Data Science

Opportunities and Risks

Patrick Valduriez

Qu'est ce qu'une donnée?

Les données en question

PAR
Stéphane Grumbach
Patrick Valduriez

NIVEAU DE LECTURE
Facile

PUBLIÉ LE
31/03/2016

Au cœur de la connaissance et de l'information, les données ont peu à peu pris une importance qui nous dépasse. Mais qu'entend-on exactement par données ? Quels sont les enjeux autour de leur gestion ou de leur analyse ? Quels impacts sur la société ?

© Fotolia - ptnphotof

Une donnée est la description élémentaire d'une réalité ou d'un fait, comme par exemple un

Stéphane Grumbach

Directeur de recherche Inria, directeur de l'Institut rhônalpin des systèmes complexes (IXXI), co-responsable de l'équipe de recherche DICE.

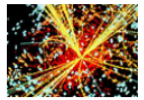
Patrick Valduriez

Directeur de recherche Inria, responsable de l'équipe de recherche ZENITH.

→ Voir tous les auteurs

SUR LE MÊME SUJET

L'Open Data, l'ouverture des données pour de nouveaux usages



Un déluge de données

Data versus Information

- See "Les données en question"
  - http://interstices.info/donnees (Grumbach & Valduriez, 2016)
- Data
  - Elementary definition of a reality or fact
    - E.g. temperature measurement, exam grade, account balance, message, photo, transaction, etc.
  - Can be very simple, and taken in isolation, not very useful
- Information
  - Obtained by interpretation and analysis of a collection of data
  - Can be very useful to understand the world
    - E.g. climate evolution, ranking of a student, etc.

Outline of the Talk

- Data science
- Big data
- The good, the bad and the ugly
- Cloud & big data
- Technologies for data science
- Opportunities and risks

Data Science

Data Science: origins and landmarks

- Statistics
  - Started in the 18th century, primarily to analyze census (*state*) data
  - Fisher: The Design of Experiments, 1935
- Data analysis
  - Tukey: Exploratory Data Analysis, 1977
- Machine learning
  - Mitchell: Machine Learning, 1997
- Data mining
  - Han, Kamber & Pei: Data Mining: Concepts and Techniques, 3rd edition, 2011
- Business intelligence
  - Dresner: The Performance Management Revolution, 2007
- Data-driven science
  - The Fourth Paradigm: Data-Intensive Scientific Discovery, 2009, inspired by Jim Gray (ACM Turing award)

Data Science: definition

- Data science
  - The use of computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze and visualize big data
  - Ultimate goal: create data products and data services
- Data scientist
  - Strong skills in statistics, data analysis and machine learning
  - AND strong knowledge of the business domain, to interpret the analysis results and draw meaningful conclusions

Data Science: definition

- Data science

Hard to find data scientists !

Training programs just starting all over the world

- Data scientist
  - Strong skills in statistics, data analysis and machine learning
  - AND strong knowledge of the business domain, to interpret the analysis results and draw meaningful conclusions

Introduction to Big Data

Big Data: what is it?

- A buzz word!
  - With different meanings depending 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 using on-hand data management tools
  - *But size is only one dimension of the problem*
- 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, $0.02 in 2015
- Very useful in a digital world!
  - Massive data => high-value information and knowledge

Big Data Dimensions: the V's

- Volume
  - Refers to massive amounts of data
  - Makes it hard to store, manage, and analyze (big analytics)
- Velocity
  - Continuous data streams are being produced
  - Makes it hard to perform online processing
- Variety
  - Different data formats, different semantics, uncertain data, multiscale data, etc.
  - Makes it hard to integrate and analyze
- Other V's
  - Validity: is the data correct and accurate?
  - Veracity: are the results meaningful?
  - Volatility: how long do you need to store this data?

