

On the conditionality principle in pattern recognition

Carey E. Priebe

cep@jhu.edu

Johns Hopkins University

Department of Applied Mathematics & Statistics

Department of Computer Science

Center for Imaging Science

February Fourier Talks

The Norbert Wiener Center for Harmonic Analysis and Applications

University of Maryland at College Park

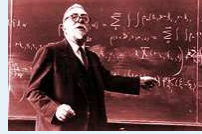
February 16 - 17, 2006

Harmonic Analysis?

Harmonic Analysis?



Harmonic Analysis?



Harmonic Analysis?



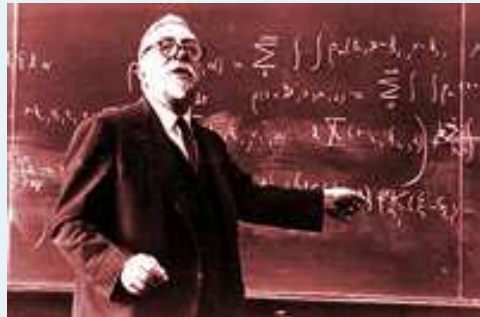
Harmonic Analysis?



Harmonic Analysis?



Harmonic Analysis?



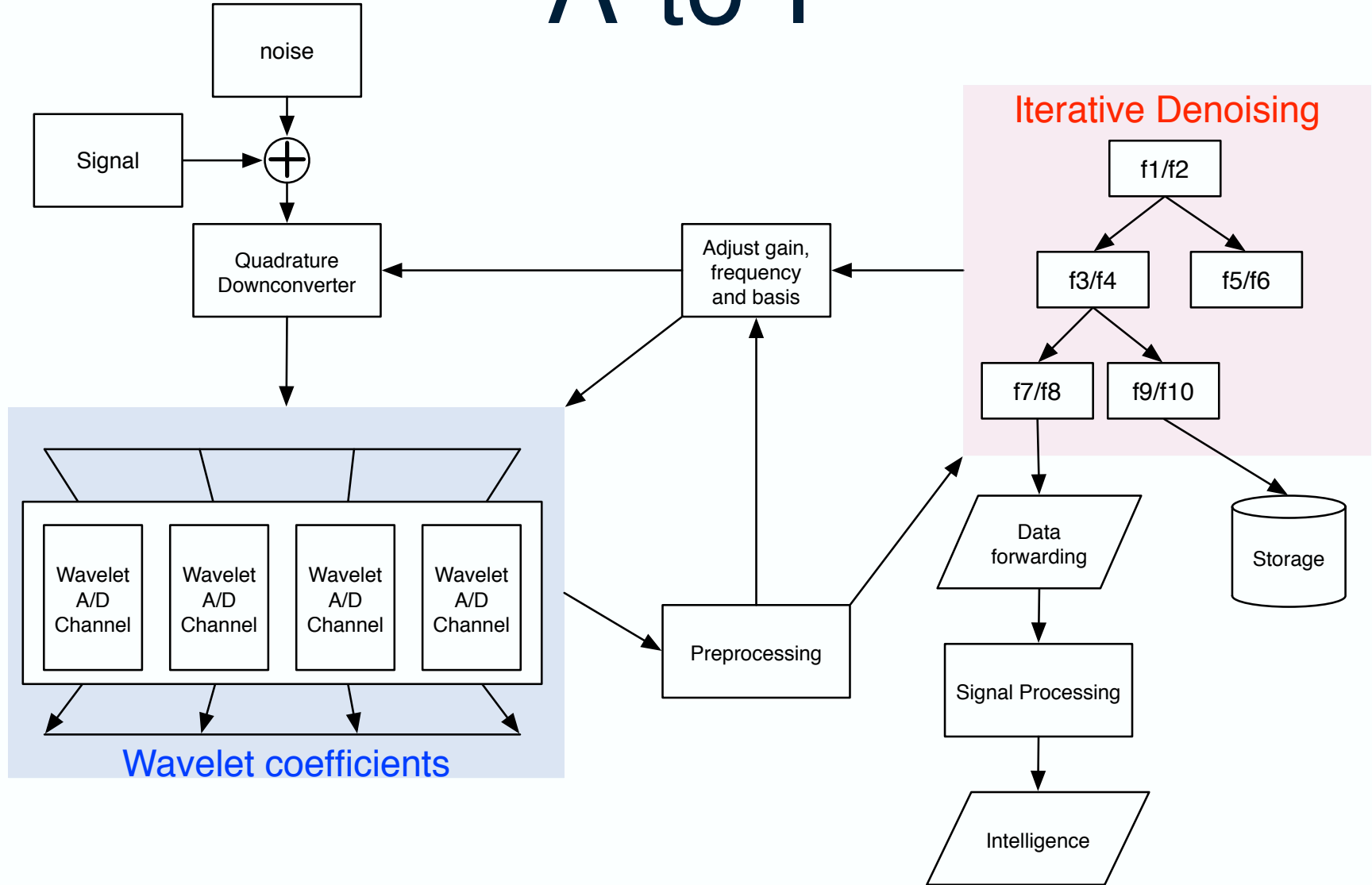
A-to-I

- Jim Clay
- Yasmin Said
- Edward J. Wegman
- Bill May

