Traditionally, we'd like to capture analog signals from the physical world and store them digitally on computers for subsequent processing, transmission, or reconstruction.

Examples of "signals" include:
- speech or audio waveforms,
- images (i.e., 2D waveforms),
- video (i.e., 3D waveforms).

The Nyquist theorem says that any bandlimited (i.e., smooth) signal can be sampled (giving a sequence) and then perfectly reconstructed.

The Nyquist rate is the minimum sampling rate (i.e., # samples per unit time) needed for perfect reconstruction.
Some signals are intrinsically simple and thus can be compressed without much loss of quality.

- Audio: MP3 gives roughly 10:1 compression relative to CD (=Nyquist)
- Images: JPEG gives roughly 25:1 compression relative to Nyquist
- Videos: MPEG gives roughly 100:1 compression relative to Nyquist

Compression facilitates efficient storage or transmission:
Compressive sensing

- Sometimes Nyquist sampling is too expensive.

- For compressible signals, Nyquist sampling is overkill.

- Can we do “compressive” sampling? Yes!

- Typical ingredients are:
  - randomly designed linear measurements
  - sparse signal representation
  - sophisticated signal reconstruction
Motivation

In some applications, measurements are costly:

- **Magnetic resonance imaging:**
  - scan time \( \approx 30 \) minutes
  - scan time proportional to \# samples taken

- **Imaging outside visible spectrum:**
  - CMOS doesn’t work
  - high cost per pixel

- **Wireless communication:**
  - pilots inserted to measure channel
  - more pilots means less payload
System architecture

- Classical approach:

\[
\{s(t)\}_{t \in [0,T)} \rightarrow \text{Nyquist sample} \rightarrow \{s_n\}_{n=1}^{N} \rightarrow \text{compress} \rightarrow \{c_k\}_{k=1}^{K} \rightarrow \text{reconstruct}
\]

- New approach:

\[
\{s(t)\}_{t \in [0,T)} \rightarrow \text{compressively sample} \rightarrow \{y_m\}_{m=1}^{M} \rightarrow \text{reconstruct}
\]

Nyquist rate \(\frac{N}{T} \gg\) compressive sampling rate \(\frac{M}{T} \gtrsim\) information rate \(\frac{K}{T}\)
Principal challenges in compressive sensing

1. Design of the compressive-measurement scheme

2. Reconstruction from the compressed measurements
   - We focus on recovering the Nyquist-rate signal samples \( \{s_n\}_{n=1}^N \)
   - Could easily reconstruct analog \( \{s(t)\}_{t \in [0,T]} \) from Nyquist samples.
For now, assume noiseless linear measurements, e.g.,

\[ y_m = \int_0^T \phi_m(t) s(t) \, dt, \quad m = 1, \ldots, M \]

Also assume signal \( s(t) \) is bandlimited, in which case Nyquist says

\[ s(t) = \sum_{n=1}^N s_n \text{sinc} \left( \frac{t}{T_s} - n + 1 \right), \quad t \in [0, T). \]

Putting these together, we get the convenient discrete representation

\[ y_m = \sum_{n=1}^N s_n \int_0^T \phi_m(t) \text{sinc} \left( \frac{t}{T_s} - n + 1 \right) \, dt \]

\[ \Delta \Phi_{m,n} \]

or, in matrix/vector form,

\[ \mathbf{y} = \Phi \mathbf{s} \]

for \( \mathbf{s} \in \mathbb{R}^N \) and \( \mathbf{y} \in \mathbb{R}^M \).
Design of linear measurements

Goal: design the matrix $\Phi \in \mathbb{R}^{M \times N}$ so that

1. any signal $s$ in class $S$ can be reconstructed from $y = \Phi s$,
2. the number of measurements $M$ is minimal.

Key challenge:

There are fewer measurements $M$ than unknowns $N$.

$\Rightarrow$ Many $s$ satisfy the equation $y = \Phi s$. How to find the correct $s$?

