Big Data Architectures principles and practice

Carlyna
Bondiombouy

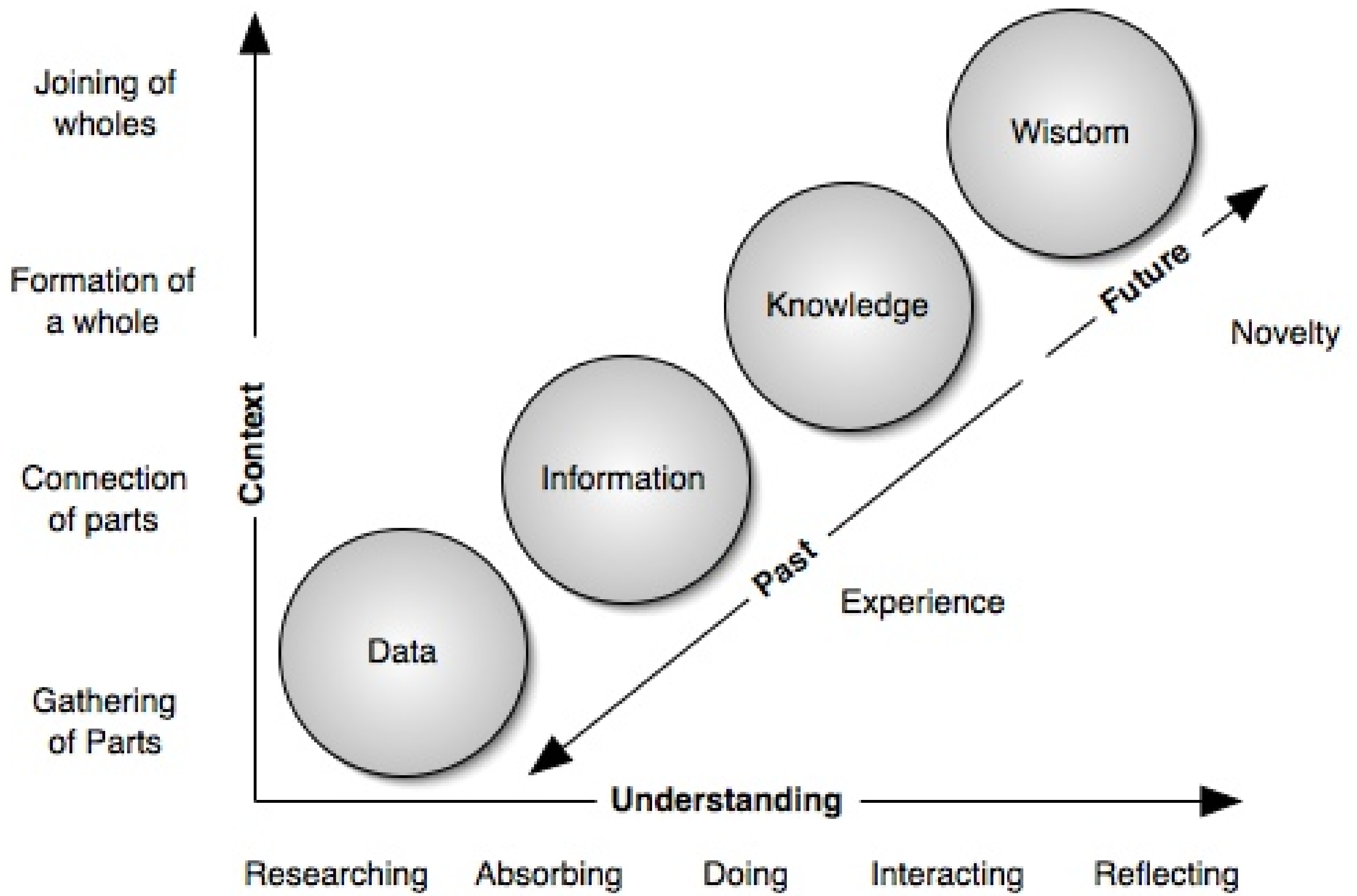
Outline

1. Big data: challenges and opportunities
2. Big data architectures
3. NoSQL & NewSQL
4. Big data frameworks
5. Big data integration
6. Data analytics with Spark

