Sparse representations: from Compression to Source Separation and Compressed Sensing

Rémi Gribonval
EPI METISS
INRIA Rennes - Bretagne Atlantique

remi.gribonval@inria.fr
http://www.irisa.fr/metiss/members/remi

Structure of the course

- Part I: Overview
- Part II: Algorithms, complexity & convergence
  - Lp minimization
  - Greedy Algorithms
- Part III: Recovery, stability, robustness
  - Null Space Properties and Lp minimization
  - Exact Recovery Condition and greedy algorithms
  - Restricted Isometry Constants, stability and robustness
- Part IV: Compressed Sensing and Random

Overview

- Sparsity and compression of large-scale data
- Sparsity for source separation and inverse problems
- Sparse decomposition algorithms
  - L1 minimisation
  - Matching Pursuits
- Sparsity and compressed sensing

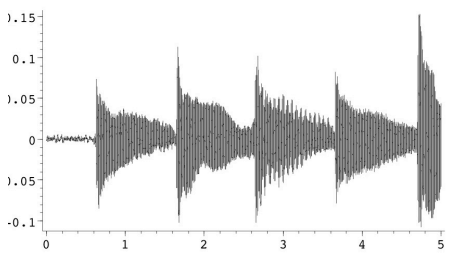
Sparsity & data compression

Large-scale data

- **Fact** : digital data = large volumes
  - 1 second stereo audio, CD quality = 1,4 Mbits
  - 1 uncompressed 10 Mpixels picture = 240 Mbits
- **Need** : «concise» data representations
  - storage & transmission (volume / bandwidth) ...
  - manipulation & processing (algorithmic complexity)

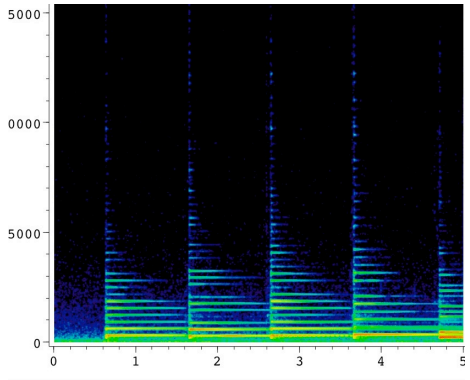
Sparse representations

- Audio : time-frequency representations (MP3)



ANALYSIS

SYNTHESIS



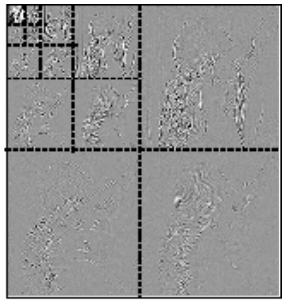
Black
= zero

- Images : wavelet transform (JPEG2000)

ORIGINAL
128, 129, 125, 64, 65,

ANALYSIS

SYNTHESIS



TRANSFORM COEFFICIENTS
4123, -12.4, -96.7, 4.5,

Gray
= zero

Mathematical expression

- Signal / image = high dimensional vector

$$y \in \mathbb{R}^N$$

- **Model** = linear combination of basis vectors (ex: *time-frequency atoms, wavelets*)

$$y \approx \sum_k x_k \varphi_k = \mathbf{\Phi} x$$

atoms
dictionary

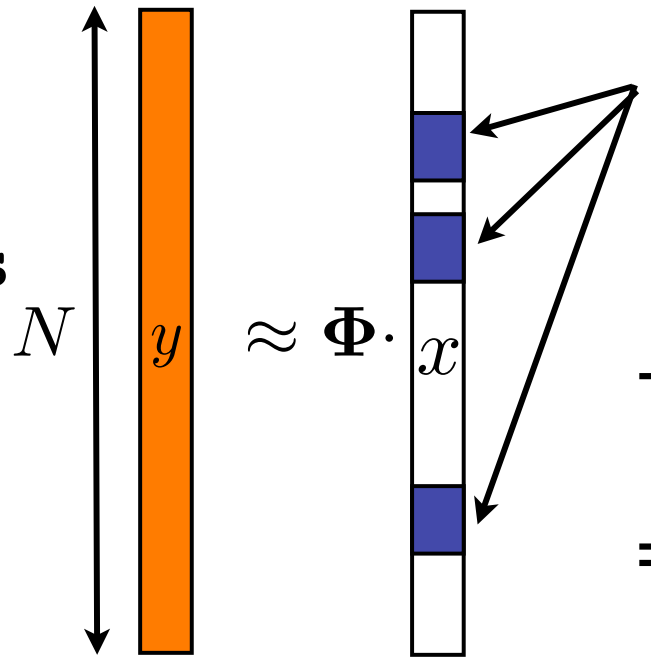
- **Sparsity** = small L0 (quasi)-norm

$$\|x\|_0 = \sum_k |x_k|^0 = \mathrm{card}\{k, x_k \neq 0\}$$

Sparsity & compression

- Full vector

N entries
= N `floats`



- Sparse vector

$k \ll N$ nonzero entries
= k `floats`

+ k positions among N

$$= \log_2 \binom{N}{k} \approx k \log_2 \frac{N}{k} \texttt{ bits}$$

Key practical issues: choose dictionary

Sparsity & inverse problems

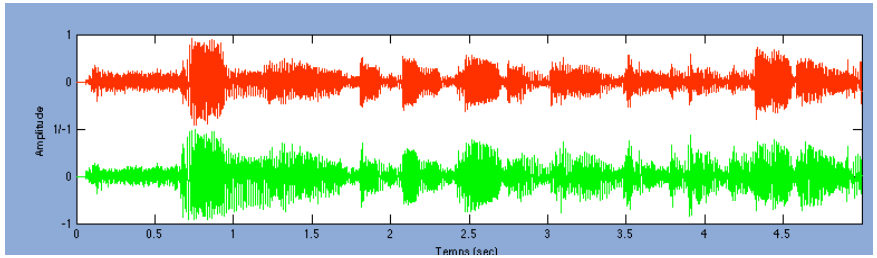
Example: image inpainting

Courtesy of: G. Peyré, Ceremade, Université Paris 9 Dauphine

Inpainting →

Example : audio source separation

- « Softly as in a morning sunrise »

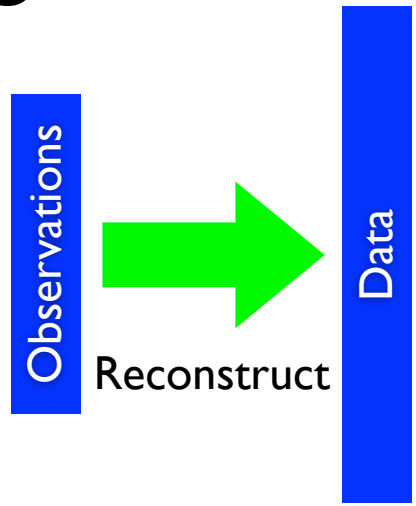


Inverse problems

- **Inverse problem** : exploit indirect or incomplete obervation to recontruct some data

$$z = \mathbf{M}\mathbf{y}$$

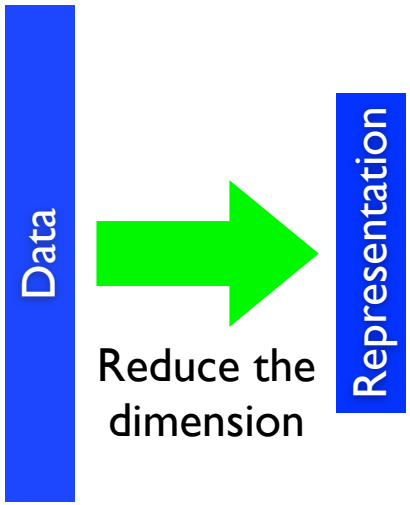
fewer equations than unknowns



- **Sparsity** : represent / approximate high-dimensional & complex data using few parameters