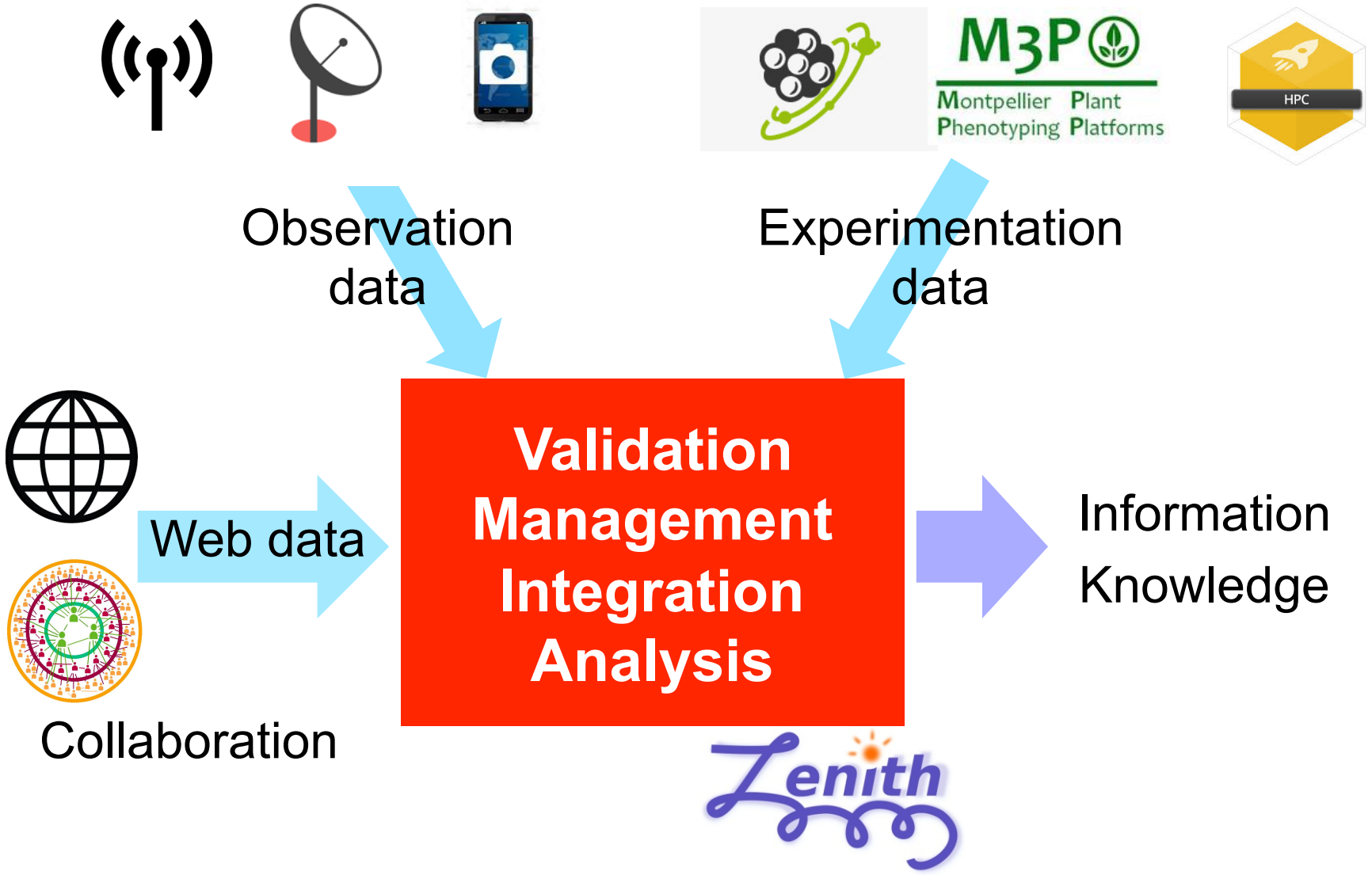
Big Data Analytics (BDA)

- Objective: find useful information and discover knowledge in data
  - Typical uses: forecasting, decision making, research, science, ...
  - Techniques: data analysis, data mining, machine learning, ...
- 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 structures
    - Unstructured text, semi-structured document, key/value, table, array, graph, stream, time series, etc.
    - Hard to integrate
  - Simple machine learning models don't work
    - See next: "When big data goes bad" stories