A COMPRESSED SENSING APPROACH TO SIGINT PROCESSING DARPA A-to-I Grant N66001-06-1-2009

“Dr. Healy manages several programs where mathematical algorithms play a central role in the optimization, control, and exploitation of microelectronic and optical systems. **The Analog-to-Information (A-to-I) program is exploring new ways to extract information from complex signals, seeking significant reduction of the sampling resources required by classical Shannon representations, effectively concentrating meaningful information into less [digitized] data.**”

A-to-I



A-to-I?



Nude Descending a Staircase. Marcel Duchamp, 1912.

A-to-I?



Lone Rock, Iowa

Catch-22

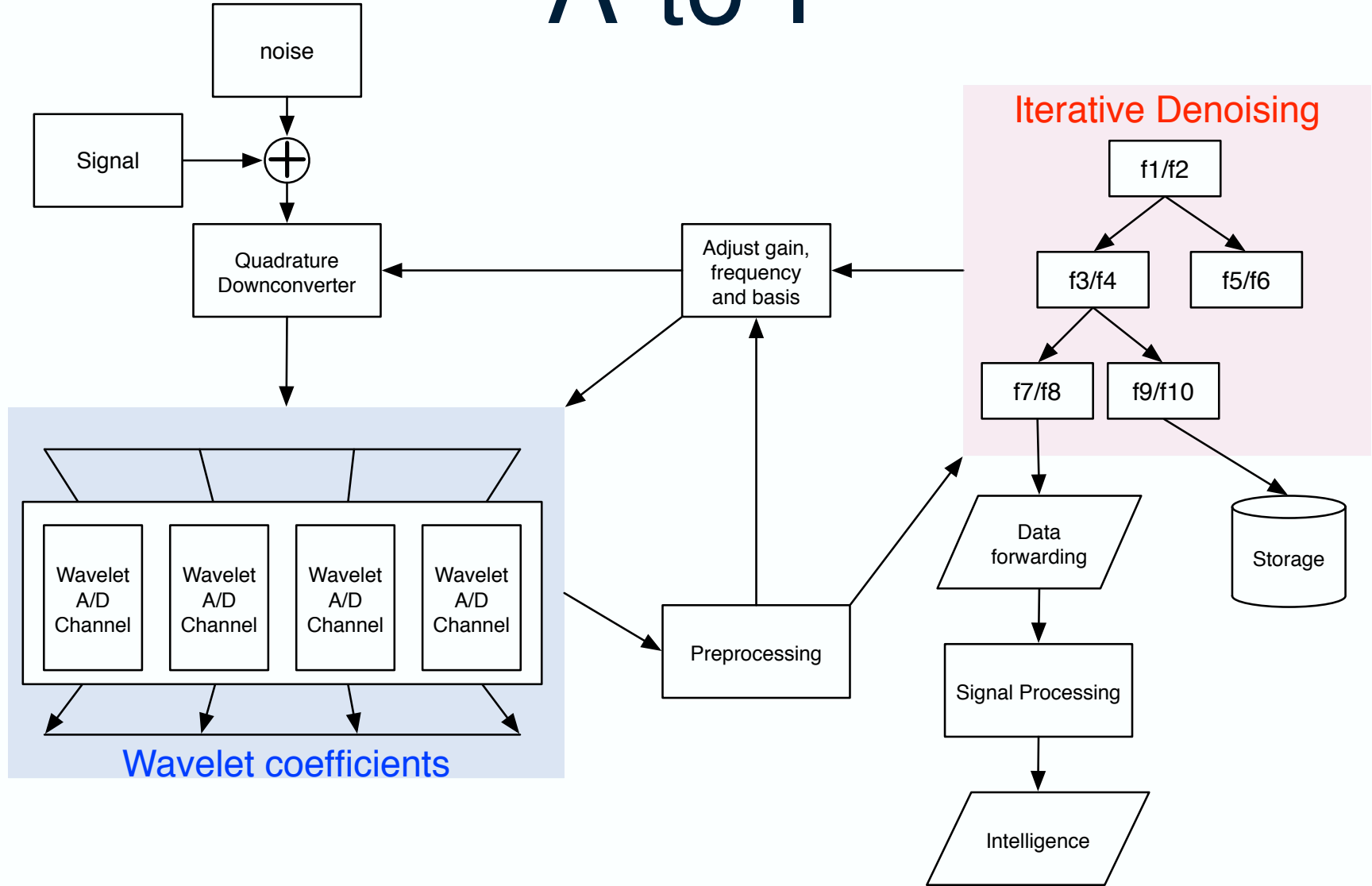
- Jim Clay, AST
- Yasmin Said
- Edward J. Wegman
- Bill May