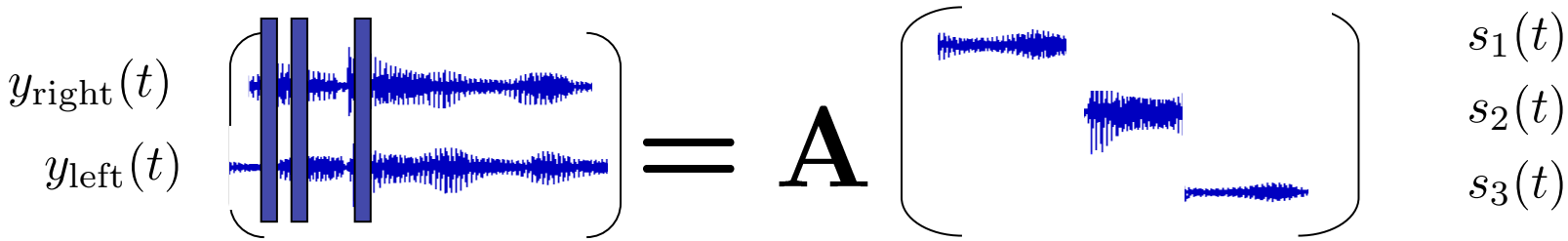
$$\mathbf{y} \approx \mathbf{\Phi} x$$

few nonzero components

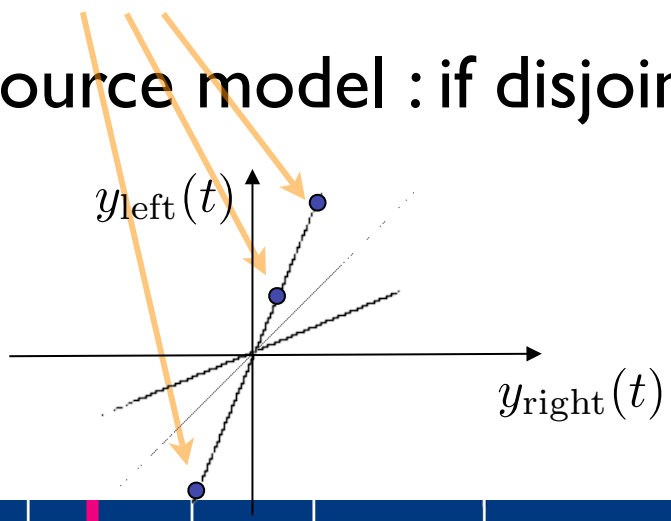


Blind Source Separation

- Mixing model : linear instantaneous mixture



- Source model : if disjoint time-supports …



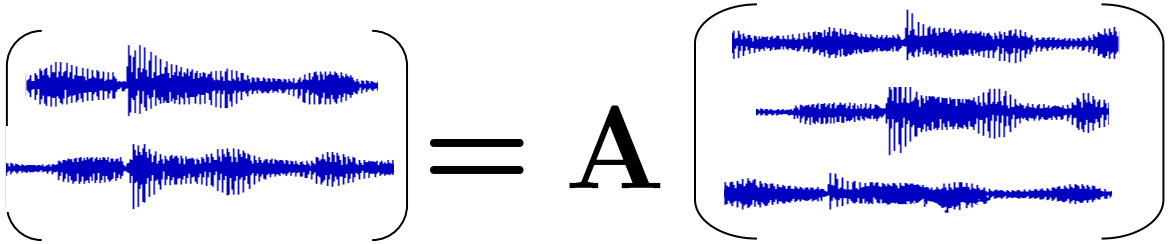
... then clustering to :
1- identify (columns of) the mixing matrix
2- recover sources

Blind Source Separation

- Mixing model : linear instantaneous mixture

$y_{\text{right}}(t)$

$y_{\text{left}}(t)$



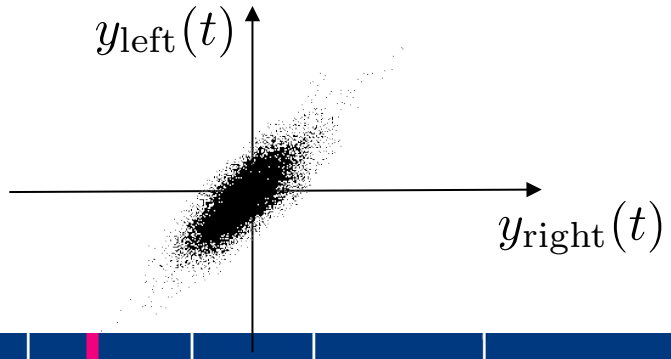
$= \mathbf{A}$

$s_1(t)$

$s_2(t)$

$s_3(t)$

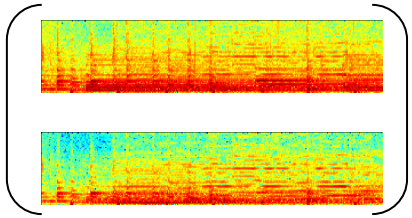
- In practice ...



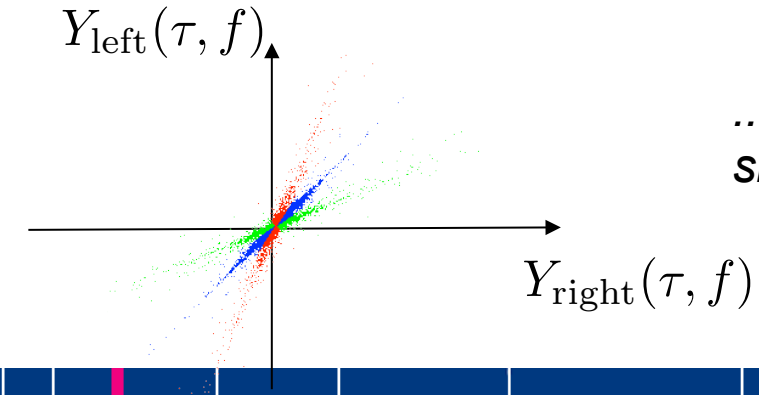
Time-Frequency Masking

- Mixing model in the time-frequency domain

$$\begin{pmatrix} Y_{\text{right}}(\tau, f) \\ Y_{\text{left}}(\tau, f) \end{pmatrix} = \mathbf{A}\,\mathbf{S}(\tau, f)$$



- And "miraculously" ...



... time-frequency representations of audio signals are (often) ***almost disjoint****.*

