Mathematical Methods in Machine Learning

Wojciech Czaja

UMD, Spring 2016

Outline

Norbert Wiener Center for Harmonic Analysis and Applications

Introduction

- There is an abundance of available data. This data is often large, high-dimensional, noisy, and complex, e.g., geospatial imagery.
- Typical problems associated with such data are to cluster, classify, or segment it; and to detect anomalies or embedded targets. Regression and dimensionality reduction are other types of typical examples of problems that we want to deal with.
- Our proposed approach to deal with these problems is by combining techniques from harmonic analysis and machine learning:
  - **Harmonic Analysis** is the branch of mathematics that studies the representation of functions and signals.
  - **Machine Learning** is the branch of computer science concerned with algorithms that allow machines to infer rules from data.

Norbert Wiener Center
for Harmonic Analysis and Applications

Machine Learning

Machine learning has many different faces. We are interested in these aspects of machine learning which are related to representation theory. However, machine learning has been combined with other areas of mathematics.

- Statistical machine learning.
- Topological machine learning.
- Computer science.

Machine Learning

Another way to classify machine learning is by the type of tasks it deals with, depending on the nature of the learning (training or feedback) available to a learning system.

- **Supervised learning:** The computer is presented with example inputs and their desired outputs, given by a ”teacher”, and the goal is to learn a general rule that maps inputs to outputs.
- **Semisupervised learning:** Between supervised and unsupervised learning is semi-supervised learning, where the teacher gives an incomplete training signal: a training set with some (often many) of the target outputs missing.
- **Unsupervised learning:** No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).
- **Reinforcement learning:** An area of machine learning inspired by behaviorist psychology, concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

Motivation for Machine Learning - Big Data

"Big Data" refers to the exponential growth data, with many challenging tasks of analyzing and efficiently finding the important information that is given in this complex setting.

- The roots of big data are in the data storage, database management, and data analytics for, both, commercial and non-profit applications.
- The integration of many large datasets is a primary source of big data problems present in the modern scientific and research environment, as is evident in applications ranging from 'omics' data analysis for cancer research, to studies of social networks.
- Another source of big data problems are large and heterogeneous dynamic data sets, such as those arising in the context of climate change analysis, or for the analysis of network traffic patterns.