Solution:

- If the signals in class $S$ are sufficiently structured, only one of the $s$ satisfying “$y = \Phi s$” will be valid!
- Examples of structured signals include sparse signals, signals on manifolds, signals that can be expressed as low-rank matrices, etc.
Many real-world signals are approximately sparse in a known basis. For example, natural images are sparse in the discrete wavelet transform (DWT) basis:

Typically: 99% signal energy captured by only 2.5% of DWT coefficients!
We say that a signal class $S$ is $K$-sparse in the dictionary $\Psi$ if each $s \in S$ can be written as

$$s = \Psi x$$

for some $K$-sparse vector $x$ (i.e., $x$ has at most $K$ nonzero elements).

Usually orthonormal dictionaries $\Psi$ are used (e.g., DWT, DCT, DFT), but overcomplete dictionaries may also be considered.

Geometrically, a $K$-sparse vector $x \in \mathbb{R}^N$ lives in a union of $\binom{N}{K}$ subspaces, each of dimension $K$. 
Merging sparsity with linear compression

Recall...

- Linear measurement model: \( y = \Phi s \) for \( \Phi \in \mathbb{R}^{M \times N} \)
- Sparse signal model: \( s = \Psi x \) for \( K\)-sparse \( x \in \mathbb{R}^N \)

Together...

- Compressive sensing model: \( y = \Phi \Psi x \) for \( A \in \mathbb{R}^{M \times N} \)
  \[ \triangleq A \]

Questions:

1. What properties of \( A \) ensure the recovery of \( x \)?
2. Given dictionary \( \Psi \), how can we design \( \Phi \) to ensure a good \( A \)?
**Restricted isometry property**

- Recall model: \( \mathbf{y} = \mathbf{A}\mathbf{x} \) for \( \mathbf{A} \in \mathbb{R}^{M \times N} \) and \( K \)-sparse \( \mathbf{x} \in \mathbb{R}^{N} \).

- Note: if signals \( \mathbf{x}_1 \neq \mathbf{x}_2 \) map to the same \( \mathbf{y} \), they can’t be recovered!

- In general, for our measurement system to be information preserving, we want that \( \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \approx \|\mathbf{A}\mathbf{x}_1 - \mathbf{A}\mathbf{x}_2\|_2 \) for all \( K \)-sparse \( \mathbf{x}_1, \mathbf{x}_2 \), or

\[
1 - \delta \leq \frac{\|\mathbf{A}\mathbf{d}\|_2^2}{\|\mathbf{d}\|_2^2} \leq 1 + \delta \quad \text{for all } 2K\text{-sparse } \mathbf{d}. \quad \text{“RIP”}
\]
Ensuring RIP with randomness

- Testing a given matrix for RIP is an NP-hard (combinatorial) problem.

- Fortunately, if $A$ is randomly drawn with independent zero-mean sub-Gaussian entries (e.g., normal, $\pm 1$), then with high probability it will satisfy RIP if

$$M \geq O\left(K \log \frac{N}{K}\right).$$

- Similarly, if $\Phi$ is constructed randomly in the same way, then $A = \Phi \Psi$ will satisfy RIP for any orthonormal $\Psi$.

- In practice, semi-random $\Phi$ are preferable, e.g.,

Create $\Phi = JFD$, where $D$ is a diagonal matrix with random $\pm 1$s, $F$ is the $N$-FFT matrix, and $J$ randomly selects $M$ outputs.
Example: Single-pixel camera (Rice Univ.)
Other examples

Random demodulator:

$$x(t) \times p_c(t) \rightarrow \int \rightarrow \text{Sample-and-Hold} \rightarrow \text{Quantizer} \rightarrow y[n]$$

Compressive multiplexor:

$$x_1(t) \times p_1(t) \rightarrow \text{ADC} \rightarrow y[n]$$

$$x_j(t) \text{ is } \frac{W}{2} \text{ Hz wide}$$

$$f_{\text{chip}} = W \text{ Hz}$$
So far we’ve talked about the design of the compressive sampler. Now we’ll shift focus to \textit{signal reconstruction from compressed $y$}.