Mathematical foundations

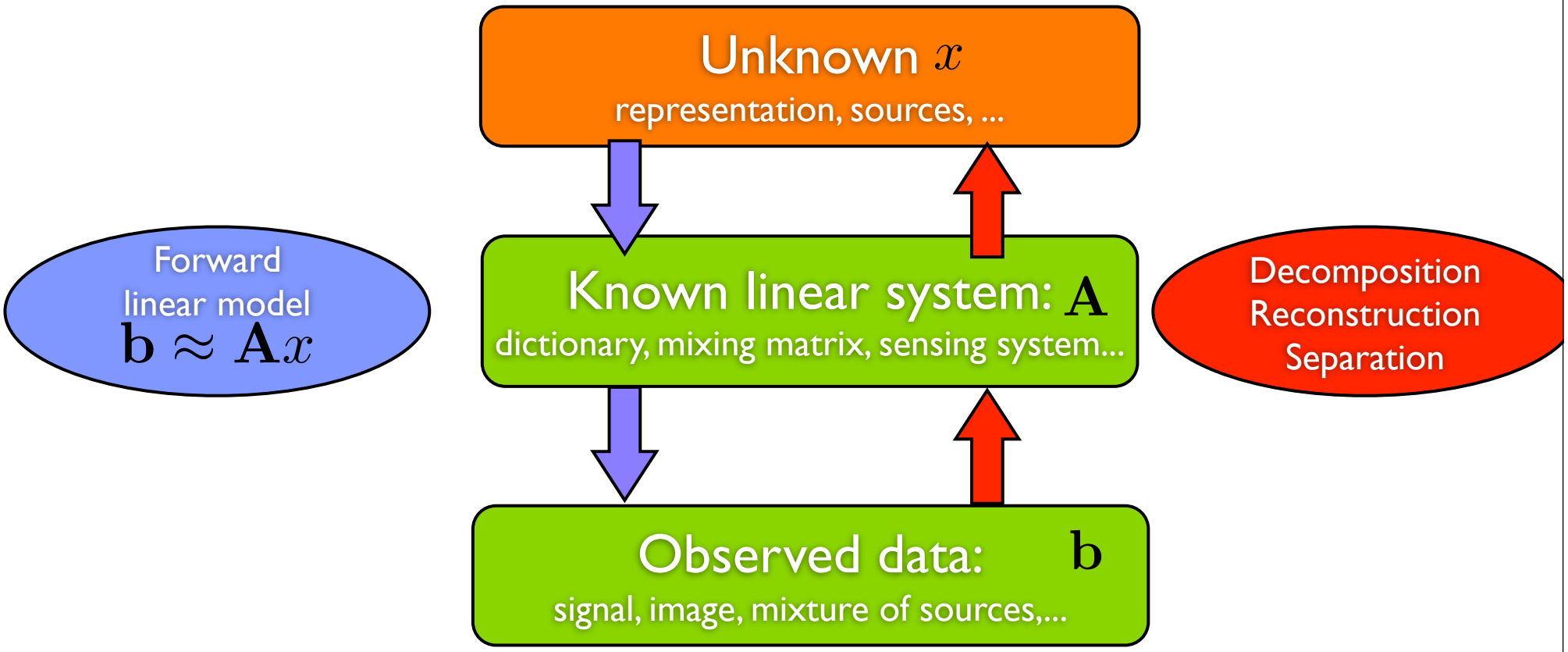
- Bottleneck 1990-2000 : fewer equations than unknowns

$$\mathbf{A}x_0 = \mathbf{A}x_1 \not\Rightarrow x_0 = x_1$$

- Novelty 2001-2006 :
  - Uniqueness of sparse solution:
    - if x_0, x_1 are "sufficiently sparse",
    - then $\mathbf{A}x_0 = \mathbf{A}x_1 \Rightarrow x_0 = x_1$
  - Recovery of x_0 with practical algorithms
    - Thresholding, Matching Pursuits, Minimisation of Lp norms p<=1,...

Algorithmic principles for sparse approximation

Vocabulary


Unknown x
representation, sources, ...
Forward
linear model
$\mathbf{b} \approx \mathbf{A}x$
Known linear system: $\mathbf{A}$
dictionary, mixing matrix, sensing system...
Decomposition
Reconstruction
Separation
Observed data: $\mathbf{b}$
signal, image, mixture of sources,...

Ideal sparse approximation

- Input:

  m x *N* matrix **A**, with *m* < *N*, *m*-dimensional vector **b**

- Possible objectives:

  find the sparsest approximation within tolerance

$$\arg\min_x \|x\|_0, \ \text{s.t.} \|\mathbf{b} - \mathbf{A}x\| \leq \epsilon$$

  find best approximation with given sparsity

$$\arg\min_x \|\mathbf{b} - \mathbf{A}x\|, \ \text{s.t.} \|x\|_0 \leq k$$

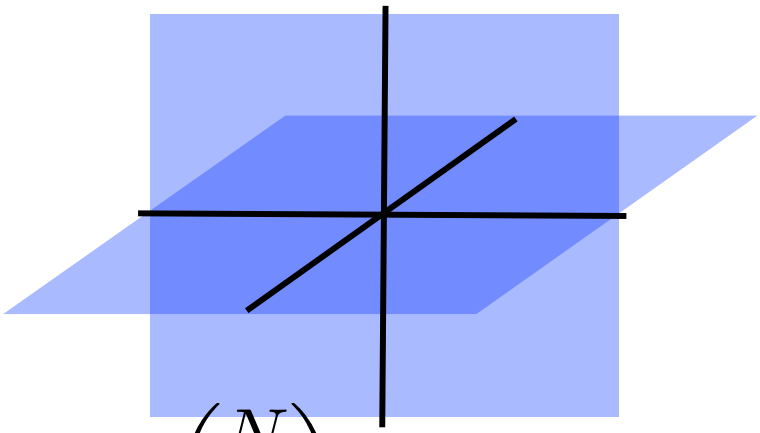
  find a solution *x* to

$$\|\mathbf{b} - \mathbf{A}x\| \leq \epsilon, \ \text{and} \ \|x\|_0 \leq k$$

Geometric interpretation of sparse approximation

- Coefficient domain $\mathbb{R}^N$:
  - ✦ set Σ_k of sparse vectors

$\|x\|_0 \leq k$



$\binom{N}{k}$ subspaces

- Set $\mathbf{A}\Sigma_k = \binom{N}{k}$ subspaces in signal domain

- Ideal sparse approximation = find nearest subspace among $\binom{N}{k}$

Combinatorial search!
Actual complexity ?
NP-complete!

Practical approaches: Optimization *principles*

Overall compromise

- Approximation quality

$$\|\mathbf{A}x - \mathbf{b}\|_2$$

- Ideal sparsity measure : ℓ^0 "norm"

