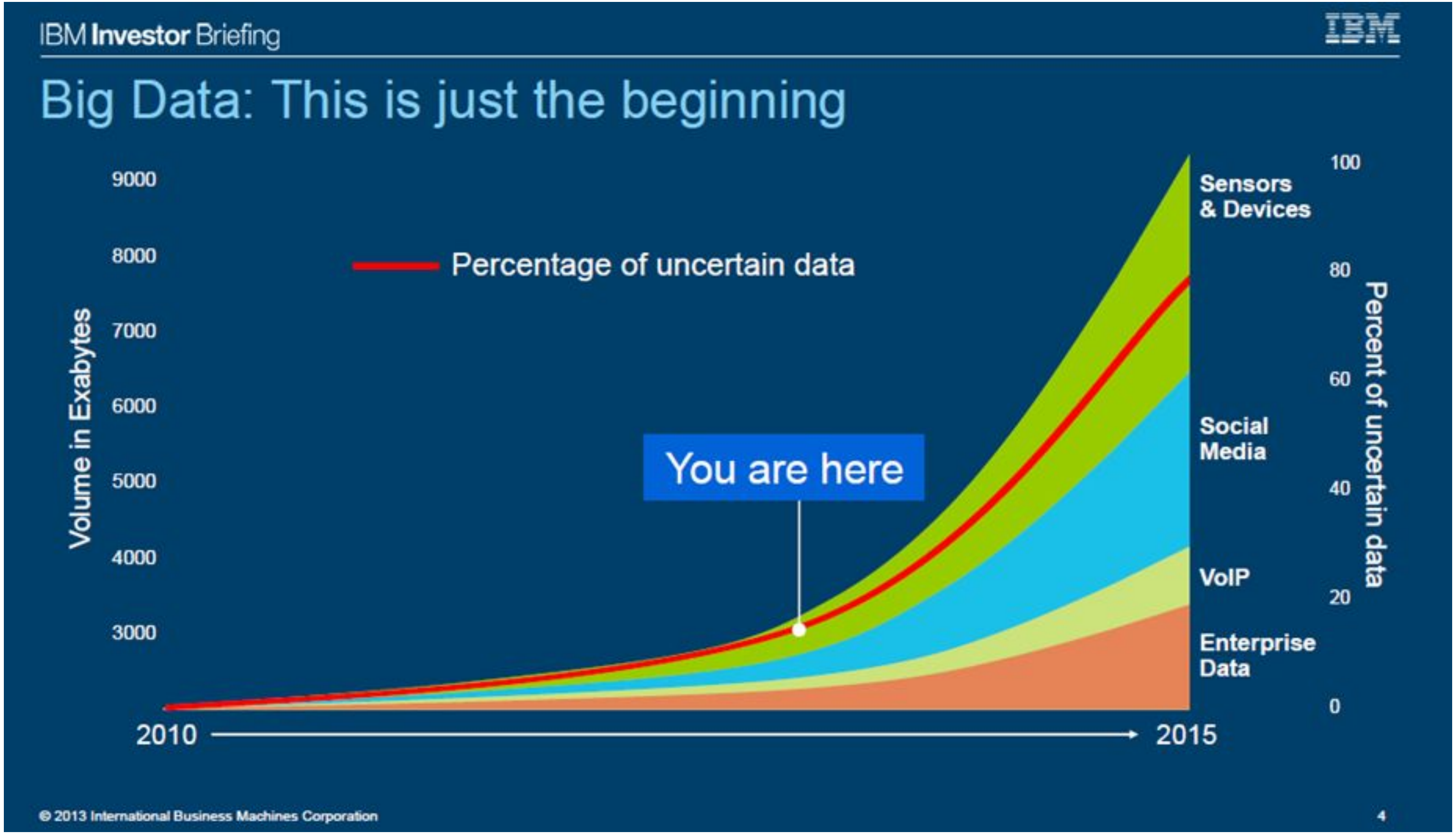
Data Visualization

Gina Lucia Muñoz Salas

Universidad Católica
San Pablo

Centro de Investigación
e Innovación en
Ciencia Computación

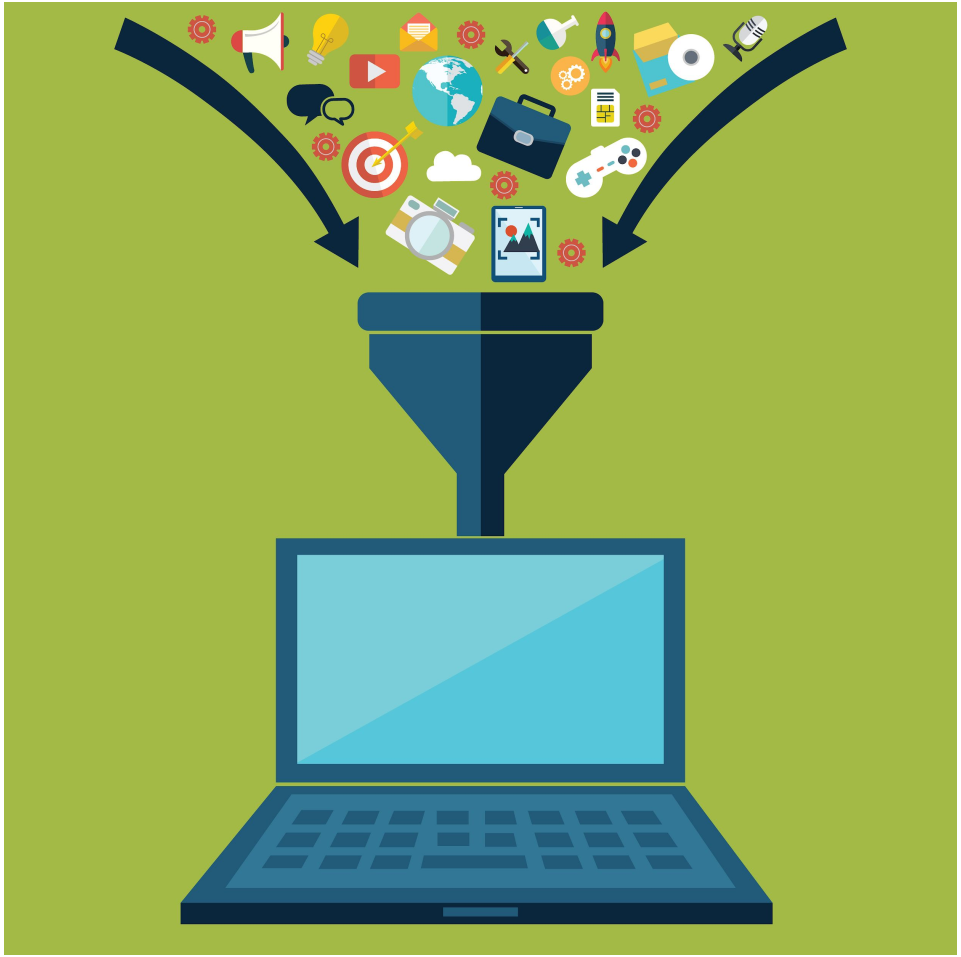
Data growth


IBM Investor Briefing
IBM
Big Data: This is just the beginning
Percentage of uncertain data
Volume in Exabytes
9000
8000
7000
6000
5000
4000
3000
You are here
Sensors & Devices
Social Media
VoIP
Enterprise Data
100
80
60
40
20
0
Percent of uncertain data
2010
2015
© 2013 International Business Machines Corporation
4

