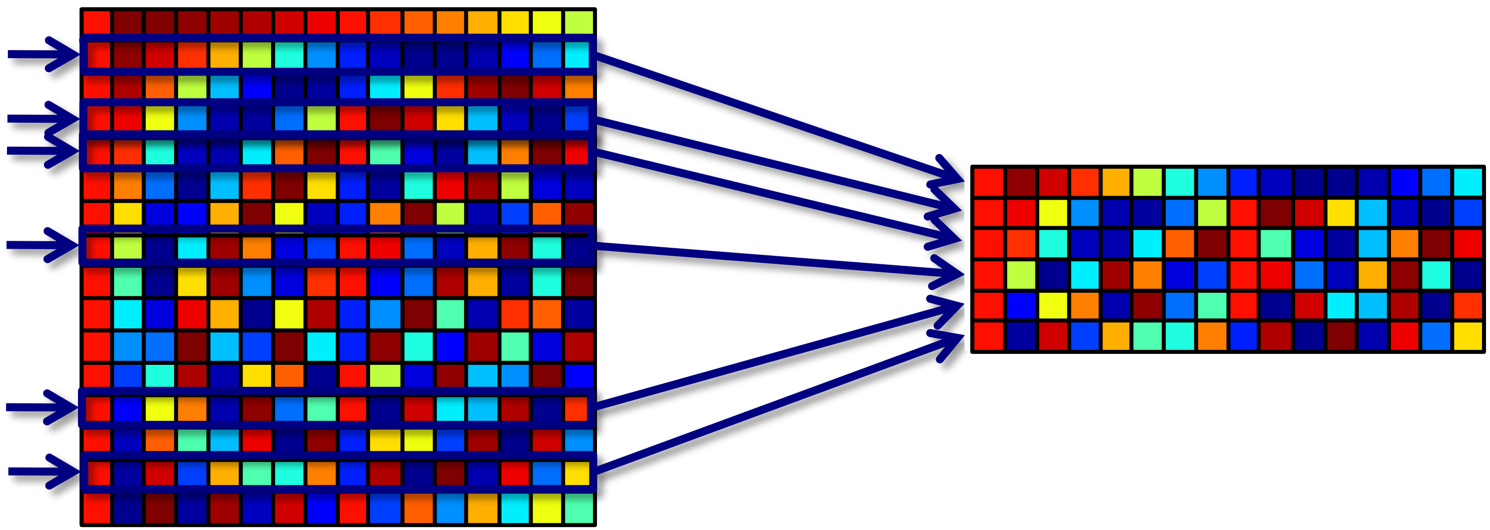


$$M = O(S \log(N/S)) \ll N$$

[Baraniuk, Davenport, DeVore, Wakin –2008]

RIP matrix: Option 2

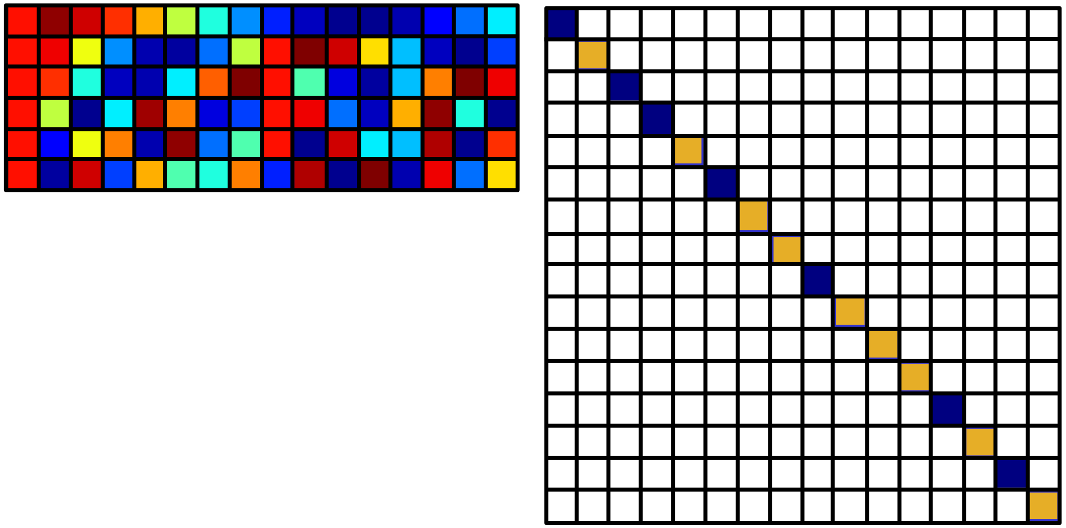
- Random Fourier submatrix



$$M = O(S \log^p (N/S)) \ll N$$

[Candès and Tao - 2006]

RIP matrix: Option 3
"Fast JL Transform"



- By first multiplying by random signs, a random Fourier submatrix can be used for efficient JL embeddings

- If you multiply the columns of *any* RIP matrix by random signs, you get a JL embedding!

[Ailon and Chazelle – 2007; Krahmer and Ward - 2010]

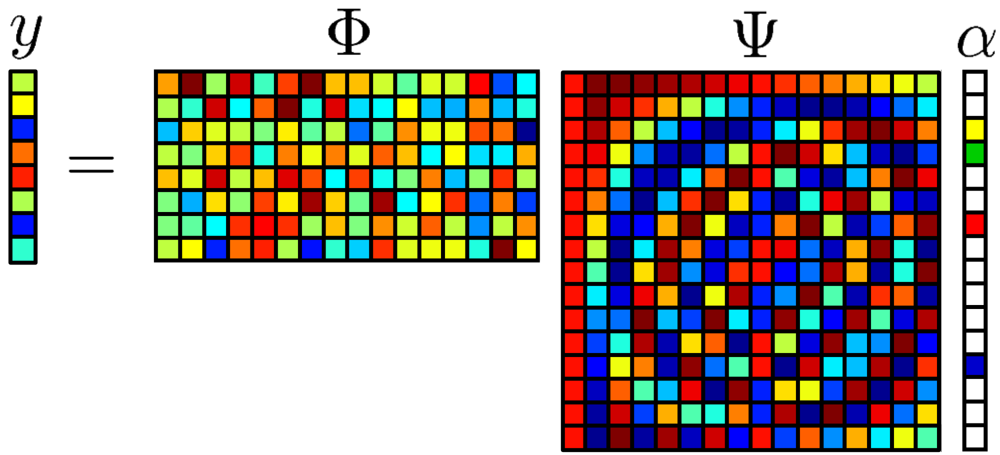
Hallmarks of random measurements

Stable

With high probability, Φ will preserve information, be robust to noise

***Universal** (Options 1 and 3)*

Φ will work with ***any*** fixed orthonormal basis (w.h.p.)

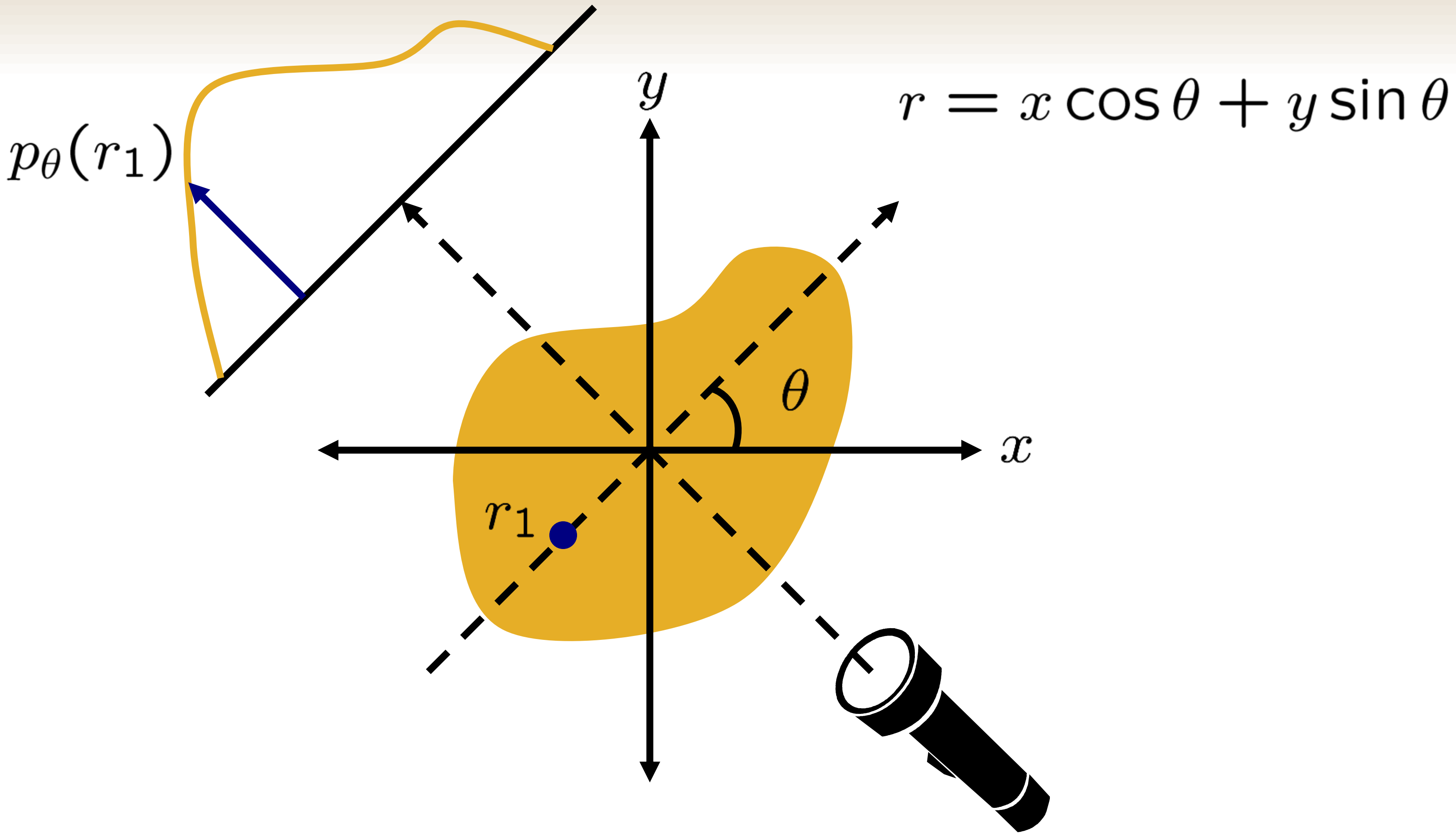


Democratic

Each measurement has "equal weight"

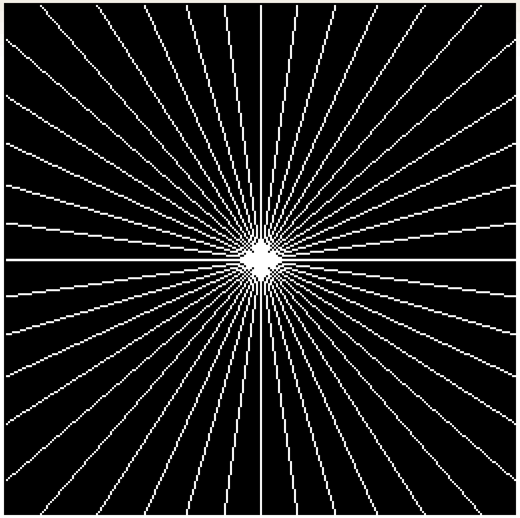
Compressive Sensors in Practice

Tomography in the abstract



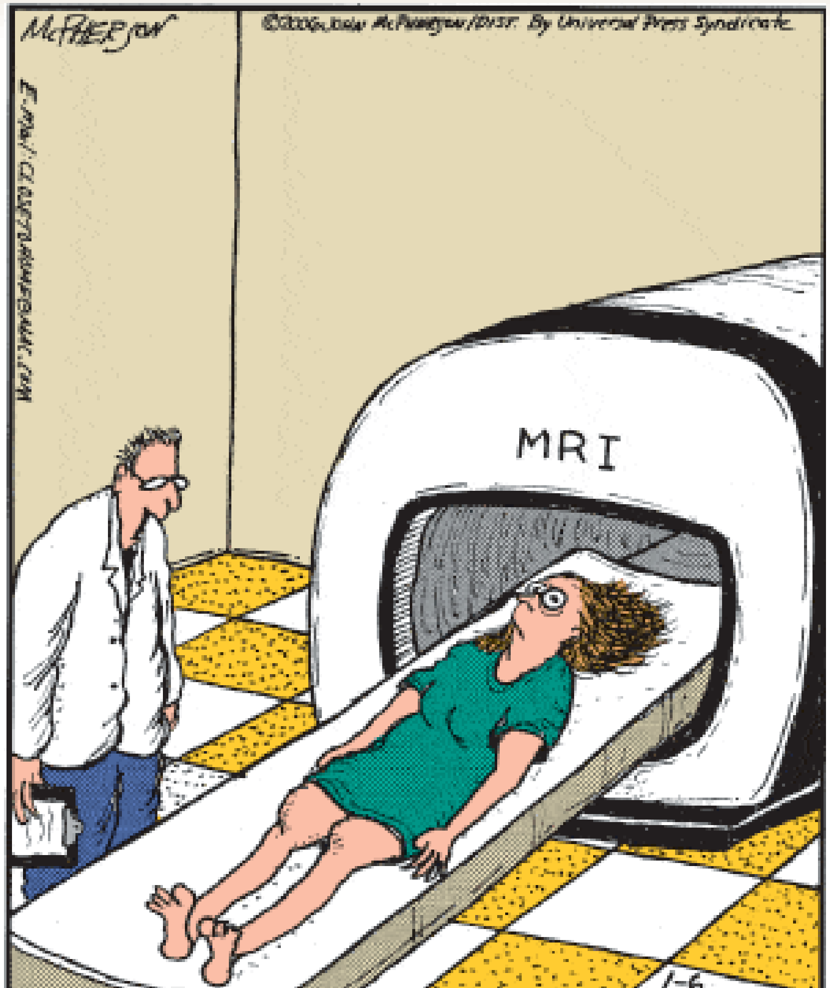
$$p_\theta(r) = \int\int f(x,y)\delta(x\cos\theta + y\sin\theta - r)dxdy$$

