Μεγάλα Δεδομένα από την Οπτική Γωνία των Ερευνητών Διαχείρισης Πληροφορίας: Μόδα ή Ευκαιρία;

Αλέξανδρος Λαμπρινίδης

Advanced Data Management Technologies Lab
Department of Computer Science
University of Pittsburgh
http://labrinidis.cs.pitt.edu

12ο Ελληνικό Συμπόσιο Διαχείρισης Δεδομένων – Αθήνα, 25 Ιουλίου 2014
You know Big Data is an important problem if...

- It is featured on the cover of Nature and the Economist!
You know Big Data is an even more important problem if...

- It has a Dilbert cartoon!
What is Big Data?

**Definition #1:**
- Big data is like teenage sex:
  - everyone talks about it,
  - nobody really knows how to do it,
  - everyone thinks everyone else is doing it,
  - so everyone claims they are doing it...

**Definition #2:**
- Anything that Won't Fit in Excel!

**Definition #3:**
- Using the Vs
The three Vs

- **Volume** - size does matter!
- **Velocity** - data at speed, i.e., the data “fire-hose”
- **Variety** - heterogeneity is the rule
Five more Vs

• **Variability** - rapid change of data characteristics over time

• **Veracity** - ability to handle uncertainty, inconsistency, etc

• **Visibility** – protect privacy and provide security

• **Value** – usefulness & ability to find the right hay-colored needle in the haystack

• **Voracity** - strong appetite for data!
Enter Moore’s Law

Microprocessor Transistor Counts 1971-2011 & Moore’s Law

The graph shows the trend of transistor counts in microprocessors from 1971 to 2011, with Moore’s Law indicating that the number of transistors doubles every two years. The data points represent various microprocessors and their corresponding transistor counts.

Key points:
- The logarithmic scale on the y-axis represents the transistor count.
- The x-axis represents the date of introduction of the microprocessors.
- The trend line confirms the doubling of transistors every two years, as predicted by Moore’s Law.

[ Wikipedia Image ]
Storage capacity increase

HDD Capacity (GB)

HDD Capacity (GB)

Insert other exponentially increasing graphs here
(e.g., data generation rates, world-wide smartphone access rates, Internet of Things, …)

[Wikipedia Data]
But

- **Human processing capacity** remains roughly the same!
We refer to this as the:

**Big Data – Same Humans Problem**
Phases in the Big Data Pipeline

- Big Data Processing goes well beyond data storage
- Must consider end-to-end and human aspect [1]

Source: [2]
• Μνημειώδη ευκαιρία!

• Big Data will not do “everything” for all businesses, BUT:

• Big Data will influence even more aspects of life, including government and science.

• There are new technical/research challenges:
  o Heterogeneity / diversity (users, data, hardware, etc.)
  o Knowledge bases (as part of a bigger pipeline)
  o People in the processing pipeline (visualization, crowdsourcing)
  o Privacy / security
References

   By Abadi et al, October 2013
   http://beckman.cs.wisc.edu

   By Jagadish, Gehrke, Labrinidis, Papakonstantinou, Patel, Ramakrishnan, and Shahabi,
   Communications of the ACM, July 2014