Some BDA Killer Apps

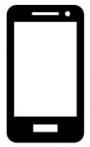
- Social network analysis
  - Modeling, simulation, visualization of large-scale networks
- Real-time processing and analysis of raw data from high-throughput scientific instruments
  - E.g. to detect changing external conditions
- Uncertainty quantification in data, models, and experiments
  - E.g. to measure the reliability of simulations involving complex numerical models
- Online fraud detection across massive databases
  - Applicable in many domains (e-commerce, banking, telephony, etc.)
- National security
  - Signal intelligence, anomaly detection, cyber analytics
  - Anti-terrorism, anti-crime
- Health care/medical science
  - Drug design, personalized medicine
  - Epidemiology
  - Systems biology

Example: data-intensive science



Example: Data-intensive Science





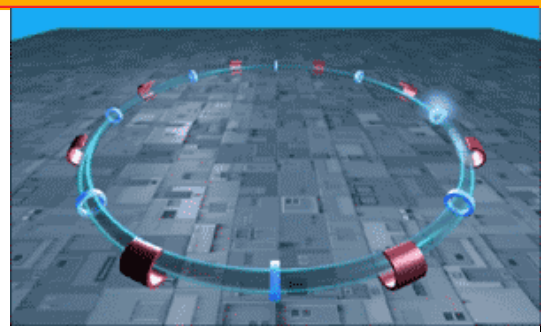

The problem

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

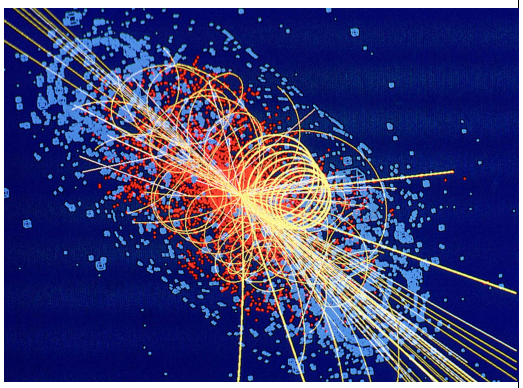
The Office Science Data Management Challenge
USA DoE 2004

Big Data
the good, the bad and the ugly

Higgs Boson Discovery @ CERN



- LHC (Large Hadron Collider)
  - Instrument to study the properties of fundamental particules in physics
  - Produces 15 petabytes / year made available through the LHC Computing Grid to several computing centers, e.g. CC-IN2P3, Lyon
  - Up to 200,000 simultaneous analyses



- 2012: CERN announces that it had discovered a particle that was probably a Higgs boson particle as predicted by the Standard Model of particle physics
- 2014: CERN confirms the discovery

Google Sponsored Search Links

- Google Adwords and Adsense programs
  - Revenue around $50 billion/year from marketing
- Sponsored search uses an auction
  - A pure competition for marketers trying to win access to consumers, i.e. a competition for models of consumers – their likelihood of responding to the ad – and of determining the right bid for the item
- There are around 30 billion search requests a month, perhaps a **trillion events** of history between search providers

Google Flu Trends

- Once of the first success stories
  - Detecting outbreaks two weeks ahead of CDC data in 2008-2009
  - New models are estimating which cities are most at risk for spread of the Ebola and Dengue viruses

20

Google Flu Trends

- Once of the first success stories

OK until 2011, but since 2012, there has been major overestimation (up to 50% of cases)

- The Parable of Google Flu: Traps in Big Data Analysis. David Lazer et al. Science, 2014.

=>

Google Flu Trends and Google Dengue Trends are no longer publishing current estimates of Flu and Dengue fever based on search patterns.