$$\|x\|_0 := \sharp\{n,\ x_n \neq 0\} = \sum_n |x_n|^0$$

- "Relaxed" sparsity measures

$$0 < p < \infty,\ \|x\|_p := \Big(\sum_n |x_n|^p\Big)^{1/p}$$

L*p* norms / quasi-norms

- **Norms** when $1 \leq p < \infty$ = convex

$$\|x\|_p = 0 \Leftrightarrow x = 0$$

$$\|\lambda x\|_p = |\lambda| \|x\|_p, \forall \lambda, x$$

Triangle inequality $$\|x + y\|_p \leq \|x\|_p + \|y\|_p, \forall x, y$$

- **Quasi-norms** when $0 < p < 1$ = nonconvex

$$\|x + y\|_p \leq 2^{1/p} \big(\|x\|_p + \|y\|_p \big), \forall x, y$$

Quasi-triangle inequality

$$\|x + y\|_p^p \leq \|x\|_p^p + \|y\|_p^p, \forall x, y$$

- "*Pseudo*"-norm for p=0

$$\|x + y\|_0 \leq \|x\|_0 + \|y\|_0, \forall x, y$$

Optimization problems

- Approximation

$$\min_x \|\mathbf{b} - \mathbf{A}x\|_2 \text{ s.t. } \|x\|_p \leq \tau$$

- Sparsification

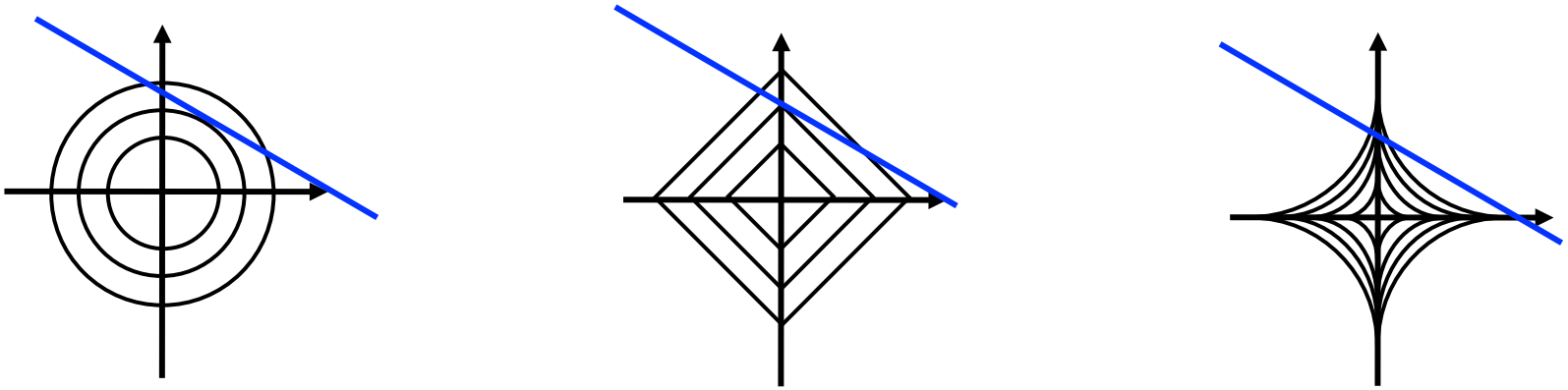
$$\min_x \|x\|_p \text{ s.t. } \|\mathbf{b} - \mathbf{A}x\|_2 \leq \epsilon$$

- Regularization

$$\min_x \frac{1}{2}\|\mathbf{b} - \mathbf{A}x\|_2 + \lambda\|x\|_p$$

L*p* "norms" level sets

- Strictly convex when $p>1$
- Convex $p=1$
- Nonconvex $p<1$

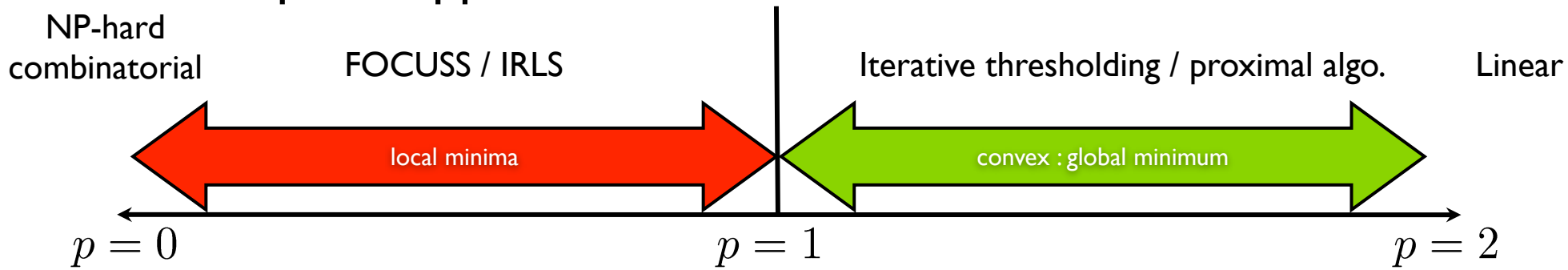


Observation: *the minimizer is sparse*

$\{x \text{ s.t.} \mathbf{b} = \mathbf{A}x\}$

Global Optimization : from Principles to Algorithms

- Optimization principle $$\min_x \frac{1}{2}\|\mathbf{A}x - \mathbf{b}\|_2^2 + \lambda \|x\|_p^p$$
  - ✦ Sparse representation $\lambda \to 0$ $\mathbf{A}x = \mathbf{b}$
  - ✦ Sparse approximation $\lambda > 0$ $\mathbf{A}x \approx \mathbf{b}$



Lasso [Tibshirani 1996], *Basis Pursuit (Denoising)* [Chen, Donoho & Saunders, 1999]
Linear/Quadratic programming (interior point, etc.)
Homotopy method [Osborne 2000] / Least Angle Regression [Efron &al 2002]
Iterative / proximal algorithms [Daubechies, de Frise, de Mol 2004, Combettes & Pesquet 2008, ...]

Greedy Algorithms

Greedy algorithms

- Observation: when **A** is orthormal,
  - the problem

$$\min_x \|\mathbf{b} - \mathbf{A}x\|_2^2 \text{ s.t. } \|x\|_0 \leq k$$

  - is equivalent to

$$\min_x \sum_n (\mathbf{A}_n^T \mathbf{b} - x_n)^2 \text{ s.t. } \|x\|_0 \leq k$$

- Let Λ_k index the *k* largest inner products

$$\min_{n \in \Lambda_k} |\mathbf{A}_n^T \mathbf{b}| \geq \max_{n \notin \Lambda_k} |\mathbf{A}_n^T \mathbf{b}|$$

  - an optimum solution is

$$x_n = \mathbf{A}_n^T \mathbf{b}, n \in \Lambda_k; \; x_n = 0, n \notin \Lambda_k$$

Greedy algorithms

- Iterative algorithm (= *Matching Pursuit*)
  - Initialize a residual to $\mathbf{r}_0 = \mathbf{b} \qquad i = 1$
  - Compute all inner products

$$\mathbf{A}^T \mathbf{r}_{i-1} = (\mathbf{A}_n^T \mathbf{r}_{i-1})_{n=1}^N$$

  - Select the largest in magnitude

$$n_i = \arg\max_n |\mathbf{A}_n^T \mathbf{r}_{i-1}|$$

  - Compute an updated residual

$$\mathbf{r}_i = \mathbf{r}_{i-1} - (\mathbf{A}_{n_i}^T \mathbf{r}_{i-1}) \mathbf{A}_{n_i}$$

  - If $i \geq k$ then stop, otherwise increment *i* and iterate

Matching Pursuit (MP)

- Matching Pursuit (*aka* Projection Pursuit, CLEAN)
  - Initialization $\mathbf{r}_0 = \mathbf{b} \qquad i = 1$
  - Atom selection:

$$n_i = \arg\max_n |\mathbf{A}_n^T \mathbf{r}_{i-1}|$$

  - Residual update

$$\mathbf{r}_i = \mathbf{r}_{i-1} - (\mathbf{A}_{n_i}^T \mathbf{r}_{i-1})\mathbf{A}_{n_i}$$

- Energy preservation (Pythagoras theorem)

$$\|\mathbf{r}_{i-1}\|_2^2 = |\mathbf{A}_{n_i}^T \mathbf{r}_{i-1}|^2 + \|\mathbf{r}_i\|_2^2$$

Summary

	Global optimization	Iterative greedy algorithms
Principle	$\min_x \frac{1}{2}\|\mathbf{A}x - \mathbf{b}\|_2^2 + \lambda\|x\|_p^p$	iterative decomposition $\mathbf{r}_i = \mathbf{b} - \mathbf{A}x_i$ • select new components • update residual
Tuning quality/sparsity	regularization parameter λ	stopping criterion (nb of iterations, error level, ...) $\|x_i\|_0 \geq k \qquad \|\mathbf{r}_i\| \leq \epsilon$
Variants	• choice of sparsity measure p • optimization algorithm • initialization	•selection criterion (weak, stagewise ...) •update strategy (orthogonal ...)

