The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964-2014

2585-29

Joint ICTP-TWAS School on Coherent State Transforms, Time-Frequency and Time-Scale Analysis, Applications

2 - 20 June 2014

Sparsity for big data

C. De Mol
ULB, Brussels
Belgium

Sparsity for Big Data

Part I - Mathematics to deal with the Data Deluge

Christine De Mol

Université Libre de Bruxelles
Dept Math. and ECARES

Joint ICTP-TWAS School on Coherent State Transforms, Time-Frequency and Time-Scale Analysis, Applications
June 2014

A tsunami of data

A tsunami of data

Finding a needle in a haystack

Quantifying the amount of data

- Unit: 1 bit (b) (0 ou 1)
- 1 byte (B) = 8 bits

Assume everything is encoded in binary form, namely by zeros and ones

Quantifying the data deluge

- 1 kB = 1 kilobyte = 1000 bytes (other definition: 1024)
- 1 MB = 1 megabyte = 1000 kB = 10^6 Byte
- 1 GB = 1 gigabyte = 1000 MB = 10^9 Byte
- 1 TB = 1 terabyte = 1000 GB = 10^{12} Byte
- 1 PB = 1 petabyte = 1000 TB = 10^{15} Byte
- The next prefixes: exa (10^{18}), zetta (10^{21}), yotta (10^{24})

Examples: the petabyte era

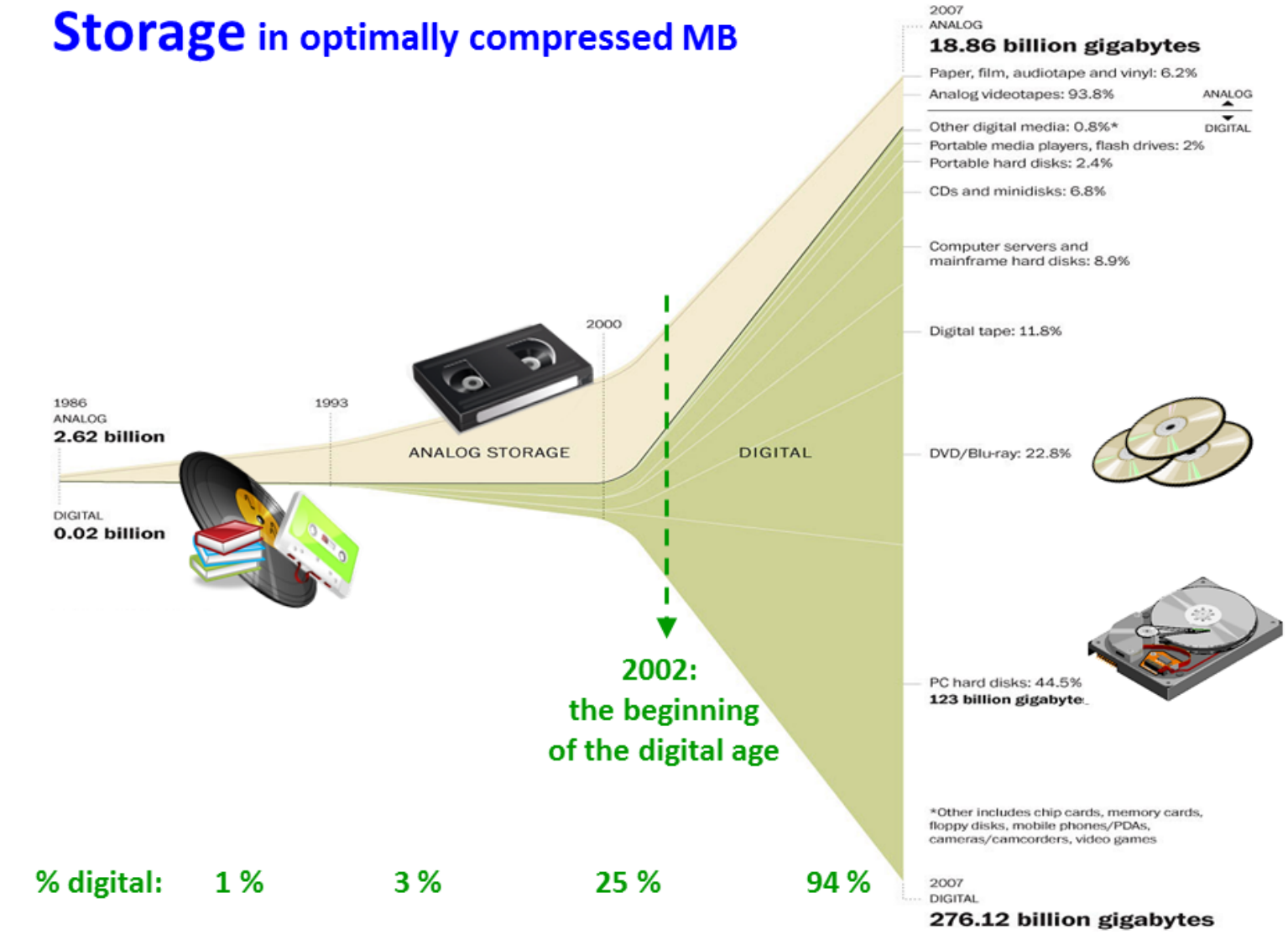
- bytes : 1 byte $\sim$ 1 letter (ascii symbol)
- kilobytes : 1 kB $\sim$ 1 page,
  1 article in pdf (50-500 kB) , 1 small image
- megabytes : 1 book
- gigabytes : 1 Audio CD (700 MB), 1 DVD (5 GB),
  a private library
- terabytes : a public library
  LOC (20 TB) (digital content of U.S. Library of Congress)
- petabytes : amount of data treated by the servers of
  Google in one hour (1 PB)

Growth

- 90 % of the recorded data have been collected during the last two years!!!
- Most data are now digital (numbers)
(1 % en 1986, 25 % en 2000, 94% en 2007)
- In 2007, $\sim$ 300 exabytes of data stored
(61 CD-ROM per person, i.e. a stack which would go beyond the moon!)
- et $\sim$ 2 zettabytes of data exchanged!