Big Data: challenges and opportunities

1. Data science and big data
2. Evolution of data
3. Use cases in industry
4. Data protection
5. Impact of hardware progress
6. Opportunities and risks

The Continuum of Understanding



- The more the data, the better the understanding
  - *If we manage to deal with the data well*

1. Data Science: definition

- Data science
  - The science of making sense of data
  - The use of data management, statistics and machine learning, visualization to collect, clean, integrate, process, analyze and visualize big data
- Goal: create data products and data services
  - "Data is the new oil of the digital economy" (Wired, 2014)
- Data scientist
  - Not to be confused with data analyst
  - Strong technical skills
  - AND good knowledge of the business domain

Data Science: definition

Hard to find data scientists !

New training programs all over the world

But many "fake" data scientists on the market

- Goal: create data products and data services
  - "Data is the new oil of the digital economy" (Wired, 2014)
- Data scientist
  - Not to be confused with data analyst
  - Strong technical skills
  - AND good knowledge of the business domain

Big Data: what is it?

- A buzz word!
  - It depends on your perspective
    - E.g. 10 terabytes is big for an OLTP system, but small for a web search engine
- A definition (Wikipedia)
  - Consists of data sets that grow so *large* that they become awkward to work with
  - *But size is only one dimension of the problem*
  - Dimensions (Vs): volume, velocity, variety, veracity, validity
- How *big* is big?
  - Moving target: terabyte (10^{12} bytes), petabyte (10^{15} bytes), exabyte (10^{18}), zetabyte (10^{21})
  - Landmarks in DBMS products
    - 1980: Teradata database machine
    - 2010: Oracle Exadata database machine

Why Big Data Today?

- Overwhelming amounts of data
  - Exponential growth, generated by all kinds of programs, networks and devices
    - E.g. Web 2.0 (social networks, etc.), mobile devices, computer simulations, satellites, radiotelescopes, sensors, etc.
- Increasing storage capacity
  - Storage capacity has doubled every 3 years since 1980 with prices steadily going down
    - 1 Gigabyte (HDD): $400K in 1980, $10K in 1990, $1K in 1995, $10 in 2000, $0.01 in 2018
- Very useful in a digital world!
  - Massive data => high-value information and knowledge

Big Data Dimensions: the five *V's*

- Volume
  - Refers to massive amounts of data
  - Makes it hard to store and manage, but also to analyze
- Velocity
  - Continuous data streams are being captured (e.g. from sensors, mobile devices, IoT) and produced
  - Makes it hard to perform online processing
- Variety
  - Different data formats, different semantics, uncertain data
  - Makes it hard to integrate and analyze
- Veracity
  - Authenticity and conformity of the data with reality
  - Altered by bias, noise, misinformation, fake news
- Validity
  - Correction and accuracy of data for the intended use

Big Data Analytics (BDA)

- Objective: find useful information and discover knowledge in data
  - Predictive analysis, decision support, research, …
- Why is this hard?
  - Low information density (unlike in corporate data)
    - Like searching for needles in a haystack
  - External data from various sources
    - Hard to verify and assess, hard to integrate
  - Different kinds of data
    - Structured data: transaction, decision-support, scientific
    - Unstructured: web document, social network, open data, IoT
    - Hard to integrate