21

Ebola Alerting System

- Echo Mobile (Nairobi, Kenya)
- Challenge
  - Alert the Sierra Leone government on new infections of the Ebola epidemic in different areas of the country
- Solution
  - A reporting SMS-based system that allows citizens and health workers to alert the Central Government Co-ordination Unit
  - The data is analyzed by a system developed by IBM Africa research lab.
- Results
  - Helped the government map the spread of Ebola and quickly respond to new infections while at the same time managing the epidemic in the affected communities

When Big Data goes bad

FORTUNE

November 5, 2013: 1:00 PM ET

How the models underlying today's supercomputing prowess are costing us its success.

By Joshua Klein

The Bad

The Making of a Fly: The Genetics of Animal Design (Paperback)
by Peter A. Lawrence

‹ Return to product information

Always pay through Amazon.com's Shopping Cart or 1-Click.
Learn more about Safe Online Shopping and our safe buying guarantee.

Price at a Glance

List Price: ~~$70.00~~
Used: from **$35.54**
New: from **$1,730,045.91**

Have one to sell? Sell yours here

All | New (2 from $1,730,045.91) | Used (15 from $35.54)

Show: New | Prime offers only (0) — Sorted by: Price + Shipping

New 1-2 of 2 offers

Price + Shipping	Condition	Seller Information	Buying Options
$1,730,045.91 + $3.99 shipping	**New**	Seller: **profnath** Seller Rating: ★★★★★ **93% positive** over the past 12 months. (8,193 total ratings) In Stock. Ships from NJ, United States. Domestic shipping rates and return policy. Brand new, Perfect condition, Satisfaction Guaranteed.	Add to Cart or Sign in to turn on 1-Click ordering.
$2,198,177.95 + $3.99 shipping	**New**	Seller: **bordeebook** Seller Rating: ★★★★★ **93% positive** over the past 12 months. (125,891 total ratings) In Stock. Ships from United States. Domestic shipping rates and return policy. New item in excellent condition. Not used. May be a publisher overstock or have slight shelf wear. Satisfaction guaranteed!	Add to Cart or Sign in to turn on 1-Click ordering.

The Bad

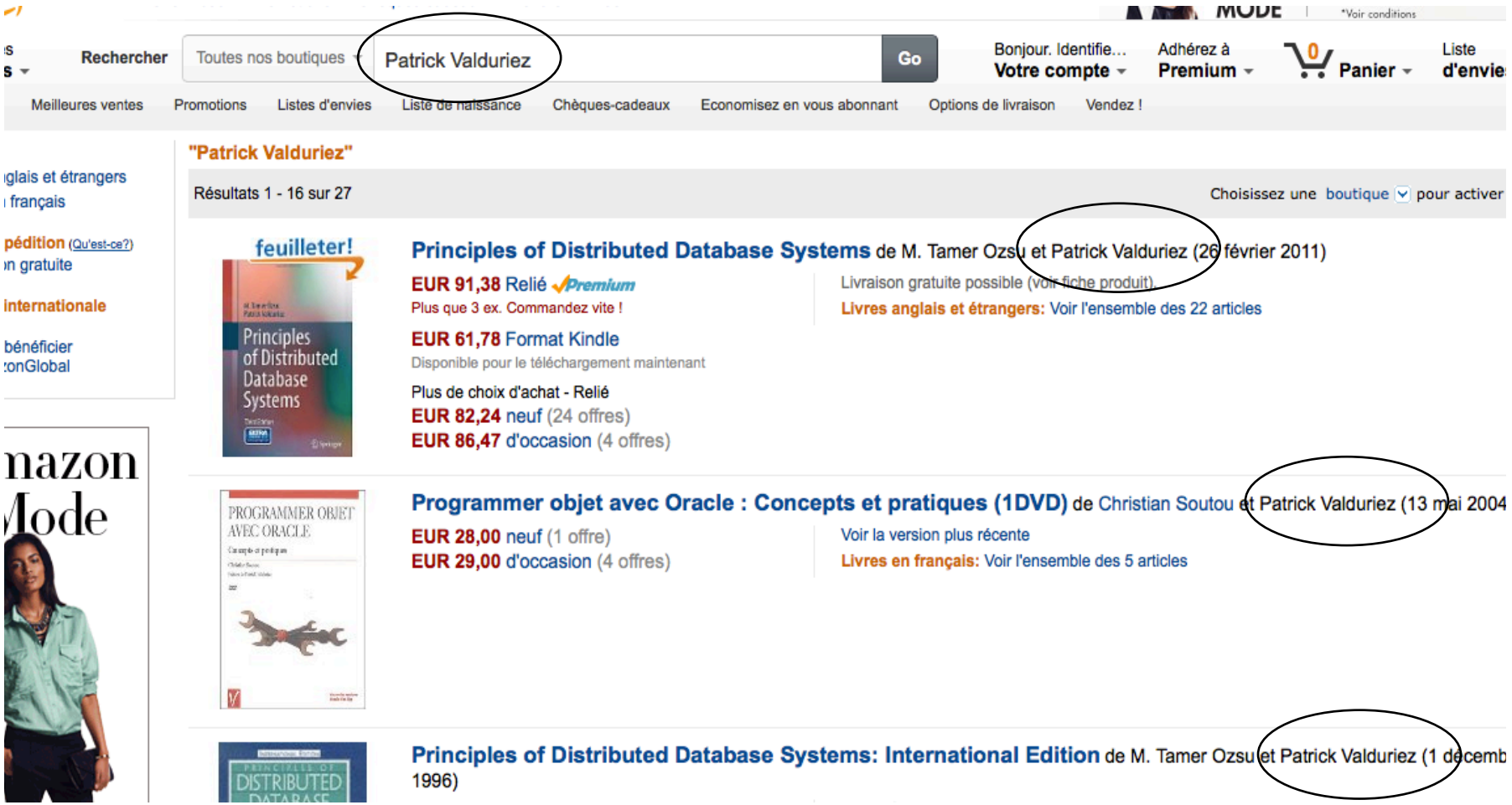
- Excerpts:

  What had happened was that two automated programs, one run by seller "bordeebook" and one by seller "profnath," were engaged in an iterative and incremental bidding war.

  Once a day profnath would raise their price to *x* times bordeebook's listed price. Several hours later, bordeebook would increase their price to *y* times profnath's latest amount.

Problem: over simplified models, but reality is complex!

The Bad (for me)



The Bad (for me)

Object Technology de Mokrane Bouzeghoub, Georges Gardarin et Patrick Valduriez (5 juin 1997)

EUR 1,46 d'occasion (5 offres)

Livres anglais et étrangers: Voir l'ensemble des 22 articles

Ozzy Osbourne - Talking. de Patrick Valduriez (31 août 2003)

EUR 28,50 neuf (1 offre)

EUR 23,74 d'occasion (2 offres)

Livres anglais et étrangers: Voir l'ensemble des 22 articles

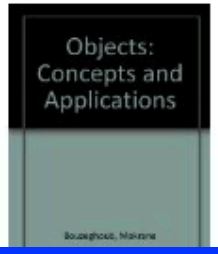
Ozzy Osbourne. Fucking Mad. Die Story zu seinen Songs. de Patrick Valduriez (30 juin 2003)

EUR 21,46 neuf (1 offre)

EUR 6,26 d'occasion (2 offres)

Livres anglais et étrangers: Voir l'ensemble des 22 articles

The Bad (for me)



Object Technology de Mokrane Bouzeghoub, Georges Gardarin et Patrick Valduriez (5 juin 1997)

EUR 1,46 d'occasion (5 offres) | **Livres anglais et étrangers:** Voir l'ensemble des 22 articles

Problem: how do I get complete deletion of wrong information?