Fourier-domain interpretation



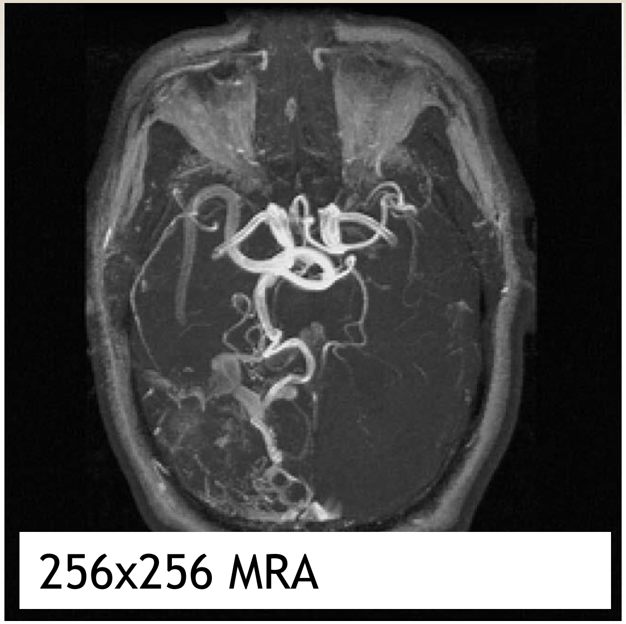
- Each projection gives us a “slice” of the 2D Fourier transform of the original image
- Similar ideas in MRI
- Traditional solution: Collect lots (and lots) of slices

Why CS?

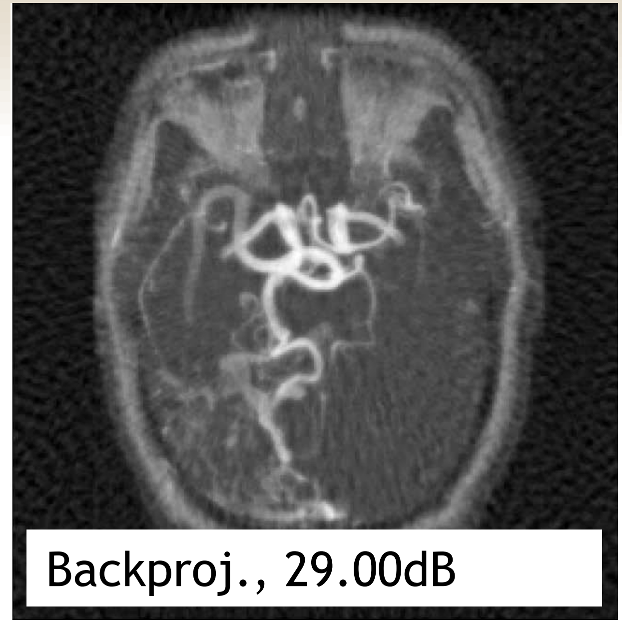


"OK, Mrs. Dunn. We'll slide you in there, scan your brain, and see if we can find out why you've been having these spells of claustrophobia."

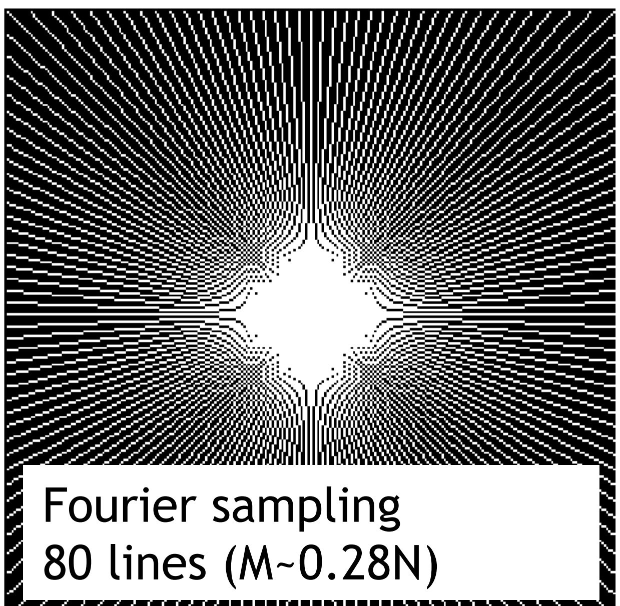
CS for MRI reconstruction



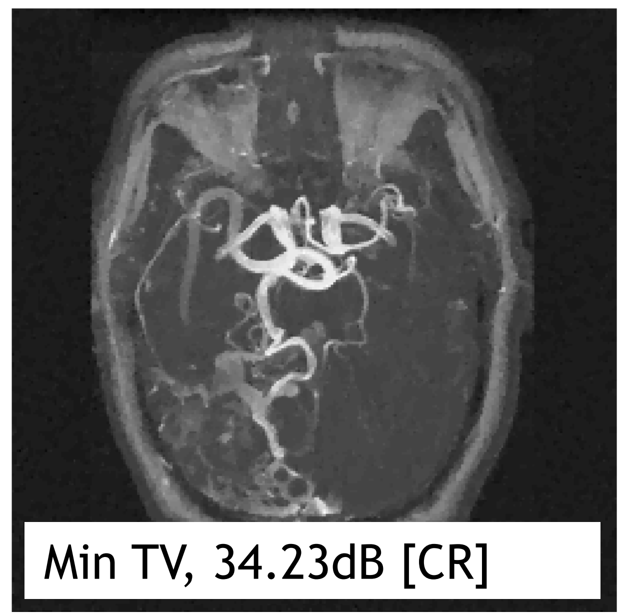
256x256 MRA



Backproj., 29.00dB

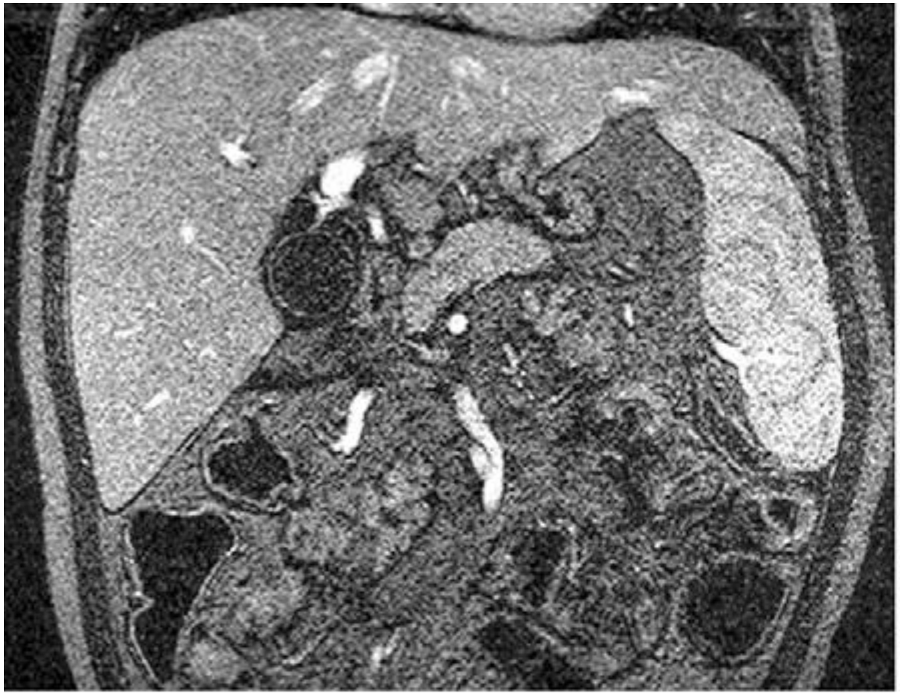


Fourier sampling
80 lines (M~0.28N)



Min TV, 34.23dB [CR]

Pediatric MRI



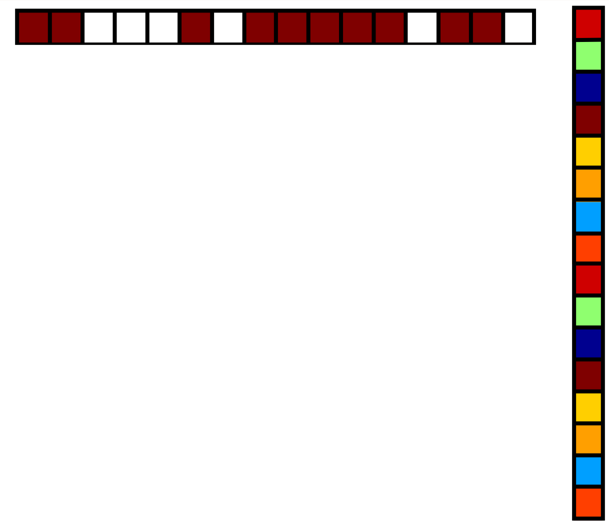
Traditional MRI

CS MRI

4-8 x faster!

[Vasanawala, Alley, Hargreaves, Barth, Pauly, Lustig - 2010]

“Single-Pixel Camera”

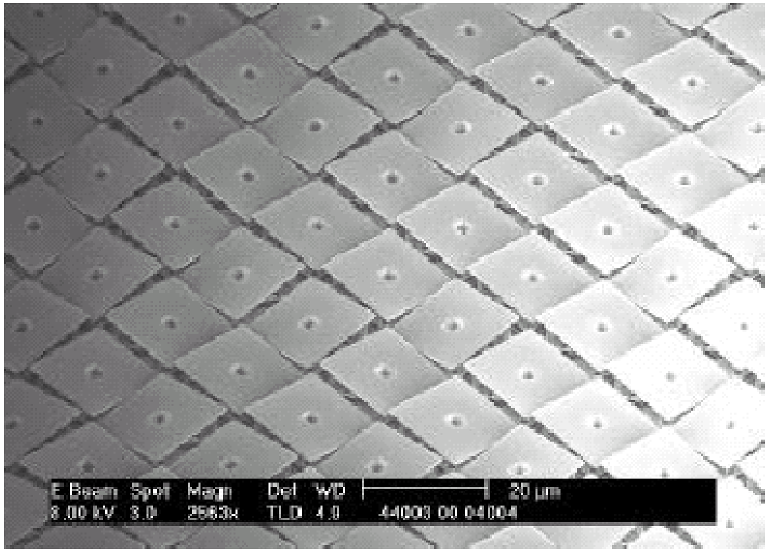


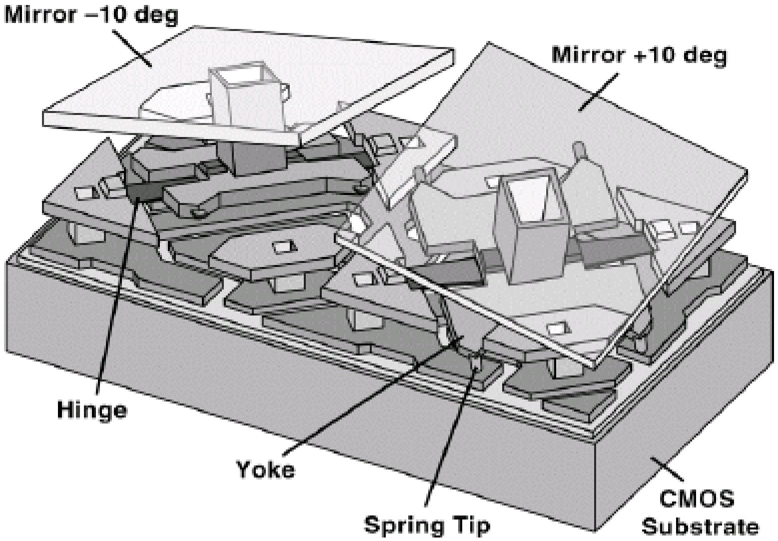
$$y[m] = \sum_{n \in I_m} x[n]$$

$$x[n] = \int\!\!\int_{\text{pixel } n} x(t_1, t_2)\, dt_1\, dt_2$$

[Duarte, Davenport, Takhar, Laska, Sun, Kelly, Baraniuk - 2008]