(M. Hilbert, P. López, Science 2011)

Storage capacity



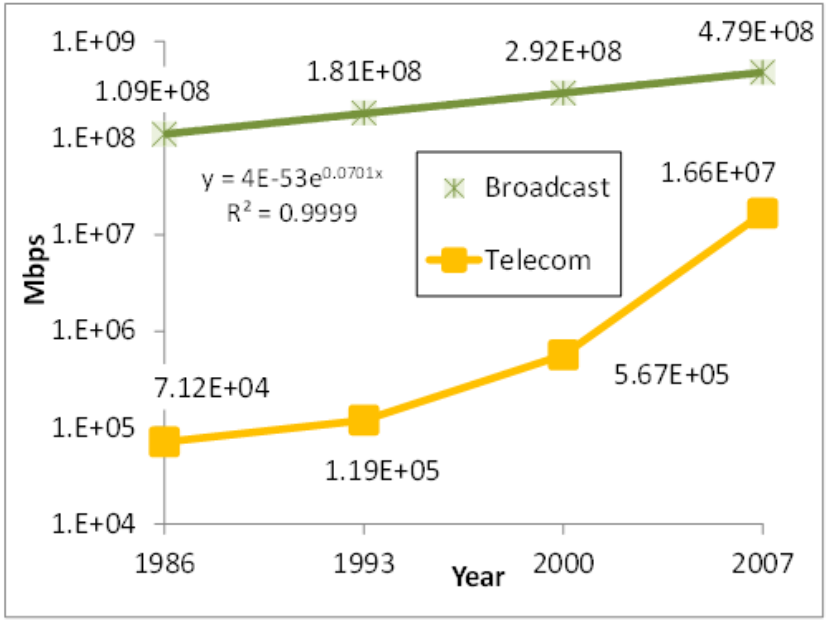
Source: Washington Post, based on Hilbert and Lopez, 2011