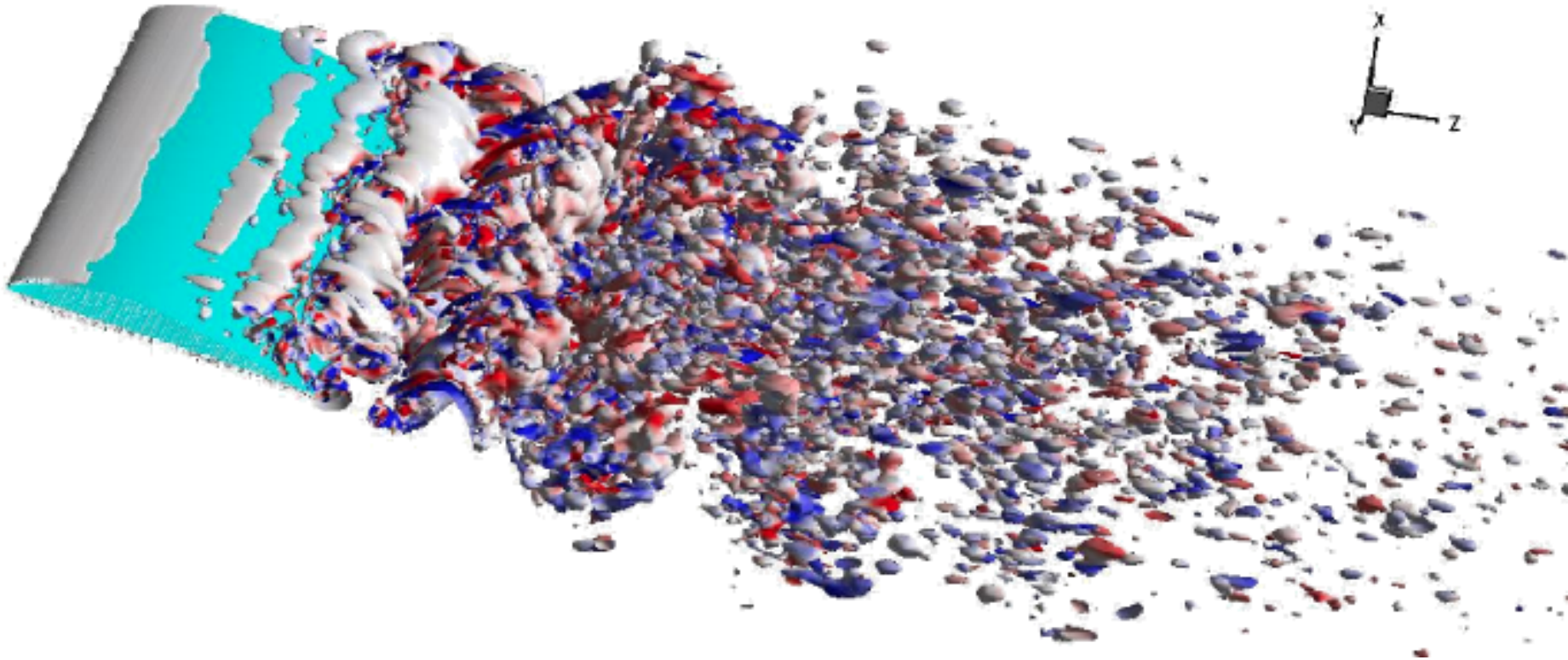
Big Data Characteristics

In view of the above, big data can be identified by the following:

- volume;
- heterogeneity;
- dynamics.

In addition to the above major characteristics, we can add: ambiguity, complexity, noise, variability, etc.