A COMPRESSED SENSING APPROACH TO SIGINT PROCESSING DARPA A-to-I Grant N66001-06-1-2009

“Dr. Healy manages several programs where mathematical algorithms play a central role in the optimization, control, and exploitation of microelectronic and optical systems. **The Analog-to-Information (A-to-I) program is exploring new ways to extract information from complex signals, seeking significant reduction of the sampling resources required by classical Shannon representations, effectively concentrating meaningful information into less [digitized] data.**”

A-to-I



Wavelet A/D

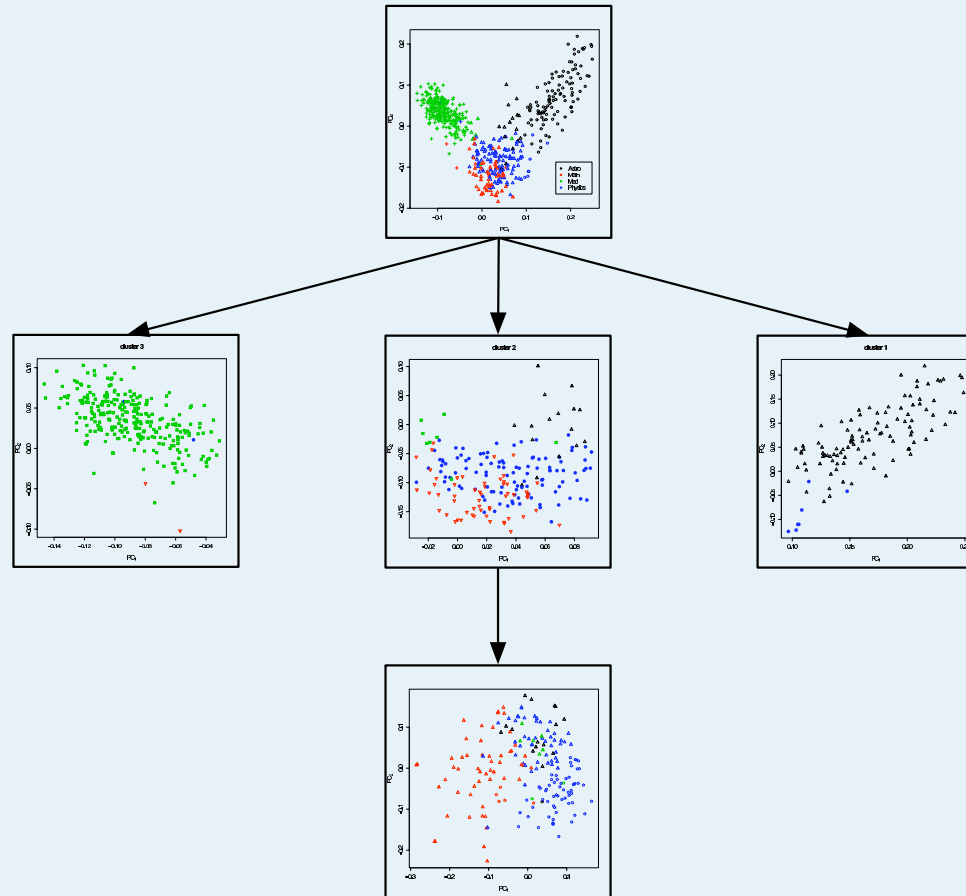
- Haar

Wavelet A/D

- Haar
- other multiresolution families . . . that have constant amplitude

Iterative Denoising

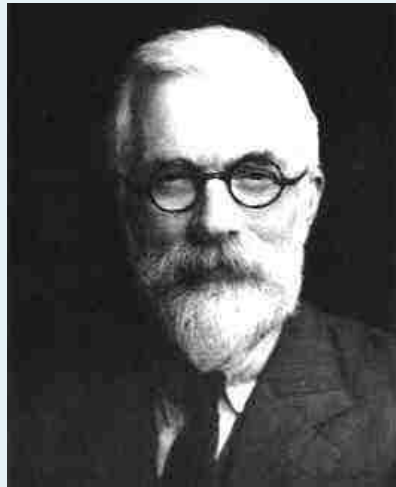
We present Iterative Denoising as (the brains of)
an Intelligent Sensor in an Analog-to-Information framework.



black: Astronomy, red: Mathematics, green: Medicine, blue: Physics

Fisher's Conditionality Principle

Foundations of Statistical Inference:
Likelihood Principle, Sufficiency Principle, Conditionality Principle



Fisher's Conditionality Principle (1950,1956)

But: Welch (1939)?