Equivalence between L0, L1, OMP

- **Theorem** : assume that $\mathbf{b} = \mathbf{A}x_0$
  - ✦ if $\|x_0\|_0 \leq k_0(\mathbf{A})$ then $x_0 = x_0^\star$
  - ✦ if $\|x_0\|_0 \leq k_1(\mathbf{A})$ then $x_0 = x_1^\star$

where $x_p^\star = \arg \min_{\mathbf{A}x = \mathbf{A}x_0} \|x\|_p$

- *Donoho & Huo 01 : pair of bases, coherence*
- *Donoho & Elad, Gribonval & Nielsen 2003 : dictionary, coherence*
- *Tropp 2004 : Orthonormal Matching Pursuit, cumulative coherence*
- *Candes, Romberg, Tao 2004 : random dictionaries, restricted isometry constants*

IRISA

State of the art tools to estimate $k_p(\mathbf{A})$

$\mathbf{A}$

N unit columns

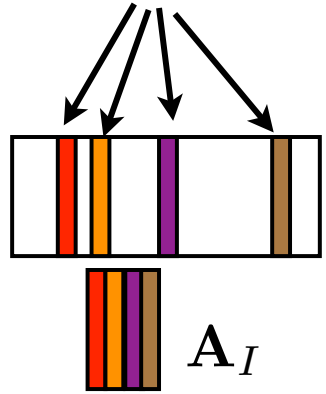
$\|\mathbf{A}_n\|_2 = 1$

$\mathbf{A}^T\mathbf{A} - \mathbf{Id}$

max over N(N-1) entries

$$\mu = \mu(\mathbf{A}) := \max_{k \neq l} |\langle \mathbf{A}_k, \mathbf{A}_l \rangle|$$

$n \in I, \sharp I \leq k$



$\mathbf{A}_I$

max over $\frac{N!}{k!(N-k)!}$ subsets *I*

$$\delta_k := \sup_{\sharp I \leq k,\ c \in \mathbb{R}^k} \left| \frac{\|\mathbf{A}_I c\|_2^2}{\|c\|_2^2} - 1 \right|$$

$$\widehat{k}(\mathbf{A}) = (1 + 1/\mu)/2$$

L0-recovery (identifiability) $\delta_{2k_0} < 1$

L1-recovery (identification) $\delta_{2k_1} < \sqrt{2} - 1$

(Cumulative) coherence
Low cost
"Coarse / pessimistic"

Common beliefs

Isometry constants
Hard to compute
"Almost sharp" ?

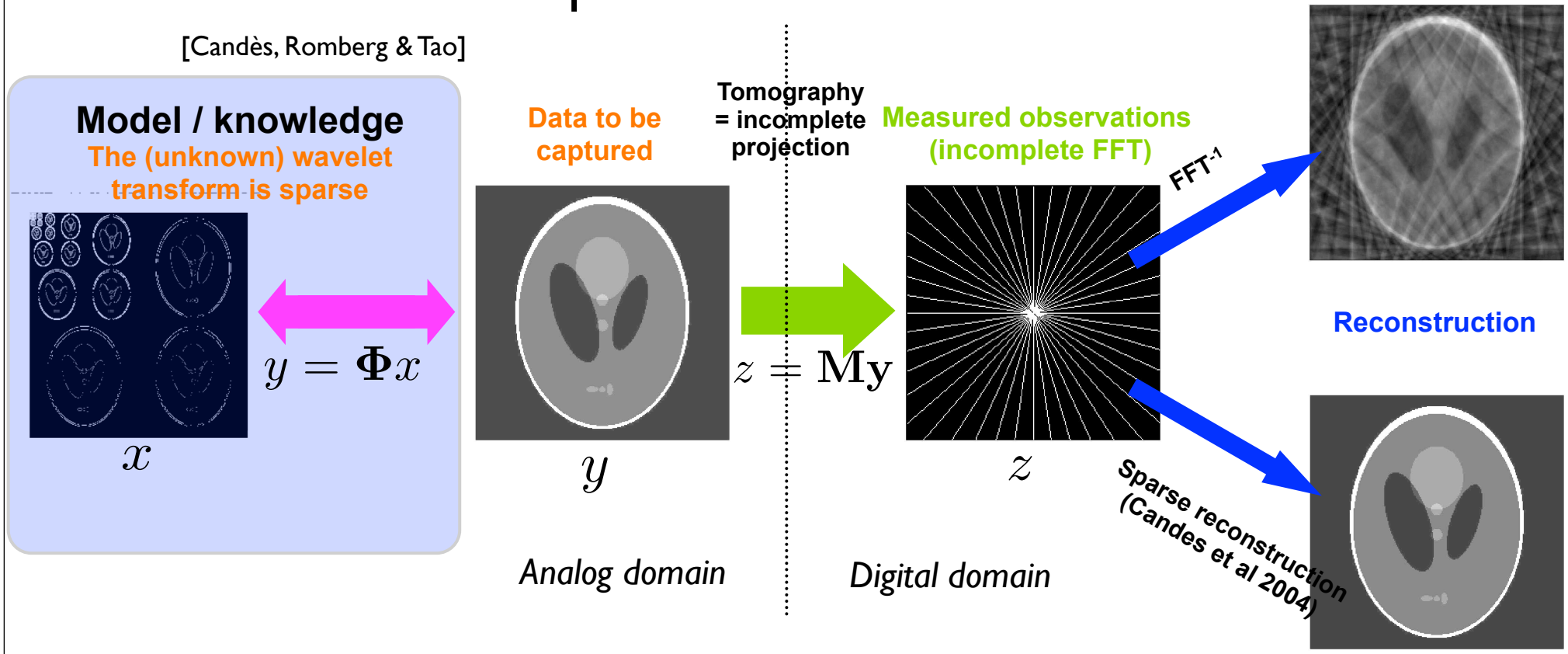
IRISA

Compressed sensing

Example: tomography

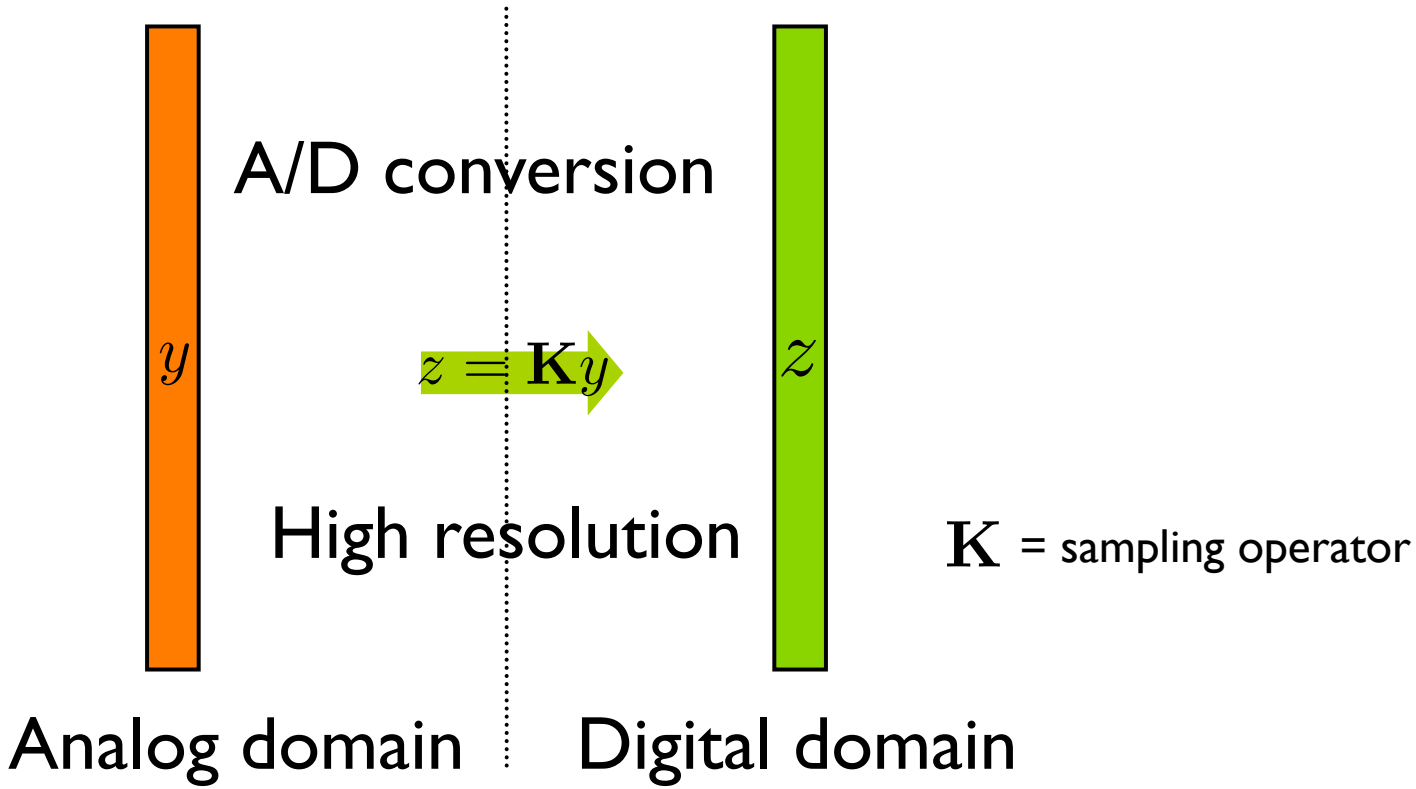
- MRI from incomplete data

[Candès, Romberg & Tao]