Data flux

Communication in optimally compressed MB

One-way broadcast

Single message for all



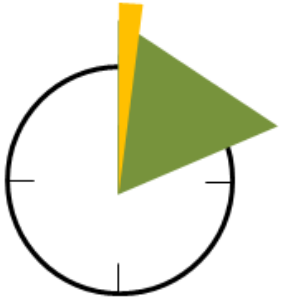
Two-way telecommunication

Individual messages

Broadcast / telecom = 1,550 / 1

TV ≈ 2.5 hours/day

Phone ≈ 14 minutes/day

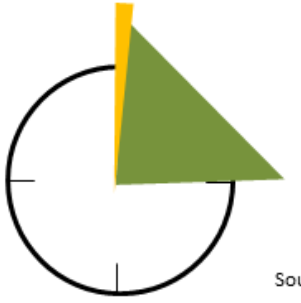


===>

30 /1

TV ≈ 3 hours/day

Internet ≈ 9 minutes/day

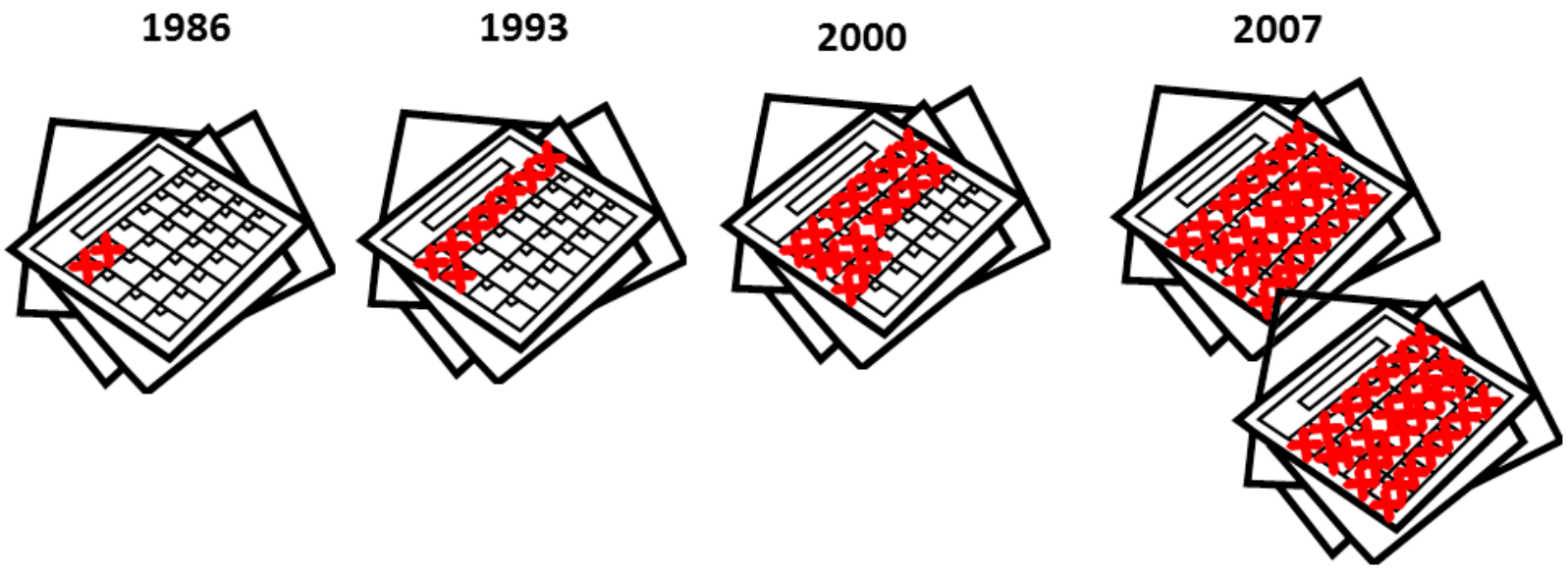


Source: Hilbert and Lopez, 2011

Flux/Storage

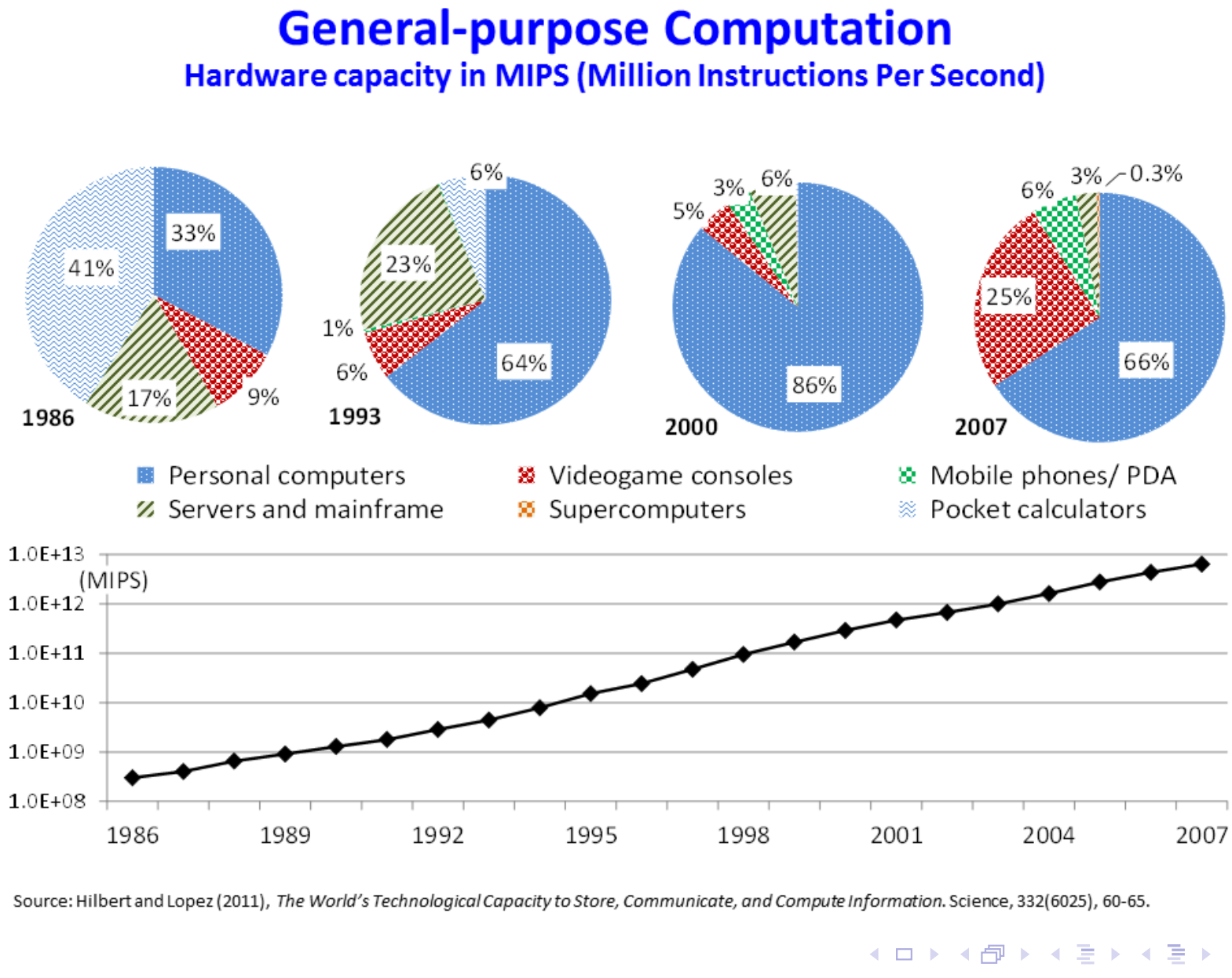
Communication vs. Storage

How quickly would our communication capacity (broadcast + telecom) fill up our available information storage?



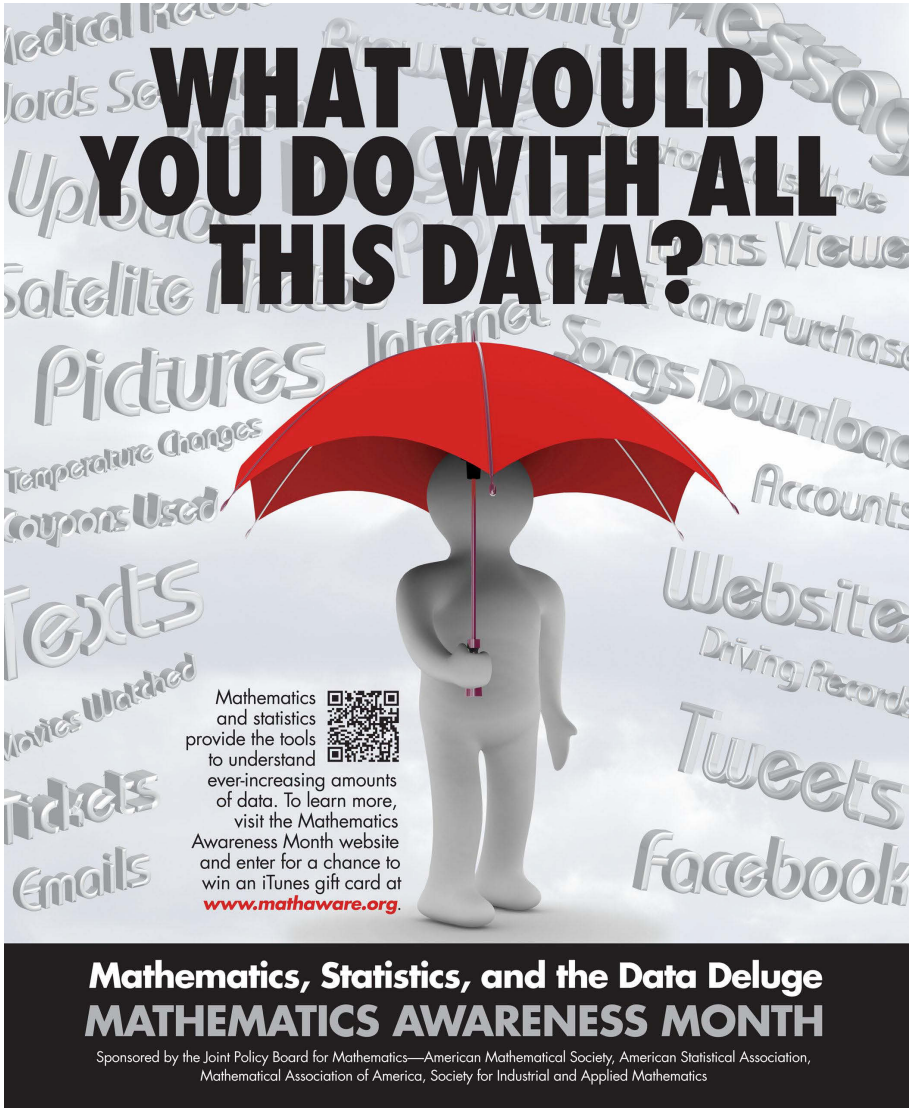
Source: Hilbert and Lopez (2011). Science, 332(6025), 60-65.