Ozzy Osbourne. Fucking Mad. Die Story zu seinen Songs. de Patrick Valduriez (30 juin 2003)

EUR 21,46 neuf (1 offre)

EUR 6,26 d'occasion (2 offres) | **Livres anglais et étrangers:** Voir l'ensemble des 22 articles

The Ugly


amazon.co.uk
Shop by Department
Search
Keep Calm and Rape a Lot (Black) Fine Jersey T-Shirt
Solid Gold Bomb
£30.00
£14.99 - £16.99
Select
Sizing info
Light Blue
KEEP CALM AND RAPE A LOT
American Apparel 2001 Fine Jersey T-Shirt
T-Shirt Made & Printed in the USA
Super Easy Solid Gold Bomb Size Guarantee!
Save up to 70% on winter fashion for men, women and children, including brands like Levi's,
selection here.
See more product promotions

The Ugly

A tiny company in Worcester, Mass., has paid the ultimate price for posting offensive T-shirts for sale online.

Fierce public backlash brought down Solid Gold Bomb, which made headlines in March for offering shirts that said "Keep Calm and Rape a Lot." The company closed its doors last week and let go its remaining three employees.

The Ugly

- Excerpts:

Solid Gold Bomb, the company that made the shirt, wasn't necessarily aware that it was even selling it. Solid Gold Bomb's business isn't in artfully designing T-shirts. Instead, it writes code that takes libraries of words that slot into popular phrases (such as "Keep Calm and Carry On," which enjoyed a brief mimetic popularity online) to make derivations that get dropped onto a template of a T-shirt and automatically get posted as an Amazon item for sale.

Their mistake was overlooking a single word in a list of 4,000 or so others.

The Ugly

- Excerpts:

Solid Gold Bomb, the company that made the shirt, wasn't necessarily aware that it was even selling it. Solid Gold Bomb's business isn't in artfully designing T-shirts. Instead, it writes code that takes libraries of words that template of a T-shirt and automatically get posted as an Amazon item for sale.

Problem: context-independent model, but context does matter!

Their mistake was overlooking a single word in a list of 4,000 or so others.

Cloud & Big Data

Why the Hype?

- To make big money with big data
  - Intel, IBM, Microsoft, Oracle, Google, Facebook, Amazon, …
- In a big market
  - $18 billion in 2013, $24 billion in 2016
  - Source: International Data Corp. (IDC)

Why the Hype?


IBM
developerWorks
Technical topics
Evaluation software
Community
Events
developerWorks > Technical topics > Big data > Technical library >
ATOS TO ACQUIRE BULL TO CREATE A EUROPEAN GLOBAL LEADER IN CLOUD, CYBERSECURITY, AND BIG DATA
Share this content
« Previous press release
Atos to acquire Bull to create a European global leader in Cloud, Cybersecurity, and Big Data
With 2013 revenue of €1,262 million and operations across more than 50 countries, Bull is a leading player in Cloud, Cybersecurity, and Big Data, and the European global leader in High-Performance Computing.
The combination will create the #1 European player in Cloud operations and a leading Cybersecurity solutions provider.
new discoveries.

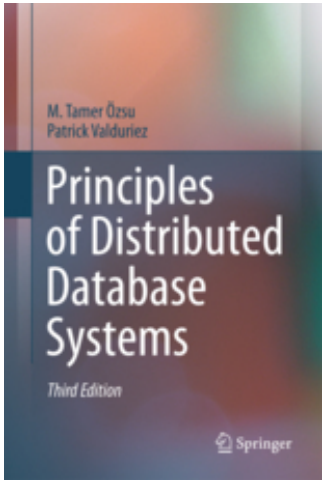
A Marriage of Convenience?

- Cloud and big data have different goals
  - Big data aims at added value and operational performance
  - Cloud targets flexibility and reduced cost
- But they can help each other by
  1. Encouraging organizations to outsource more and more strategic internal data in the cloud
  2. Get value out of it, e.g. by integrating their data with external data, through big data analytics at affordable cost

Technologies for Data Science

Big Data Management

- Principle: exploit data distribution and parallelism, typically in cluster grid and cloud environments
  - Semantic data integration (using ontologies)
  - Data partitioning (sharding) and indexing
  - Parallel & in-memory data processing
  - Replication and failover
  - Exploiting new technologies: multicore CPU/ GPU, virtual machines, flash memory, RDMA, etc.