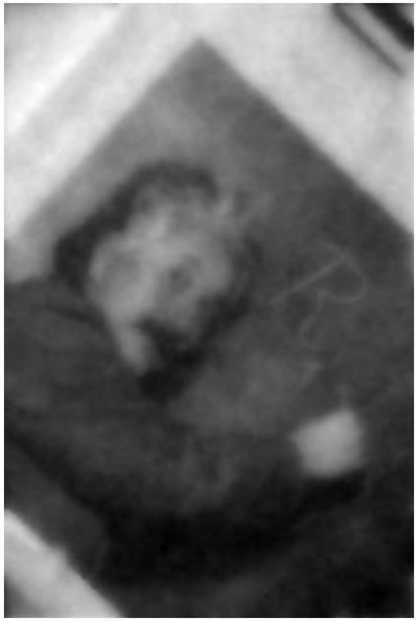
TI Digital Micromirror Device





Single-Pixel Camera

256×384 pixels



10%

20%

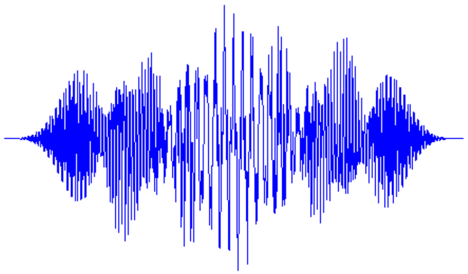
30%

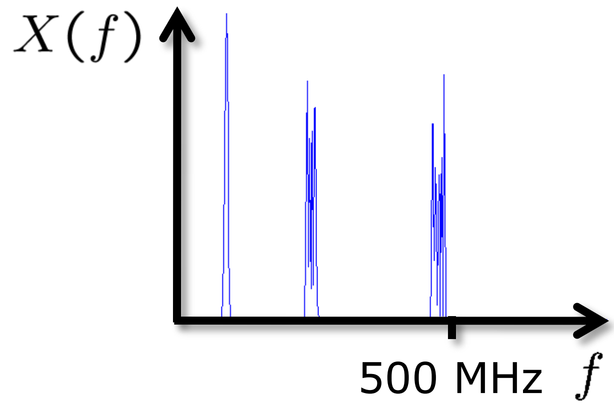
40%

InView

Compressive ADCs

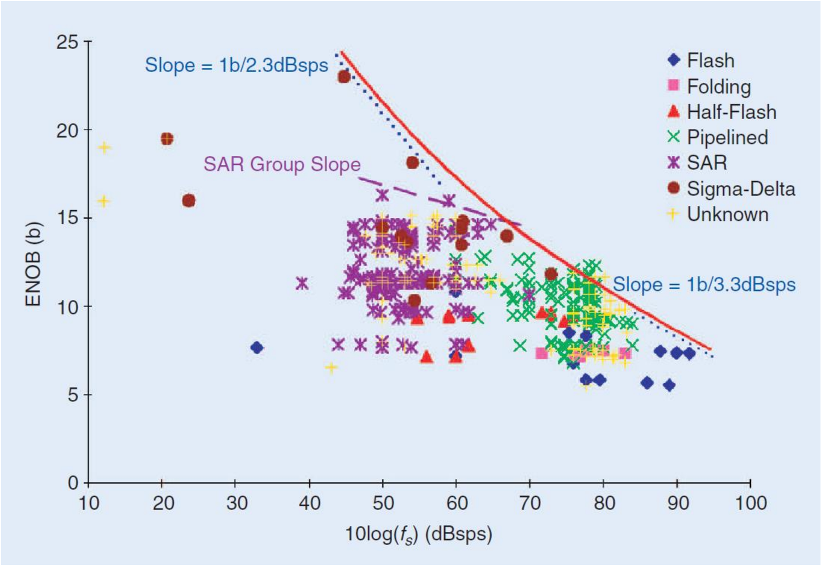
DARPA "Analog-to-information" program:
Build high-rate ADC for signals with sparse spectra

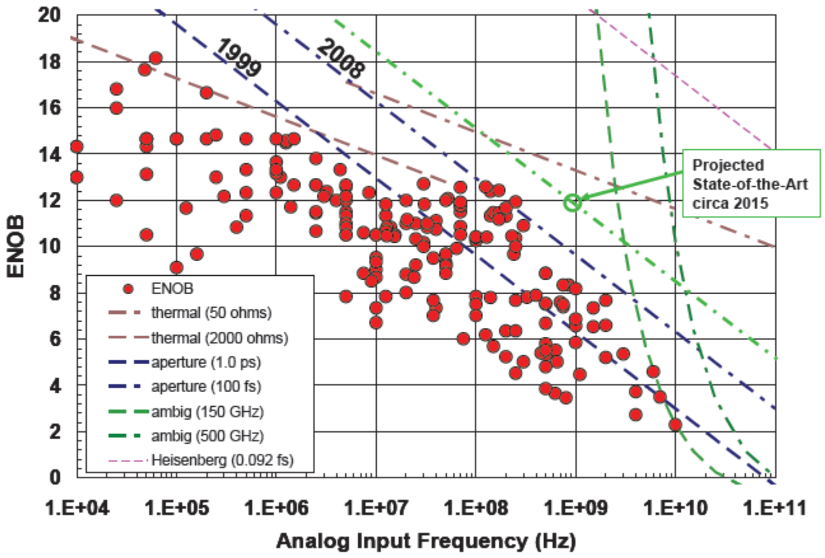




Compressive ADCs

DARPA "Analog-to-information" program:
Build high-rate ADC for signals with sparse spectra





[Le – 2005; Walden – 2008]

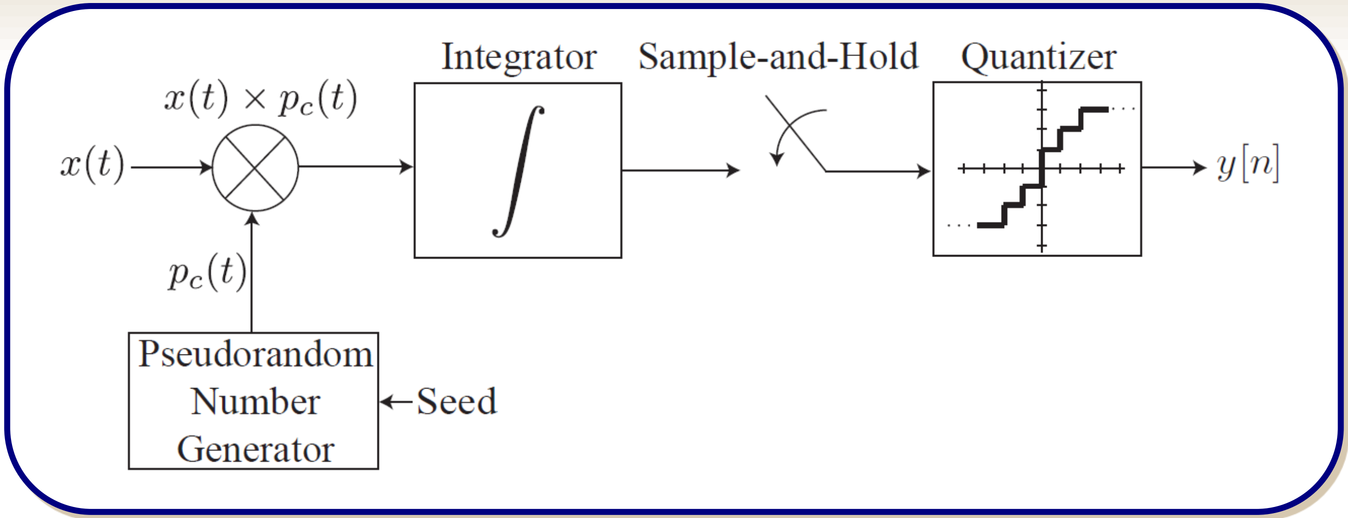
Compressive ADC approaches

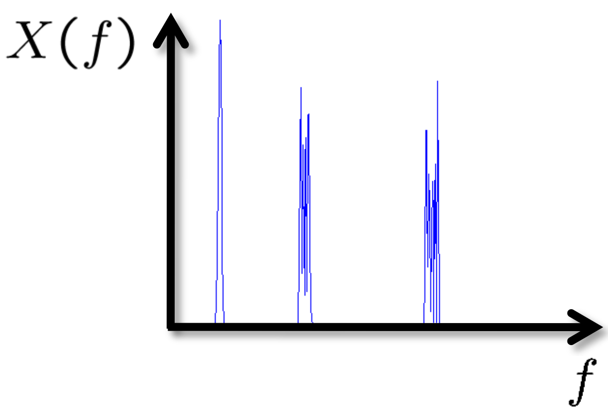
- Random sampling
  - long history of related ideas/techniques
  - random sampling for Fourier-sparse data equivalent to obtaining random Fourier coefficients for sparse data

- Random demodulation
  - CDMA-like spreading followed by low-rate uniform sampling
  - modulated wideband converter
  - compressive multiplexor, polyphase random demodulator

- Both approaches are specifically tailored for Fourier-sparse signals

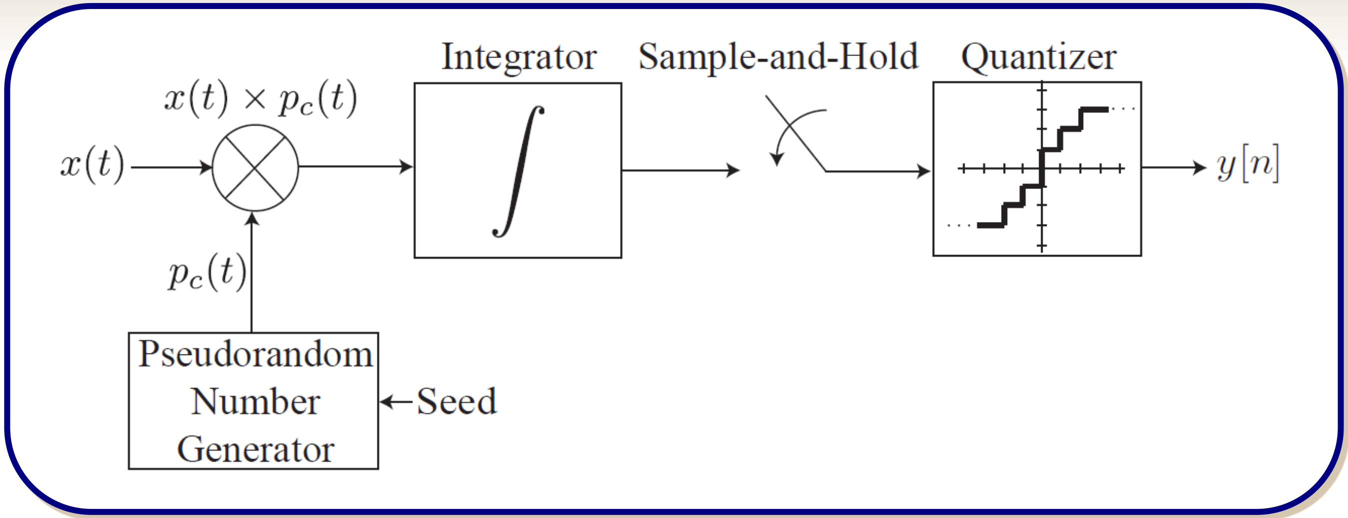
Random demodulator

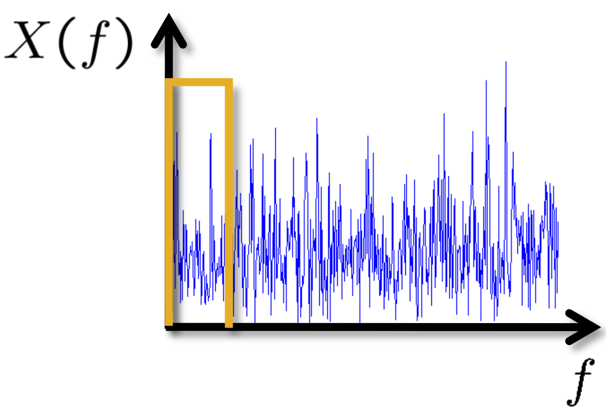


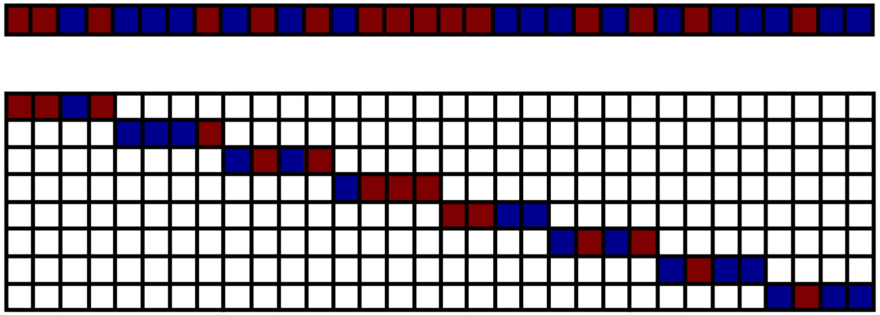


[Tropp, Laska, Duarte, Romberg, Baraniuk – 2010]