Computing Power



Source: Hilbert and Lopez (2011), *The World's Technological Capacity to Store, Communicate, and Compute Information.* Science, 332(6025), 60-65.

Mathematics Awareness Month April 2012



Intelligent design to exploit them!

Big challenges ...

for

- mathematicians
- statisticians
- computer scientists, engineers, etc.

in order to develop automatic procedures for extracting useful information from huge amounts of data.

$\rightarrow$ rapid development of (new) research fields:

- Computer vision
- Data Mining
- Statistical Learning (“Machine Learning”)
- Bioinformatics, etc.

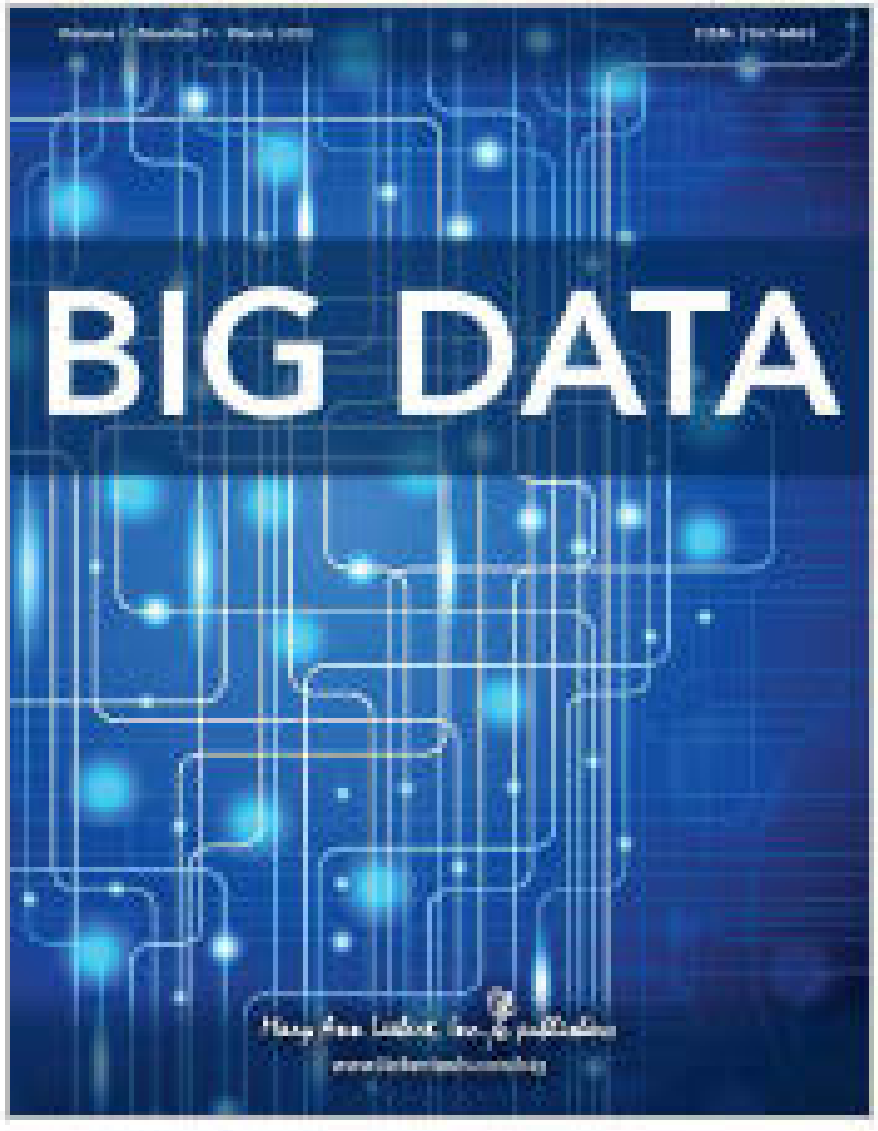
... in all scientific areas

- Physics
  The LHC (Large Hadron Collider) (CMS) at CERN collects per seconde 1 Petabyte of raw data (600 million collisions at 1 MB each), reduced to $\sim$ 100 000 by the “trigger”, then to $\sim$ 100 per second by the “grid” ($\rightarrow$ 15 PB of data stored per year)
- Astronomy
  The “Sloane Digital Sky Survey” (SDSS), with 200 GB per night, has recorded in a few weeks more data than all astronomical observations in history (already 71 TB)!
- Geophysics
  Arrays of seismographs permanently monitoring the Earth. For example, the “US Array” records per day 4.9 GB (already for a total of 12 TB)

... in all scientific areas

- Biology (chiefly genomics, proteomics)
  Your own genome as a birthday present?
- Economics and finance
  (“tick-by-tick data”)
- Social sciences
  Studies, “Data Journalism”
- Politics
  More than 100 statisticians, mathematicians, engineers, etc. worked for Obama’s campaign