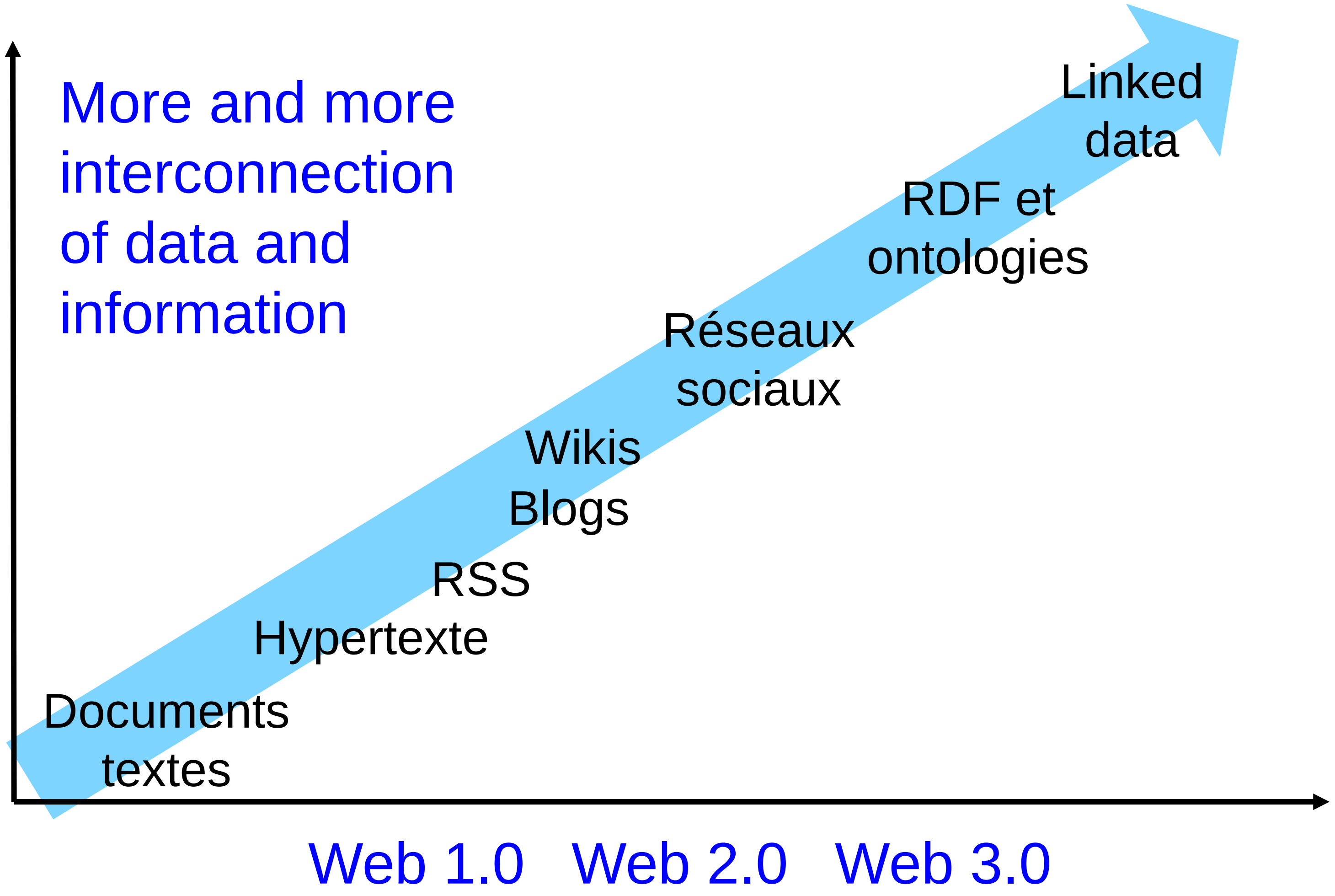
Some BDA Killer Apps

- 360° view of customers
  - Marketing, recommendation
    - Requires combining corporate (structured) data with external (unstructured) data (web, social networks, phone recordings, …)
- Online fraud detection across massive databases
  - E-commerce, banking, telephony, etc.
- National security
  - Signal intelligence, cyber analytics
- Medical science
  - Personalized medicine, with major investment from the GAFAM

2. Evolution of Data

- Data interconnection
- Data streams
- Internet of Things (IoT)
- Data-intensive science