Random demodulator

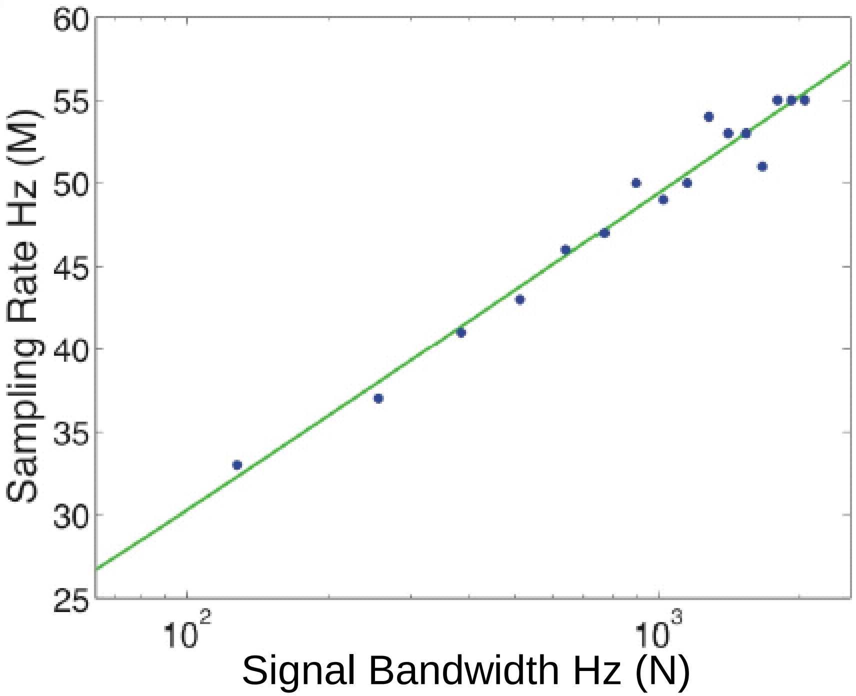


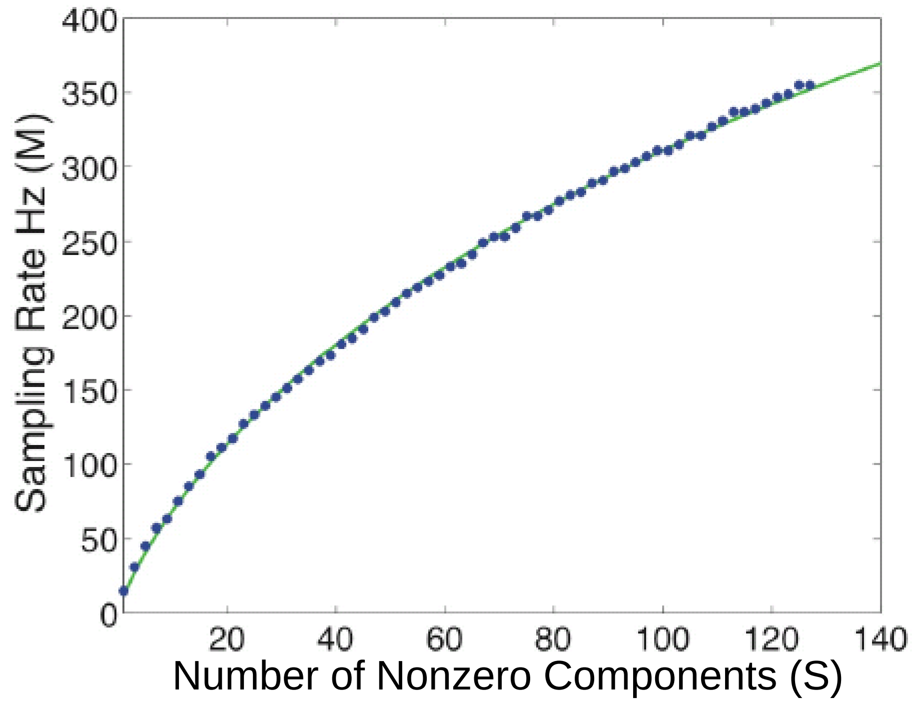




[Tropp, Laska, Duarte, Romberg, Baraniuk – 2010]

Empirical results





$$1.69S\log(N/S+1)+4.51$$

$$1.71S\log(N/S+1)+1$$

$$M \approx 1.7S\log(N/S+1)$$

[Tropp, Laska, Duarte, Romberg, Baraniuk – 2010]

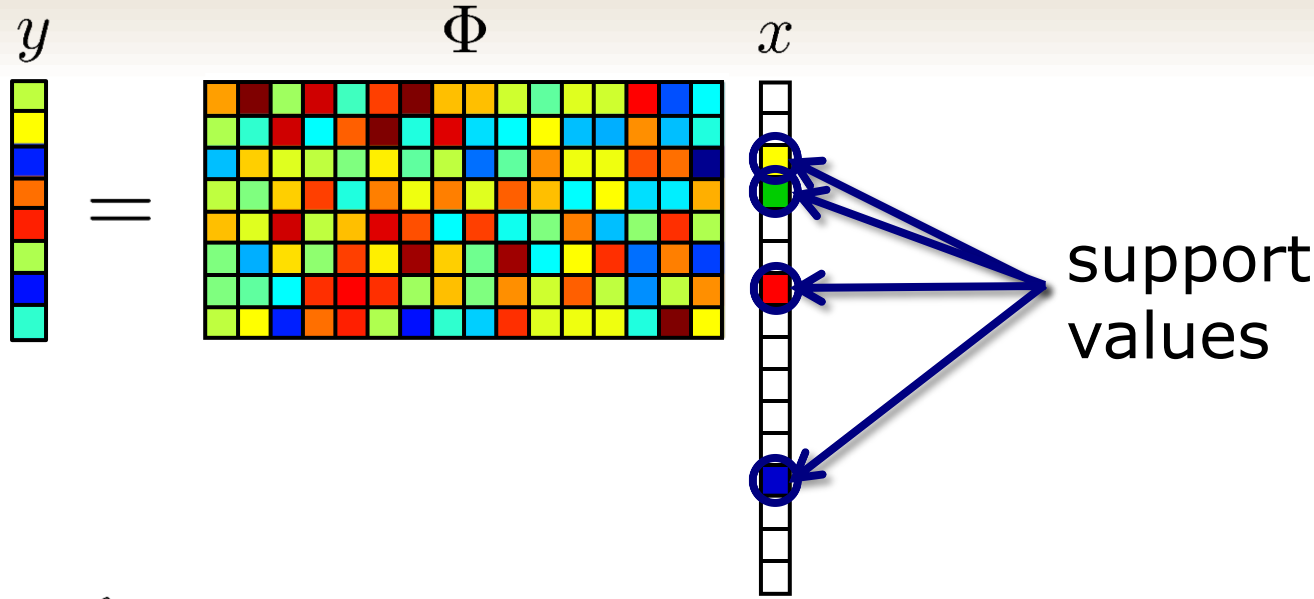
Compressive sensors wrap-up

- CS is built on a theory of *random measurements*
  - Gaussian, Bernoulli, random Fourier, fast JLT
  - stable, universal, democratic

- Randomness can often be built into real-world sensors
  - tomography
  - cameras
  - compressive ADCs
  - microscopy
  - astronomy
  - sensor networks
  - DNA microarrays and biosensing
  - radar
  - …

Sparse Signal Recovery

Sparse signal recovery



- Optimization / ℓ_1 -minimization
- Greedy algorithms
  - matching pursuit
  - orthogonal matching pursuit (OMP)
  - Stagewise OMP (StOMP), regularized OMP (ROMP)
  - CoSaMP, Subspace Pursuit, IHT, …

Sparse recovery: Noiseless case

given $y = \Phi x$
find x

- ℓ_0-minimization: $\widehat{x} = \arg\min_{x \in \mathbb{R}^N} \|x\|_0$ s.t. $y = \Phi x$ ← *nonconvex* *NP-Hard*

- ℓ_1-minimization: $\widehat{x} = \arg\min_{x \in \mathbb{R}^N} \|x\|_1$ s.t. $y = \Phi x$ ← *convex* *linear program*

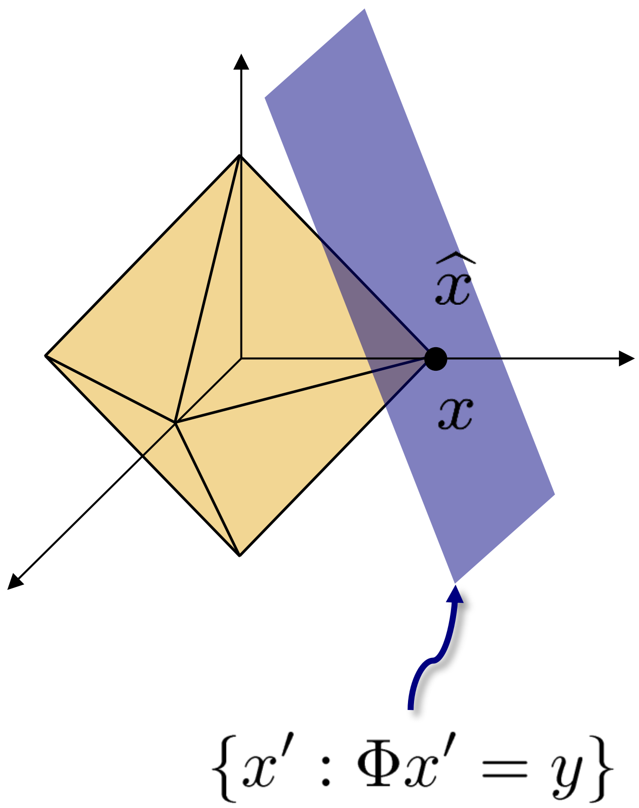
- If Φ satisfies the RIP, then ℓ_0 and ℓ_1 are equivalent!