A new scientific journal



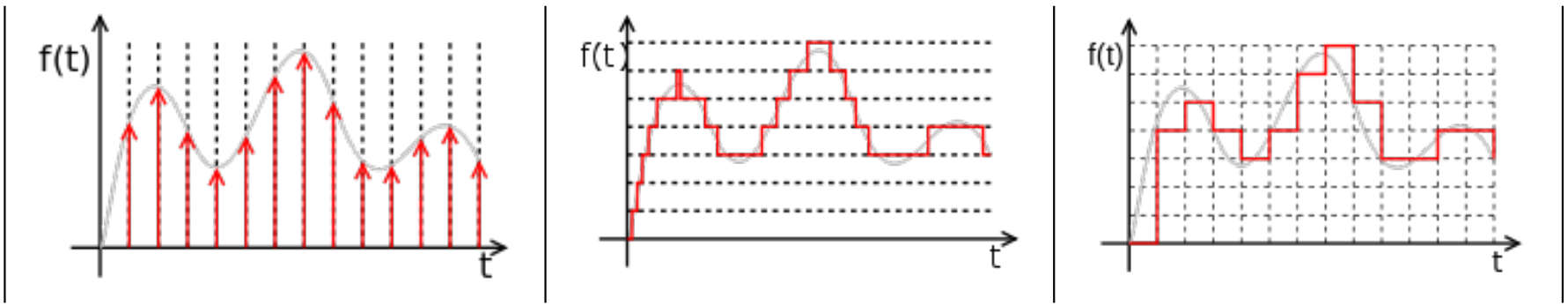
Some mathematical tools for Big Data

- Digitization (sampling and quantization)
- Compression
- Sparsity
- Compressed sensing
- Dimension Reduction
- Algorithms

Digitization

an analog signal, e.g. sound or speech

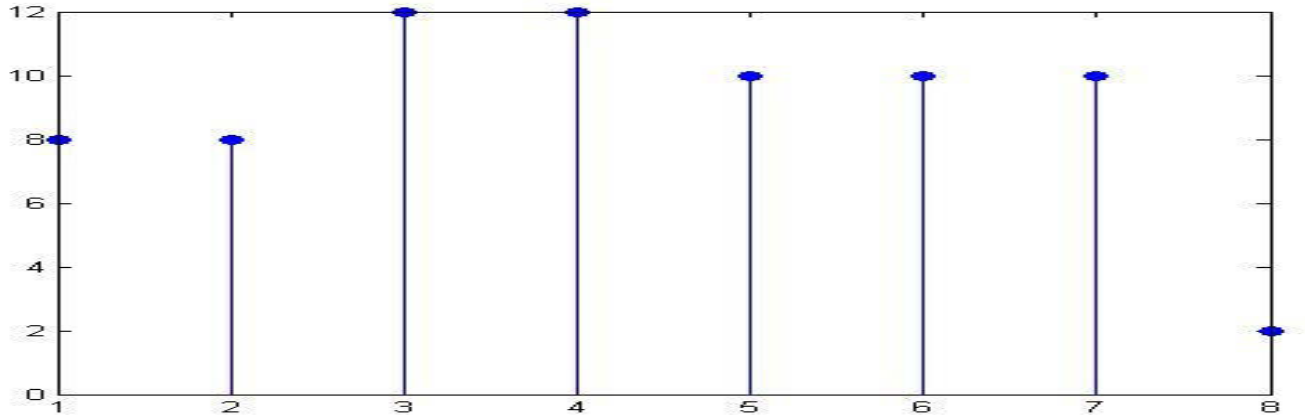
• Sampling: recording the values of a signal (samples) in discrete points, in general equidistant and separated - at most - by the Nyquist distance $1/2\nu_{max}$, where ν_{max} is the maximal frequency in the signal) (Shannon's theorem)



• Quantizing: expressing the values of a signal as integer multiples of some finite quantity (e.g. 256 grey levels in an image) – and transforming them in fine into binary numbers

$\rightarrow$ interesting mathematical problems

Data compression



Coding by sums and differences:

a et b = values of 2 successive samples

$$c = \frac{a+b}{2} \qquad d = \frac{a-b}{2}$$

One can always recover: $a = c + d$ et $b = c - d$

8 8 12 12 10 10 10 2

8 12 10 6

0 0 0 4

10 8

-2 2

9

1

9 0 -2 0 1 0 2 4

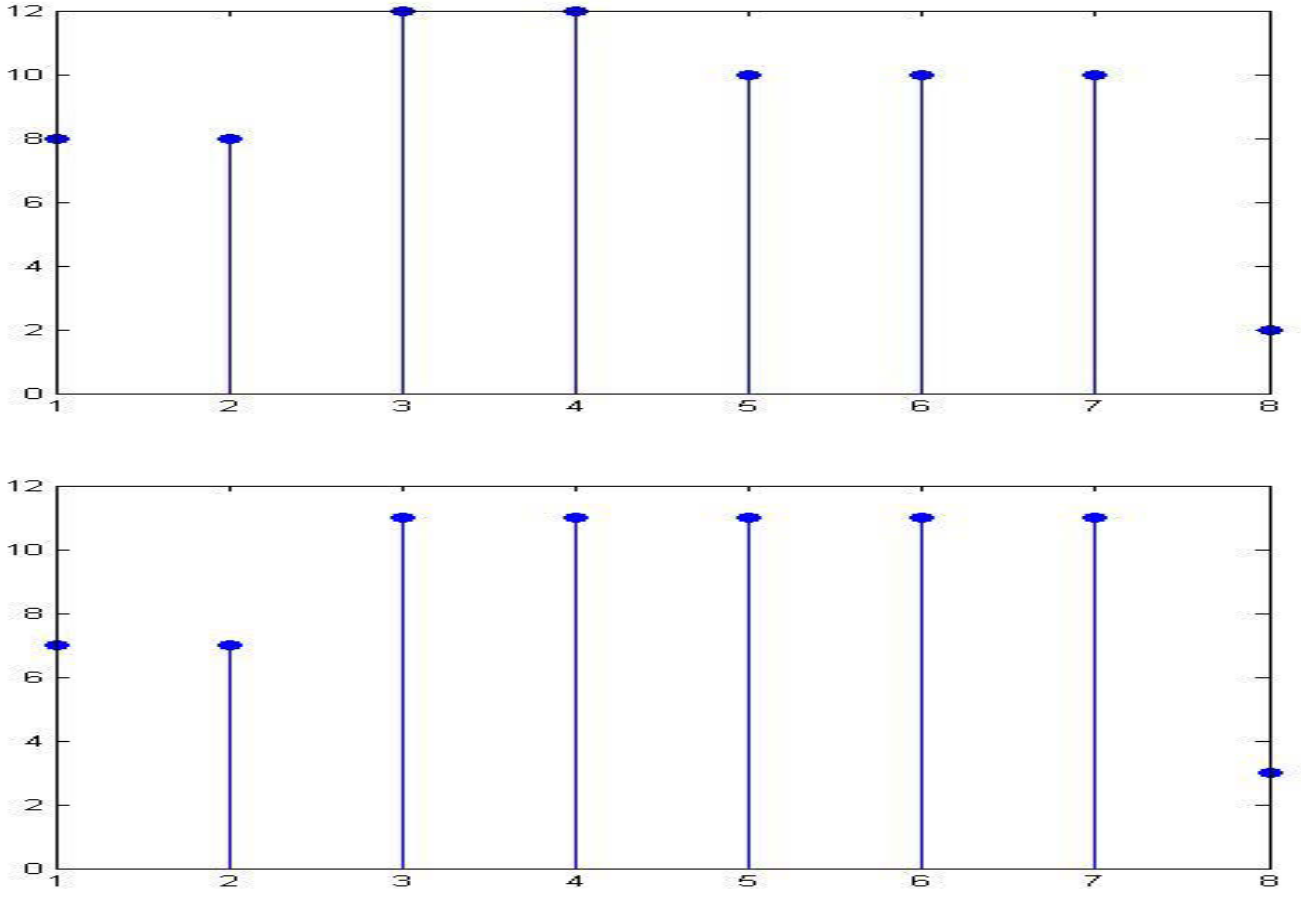
Data compression

- Note that in the new coding many coefficients vanish
$\rightarrow$ no need to store/transmit those values

- One can always recover the original signal
("lossless" compression)

- One can also set to zero the smallest coefficients
(thresholding) and reconstruct an approximation of the original signal
("lossy" compression)

Data compression



reconstructed signal: (1) lossless compression (rate 5/8)
(2) lossy compression (rate 4/8)

Data compression

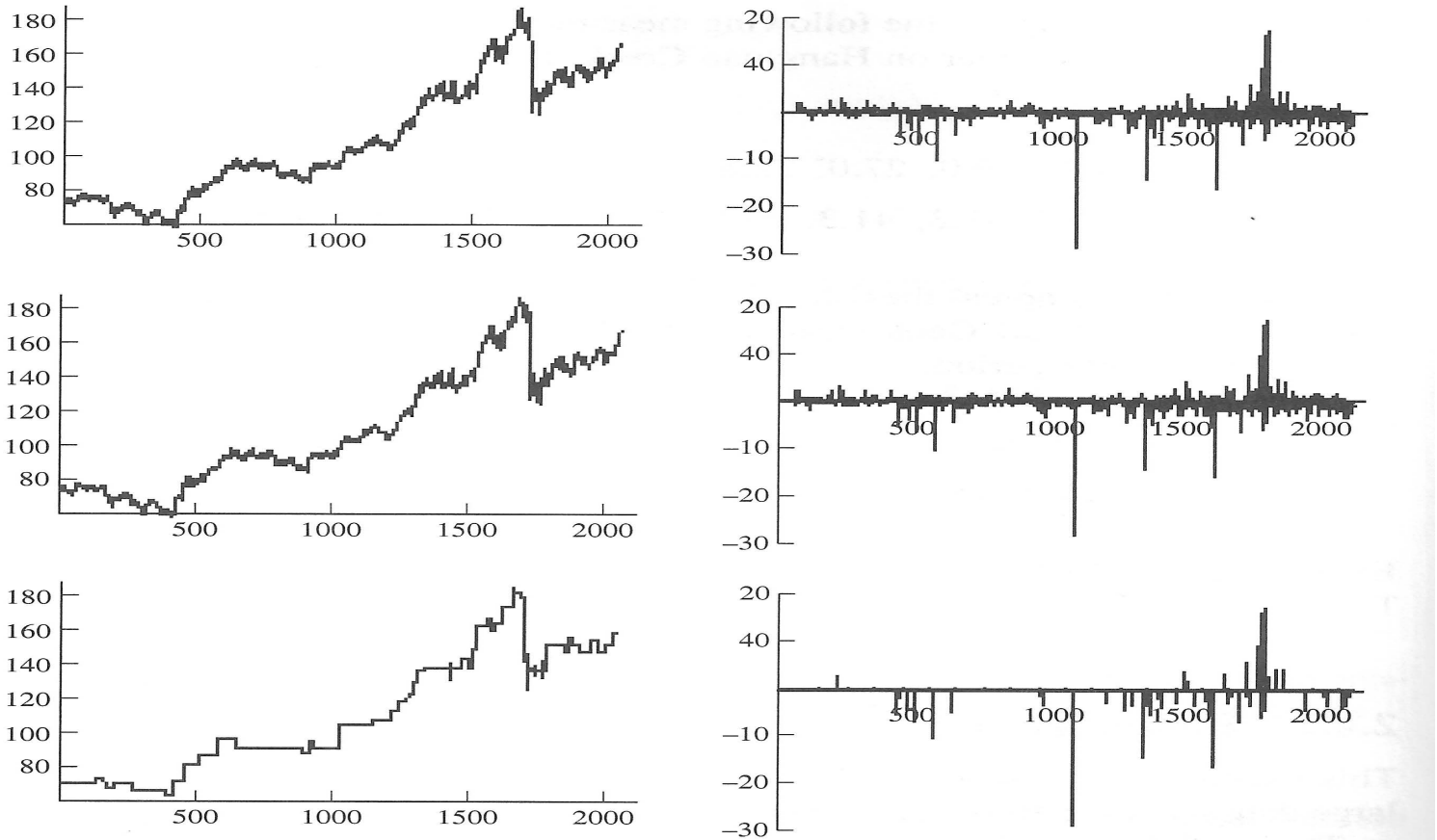


Figure 2.13 New York Stock Exchange Composite Index for 1981–1987. *Top left.* Data:

Example of compression (in a Haar wavelet basis):
original (top); 56% (middle); 2 % (bottom)
(from: Y. Nievergelt, “Wavelets Made Easy”)

Data compression

- A generalization of such coding (in a basis of Haar wavelets) has been devised by Ingrid Daubechies (Daubechies' wavelets), in which the sums and differences are more refined and include more points

- This can be applied to digital images (pixels), line by line and column by column

- These ideas gave rise to a new image compression standard (JPEG 2000) implemented nowadays e.g. in digital cinema and for medical images

Sparsity

“Entia non sunt multiplicanda sine necessitate”