Amari (1985)

Consider Amari's 1985 statement of the Conditionality Principle:

"When there exists an exact ancillary statistic r , the conditionality principle requires that the statistical inference should be performed by conditioning on r"

Shun-ichi Amari,
Differential Geometric Methods in Statistics,
Lecture Notes in Statistics, Vol 28, 1985,
page 217.

Amari continues ...

(inference about u is the goal)

"... A statistical problem then is decomposed into subproblems in each of which r is fixed at its observed value, thus dividing the whole set of the possible data points into subclasses. It is expected that each subclass consists of relatively homogeneous points with respect to the informativeness about u . We can then evaluate our conclusion about u based on r , and it gives a better evaluation than the overall average one. This is a way of utilizing information which ancillary r conditionally carries."

Theorem

Given $\epsilon > 0$, we construct F_{XY} with feature vectors $X = [X_1, \dots, X_d]' \in \mathbb{R}^{d=d(\epsilon)}$ and class labels $Y \in \{0, 1\}$ such that for $(X, Y) \sim F_{XY}$

$$\min_{g; i, j} P[g(X_i, X_j) \neq Y] \geq \frac{1}{2} - \epsilon$$

while

$$\exists g \text{ with } P[g(X_1, X_{X_1}) \neq Y] = 0,$$

and X_1 is ancillary for the classification task at hand.

That is, there is no pair of features X_i, X_j which work, while conditioning on the ancillary X_1 — using X_1, X_{X_1} — works.

P-M-H, IEEE PAMI, 2004.

Theorem

Given $\epsilon > 0$, we construct F_{XY} with feature vectors $X = [X_1, \dots, X_d]' \in \mathbb{R}^{d=d(\epsilon)}$ and class labels $Y \in \{0, 1\}$ such that for $(X, Y) \sim F_{XY}$