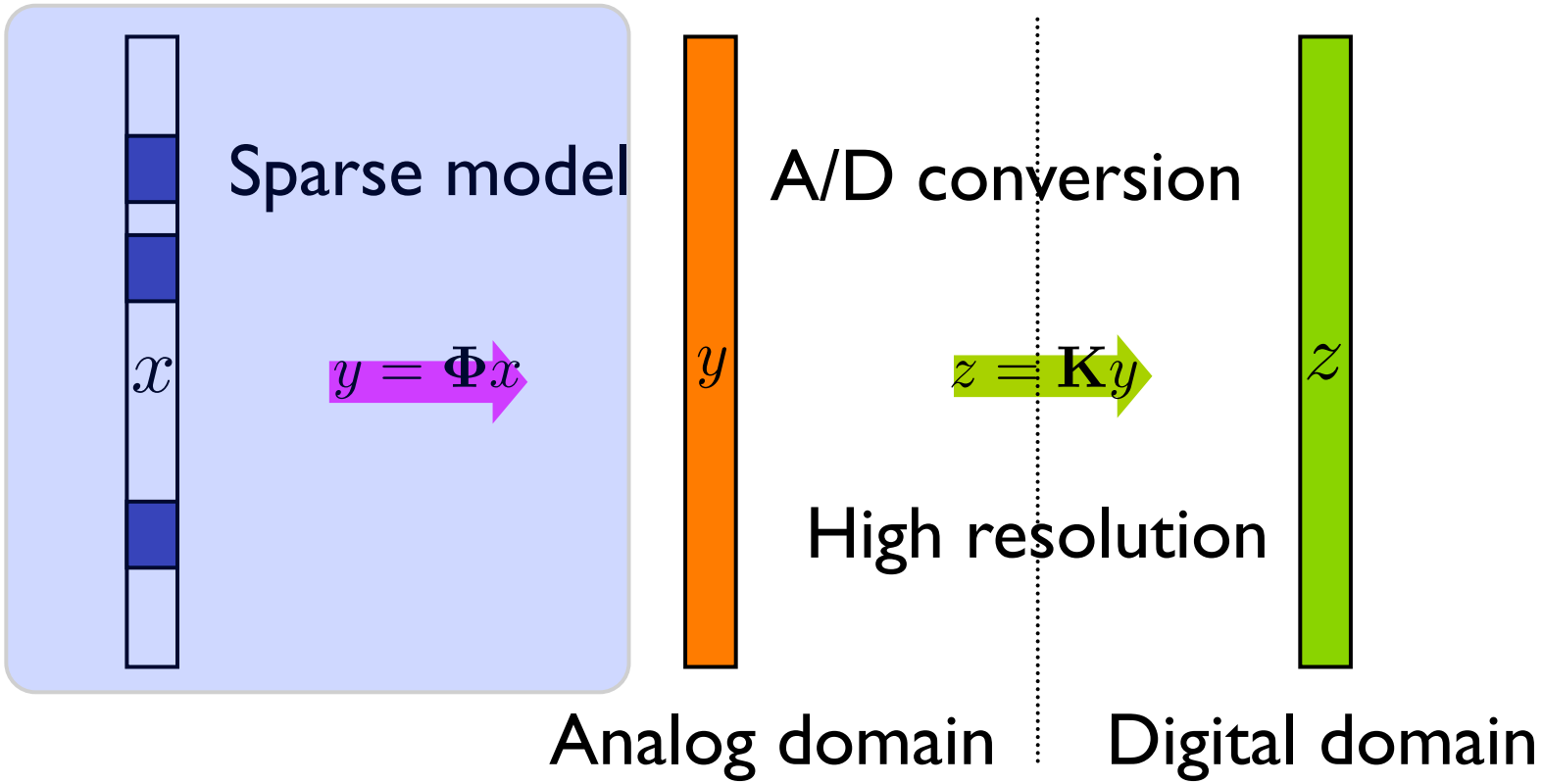
Classical Shannon Sampling

- « Sample first, think and compress afterwards »



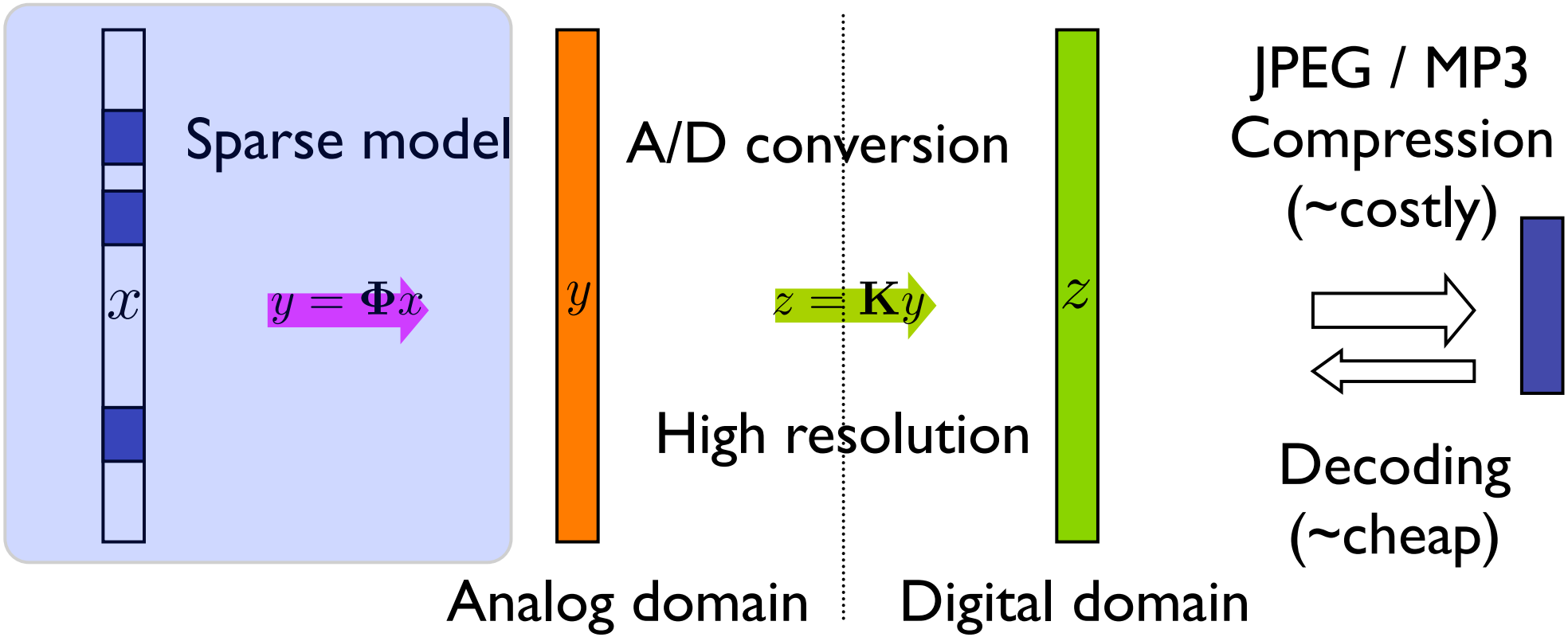
Classical Shannon Sampling

- « Sample first, think and compress afterwards »