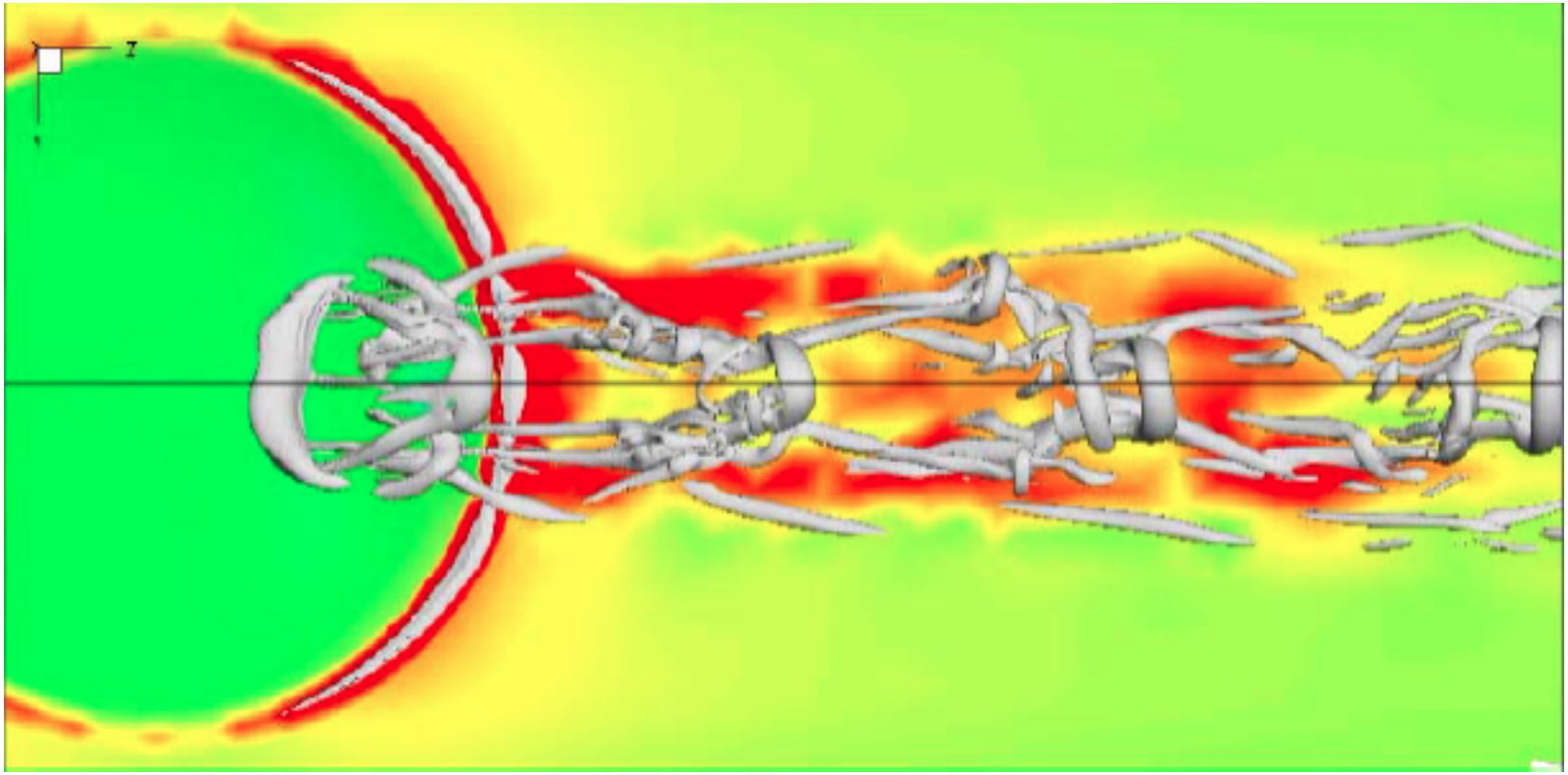
Big Data Example 1



Large Eddy simulation (LES) around an Eppler foil at Re=10,000. A series of high fieldity LES of the flow around Eppler airfoils has been conducted to generate a comprehensive data base. Reynolds numbers vary from 10,000 to 120,000 and the angle of attach varies from 0 to 20 degrees.

Courtesy of Prof. Elias Balaras (GWU), via US Air Force contract FA9550-12-C-058 (2012): Learning from Massive Data Sets Generated by Physics Based Simulations

Big Data Example 2



A simplified case of the previous LES for a 3-dimensional flow over a dimpled plate.

Courtesy of Prof. Elias Balaras (GWU), via US Air Force contract FA9550-12-C-058 (2012): Learning from Massive Data Sets Generated by Physics Based Simulations

Big Data Example Estimation

Let us provide a small numerical estimation:

- 2,000 x 1,000 x 1,000 = 2 x 10^9 grid points;
- Each grid point characterized by 3 spatial coordinates and 3 velocity components, pressure, plus possibly some other parameters;
- Flow simulation for 200 time steps;
- One way to look at it: 2 x 10^9 points in a space of dimension 1,400;
- As an example, think of computing PCA for M points in N dimensional space. The cost is $O(MN^2) + O(N^3)$;
- In our case this results in a problem with complexity on the order of 4 x 10^{15} = 4 petaFLOPs;
- Lawrence Livermore National Laboratory's IBM Sequoia reaches 16 petaFLOPS (16 x 10^{15} floating point operations per second) - it was considered to be the fastest computer in 2012, it runs 1.57 million PowerPC cores, costs approx. 250M USD.

Another Big Data Example

- Consider the human genome. First estimates pointed at 100,000 genes. Nowadays this number has been scaled down to appprox. 45,000.
- There are many ways of representing genes. One of the more popular is by means of base pairs: approx. three billion DNA base pairs represent human genome.
- Alternatively, we could consider gene expressions (think of it as a function). There are many ways of assembling such expressions, and they are different for different individuals. Hence resulting in a much larger data set.

HA and Big Data to-date

Harmonic analysis ideas have been used in many problems dealing with large and heterogeneous data. Some relevant examples include:

- Multiscale methods
- Compressive sensing
- Sparse representations
- Geometric and graph-based methods
- Scattering transforms

Among those listed, multiscale methods are historically the oldest (though not old) class of approaches. They have been successfully used in image compression applications. We can view the JPEG 2000 as a prototypical “dimension reduction” attempt for a large data class.

Multiscale representations

- **Multiscale representation** (Multiresolution analysis (MRA), pyramid algorithms) can be described as a class of design methods in representation theory, where the input is subject to repeated transformation (filtering) in order to extract features associated with different scales.
- In image processing and computer graphics the concept of multiscale representations can be traced back to P. Burt and E. Adelson, and J. Crowley.
- In mathematics, it is associated with **wavelet** theory and MRA as introduced by Y. Meyer and S. Mallat.

S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", IEEE TPAMI, 1989, Vol. 11, pp. 674–693.

- Multiscale representations found many applications to image processing and remote sensing: compression, feature detection, segmentation, classification, but also in registration and image fusion.