$$\min_{g; i, j} P[g(X_i, X_j) \neq Y] \geq \frac{1}{2} - \epsilon$$

while

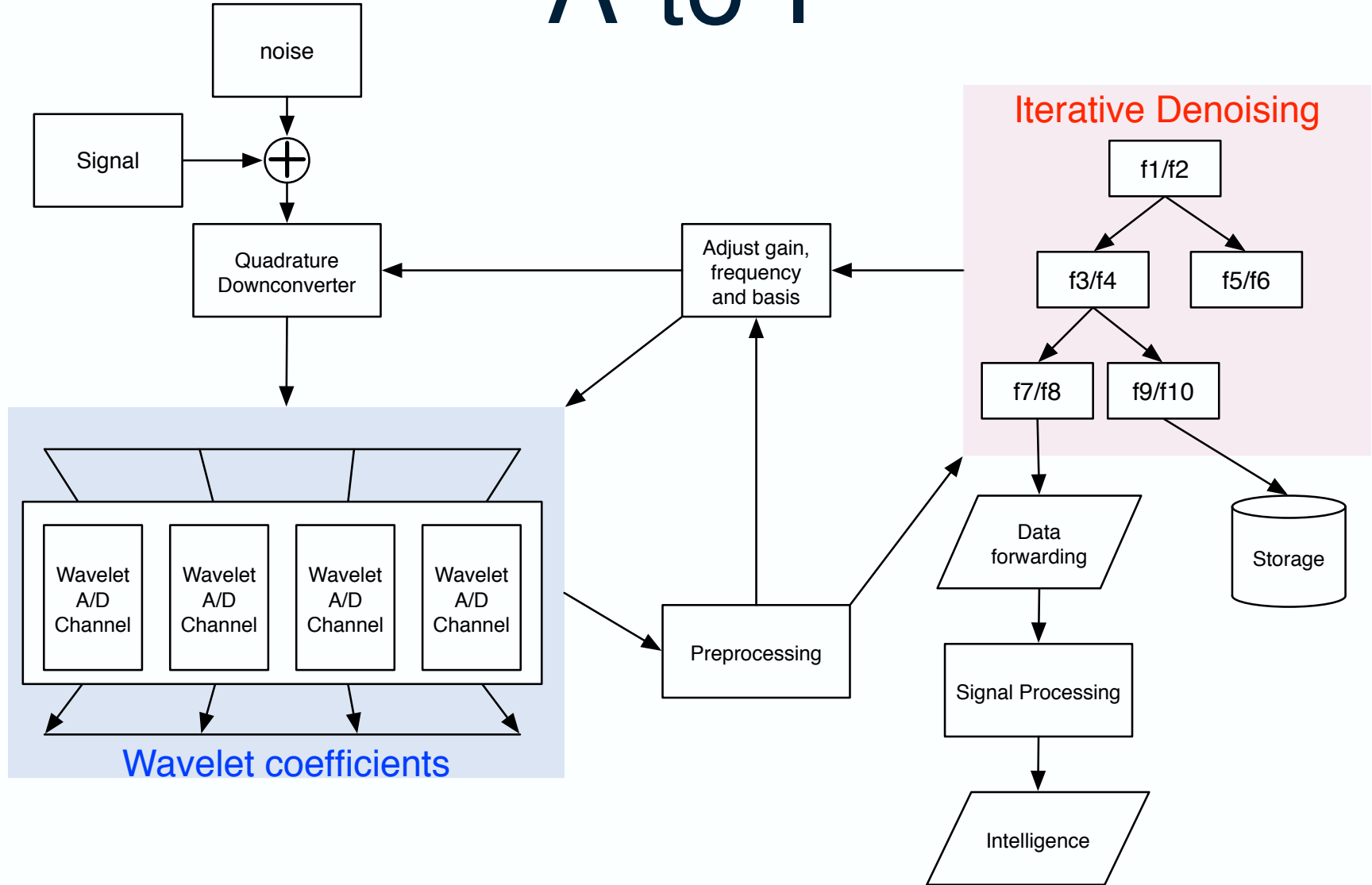
$$\exists g \text{ with } P[g(X_1, X_{X_1}) \neq Y] = 0,$$

and X_1 is ancillary for the classification task at hand.

That is, there is no pair of features X_i, X_j which work, while conditioning on the ancillary X_1 — using X_1, X_{X_1} — works.

P-M-H, IEEE PAMI, 2004.

A-to-I



Example 1



Task #1

A-to-I determination of carrier frequency in broadband applications

“What if for some reason it is difficult to identify the carrier frequency?
For example, what would be done if there are multiple signals at
several carriers present at the same time?”

Example 1

detection & classification

$$g_j(t) = \sum_{k=1}^{K_j} a_{j,k} \sin(\omega_{j,k}t - \phi_{j,k}) + n_j(t)$$

Constraint 1: we can sense, at any one time, via only a simple Haar wavelet (due to anticipated hardware constraints).

Constraint 2: we have time for only a fixed number of these Haar senses (due to anticipated signal duration).

Example 2

We present here a simplified two-class, two-stage theoretical scenario, after P-M-H IEEE PAMI 2004 Theorem 2.1.

For $p \in [0, 1]$,

let $B(p)$ denote the Bernoulli distribution with parameter p .

Let the distribution F_{XY} for three-dimensional random vector $X = [X_0, X_1, X_2]'$ and class label Y be given as follows.

Let $X_0 \sim B(1/2) + 1$.

Let $X_1|X_0 = 1, Y = 0 \sim B(1)$ and $X_1|X_0 = 2, Y = 0 \sim B(1/2)$
and $X_1|X_0 = 1, Y = 1 \sim B(0)$ and $X_1|X_0 = 2, Y = 1 \sim B(1/2)$.

Let $X_2|X_0 = 1, Y = 0 \sim B(1/2)$ and $X_2|X_0 = 2, Y = 0 \sim B(0)$
and $X_2|X_0 = 1, Y = 1 \sim B(1/2)$ and $X_2|X_0 = 2, Y = 1 \sim B(1)$.

Example 2 (continued)

Then $L^* = L_{X_0, X_1, X_2}^* = 0$.

Optimal performance using one and only one canonical dimension is

$$L_{X_0}^* = 1/2 \text{ and } L_{X_1}^* = L_{X_2}^* = 1/4.$$

Optimal performance using two canonical dimensions is

$$L_{X_0, X_1}^* = L_{X_0, X_2}^* = 1/4 \text{ and } L_{X_1, X_2}^* = 1/4.$$

But optimal performance using two dimensions

conditionally on X_0 is

$$L_{X_0, X_{X_0}}^* = 0$$

... and X_0 is ancillary!