Interconnection



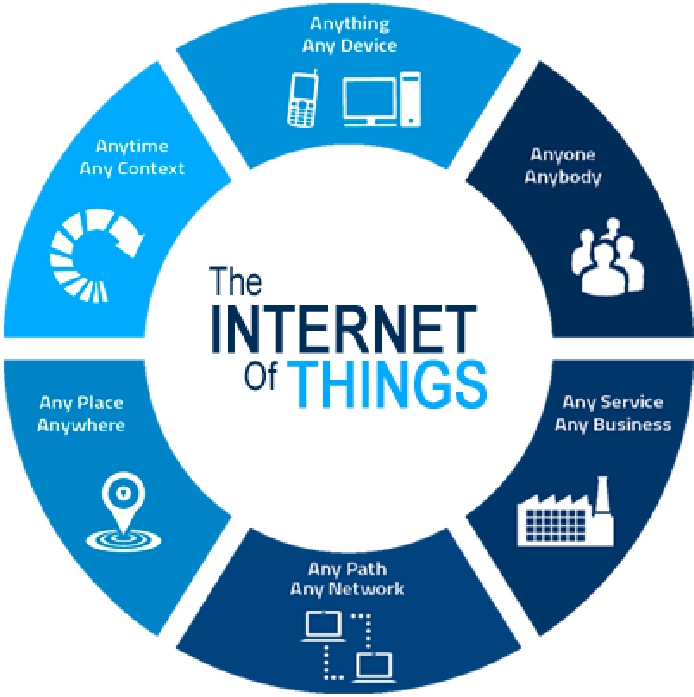
Data Streams

- Continuous, unlimited, fast, and time-varying data streams
  - User session information, trading data, stock prices, news, etc.

- Problem
  - How to analyze data in real time, when you don't have the time to store them in a database?
  - How to integrate with company data, e. g. a user profile?
  - How to make the IS reactive?

IoT

- Interconnection of all kinds of digital objects via the Internet
  - Sensors, smart meters, connected watches, etc.
  - In general, excludes smartphones, tablets and PCs