[Donoho; Candès, Romberg, Tao - 2004]

Why ℓ_1-minimization works

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^N} \|x\|_1$$

$$\text{s.t.} \quad y = \Phi x$$



Sparse recovery: Noisy case

Suppose we observe $y = \Phi x + e$, where $\|e\|_2 \leq \epsilon$

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^N} \|x\|_1$$

$$\text{s.t.} \quad \|y - \Phi x\|_2 \leq \epsilon$$

$$\boxed{\|\widehat{x} - x\|_2 \leq C_0 \epsilon}$$

Similar approaches can handle Gaussian noise added to either the signal or the measurements

Sparse recovery: Non-sparse signals

In practice, x may not be exactly S-sparse

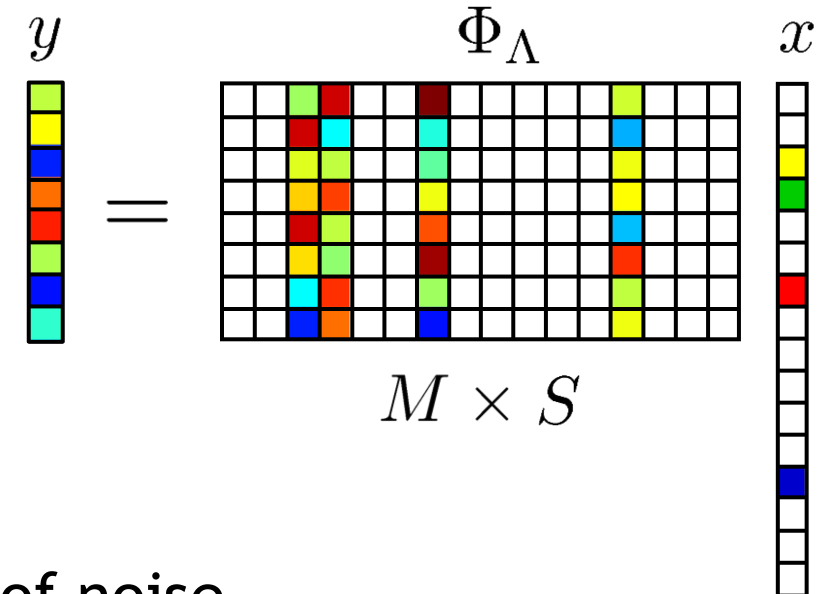
$$\widehat{x} = \arg\min_{x \in \mathbb{R}^N} \|x\|_1$$

$$\text{s.t.} \quad \|y - \Phi x\|_2 \leq \epsilon$$

$$\boxed{\|\widehat{x} - x\|_2 \leq C_0 \epsilon + C_1 \frac{\|x - x_S\|_1}{\sqrt{S}}}$$

Greedy algorithms: Key idea

If we can determine $\Lambda = \operatorname{supp}(x)$, then the problem becomes *over*-determined.



In the absence of noise,

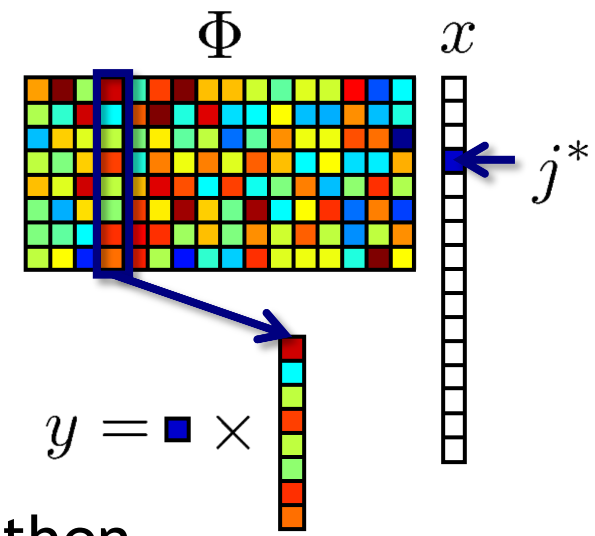
$$\begin{aligned}
\Phi_\Lambda^\dagger y &= (\Phi_\Lambda^T \Phi_\Lambda)^{-1} \Phi_\Lambda^T y \\
&= (\Phi_\Lambda^T \Phi_\Lambda)^{-1} \Phi_\Lambda^T \Phi_\Lambda x \\
&= x
\end{aligned}$$

Matching Pursuit

Select one index at a time using a simple *proxy* for x

$$p = \Phi^T y$$

$$\boxed{j^* = \arg\max_j |p_j|}$$



If Φ satisfies the RIP of order $\|u \pm v\|_0$, then

$$|\langle \Phi u, \Phi v \rangle - \langle u, v \rangle| \leq \delta \|u\|_2 \|v\|_2$$

Set $u = x$ and $v = e_j$

$$|p_j - x_j| \leq \delta \|x\|_2$$

Matching Pursuit

Obtain initial estimate of x

$$x^{(1)} = p_{j^*} e_{j^*}$$

Update proxy and iterate

$$p = \Phi^T(y - \Phi x^{(j-1)})$$

$$j^* = \arg\max_j |p_j|$$

$$x^{(j)} = x^{(j-1)} + p_{j^*} e_{j^*}$$

Iterative Hard Thresholding (IHT)

step size

$$x^{(j)} = \underbrace{H_S}_{\text{hard thresholding}} \left(x^{(j-1)} + \overset{\text{step size}}{\mu} \underbrace{\Phi^T \left(y - \Phi x^{(j-1)} \right)}_{\text{proxy vector}} \right)$$

hard thresholding

proxy vector

RIP guarantees convergence and accurate/stable recovery

[Blumensath and Davies – 2008]

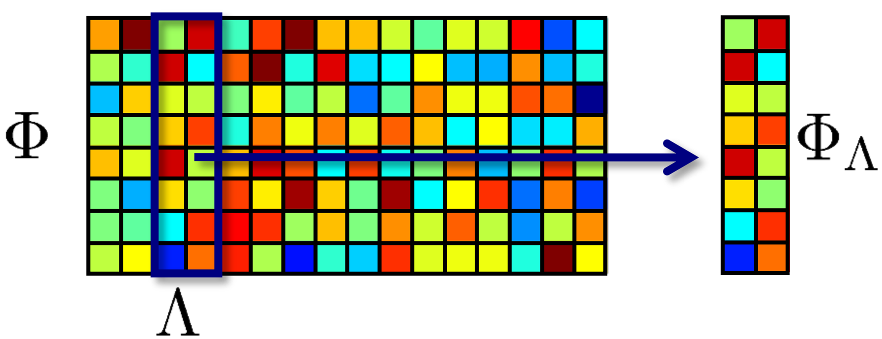
Orthogonal Matching Pursuit

Replace $x^{(j)} = x^{(j-1)} + p_{j*} e_{j*}$ with

$$x^{(j)} = \arg\min_x \|y - \Phi_\Lambda x\|_2$$

where Λ is the set of indices selected up to iteration j

$$\boxed{j^* = \arg\max_j |\langle Py, P\Phi_j \rangle|}$$



$$P = I - \underbrace{\Phi_\Lambda \Phi_\Lambda^\dagger}_{\text{Projection onto } \mathcal{R}(\Phi_\Lambda)}$$

$$P\Phi_\Lambda = 0 \quad \Longrightarrow \quad P\Phi x = P\Phi x_{\Lambda^c}$$

Orthogonal Matching Pursuit

Suppose x is S-sparse and $y = \Phi x$.
If Φ satisfies the RIP of order $S+1$ with constant $\delta < 1/3\sqrt{S}$, then the j^* identified at each iteration will be a nonzero entry of x.

➡ Exact recovery after S iterations

[Davenport and Wakin – 2010]

Extensions of OMP

- StOMP, ROMP
  - select many indices in each iteration
  - picking indices for which p_j is "comparable" leads to increased stability and robustness

- CoSaMP, Subspace Pursuit, …
  - allow indices to be discarded
  - strongest guarantees, comparable to ℓ_1-minimization

$$\|x - x^{(j+1)}\|_2 \leq \frac{1}{2}\|x - x^{(j)}\|_2 + C\|e\|_2$$

$$\|x - x^j\|_2 \leq 2^{-j}\|x\|_2 + 2C\|e\|_2$$

[Needell and Tropp – 2010]

Beyond Sparsity

Beyond sparsity

- Not all signal models fit neatly into the "sparse" setting

- The concept of "dimension" has many incarnations
  - "degrees of freedom"
  - constraints
  - parameterizations
  - signal families

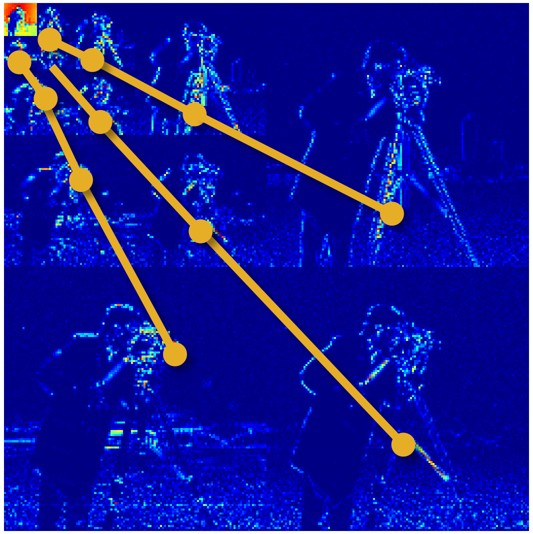
- How can we exploit these low-dimensional models?

- I will focus primarily on just a few of these
  - *structured* sparsity, finite-rate-of-innovation, manifolds, low-rank matrices

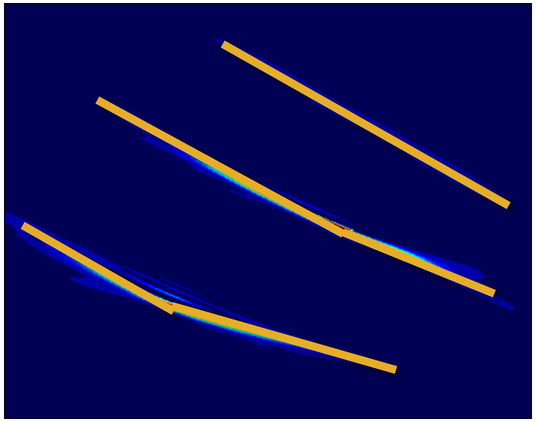
Structured sparsity

- Sparse signal model captures
  simplistic primary structure

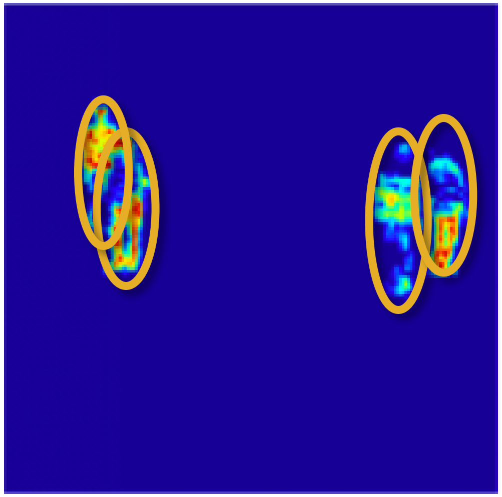
- Modern compression/processing algorithms capture
  richer secondary coefficient structure