Example 2 (continued)

Then $L^* = L_{X_0, X_1, X_2}^* = 0$.

Optimal performance using one and only one canonical dimension is

$$L_{X_0}^* = 1/2 \text{ and } L_{X_1}^* = L_{X_2}^* = 1/4.$$

Optimal performance using two canonical dimensions is

$$L_{X_0, X_1}^* = L_{X_0, X_2}^* = 1/4 \text{ and } L_{X_1, X_2}^* = 1/4.$$

But optimal performance using two dimensions

conditionally on X_0 is

$$L_{X_0, X_{x_0}}^* = 0$$

... and X_0 is ancillary!

Example 1 Revisited

Constraint 1: we can sense, at any one time, via only a simple **two-level** Haar wavelet (due to anticipated hardware constraints).

Constraint 2: we have time for only a fixed number $m = 2$ of these Haar senses (due to anticipated signal duration).

Example 1 Revisited

$$g_j(t) = \sum_{k=1}^{K_j} a_{j,k} \sin(\omega_{j,k}t - \phi_{j,k}) + n_j(t)$$

For the two class, two-stage version of our example, analogous to our simplified theoretical scenario, we use $K_0 = K_1 = 3$ with

$$a_{0,\cdot} = a_{1,\cdot} = [r_a r_a r_a]',$$

$$\phi_{0,\cdot} = \phi_{1,\cdot} = [r_\phi r_\phi r_\phi]'$$

— the random amplitudes r_a and random phases r_ϕ are ancillary to the detection/classification task — and

$\omega_{0,\cdot} = [X|Y = 0]'$, $\omega_{1,\cdot} = [X|Y = 1]'$. The random vector

$X = [X_1 X_2 X_3]'$ is specified via its class-conditional distributions

$X|Y = 0$ and $X|Y = 1$ in analogy with the simplified theoretical scenario presented above with appropriate signal frequency pairs in the role of the binary range of the Bernoullis.

(The prior probabilities of class membership, $P[Y = j]$, are given to be $P[Y = 0] = P[Y = 1] = 1/2$ for this example.)