G. Pajares and J. Cruz, "A wavelet-based image fusion tutorial", Pattern Recognition, 2004, Vol. 37(9), pp. 1855–1872.

Filters

- In signal processing, a **filter** is typically understood as a device or a process that removes from a given signal an unwanted component or feature.
- Originally, **electronic filters** were entirely analog and passive and consisted of resistance, inductance and capacitance. Nowadays, **digital filters** are much more common. They operate on signals represented in digital form. The essence of a digital filter is that it directly implements a mathematical algorithm, corresponding to the desired filter transfer function.
- In practice, a digital filter system often contains an analog-to-digital and a digital-to-analog converter together with a microprocessor and some peripheral components (such as memory to store data). In this talk we shall consider digital filtering as a signal transform, i.e., a mathematical procedure.

Filter characteristics

Digital filters can be discrete or continuous.
Digital filters may be linear or nonlinear.
Digital filters may be time-independent or may depend on time.
Digital filters may depend of the Fourier transform, the Laplace transform, a state-space representation, or any other representation system.

- The filter should have a specific impulse response.
- The filter should be causal.
- The filter should be stable.
- The computational complexity of the filter should be low.
- The filter should be hardware or software implementable.

Example of a filter design

Let $\{x(n), n \in \mathbb{Z}\}$ denote the input signal and let $\{y(n), n \in \mathbb{Z}\}$ denote the output. A filter F is a transformation:

$$F : x \mapsto y.$$

If we assume that the principle of superposition holds, i.e., that the filter is linear, then combining any two inputs x_1 and x_2 (with individual outputs y_1 and y_2, resp.) as $\alpha x_1 + \beta x_2$, results in an output of the form:

$$F : \alpha x_1 + \beta x_2 \mapsto \alpha y_1 + \beta y_2.$$

If, in addition, we assume that our filter is time-independent, then the behavior of the filter does not change with time, i.e., a delayed version of any input $x_d(n) = x(n - d)$, results in an output with a corresponding delay $y_d(n) = y(n - d)$:

$$F : x_d \mapsto y_d.$$

Example of a filter design, cont'd

Let δ denote the unit impulse at the origin ($\delta(0) = 1$ and $\delta(n) = 0$ for $n \neq 0$). Let h denote the response of δ ($F(\delta) = h$).
Under the above assumptions, we can now assert that the output of a general input signal:

$$x(n) = \sum_{k \in \mathbb{Z}} x(k)\delta(n-k)$$

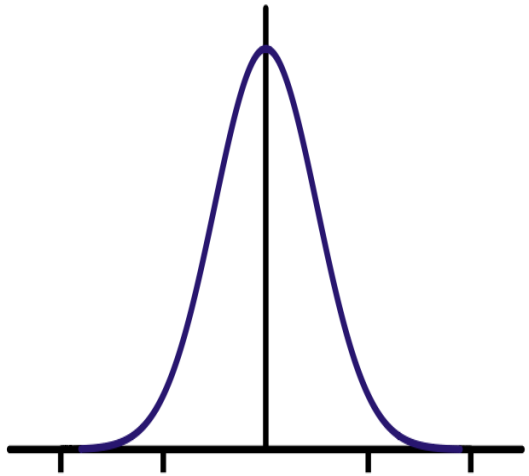
takes the form of:

$$F(x) = \sum_{k \in \mathbb{Z}} x(k)h(n-k) = x \star h(n).$$

This is a **convolution**.

Example: Gaussian filter

Gaussian filter is a filter whose impulse response is a Gaussian function, or an approximation to it.
Mathematically, a Gaussian filter modifies the input signal by convolution with a Gaussian function; this transformation is also known as the Weierstrass transform.



Source of imagery: Wikipedia

FIlter Banks

A **filter bank** is an array (collection) of band-pass filters that splits the input signal into multiple components, each one carrying a single frequency sub-band of the original signal.
A complete filter bank consist of the analysis and synthesis side. The analysis filter bank divides an input signal to different subbands with different frequency spectrums. The synthesis part reassembles the different subband signals and generates a reconstruction signal.

$$F : x \mapsto H_1(x), \ldots, H_n(x) \mapsto G_1(H_1(x)), \ldots G_n(H_n(x)) = F(x)$$

In filter bank design one often makes use of properties of decimation (downsampling) and interpolation (expansion).
The filter bank has **perfect reconstruction** if $F(x) = x$ for all input signals x. Equivalently, imperfect reconstruction means that the synthesis bank is the left inverse of the analysis bank, $GH = Id$.

