Big Data

Dan Vodislav

ETIS, Université de Cergy-Pontoise

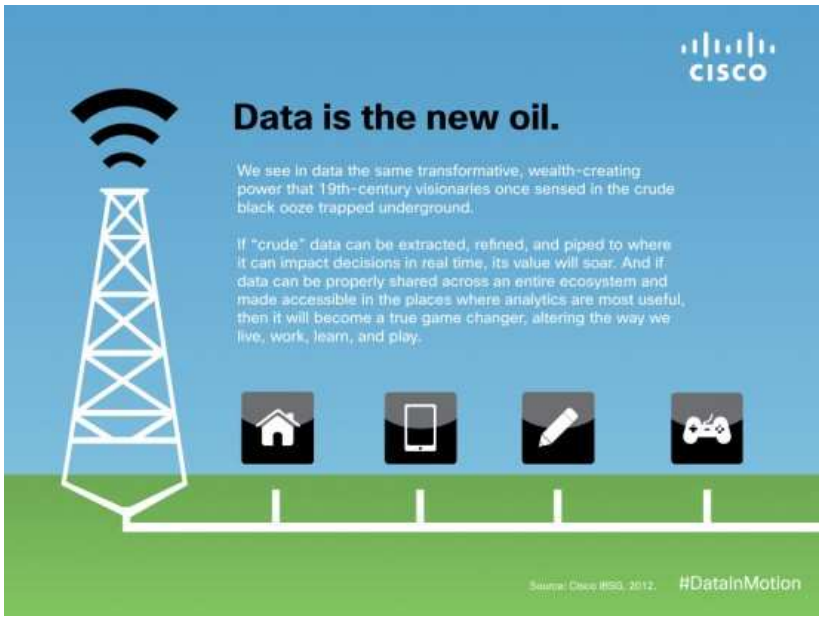
The Big Data challenge

- Data is continuously produced and used in the decision process
  - Instruments : satellites, microscopes, particle accelerators, telescopes, …
  - Simulations: climate, materials, chemistry, …
  - Imaging: medical, visualization, …
  - Metadata: descriptions, publications, knowledge bases, …

Source: Sajal Das, Keith Marzullo

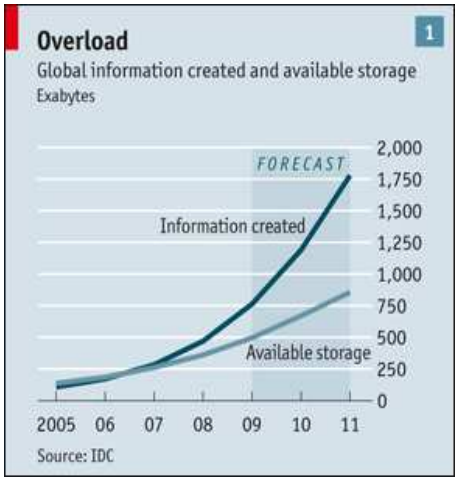
Economical importance

- Data: new raw ressource to exploit



Big Data Technologies

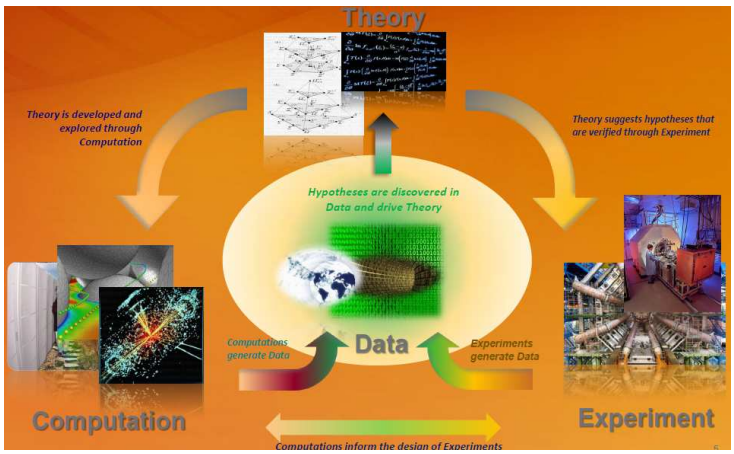
- New techniques et architectures for extracting knowledge from data
  - Huge volumes
    - Terabytes (10^{12}) → Zettabytes (10^{21})
    - 2014: 4,4 zettabytes
    - 2020: 44 zettabytes
  - Various natures
    - Structured (DB) → semi-structured (XML) → unstructured (text, images, …)
  - Continuously produced
    - Impossible to store everything
    - Batch → streaming



Data and scientific research

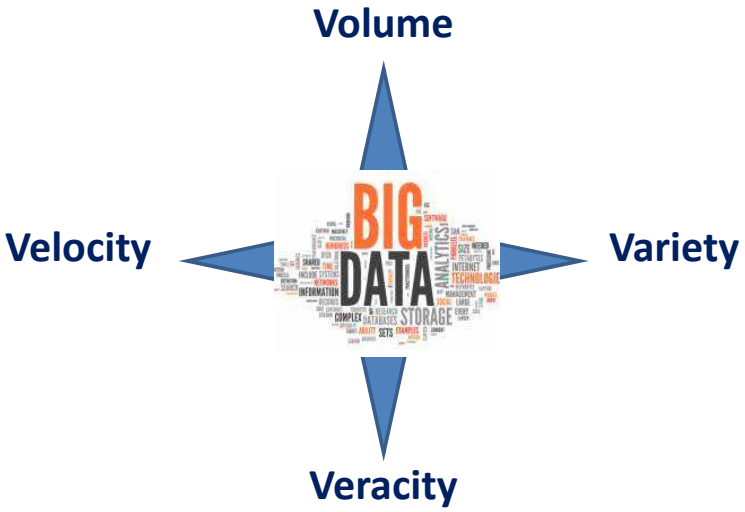
- Evolution of the scientific research towards massive exploitation of data measures
  - Physics, chemistry, medicine, engineering
  - … but also social sciences, economics, etc.
- Challenge →

  Extracting knowledge from very large data sets



Source: JOHN R. JOHNSON

The 4 « V » of Big Data



+ a 5th one: **Value**

+ also: **Variability, Validity, Vulnerability, Visualization, Volatility**

Big Data module

- In relation with the work of the MIDI team of the ETIS lab
  - Volume
    - New storage and processing models
      → ***cloud computing, Map-Reduce, NoSQL***
    - Algorithms adapted to these models
      → ***data mining, OLAP queries in the cloud***
  - Velocity
    - Continuous processing of information
      → ***social networks and information streams***
  - Variety
    - Homogeneous description of data
      → ***RDF, open data, semantic web***
- Evaluation : presentation on a given subject

Sessions

- Introduction + « Linked open data » (D. Vodislav)
- Information streams, social networks (D. Vodislav)
- Cloud computing 1 (D. Kotzinos)
- Cloud computing 2 (D. Kotzinos)
- Data mining (T. Y. Jen)
- Data warehouses (T. Y. Jen)
- Student presentations