Data collection

Data : $(r_1,r_2,r_3,...,r_n)$

Attributes : $(v_1, v_2, v_3,..., v_m)$



Data types

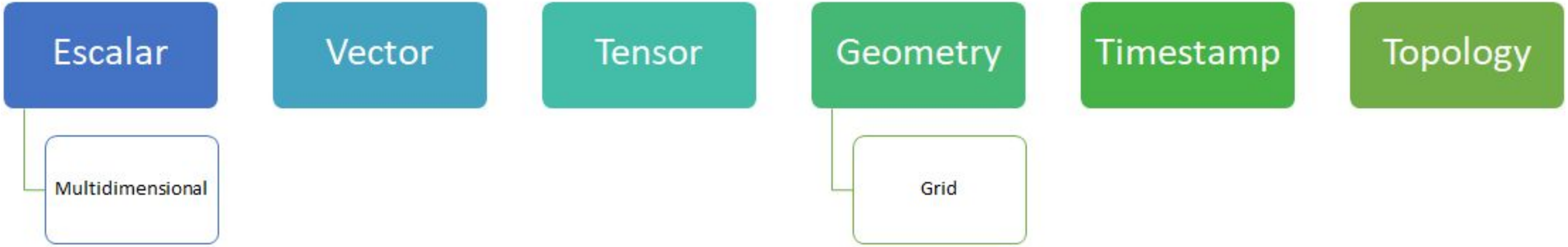
Ordinal

- Binary
- Discretes
- Continuous

Nominal

- Categorical
- Ranked
- Text

Data structures



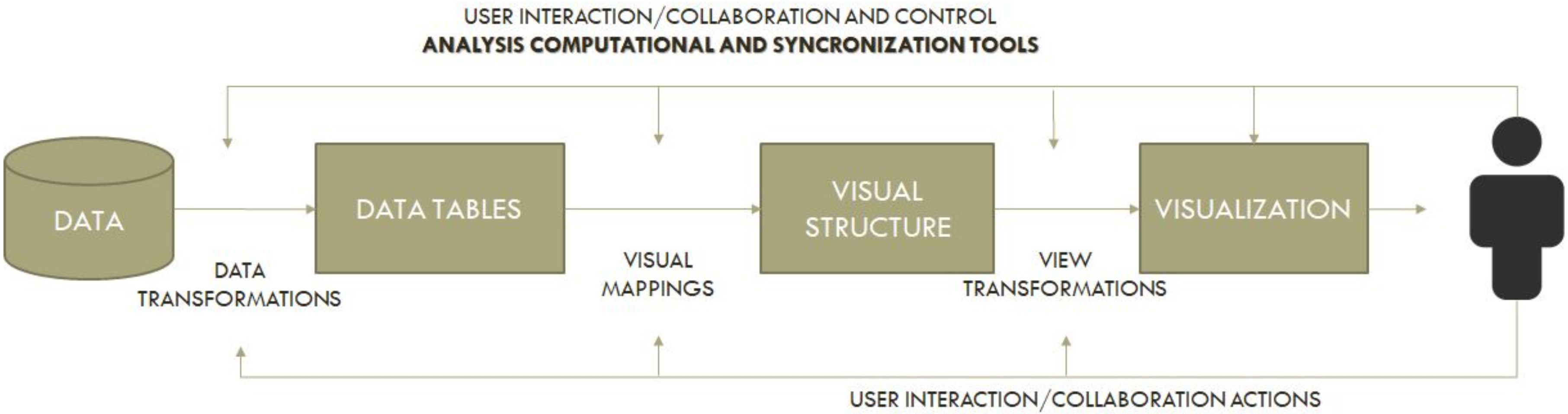
MRI: Density - 3 spatial attributes - 3D grid

Financial data : n attributes - temporal attribute - no geometric structure

Census: n attributes - spatial attribute - temporal attribute

Social media - ?