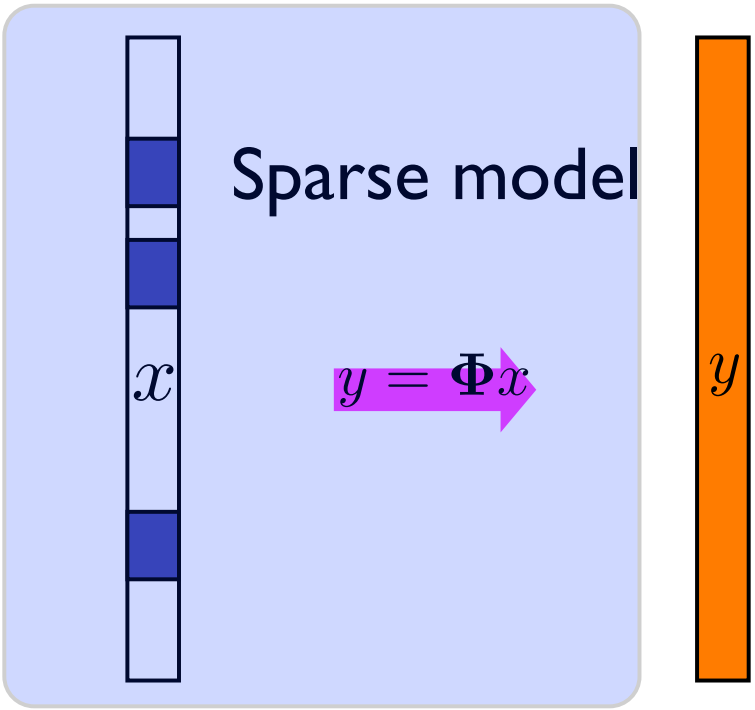
Classical Shannon Sampling

- « Sample first, think and compress afterwards »



Compressed Sensing

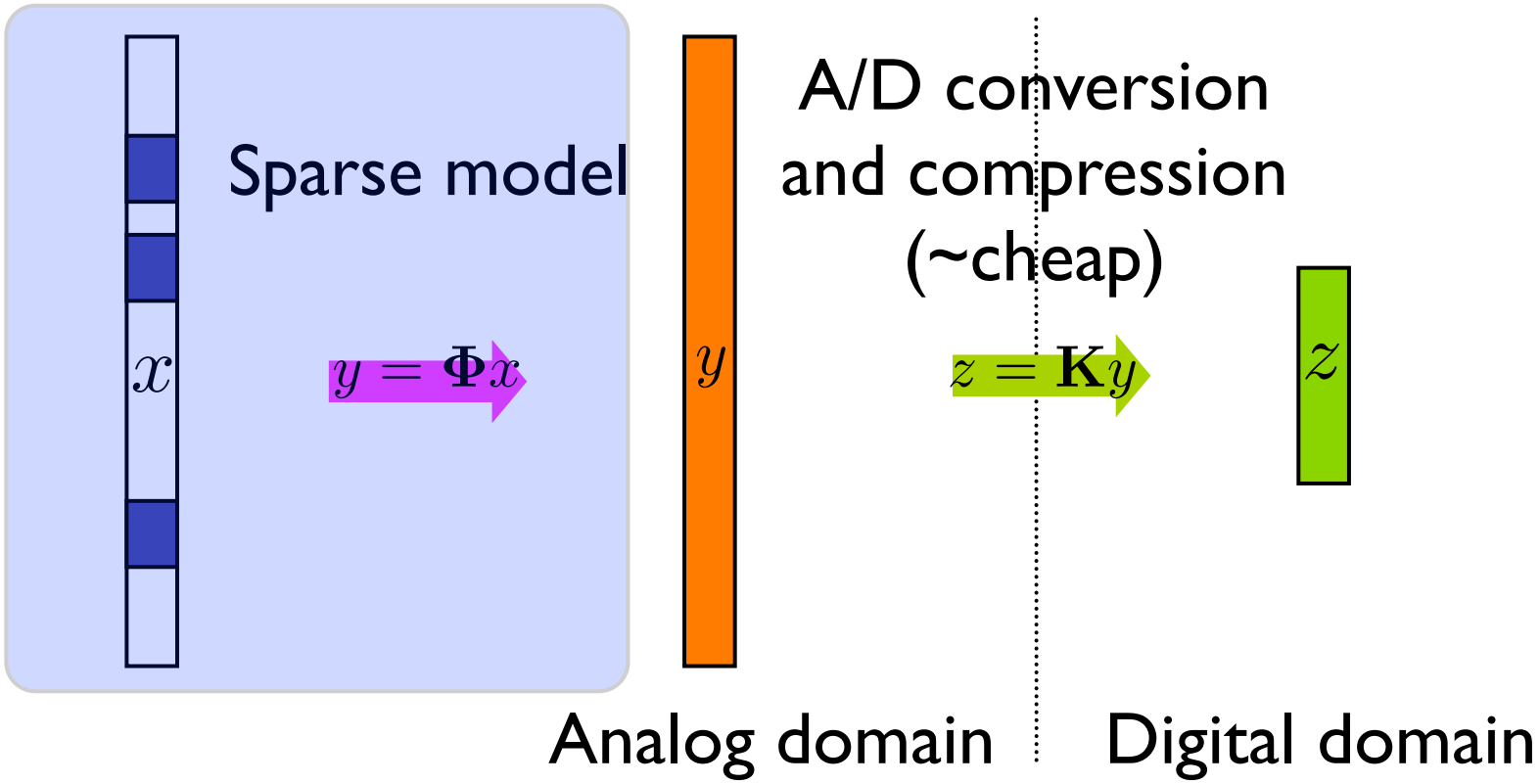
- First model the data, then sample & compress