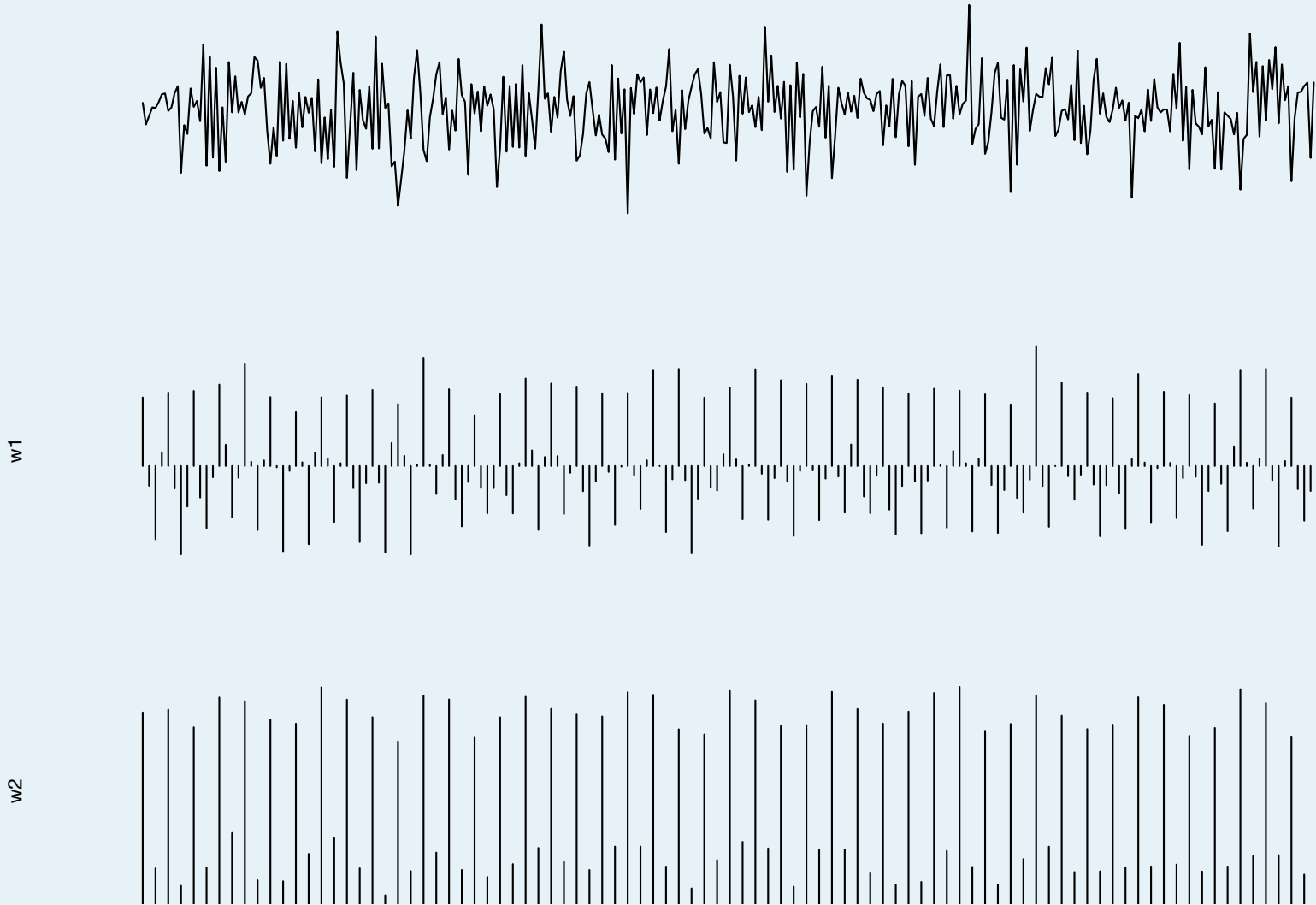
Example 1 Revisited

For signals sampled at rate 2^{sigp} — 2^{sigp} samples in $[0, 2\pi]$ — two features are extracted at each stage of the iterative denoising tree via the two-level Haar with downsampling rate 2^{dsp} . The two features obtained are (f_1, f_2) , where feature f_ℓ detects frequency $2^{sigp-dsp-\ell}$. In our example, with $sigp = 14$, we have time for just two feature extraction stages (due to anticipated signal duration). The features extracted at the first stage of the tree, with $dsp = 0$, are (f_1, f_2) with f_1 detecting frequency 2^{13} and f_2 detecting frequency 2^{12} . These two features — in and of themselves — are of no value in the classification task; they are ancillary! However, conditional upon the observed value of these first two features, the second two-level Haar, performed with either $dsp = 2$ or $dsp = 4$ and detecting frequencies $(2^{11}, 2^{10})$ or $(2^9, 2^8)$, yields perfect detection/classification performance.

Example 1 Revisited

For signals sampled at rate 2^{sigp} — 2^{sigp} samples in $[0, 2\pi]$ — two features are extracted at each stage of the iterative denoising tree via the two-level Haar with downsampling rate 2^{dsp} . The two features obtained are (f_1, f_2) , where feature f_ℓ detects frequency $2^{sigp-dsp-\ell}$. In our example, with $sigp = 14$, we have time for just two feature extraction stages (due to anticipated signal duration). **The features extracted at the first stage of the tree**, with $dsp = 0$, **are** (f_1, f_2) with f_1 detecting frequency 2^{13} and f_2 detecting frequency 2^{12} . These two features — in and of themselves — are of no value in the classification task; they are ancillary! However, conditional upon the observed value of these first two features, the second two-level Haar, performed with either $dsp = 2$ or $dsp = 4$ and detecting frequencies $(2^{11}, 2^{10})$ or $(2^9, 2^8)$, yields perfect detection/classification performance.

Example 1 Revisited



Example 1 Revisited

For signals sampled at rate 2^{sigp} — 2^{sigp} samples in $[0, 2\pi]$ — two features are extracted at each stage of the iterative denoising tree via the two-level Haar with downsampling rate 2^{dsp} . The two features obtained are (f_1, f_2) , where feature f_ℓ detects frequency $2^{sigp-dsp-\ell}$. In our example, with $sigp = 14$, we have time for just two feature extraction stages (due to anticipated signal duration). **The features extracted at the first stage of the tree**, with $dsp = 0$, **are (f_1, f_2)** with f_1 detecting frequency 2^{13} and f_2 detecting frequency 2^{12} . **These two features — in and of themselves — are of no value in the classification task; they are ancillary!** However, **conditional upon the observed value of these first two features, the second two-level Haar**, performed with either $dsp = 2$ or $dsp = 4$ and detecting frequencies $(2^{11}, 2^{10})$ or $(2^9, 2^8)$, **yields perfect detection/classification performance.**

Example 1 Revisited

The first sense uses $p_{ds} = 0$ to determine ω_0 .

The second sense is conditional on first sense:

if $f_{01} > t_{01}$

then $\omega_0 = \omega_{01}$ and second sense uses $p_{ds} = 2$ to determine ω_1 ;

if $f_{02} > t_{02}$

then $\omega_0 = \omega_{02}$ and second sense uses $p_{ds} = 4$ to determine ω_2 .