wavelets:
natural images



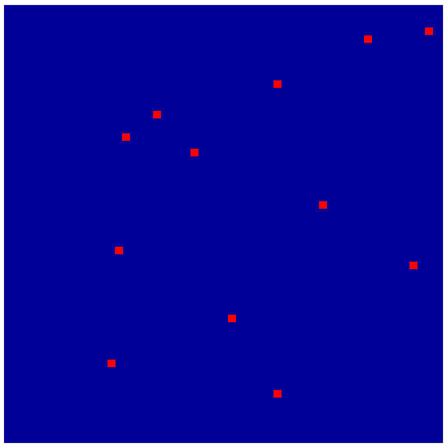
Gabor atoms:
chirps/tones

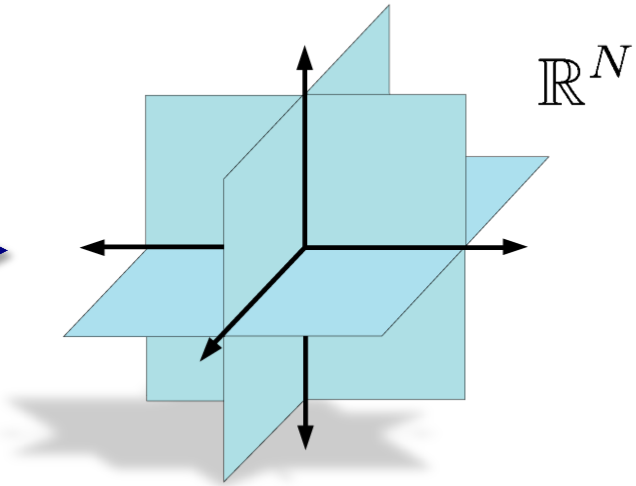


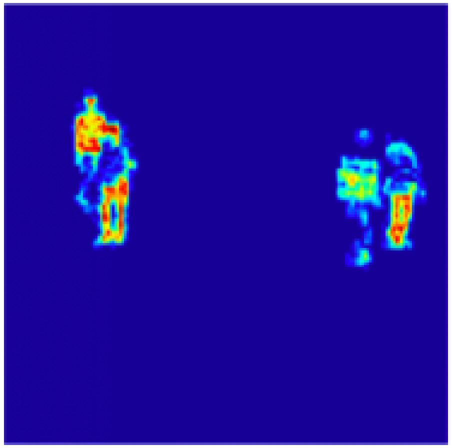
pixels:
background subtracted
images

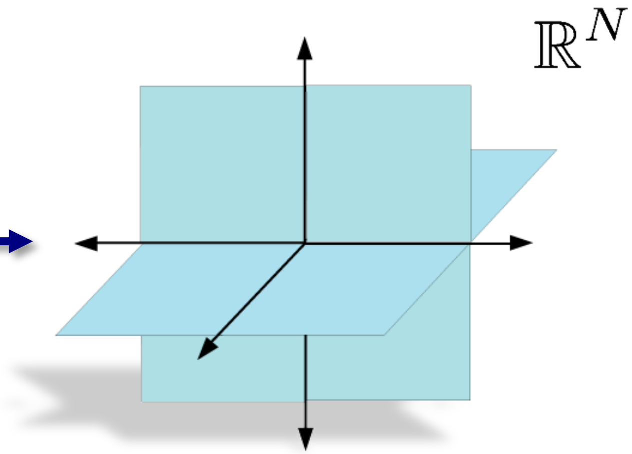
Sparse signals

Traditional sparse models allow all possible S-dimensional subspaces



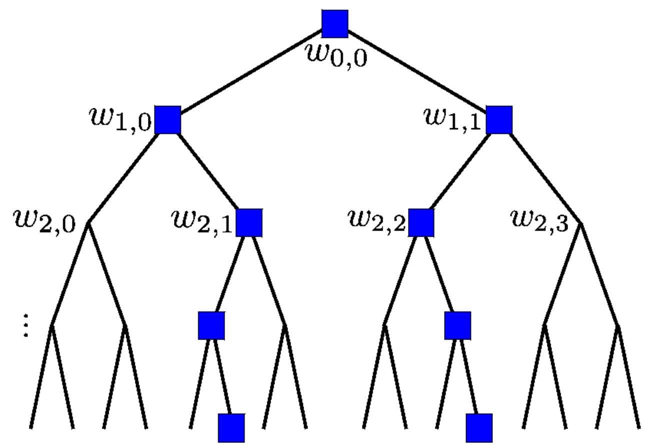


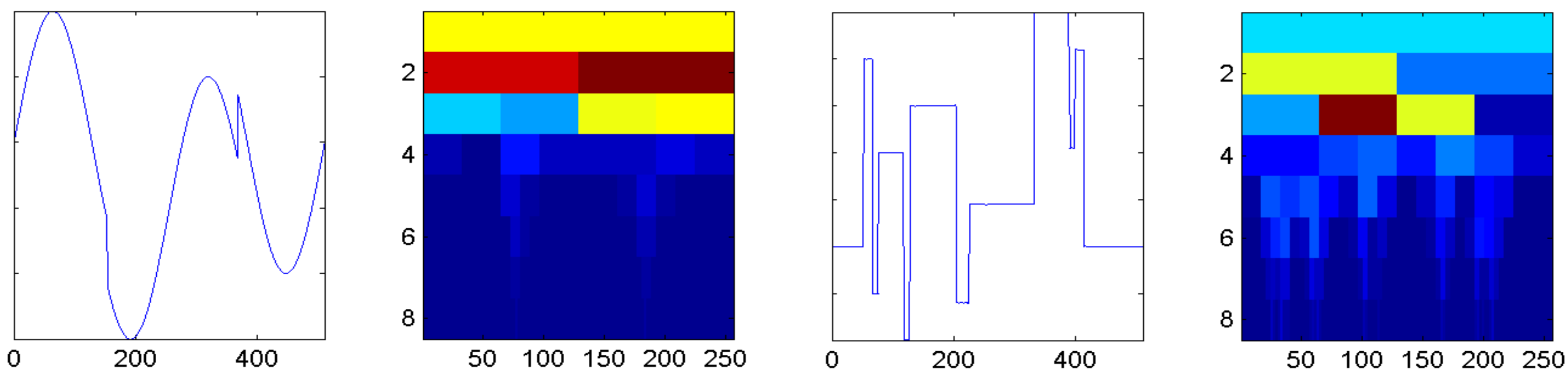




Wavelets and tree-sparse signals

Model: S nonzero coefficients lie on a connected tree

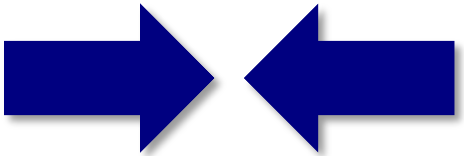




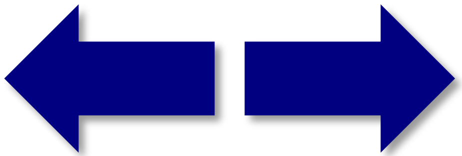
[Baraniuk, Cevher, Duarte, Hegde – 2010]

Other useful models

- Clustered coefficients
  - tree sparse
  - block sparse
  - Ising models



- Dispersed coefficients
  - spike trains
  - pulse trains

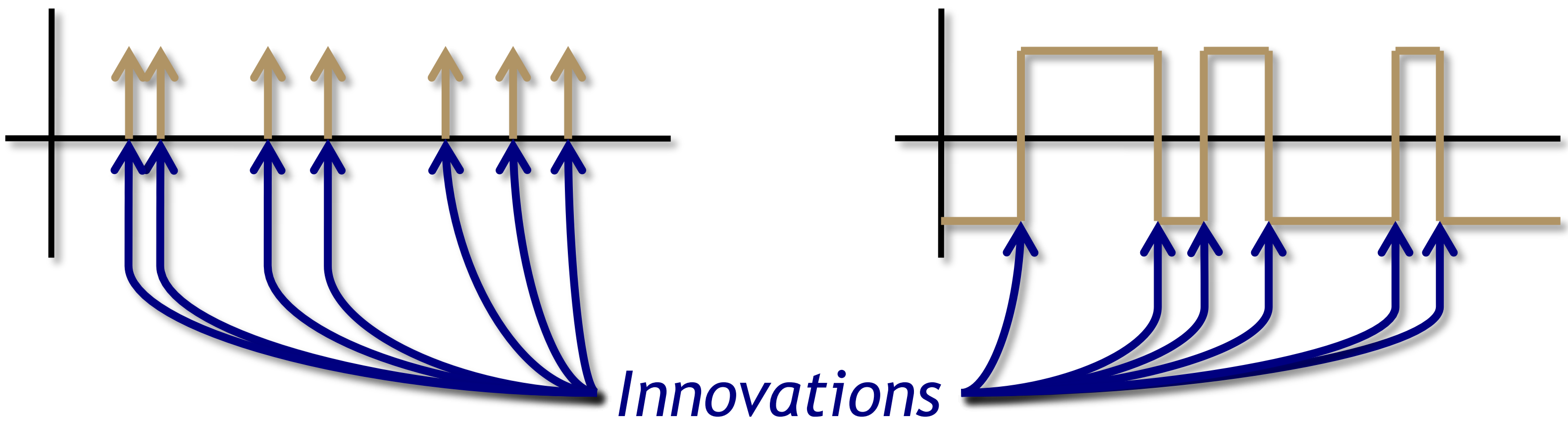


[Baraniuk, Cevher, Duarte, Hegde – 2010]

Finite rate of innovation

Continuous-time notion of sparsity: “rate of innovation”

Examples:



Rate of innovation:

Expected number of innovations per second

[Vetterli, Marziliano, Blu – 2002; Dragotti, Vetterli, Blu - 2007]

Sampling signals with FROI

We would like to obtain samples of the form

$$y[m] = \phi(t) * x(t)|_{t=mT_s} = \langle \phi(mT_s - t), x(t) \rangle$$

where we sample at the *rate of innovation.*

Requires *careful construction of sampling kernel* $\phi(t)$.

Drawbacks:

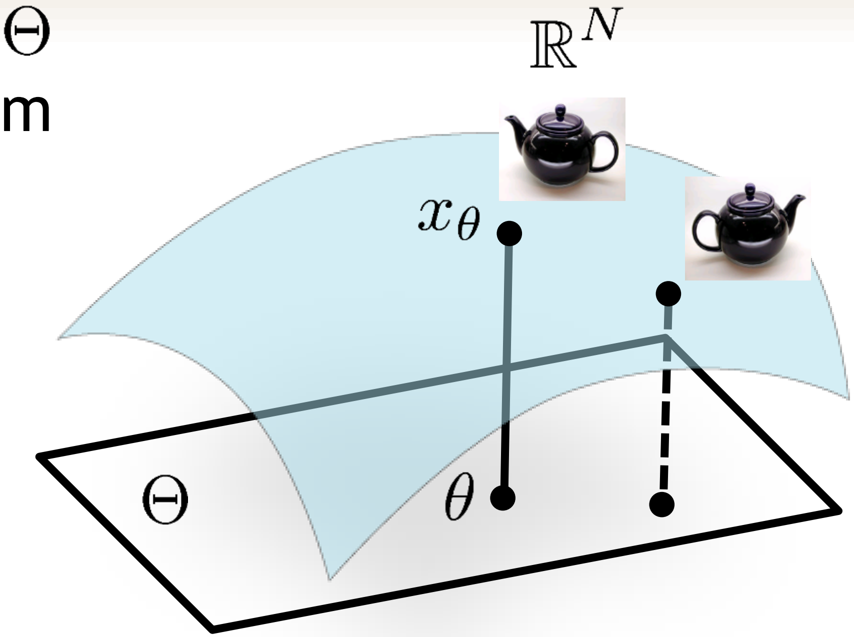
- need to repeat process for each signal model
- stability

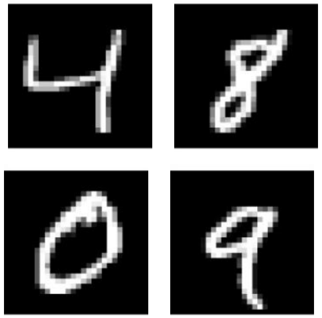
[Vetterli, Marziliano, Blu – 2002; Dragotti, Vetterli, Blu - 2007]

Manifolds

- S-dimensional *parameter* $\theta \in \Theta$ captures the degrees of freedom of signal

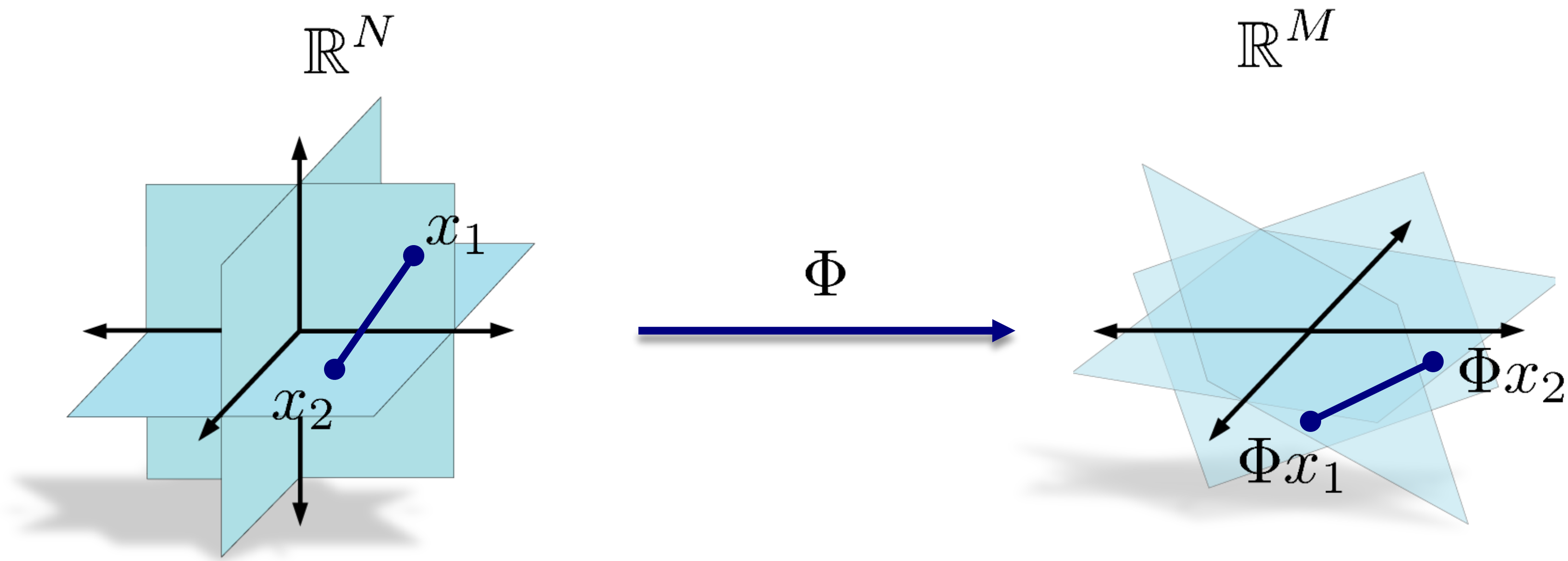
- Signal class forms an S-dimensional *manifold*
  - rotations, translations
  - robot configuration spaces
  - signal with unknown translation
  - sinusoid of unknown frequency
  - faces
  - handwritten digits
  - speech
  - …





Random projections

- For sparse signals, random projections preserve geometry



- What about manifolds?