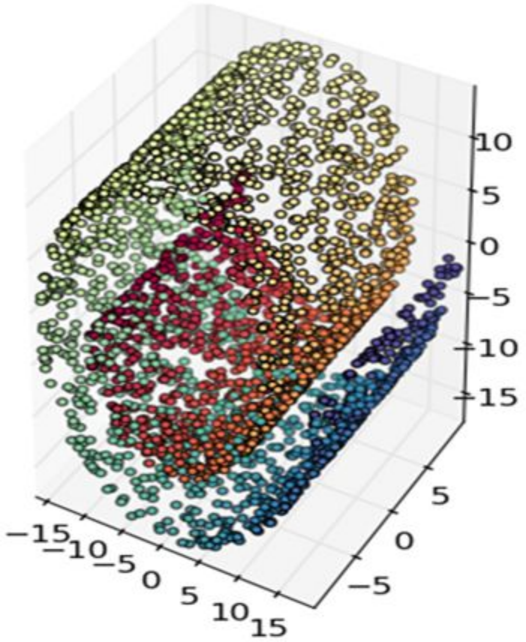
Data processing

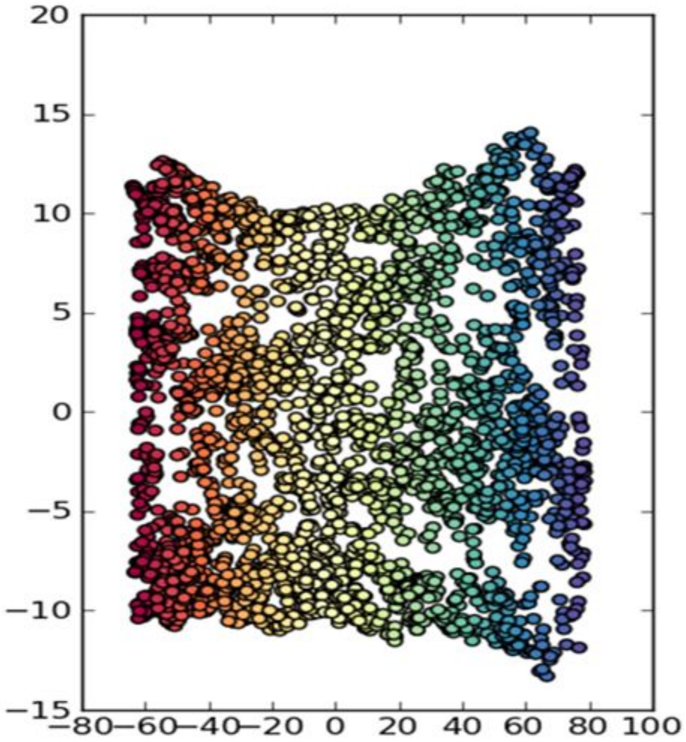


Data processing

- **Metadata**
  - Info for interpretation (references, units, symbols, resolution)
- **Statistical analysis**
  - Outliers, similar groups, redundance
  - Mean, SD
- **Incomplete data**
  - Remove instances
  - Sentry value
  - Substitute value
- **Normalization**
  - Follow statistical property
  - Unit data
  - [0,1]
  - Standardization
- **Re Sampling**
  - Interpolation
  - Reduction

Dimensionality Reduction



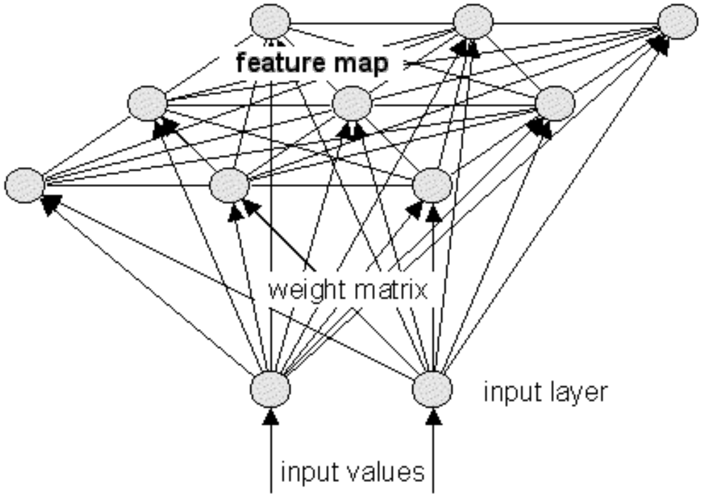


Dimensionality Reduction techniques

- Principal Component Analysis - PCA

$$Y_{ij} = \mathbf{e}_j^T \mathbf{X} = e_{1j}X_{i1} + e_{2j}X_{i2} + \ldots + e_{pj}X_{ip}$$

- Multidimensional Scaling - MDS
  - Distance information between instances in the original space for mapping data into a Cartesian space
- Self Organizing maps



Dimensionality Reduction techniques

- **Least Square Projection (LSP)**
  - Initially maps control points into a visual space
  - Projects all the remaining points by a Laplacian mapping

- **t-Distributed Stochastic Neighbor Embedding (t-SNE)**
  - Probability distributions over pairs of instances in high-dimensional and visual space
  - Minimize Kullback Leibler divergence between the two distributions with respect to positions of points in the mapping.