William of Ockham ($\sim$ 1288 - 1348)

Sparse representations

The digital signal s (vector) can be expressed as a linear combination of elements of a dictionary (basis vectors) $\{\phi_1, \phi_2, \ldots, \phi_p\}$: $s = \sum_{j=1}^{j=p} a_j \, \phi_j$ or equivalently

$$\begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix} = a_1 \begin{pmatrix} \phi_{11} \\ \phi_{12} \\ \vdots \\ \phi_{1n} \end{pmatrix} + a_2 \begin{pmatrix} \phi_{21} \\ \phi_{22} \\ \vdots \\ \phi_{2n} \end{pmatrix} + \cdots + a_p \begin{pmatrix} \phi_{p1} \\ \phi_{p2} \\ \vdots \\ \phi_{pn} \end{pmatrix}$$

with few nonzero coefficients a_j (without knowing in advance which ones)

Special case: orthogonal basis of $p = n$ elements
(but one can also have $p > n$ – redundant or overcomplete basis)

Sparse representations

In matrix form: $\boxed{s = \Phi a}$ where Φ is a $n \times p$ matrix, i.e.

$$\begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{21} & \cdots & \cdots & \phi_{p1} \\ \phi_{21} & \phi_{22} & \cdots & \cdots & \phi_{p2} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & \cdots & \phi_{pn} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_p \end{pmatrix}$$

If many elements of the vector a are zero, one says that the vector a is "sparse"

Sparse representations

For our example of the (orthonormal) Haar basis : $\Phi =$

$$\begin{pmatrix}
1/\sqrt{2} & 0 & 0 & 0 & 1/2 & 0 & 1/2\sqrt{2} & 1/2\sqrt{2} \\
-1/\sqrt{2} & 0 & 0 & 0 & 1/2 & 0 & 1/2\sqrt{2} & 1/2\sqrt{2} \\
0 & 1/\sqrt{2} & 0 & 0 & -1/2 & 0 & 1/2\sqrt{2} & 1/2\sqrt{2} \\
0 & -1/\sqrt{2} & 0 & 0 & -1/2 & 0 & 1/2\sqrt{2} & 1/2\sqrt{2} \\
0 & 0 & 1/\sqrt{2} & 0 & 0 & 1/2 & -1/2\sqrt{2} & 1/2\sqrt{2} \\
0 & 0 & -1/\sqrt{2} & 0 & 0 & 1/2 & -1/2\sqrt{2} & 1/2\sqrt{2} \\
0 & 0 & 0 & 1/\sqrt{2} & 0 & -1/2 & -1/2\sqrt{2} & 1/2\sqrt{2} \\
0 & 0 & 0 & -1/\sqrt{2} & 0 & -1/2 & -1/2\sqrt{2} & 1/2\sqrt{2}
\end{pmatrix}$$

Sparse representations: examples

- Typically, a natural image admits a sparse representation in a wavelet basis (or other $\star$-lets members of the family)

- An astronomical image made of isolated stars is sparse in the pixel representation

- Some signals contain only certain frequencies and are sparse in a Fourier basis

Sparse representations: examples

- In statistics (linear regression), the matrix Φ is made of data: one measures/observes the values of p variables for n instances/individuals and one assumes that the observed response y is a linear combination of the values of these variables

The standard notation in statistics for such relation is

$$y = X\beta$$

In some cases, one can assume that the vector β is sparse and use – to infer it from y and X – (least-squares) regression methods which enforce sparsity ("lasso" regression) and hence select a few variables (corresponding to the nonzero elements in β)

A new statistics

"The coming century is surely the century of data. A combination of blind faith and serious purpose makes our society invest massively in the collection and processing of data of all kinds, on scales unimaginable until recently. Hyperspectral Imagery, Internet Portals, Financial tick-by-tick data, and DNA Microarrays are just a few of the better known sources, feeding data in torrential streams into scientific and business databases worldwide. In traditional statistical data analysis, we think of observations of instances of particular phenomena (e.g. instance, human being), these observations being a vector of values we measured on several variables (e.g. blood pressure, weight, height,...). In traditional statistical methodology, we assumed many observations and a few, well chosen variables. The trend today is towards more observations but even more so, to radically larger numbers of variables - voracious, automatic, systematic collection of hyper-informative detail about each observed instance."
David L. Donoho

A new statistics

- Traditionally, many observations (n) relative to a small number of variables (p)

- "Modern" statistics : relatively few observations (n) relative to a large number of variables (p)

("large p, small n paradigm")

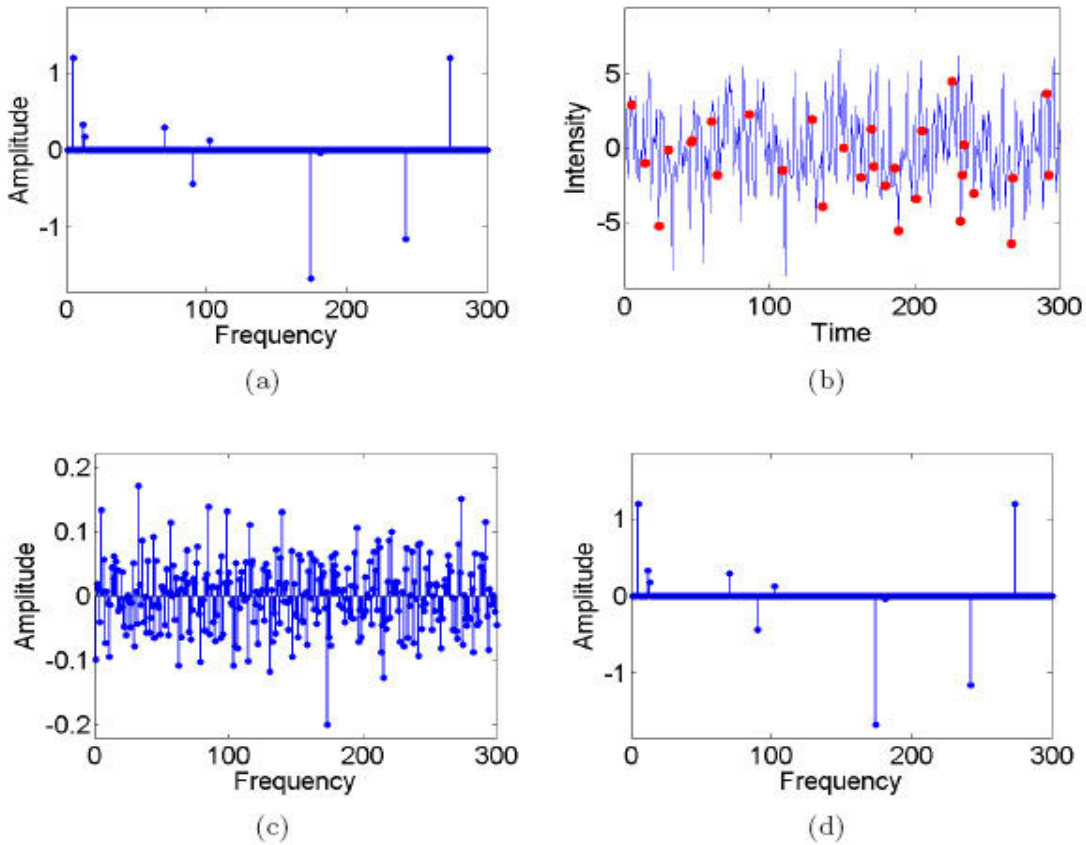
- This calls for the development of new methodologies (variable selection, regularization, etc.)