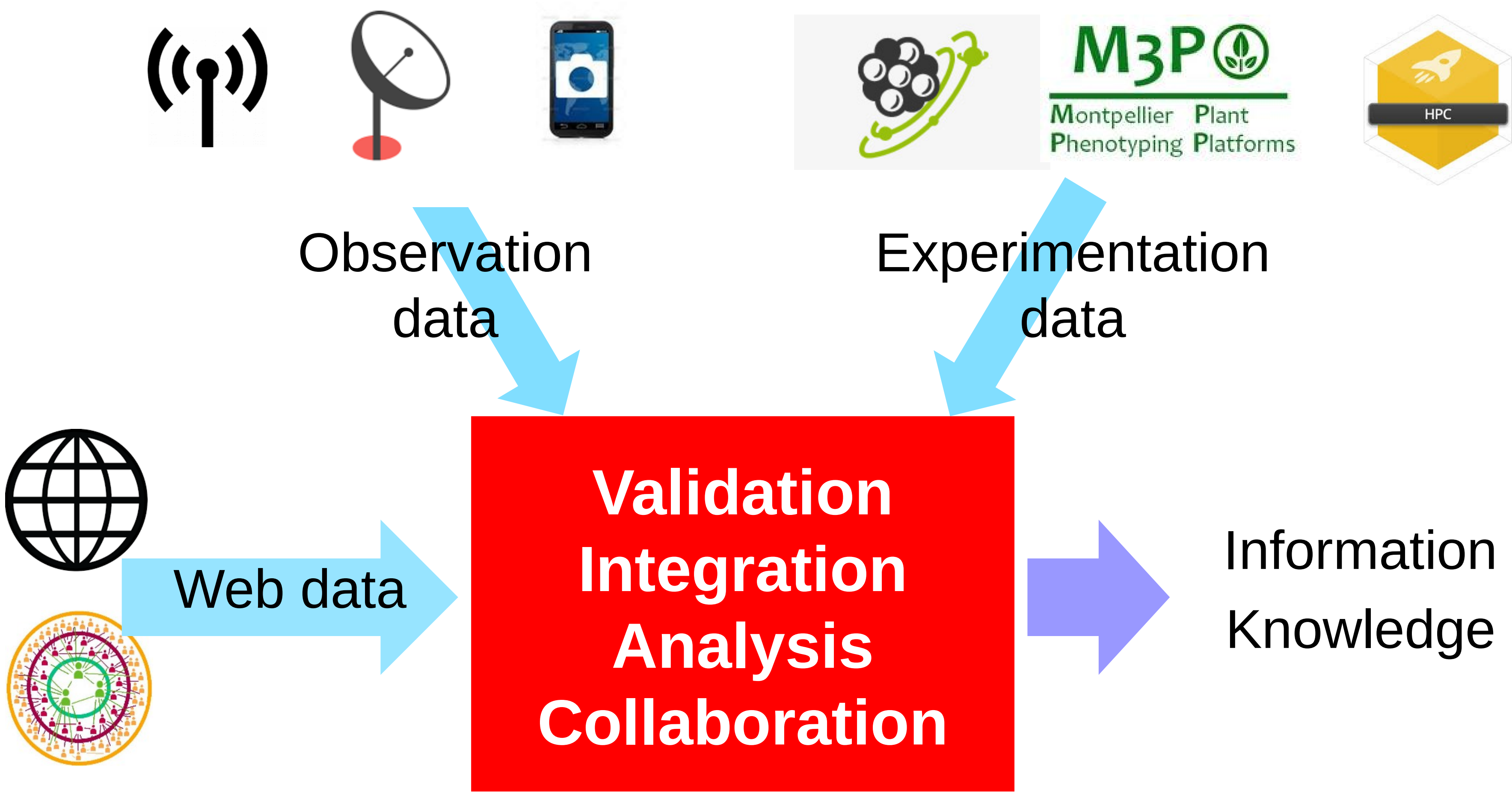


- An object must have the following capabilities :
  - Identification (IP address)
  - Data transfer (SMTP, http, …)
  - Wireless communication (Wifi, Bluetooth, RFID, …)

IoT

- Wide range of applications
  - Transport, logistics, health, home automation, intelligent city, quantified self
- Exponential growth in the number of objects
  - 50 billion by 2020 (according to Cisco) versus some billion for smartphones, tablets and PCs
- Many data processing services (big data)
  - GAFAM, network operators
- Major issues
  - Security and privacy protection
  - Scalability
  - Real-time processing

Data-intensive Science



Data-intensive Science

The problem

"Scientists are spending most of their time manipulating, organizing, finding and moving data, instead of researching. And it's going to get worse"

The Office Science Data Management Challenge USA DoE 2004

3. Use Cases in Industry

The 5 Top Use Cases (IBM)

1. Big data exploration
   - Find, visualize and understand all the data stored in different systems and silos of the company, for decision support
2. Real-time security
   - Reduce risks and detect fraud in real time, by extending security intelligence platforms with new data (e. g. social networks, emails, sensors, telco)

4. Data Protection

- GDPR (General Data Protection Regulation)
  - Applicable since 25 May 2018 throughout the EU
  - Concerns any company, even very small ones
  - Binding: fine of up to 4% of revenue and €20 million
- Objective
  - Strengthening the rights of persons whose data are processed
    - Right to forget, reinforced consent, easy access,...
  - Accountability of data controllers
- Principles
  - Obligation to ensure that the processing operation complies with the rules laid down
    - Involves data traceability at all stages of its life cycle
  - Privacy by design: confidentiality and security requirements are taken into account from the design of products and services
    - Requires good data governance