Type Mapping

Id	Age	Chest Pain	Chol.
1	63	Typical	233
2	67	Asymptom	286
3	37	Non anginal	250
4	44	Non typical	263

Id	Age	Typical	Asymptom	Non anginal	Non typical	Chol
1	63	1	0	0	0	233
2	67	0	1	0	0	286
3	37	0	0	1	0	250
4	44	0	0	0	1	263

Id	Age	Chest Pain	Chol
1	63	Typical	233
2	67	Asymptomatic	286
3	37	Non anginal	250
4	44	Non typical	263

Id	Age	Age Range	Chest Pain	Chol
1	63	60+	Typical	233
2	67	60+	Asymptomatic	286
3	37	25-39	Non anginal	250
4	44	40-59	Non typical	263

Distances and Similarities

Ordinal Data

Minkowski family

$$L_p(\mathbf{x}_i, \mathbf{x}_j) = (\sum_{k=1}^{m} |x_{ik} - x_{jk}|^p)^{\frac{1}{p}}$$

Non negative $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X},\ \delta(\mathbf{x}_i, \mathbf{x}_j) \geq 0$

Identity $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X},\ \mathbf{x}_i = \mathbf{x}_j \Leftrightarrow \delta(\mathbf{x}_i, \mathbf{x}_j) = 0$

Simmetry $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X},\ \delta(\mathbf{x}_i, \mathbf{x}_j) = \delta(\mathbf{x}_j, \mathbf{x}_i)$

Triangle inequality $\forall \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \in \mathbf{X},\ \delta(\mathbf{x}_i, \mathbf{x}_k) \leq \delta(\mathbf{x}_i, \mathbf{x}_j) + \delta(\mathbf{x}_j, \mathbf{x}_k)$

Binary distances

	1	0
1	p	q
0	r	s

$d_{ij} = (q+r)/t$ (simple matching)
$d_{ij} = (q+r)/(p+q+r)$ (Jaccard's distance)
$d_{ij} = (q+r)$ (Hamming distance)
$d_{ij} = (p+s)/t$ (simple matching coefficient)
$d_{ij} = p/t$
$d_{ij} = p/(p+q+r)$ (Jaccard's corfficient)
$d_{ij} = 2p/(2p+q+r)$
$d_{ij} = 2(p+s)/(2(p+s)+q+r)$
$d_{ij} = p/(q+r)$
$d_{ij} = (p+s)/(q+r)$

Mixed distances - Gower

CATEGORICAL

$$D(p_k, q_k) = \begin{Bmatrix} 0, p_k = q_k \\ 1, p_k \neq q_k \end{Bmatrix}$$

NUMERICAL

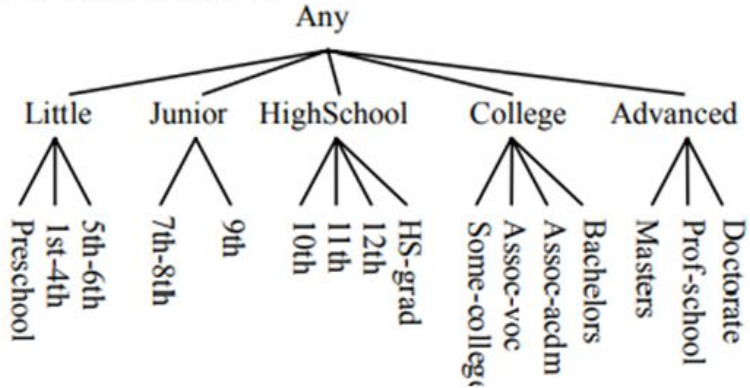
$$D(p_u, q_u) = \frac{|p_u - q_u|}{R_u}$$

AGGREGATION

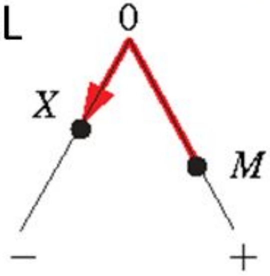
$$D(p, q) = \frac{1}{n} \sum_{i=1}^{n} D(p_i, q_i)$$

Mixed distance - Hierarchy

CATEGORICAL



NUMERICAL



$$D(p,q) = d_p + d_q - 2d_{LCP(p,q)}$$

AGGREGATION

$$D(p,q) = \left(\sum_{i=1}^{n} w_i (D(p_i, q_i))^L \right)^{1/L}$$

Data Visualization

Gina Lucia Muñoz Salas

Universidad Católica San Pablo

Centro de Investigación e Innovación en Ciencia Computación