Stable manifold embedding

Theorem

Let $\mathcal{M} \subseteq \mathbb{R}^N$ be a compact S-dimensional manifold with

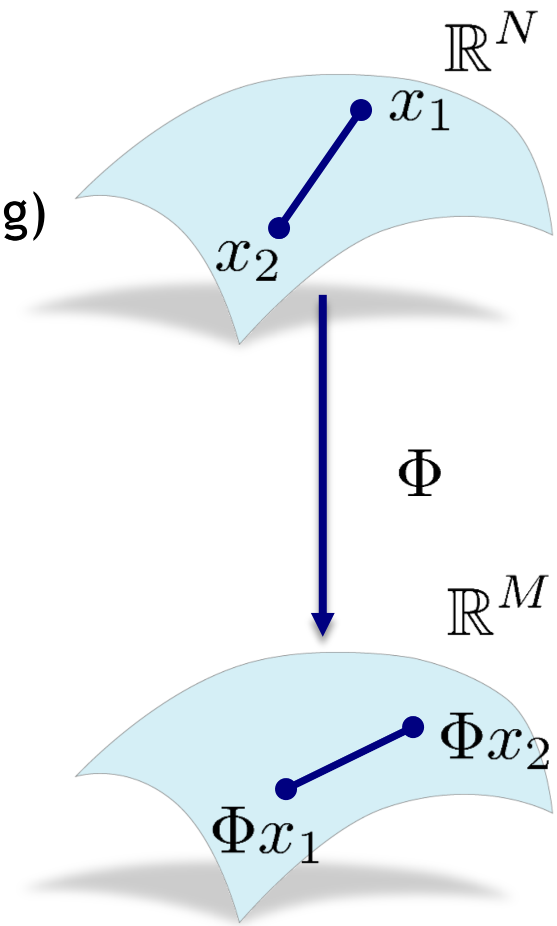
- condition number $1/\tau$ (curvature, self-avoiding)
- volume V

Let Φ be a random $M \times N$ projection with

$$M = O(S \log(NV/\tau))$$

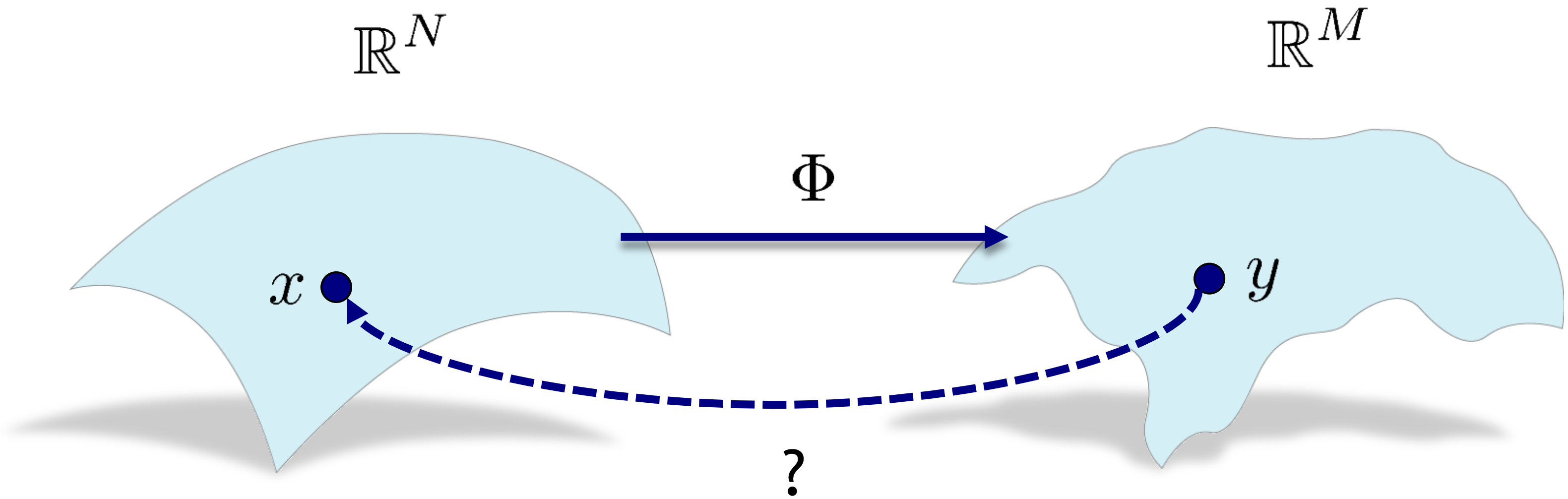
Then with high probability, and any $x_1, x_2 \in \mathcal{M}$

$$1 - \delta \leq \frac{\|\Phi x_1 - \Phi x_2\|_2^2}{\|x_1 - x_2\|_2^2} \leq 1 + \delta$$



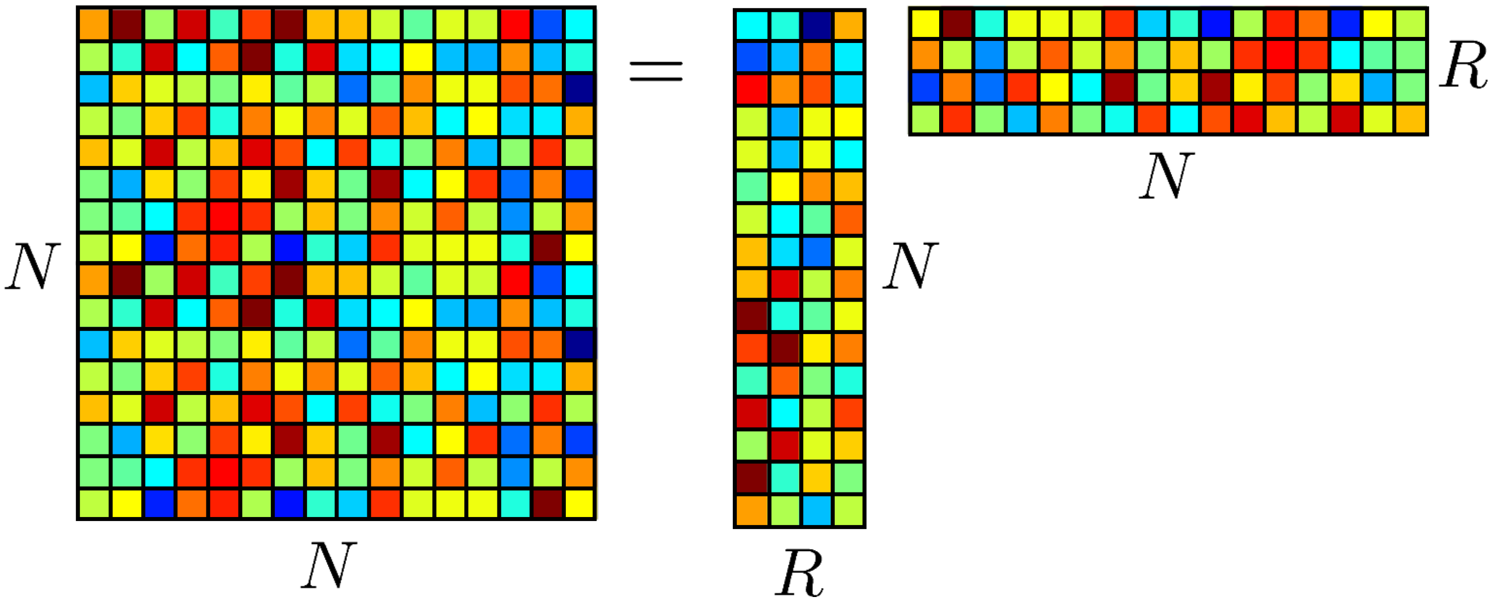
[Baraniuk and Wakin – 2009]

Compressive sensing with manifolds



- Same sensing protocols/devices
- Different reconstruction models
- Measurement rate depends on *manifold dimension*
- Stable embedding guarantees robust recovery

Low-rank matrices

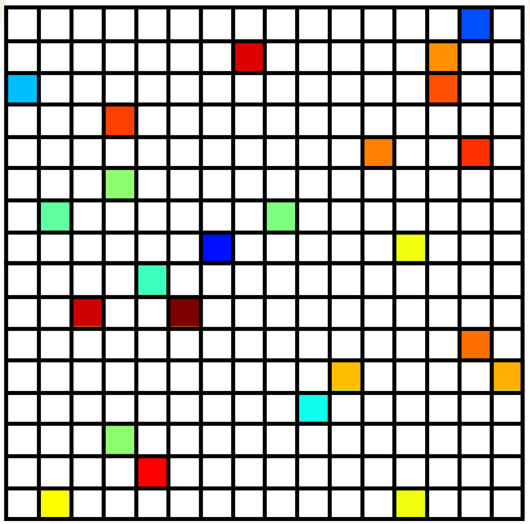


Singular value decomposition:

$$X = U\Sigma V^* \quad \Rightarrow \quad \approx NR \ll N^2$$

degrees of freedom

Matrix completion



- Collaborative filtering ("Netflix problem")
- How many samples will we need?

$$M \geq CNR$$

- Coupon collector problem

$$M \geq N \log N$$

Application: Collaborative filtering

The "Netflix Problem"

$$\boxed{X_{i,j} = \text{how much user } i \text{ likes movie } j}$$

Rank 1 model: $u_i =$ how much user i likes romantic movies

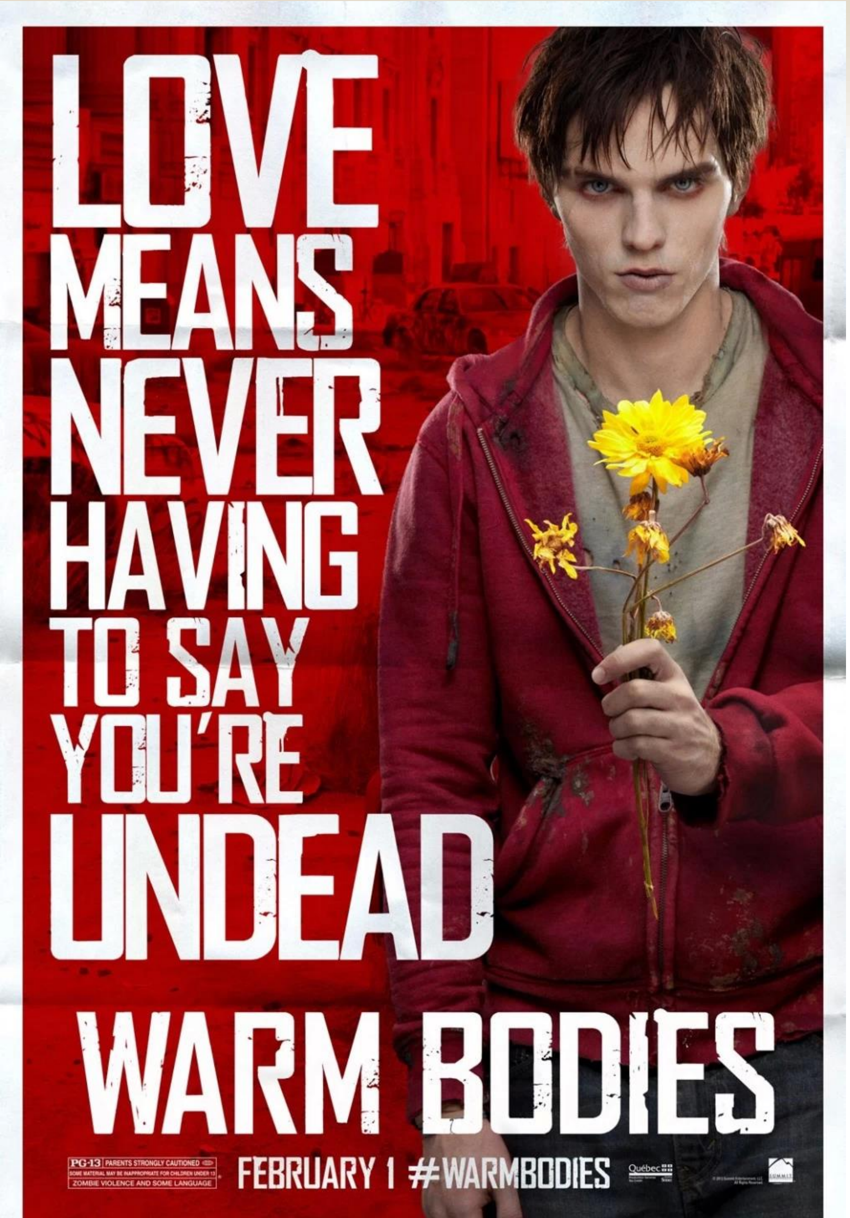
$v_j =$ amount of romance in movie j

$$X_{i,j} = u_i v_j$$

Rank 2 model: $w_i =$ how much user i likes zombie movies

$x_j =$ amount of zombies in movie j

$$X_{i,j} = u_i v_j + w_i x_j$$


LOVE
MEANS
NEVER
HAVING
TO SAY
YOU'RE
UNDEAD
WARM BODIES
PG-13 PARENTS STRONGLY CAUTIONED
SOME MATERIAL MAY BE INAPPROPRIATE FOR CHILDREN UNDER 13
ZOMBIE VIOLENCE AND SOME LANGUAGE
FEBRUARY 1 #WARMBODIES
Québec

Low-rank matrix recovery

Given:

- an $N \times N$ matrix X of rank R
- linear measurements $y = \mathcal{A}(X)$

How can we recover X ?

$$\boxed{\widehat{X} = \underset{X : \mathcal{A}(X) = y}{\arg\inf} \ \operatorname{rank}(X)}$$

Can we replace this with something computationally feasible?