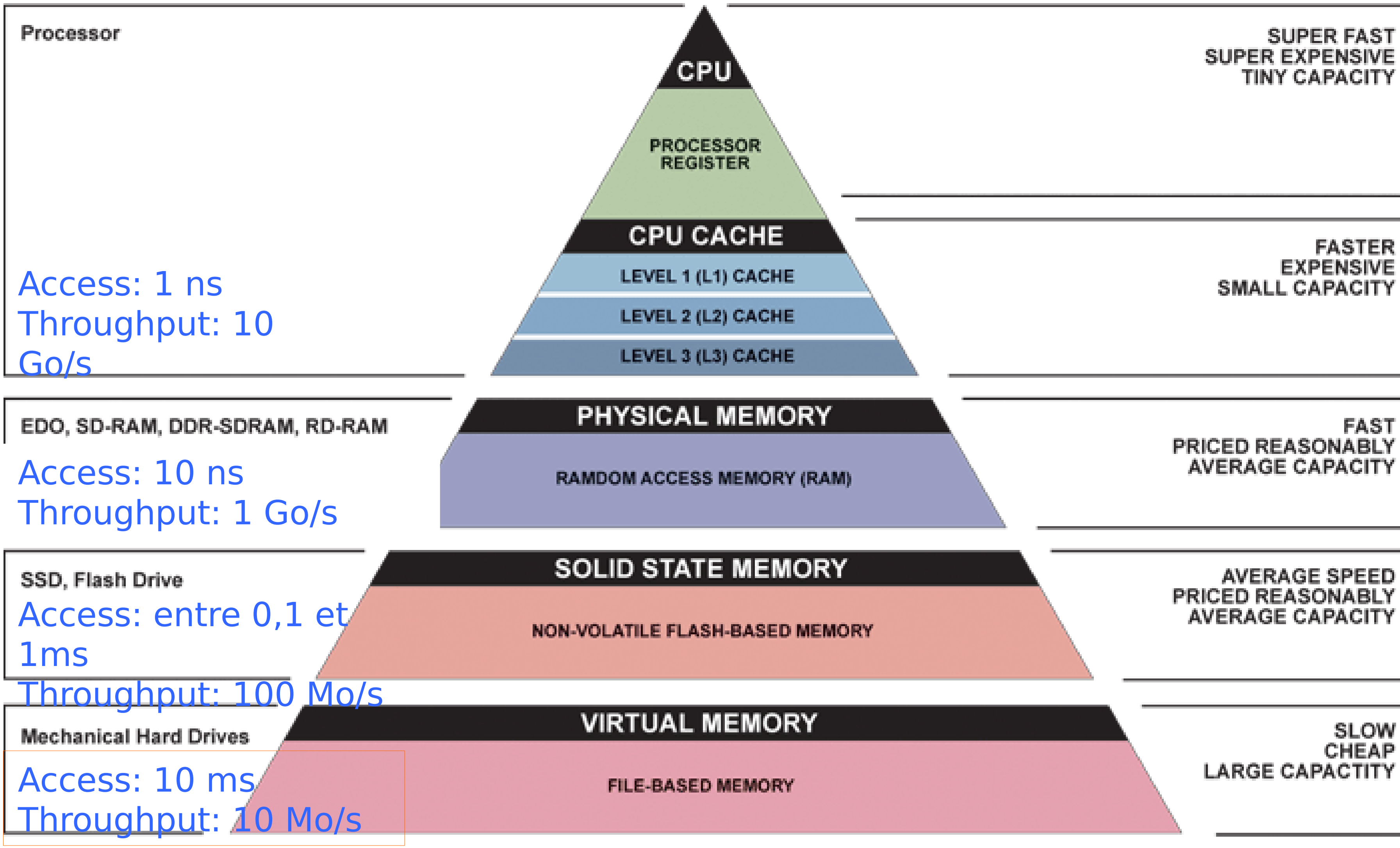
Impact on Big Data

- Advantages
  - Obligation to ensure that the processing is in conformity with big data
    - Progress in relation to the Data Protection Act of 6 January 1978 based on the principle of formality prior to collection
  - Building trust with users
  - Argument taken up by Apple, Cisco and Microsoft for a GDPR in the USA
- Drawbacks
  - Additional cost and complexity for companies
  - Limits of anonymization
    - Big data processing on anonymous data can lead to the identification of persons via quasi-identifiers, such as place of residence, occupation, gender and age