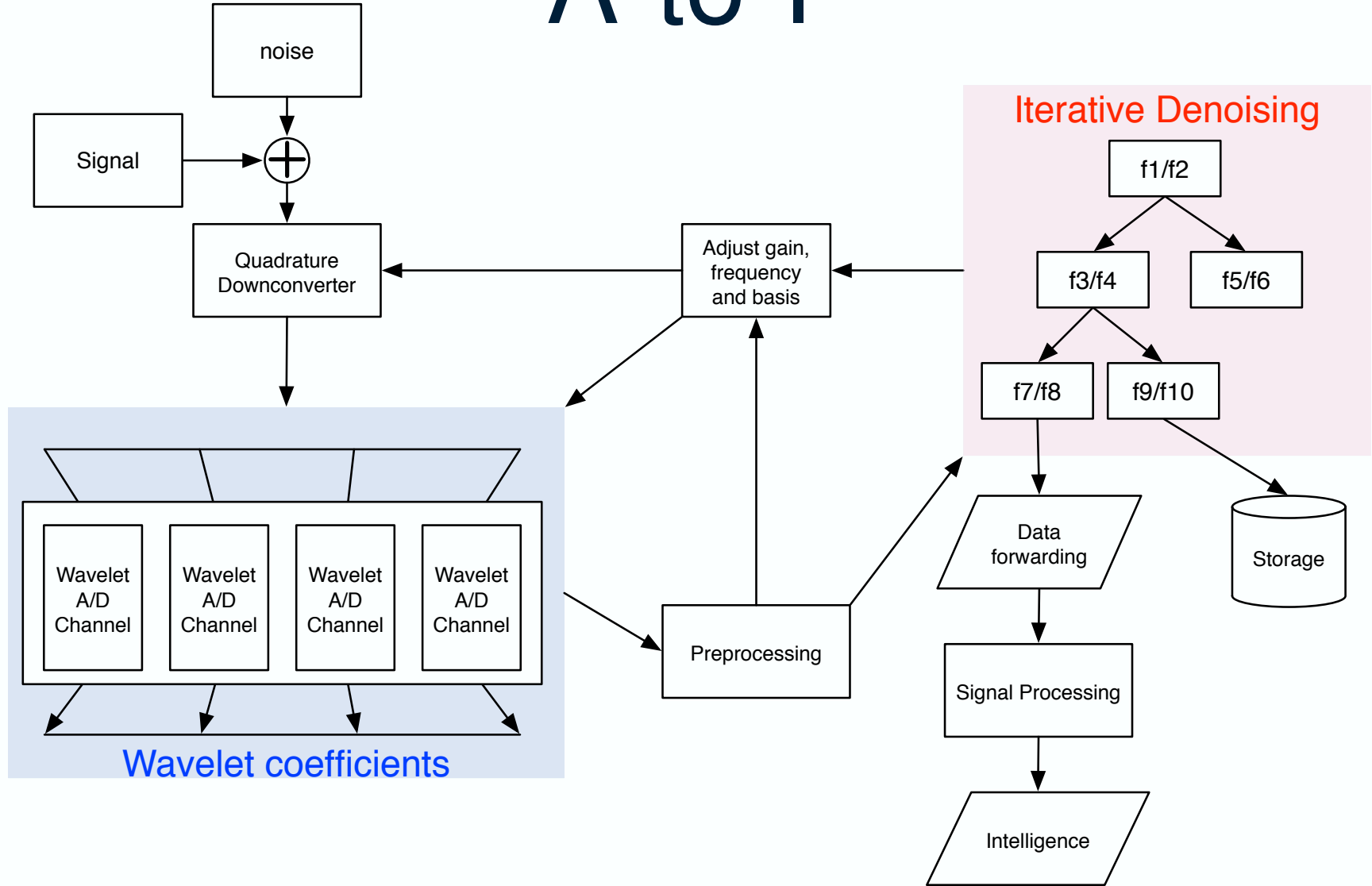
This scheme yields $L = 0$ (when there is no noise).

No unconditional scheme

with just two two-level Haar feature extraction stages

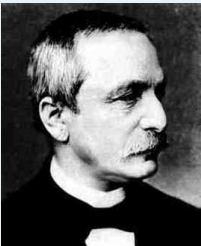
can yield $L < 0.25$.

A-to-I



Kronecker Quote

*“The wealth of your practical experience
with sane and interesting problems
will give to mathematics
a new direction and a new impetus.”*



– Leopold Kronecker to Hermann von Helmholtz –