Nuclear norm minimization

Convex relaxation!

Replace $\mathrm{rank}(X)$ with $\|X\|_* = \sum_{j=1}^{N} |\sigma_j|$

The "nuclear norm" is just the ℓ_1-norm of the vector of singular values

$$\widehat{X} = \underset{X : \mathcal{A}(X) = y}{\arg\inf} \ \mathrm{rank}(X)$$

[Candès, Fazel, Keshavan, Li, Ma, Montanari, Oh, Parrilo, Plan, Recht, Tao, Wright, …]

Nuclear norm minimization

Convex relaxation!

Replace $\mathrm{rank}(X)$ with $\|X\|_* = \sum_{j=1}^{N} |\sigma_j|$

The “nuclear norm” is just the ℓ_1-norm of the vector of singular values

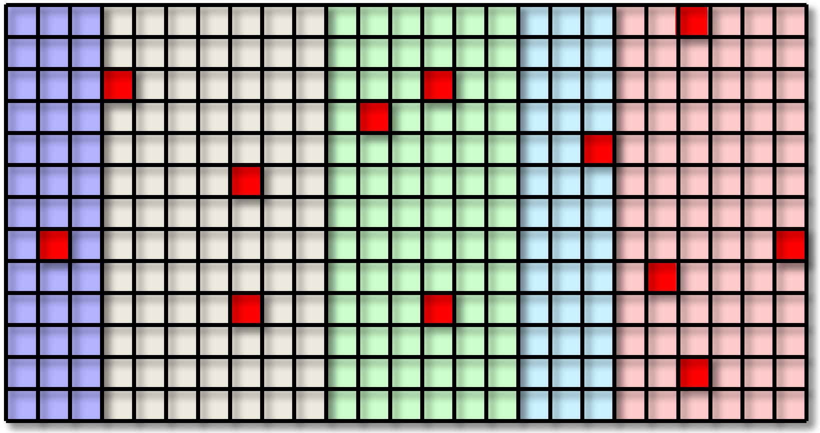
$$\boxed{\widehat{X} = \operatorname*{arg\,inf}_{X:\mathcal{A}(X)=y} \|X\|_*}$$

$$M = O(NR \log N)$$

[Candès, Fazel, Keshavan, Li, Ma, Montanari, Oh, Parrilo, Plan, Recht, Tao, Wright, …]

Robust PCA

In the presence of outliers, our data matrix $\mathbf{X}$ is no longer low-rank because some of the entries have been corrupted



$$\mathbf{X} = \underbrace{\mathbf{L}}_{\text{low-rank}} + \underbrace{\mathbf{S}}_{\text{corruptions}}$$

How to perform separation?

$$\min_{\mathbf{L},\mathbf{S}} \ \text{rank}(\mathbf{L}) + \lambda\|\mathbf{S}\|_0$$

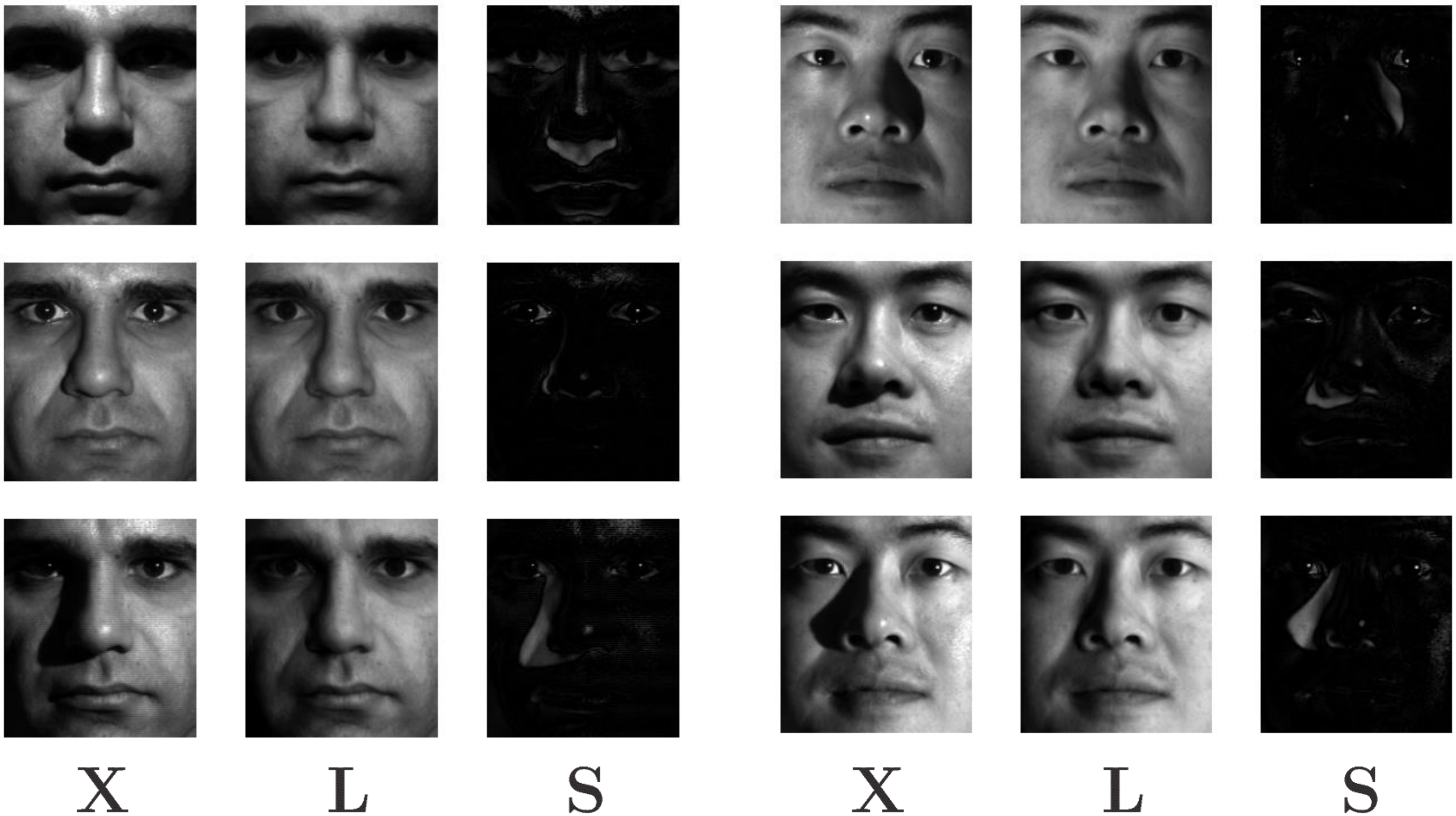
$$\text{s.t. } \mathbf{L} + \mathbf{S} = \mathbf{X}$$

$$\Downarrow$$

$$\min_{\mathbf{L},\mathbf{S}} \ \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1$$

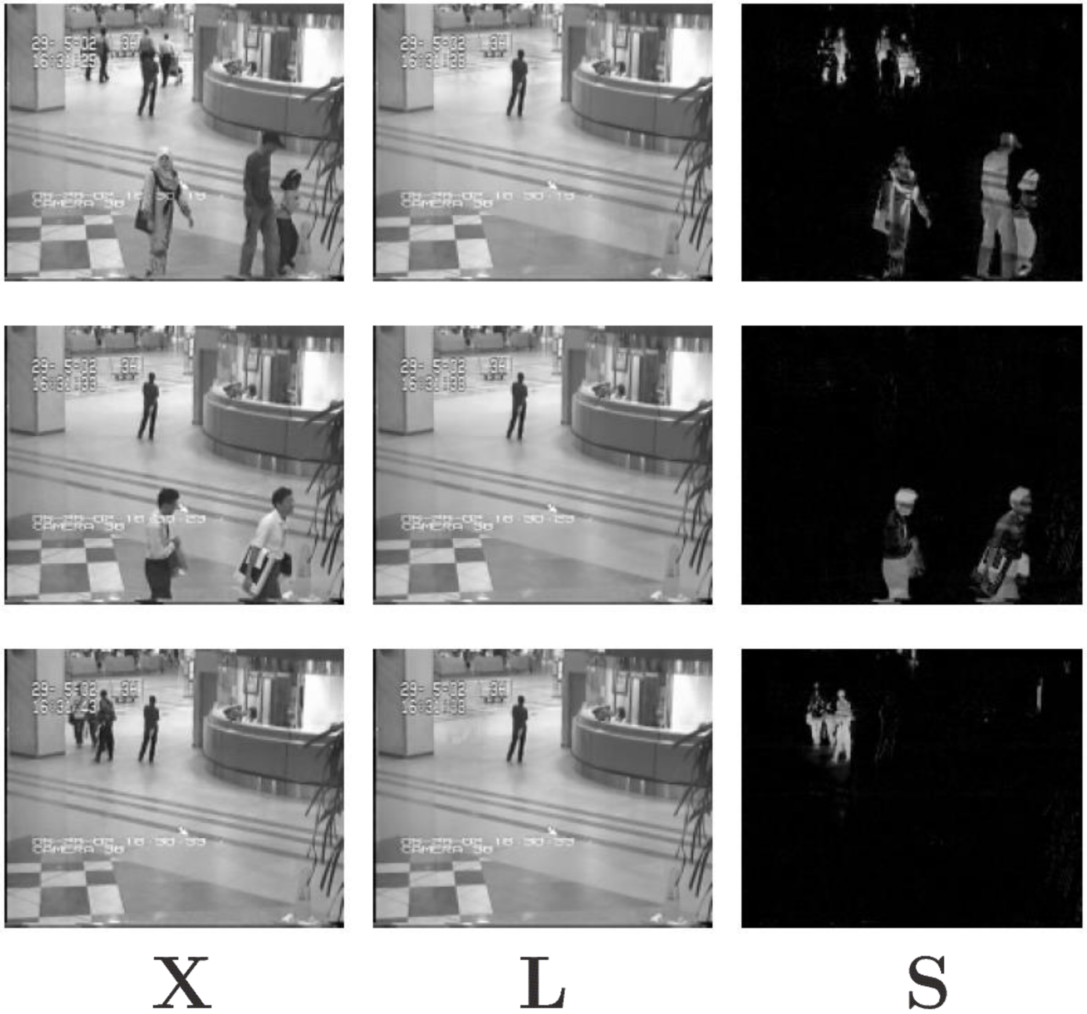
$$\text{s.t. } \mathbf{L} + \mathbf{S} = \mathbf{X}$$

Application: Removing face illumination



[Candès et al., 2009]

Application: Background subtraction



[Candès et al., 2009]

Conclusions

Conclusions

- The theory of compressive sensing allows for new sensor designs, but requires new techniques for signal recovery

- “Conciseness” has many incarnations
  - structured sparsity
  - finite rate of innovation, manifold, parametric models
  - low-rank matrices

- We can still use compressive sensing even when signal recovery is not our goal

- The theory/techniques from compressive sensing can be tremendously useful in a variety of other contexts