Analog domain | Digital domain

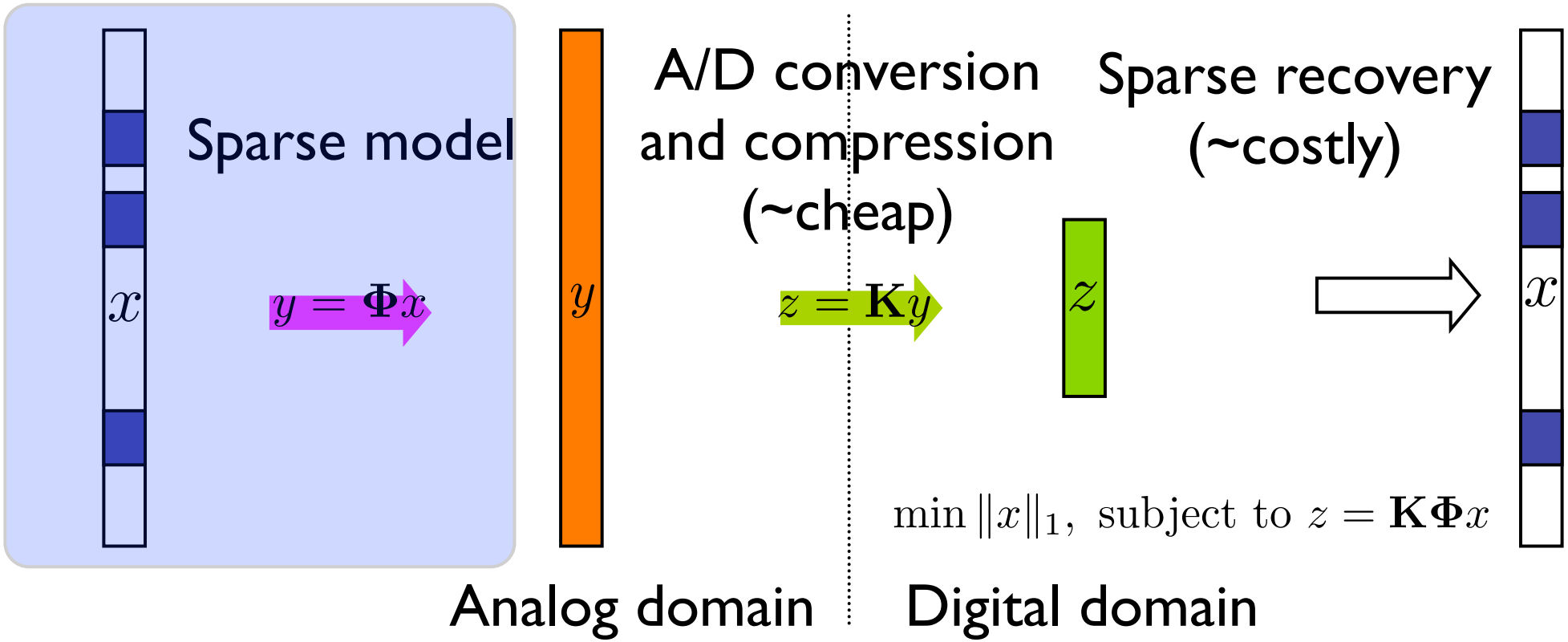
Compressed Sensing

- First model the data, then sample & compress



Compressed Sensing

- First model the data, then sample & compress



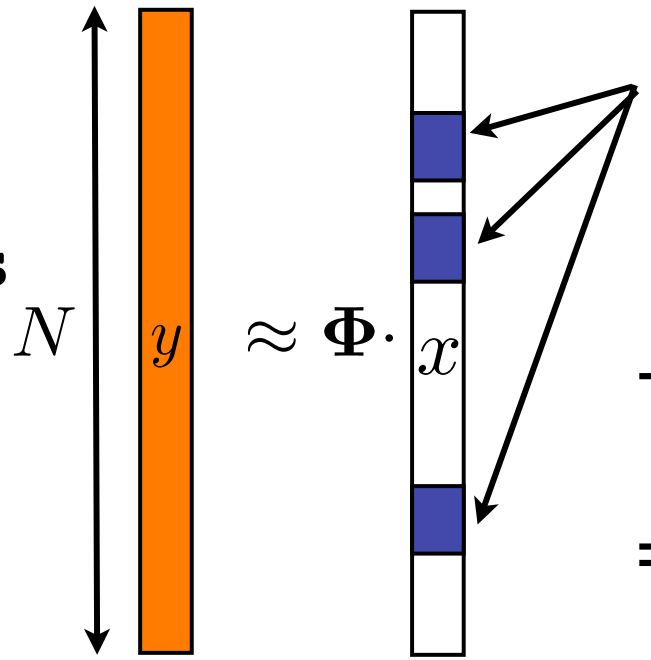
Conditions of success

- Knowledge: transform domain where data is sparse
- «Incoherence» between measurement domain and sparse domain domaine de (uncertainty principle *à la* Heisenberg), for example:
  - time domain / frequency domain
  - spatial domain / frequency domain
  - **random measurements!**
- Sufficiently many measures $$m \geq Ck \log_2 \frac{N}{k}$$
  - necessary
  - sufficient (with random Gaussian measures)

Why the log factor ?

- Full vector

N entries
= N `floats`



- Sparse vector

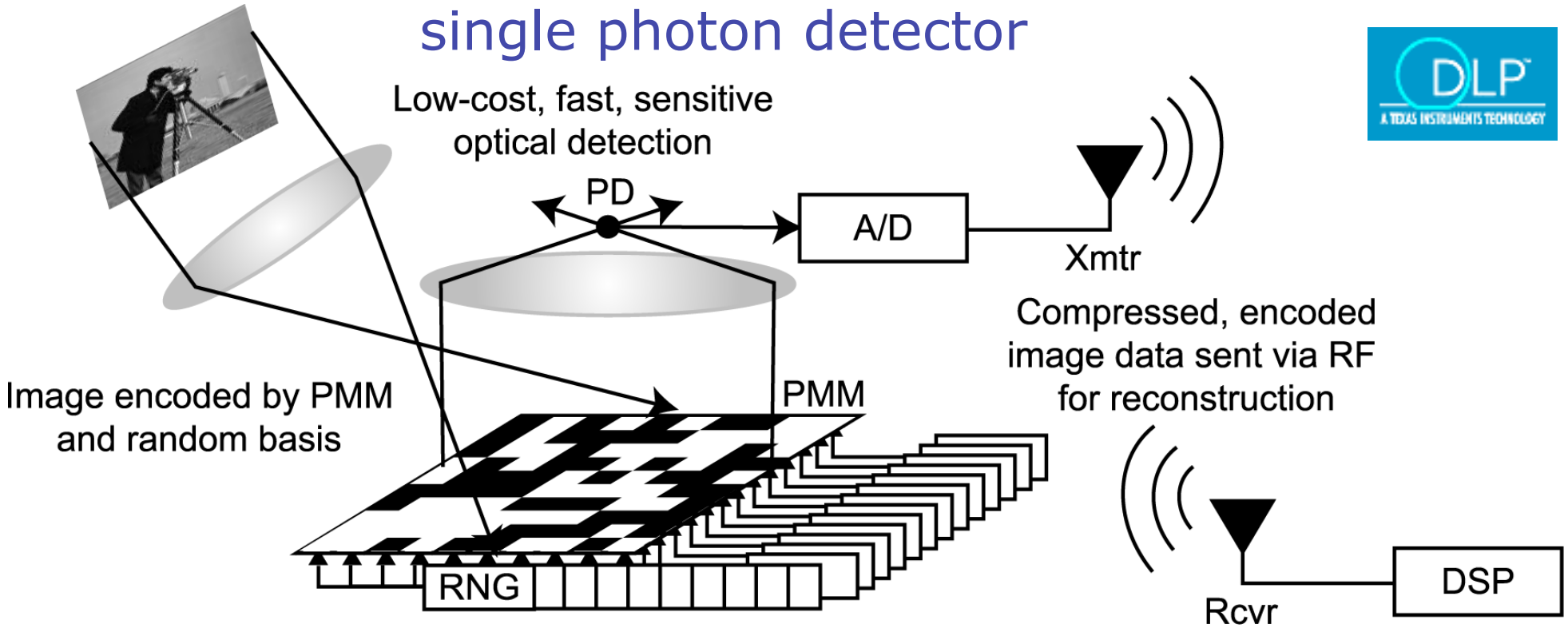
$k \ll N$ nonzero entries
= k `floats`

+ k positions among N

$$= \log_2 \binom{N}{k} \approx k \log_2 \frac{N}{k} \texttt{ bits}$$

Key practical issues: choose dictionary

Example : single-pixel camera, Rice University



Random pattern on DMD array

image reconstruction

Summary

Notion of sparsity
(Fourier, wavelets, ...)

Compression
Representation
Description
Classification

Sparsity

Denoising
Blind source separation
Compressed sensing
...

Natural / traditional role

Novel indirect role

Sparsity = low cost (bits, computations, ...)
Direct objective

Sparisty = prior knowledge, regularization
Tool for inverse problems