In particular, we’ll talk about how to reconstruct the Nyquist-rate signal samples $s$ from

$$y = \Phi s + w$$

with additive measurement noise $w$!

$$= \Phi \Psi x + w$$

$$= Ax + w$$

where $x$ is approximately $K$-sparse.

In fact, recovering $x$ is enough, since we can then construct $s = \Psi x$. 
Sparse reconstruction

Goal: estimate \( \mathbf{x} \in \mathbb{R}^N \) from \( \mathbf{y} = A\mathbf{x} + \mathbf{w} \in \mathbb{R}^M \) where

- \( \mathbf{x} \) is approximately \( K \)-sparse (although \( K \) is unknown)
- \( M \ll N \) but \( M \geq K \)
- \( A \) is RIP-like (all subsets of \( K \) columns are nearly orthonormal)

Popular methods:

- Convex methods based on \( \ell_1 \)-regularization
- Greedy search
- Bayesian inference
Best sparse fit — the $\ell_0$ technique

Find the sparsest $x$ that explains $y$ up to a specified tolerance of $\epsilon$:

$$\hat{x} = \arg \min_x \|x\|_0 \quad \text{s.t. } \|y - Ax\|_2 \leq \epsilon.$$  

Unfortunately, this is \textbf{NP-hard}; we’d need to check all $\binom{N}{K} \approx N^K$ possible supports!

Let’s think about this problem geometrically…
A toy example

Consider \( y = A x + w \) with 1-sparse \( x \).

\[
\begin{bmatrix}
\bullet \\
\bullet \\
\bullet \\
\bullet \\
\end{bmatrix}
\begin{bmatrix}
\bullet \\
\bullet \\
\bullet \\
\bullet \\
\bullet \\
\bullet \\
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
\end{bmatrix}
\begin{bmatrix}
= 2 \\
= 3 \\
= 1 \\
\end{bmatrix}
\]

- The set of \( x \) such that \( \| y - A x \|_2 \leq \epsilon \) is described by an \( \epsilon \)-thin rod.

- The \( \ell_0 \) technique would check increasingly large support hypotheses until it finds one whose signal subspace intersects the \( \epsilon \)-rod. In this example, it would recover the true \( x \) if \( \epsilon = 0 \).
The geometry of constrained $\ell_p$-minimization

Now consider, for some fixed $p > 0$, the optimization problem:

$$\hat{x} = \arg \min_x \|x\|_p \quad \text{s.t.} \quad \|y - Ax\|_2 \leq \epsilon.$$ 

The solution can be found by growing the $\ell_p$-ball until it touches the $\epsilon$-rod:

- $p \ll 1$: Solution definitely sparse but problem is **NP hard**.
- $p = 1$: Solution usually sparse and problem is **convex**!
- $p = 2$: Solution is **not sparse**; $\Leftrightarrow$ LS when $\epsilon = 0$.

*This suggests to use the $\ell_1$ norm as a surrogate for the $\ell_0$ norm!*
LASSO

\[
\hat{x} = \arg \min_x \|x\|_1 \quad \text{s.t.} \quad \|y - Ax\|_2 \leq \epsilon
\]

- Convex! Can be solved very efficiently.

- For \( A \) satisfying 2\( K \)-RIP, LASSO guarantees that

\[
\|\hat{x} - x\|_2 \leq \frac{C_1}{\sqrt{K}} \|x - x_K\|_1 + C_2 \|w\|_2
\]

where \( x_K \) is the best \( K \)-sparse approximation of \( x \) and \( C_1, C_2 \) are constants that depend on the RIP \( \delta \). Wow!

- In the special case when \( x \) is \( K \)-sparse, this simplifies to

\[
\|\hat{x} - x\|_2 \leq C_2 \|w\|_2.
\]
Greedy search

Main ideas:

- If we can correctly recover the support $\Lambda$ of $x$ (i.e., the locations of nonzeros), then determining the non-zero amplitudes is easy, e.g.,

$$x_\Lambda = (A_\Lambda^H A_\Lambda)^{-1} A_\Lambda^H y$$

(least squares)

- Estimate the support sequentially:
  - Find the column of $A$ most “similar” to $y$ and store its index.
  - Subtract the effect of this column from $y$.
  - Repeat (until residual is sufficiently small)!