Orthogonality and Conjugate Quadrature FIlters

- Filter $F = (G, H)$ is **orthogonal** if the transformation it generates is orthogonal, i.e., $FF^T = F^T F = Id$.
- A finitely supported filter F is a **Conjugate Quadrature Filter** is a filter that satisfies for every $m \in \mathbb{Z}$

$$2 \sum_{n \in \mathbb{Z}} F_n \overline{F_{n+2m}} = \delta(m).$$

- Orthogonal Conjugate Quadrature FIlters are, in mathematical nomenclature, **MRA wavelets**.

Norbert Wiener Center
for Harmonic Analysis and Applications

Wavelets

We say that a function $\psi \in L^2(\mathbb{R})$ is an **orthonormal wavelet** if it can be used to define a basis, that is a complete orthonormal system, for the Hilbert space $L^2(\mathbb{R})$, of the form

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k),$$

where $j, k \in \mathbb{Z}$. We call these operations **dyadic dilations and translations**.
Wavelet transform is an operation of convolving input signals with the elements of the wavelet basis.
Wavelet transforms can be discrete or continuous. We shall focus on the latter one.

Discrete wavelet transform

- The first DWT was discovered by Hungarian mathematician Alfréd Haar in 1909. We now know it as the Haar wavelet ψ s.t.:

$$\psi(x) = \begin{cases} 1 & x \in [0, 0.5) \\ -1 & x \in [0.5, 1) \\ 0 & \textit{otherwise} \end{cases}$$

Alfréd Haar, “Zur Theorie der orthogonalen Funktionensysteme”:

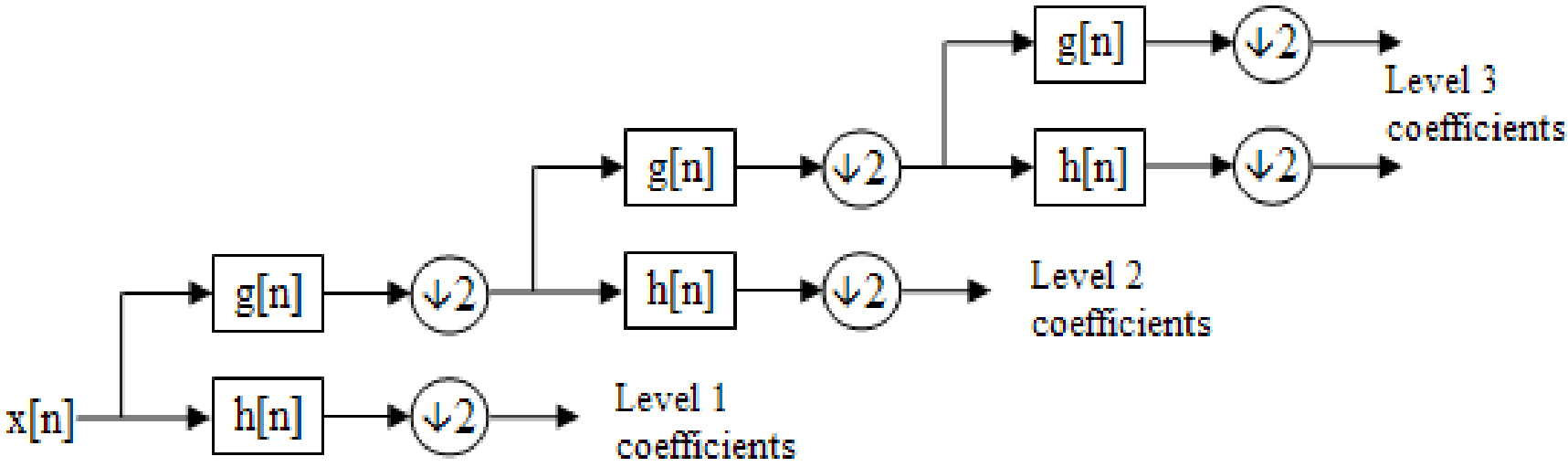
Ph.D. Thesis at Georg-August-Universitaet Goettingen 1909; published in Mathematische Annalen 69 (3), pp. 331–371.

- The concept of wavelets (derived from a French word ondelette, meaning ”small wave”) was introduced by Morlet and Grossmann in the early 1980s. The theory was then developed by Y. Meyer.
- The most commonly used set of discrete wavelet transforms was formulated by the Belgian mathematician Ingrid Daubechies in 1988.