Compressed Sensing = Compressive Sampling

- Parsimony in the way of collecting data relative to a sparse model/vector
- To sense/reconstruct a quantity (vector) of interest which is sparse (with k nonzero elements on p, and $k \ll p$), it seems indeed absurd to have to collect a number of (linear) measurements or samples n of the order of p and subsequently to compress the acquired data!
- Many recent mathematical papers on the subject have shown that, in general, a number of measurements of the order of $k\ log(p/k)$ are sufficient provided that they are acquired in a properly randomized fashion

(Candès, Tao 2006; Donoho 2006, etc.)

Compressed sensing



Reconstruction (d) of a signal which is sparse in the Fourier domain (10 frequencies) (a) from 30 random samples (in place of 300) in the time domain (b)

(Fornasier, Rauhut 2009)

Dimension reduction

A data cloud can be represented by a $n \times p$ matrix X (n points living in $\mathbb{R}^p$) and one wants a (reliable) lower-dimensional representation of this cloud.
Some examples of methods to do so:

- Principal Component Analysis (PCA)
  Project the data on the directions of maximal variance, i.e. on the k top eigen-directions of the $p \times p$ sample covariance matrix $\frac{1}{n} X^T X$ (after centering)
- Factor Analysis
  Represent the matrix X (approximately) as a product ΛF of matrices of low rank: Λ, the matrix of the "loadings", is $n \times k$ and F, the matrix of the "factors", is $k \times p$

Dimension reduction

- Nonnegative Matrix Factorization (NMF)
  Represent a matrix X with nonnegative elements (approximately) as a product AB of a $n \times k$ matrix A and a $k \times p$ matrix B, both with positive elements
- Independent Component Analysis (ICA)
- Random (linear) projections
  Projections of a data cloud of n points in $\mathbb{R}^p$ onto lower k-dimensional subspaces approximately preserve the Euclidean structure (i.e. do not modify the pairwise distances by more than $1 \pm \varepsilon$) provided that these subspaces are chosen randomly and $k \geq (\log\ n)/\varepsilon^2$ (Johnson-Lindenstrauss Lemma)

Algorithms for Big Data

- Need for algorithms that are tailored for Big Data, i.e. are sufficiently fast, do not require too much memory and scale well with the dimensionality
  (e.g. Page Rank used in Google)

- A new trend: probabilistic algorithms
  which trade accuracy/reliability for speed and are only expected to deliver the correct answer “with high probability”

The end of Theory?

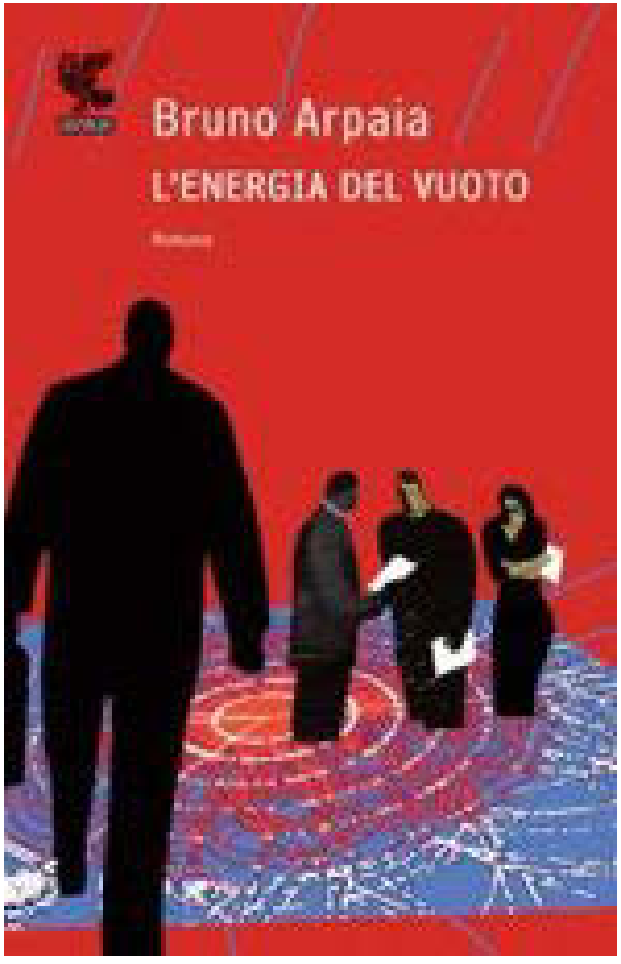
The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

[...] The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?

Chris Anderson (Wired Magazine 2008)

Security issues



Risk of theft, manipulation, leaks, deterioration, etc.

Sherlock Holmes

“Data! Data! Data!” he cried impatiently. “I can’t make bricks without clay”
(Arthur Conan Doyle, The Adventure of the Copper Beeches, 1892)

“It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”
(Arthur Conan Doyle, A Scandal in Bohemia, 1892)