4. Impact of Hardware Progress

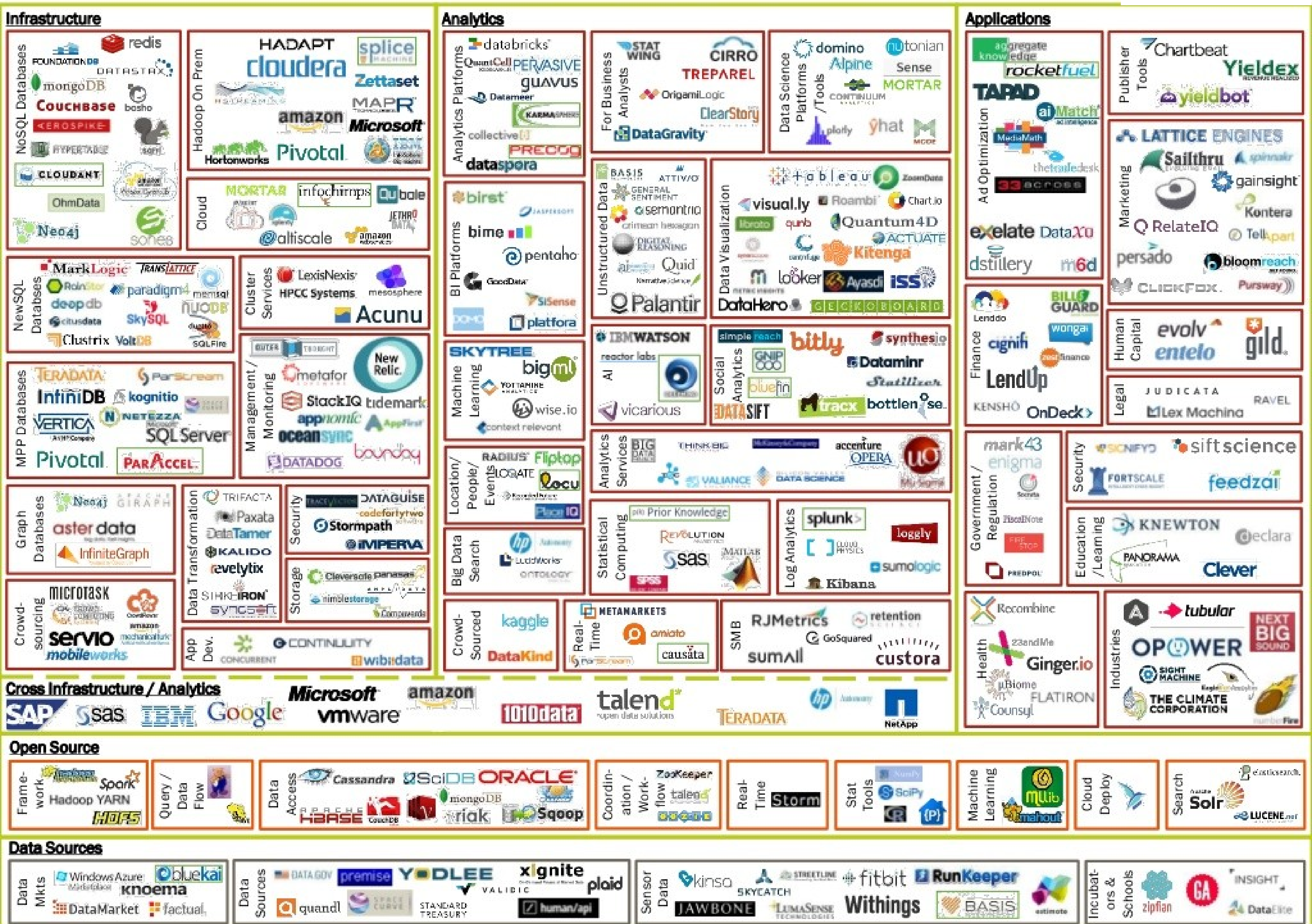
- Storage
  - Flash memory as a cache between disk and RAM
  - Solid State Disk (SSD) as a replacement for HDD
- Very large RAM memories
  - The advent of in-memory?
- Multiprocessing
  - Multi-core processors
  - CPU-GPU combination
- Broadband networks
  - Tree architectures based on switches
    - Ex. Infiniband up to 100 megabits/s (Mellanox)

New Memory Hierarchy



▲ Simplified Computer Memory Hierarchy
Illustration: Ryan J. Leng

BIG DATA LANDSCAPE, VERSION 3.0



© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

BIG DATA LANDSCAPE, VERSION 3.0

Easy to get lost

Many solutions

No standard

In constant evolution

© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

5. Opportunities

- Cost reduction (vs. data warehouse)
  - Open source Technologies (Hadoop, Spark, etc.)
  - Cloud services
- Faster, better decision
  - Online processing, e.g. fraud detection
- Meilleure découverte de connaissances
  - Virtuous circle between machine learning and big data
- New data products and services
  - Bi-sided markets (ex. Uber, AirB&B, Leboncoin, …)
  - Personalized medecine, digital agriculture, etc.

Risks

- Safety and security
  - The larger the data, the larger the target for attackers
- Privacy
  - Personal data may be misused by data scientists or other users, and may violate the law
- Cost
  - Data collection, aggregation, storage, analysis, and reporting
  - Support of security and privacy
- Incorrect analyses
  - Models too simple or false (see "when big data goes bad")
  - Misinterpretation of the reasons represented by the data and erroneous conclusions
- Bad data
  - Many projects start off wrong, collecting all kinds of useless, outdated or erroneous data