I. Daubechies, “Orthonormal bases of compactly supported wavelets”, Comm. Purr Appl. Math.. vol. 41 (1988), pp. 909–996.

DWT as a filter bank

Example of the analysis stage of 1D DWT up to level 3 decomposition with low-pass filter (g) and high-pass filter (h). The synthesis stage is symmetric and is automatically derived from the OCQF conditions.

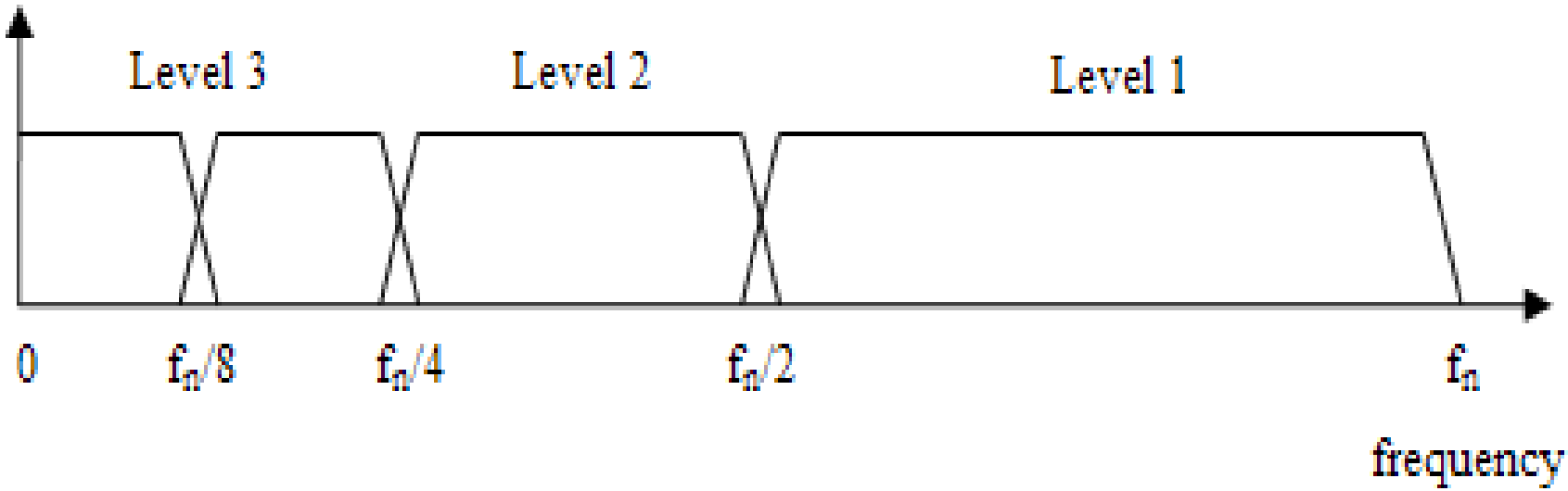


Source of imagery: Wikipedia

Norbert Wiener Center for Harmonic Analysis and Applications

Advantages of wavelets

The major advantage of wavelets over Fourier techniques in general is that wavelets are localized in both time and frequency whereas the standard Fourier transform is only localized in frequency.
The following is an illustration of the frequency domain decomposition corresponding to the above DWT.



Source of imagery: Wikipedia

Norbert Wiener Center for Harmonic Analysis and Applications

Limitations of traditional wavelet representations

- Wavelets provide optimal representations for 1-dimensional signals in the sense of measuring asymptotic error with N largest coefficients in wavelet expansion, and are superior to Fourier-type representations.
- However, in dimensions higher than 1, wavelets are known to be suboptimal for representing objects with curvilinear singularities (edges), even though they outperform Fourier methods.

  D. Donoho et al., "Data compression and harmonic analysis", IEEE TIT, 1998, Vol. 44, pp. 2435–2476.
- A number of techniques have been proposed since the introduction of wavelets to address this issue, and to find better description of geometric features in images.

  L. Jacques et al., "A panorama on multiscale geometric representations, intertwining spatial, directional and frequency selectivity", Signal Processing, 2011, Vol. 91, pp. 2699–2730.

Sumary

- We have generally described the area of machine learning that will be of interest to us in this lecture.
- We have motivated the need for machine learning using the concept of Big Data.
- We have given a brief overview of traditional multiscale/wavelet techniques used for data compression.