Famous algorithms include MP, OMP, IHT, CoSaMP, Subspace Pursuit
Bayesian Methods

In the Bayesian approach, one...

- models the signal using a prior pdf $p(x)$,
- models the measurement process using a likelihood function $p(y|x)$,
- performs inference via Bayes rule, yielding the posterior pdf
  $$p(x|y) = Z^{-1} p(y|x)p(x)$$  where $Z$ is a scaling constant,
- often summarizes the posterior pdf by a point estimate like
  $$\hat{x} = \int x \, p(x|y) \, dx$$  MMSE estimate
  $$\hat{x} = \arg \max_x p(x|y)$$  MAP estimate

and possible other statistics that quantify estimate uncertainty.
Bayesian interpretation of LASSO

If we assume...

- additive white Gaussian noise of variance $\sigma^2$
- i.i.d Laplacian signal with rate $\lambda/\sigma^2$

then

- likelihood: $p(y|x) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2} \|y - Ax\|_2^2\right)$
- prior: $p(x) = \frac{1}{(2\sigma^2/\lambda)^M} \exp\left(-\frac{\lambda}{\sigma^2} \|x\|_1\right)$

for which the maximum aposteriori (MAP) estimate is

$$\hat{x} = \arg \max_x p(x|y) = \arg \max_x \log \left(Z^{-1} p(y|x)p(x)\right)$$

$$= \arg \min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1$$

which is an unconstrained version of the LASSO problem.
The relevance vector machine (RVM)

The RVM is based on the *conditionally Gaussian* priors
\[
p(x|\alpha) = \prod_{n=1}^{N} \mathcal{N}(x_n; 0, \alpha_n^{-1}) \quad \text{and} \quad p(\alpha) = \prod_{n=1}^{N} \Gamma(\alpha_n; 0, 0)
\]
\[
p(w|\beta) \sim \prod_{m=1}^{M} \mathcal{N}(w_m; 0, \beta^{-1}) \quad \text{and} \quad \beta \sim \Gamma(0, 0)
\]
Note that, as “precision” \( \alpha_n \to \infty \), the coefficient \( x_n \) is zeroed.

The *conditional* posterior is (due to Gaussianity) simply
\[
p(x|y, \alpha, \beta) \sim \mathcal{N}(\mu, \Sigma) \quad \text{for} \quad \begin{cases} 
\mu = \beta \Sigma A^T y \\
\Sigma = (\beta A^T A + D(\alpha))^{-1}
\end{cases}
\]

In practice, \( \{\alpha, \beta\} \) are estimated using the EM algorithm and then plugged into \( \mu \) and \( \Sigma \) to approximate the posterior \( p(x|y) \).

The RVM (also known as “SBL” and “BCS”) is relatively slow.
Other Bayesian methods

- **Bayesian matching pursuits:**
  - Greedy methods that use probabilistic support selection.

- **Approximate message passing (AMP):**
  - Inspired by methods from statistical physics and information theory.
  - Near-optimal in terms of speed and accuracy if $A$ is large & random.
Phase transition curve (PTC) under large random $A$

When examining a given algorithm’s performance as a function of sampling ratio $\frac{M}{N}$ and sparsity ratio $\frac{K}{M}$, one finds a very sharp transition between perfect success and complete failure as $N, M, K \to \infty$.

In some cases (e.g., LASSO), the PTC can be determined analytically.
Algorithm comparison 1

Recall: higher PTC = better algorithm.

Here, the non-zero elements of $\mathbf{x}$ were drawn independent zero-mean Gaussian.
Algorithm comparison 1

Recall: higher PTC = better algorithm.

Here, the non-zero elements of $x$ were $= 1$.
More structure $\Rightarrow$ possibility for better performance.
Conclusions

Compressive sensing . . .

- merges sampling and signal compression into a single operation
- is motivated by applications where cost-per-sample is high
- uses random linear measurements
- exploits the inherent sparsity of natural signals
- requires sophisticated algorithms for signal reconstruction.