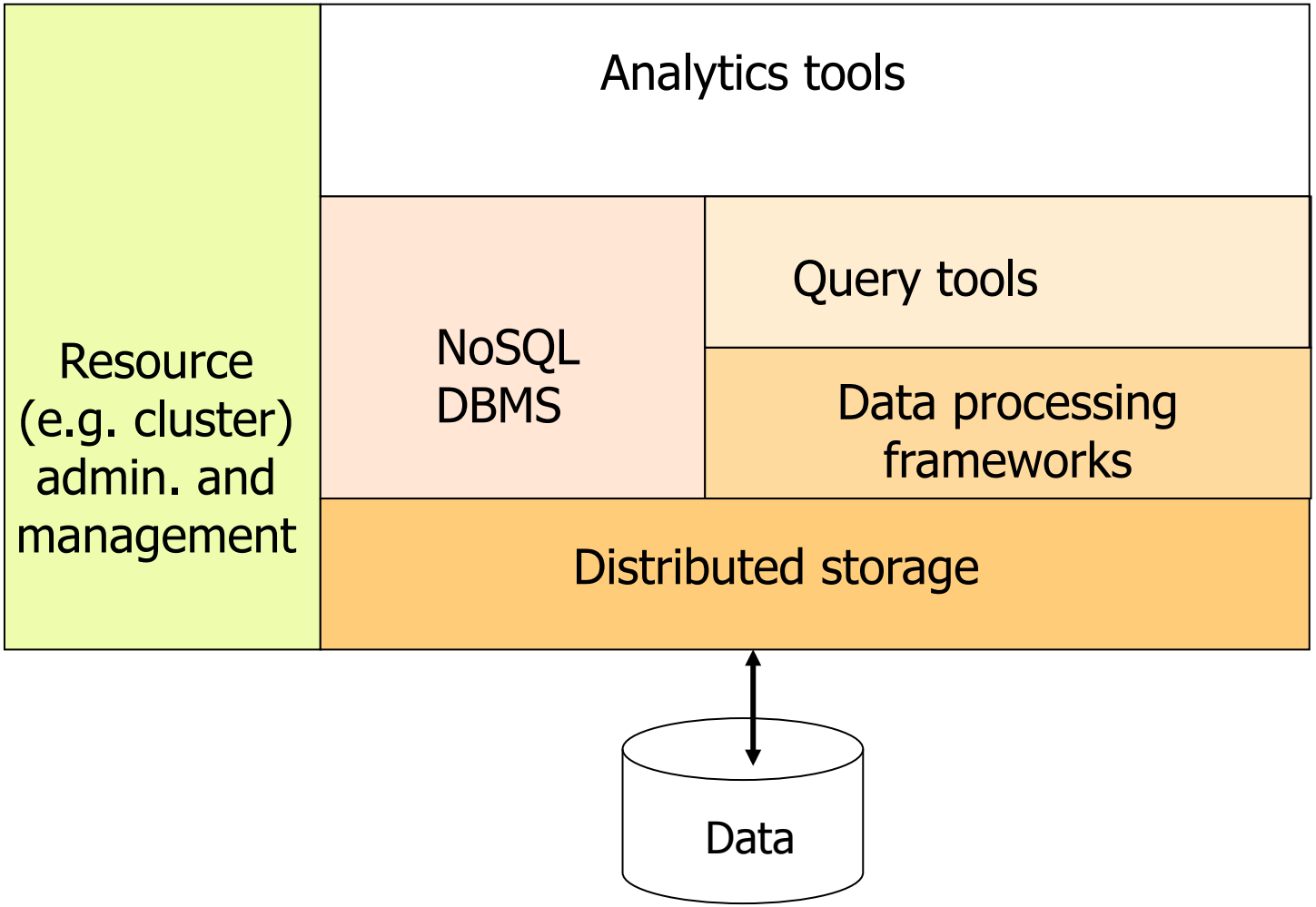


Data Analytics

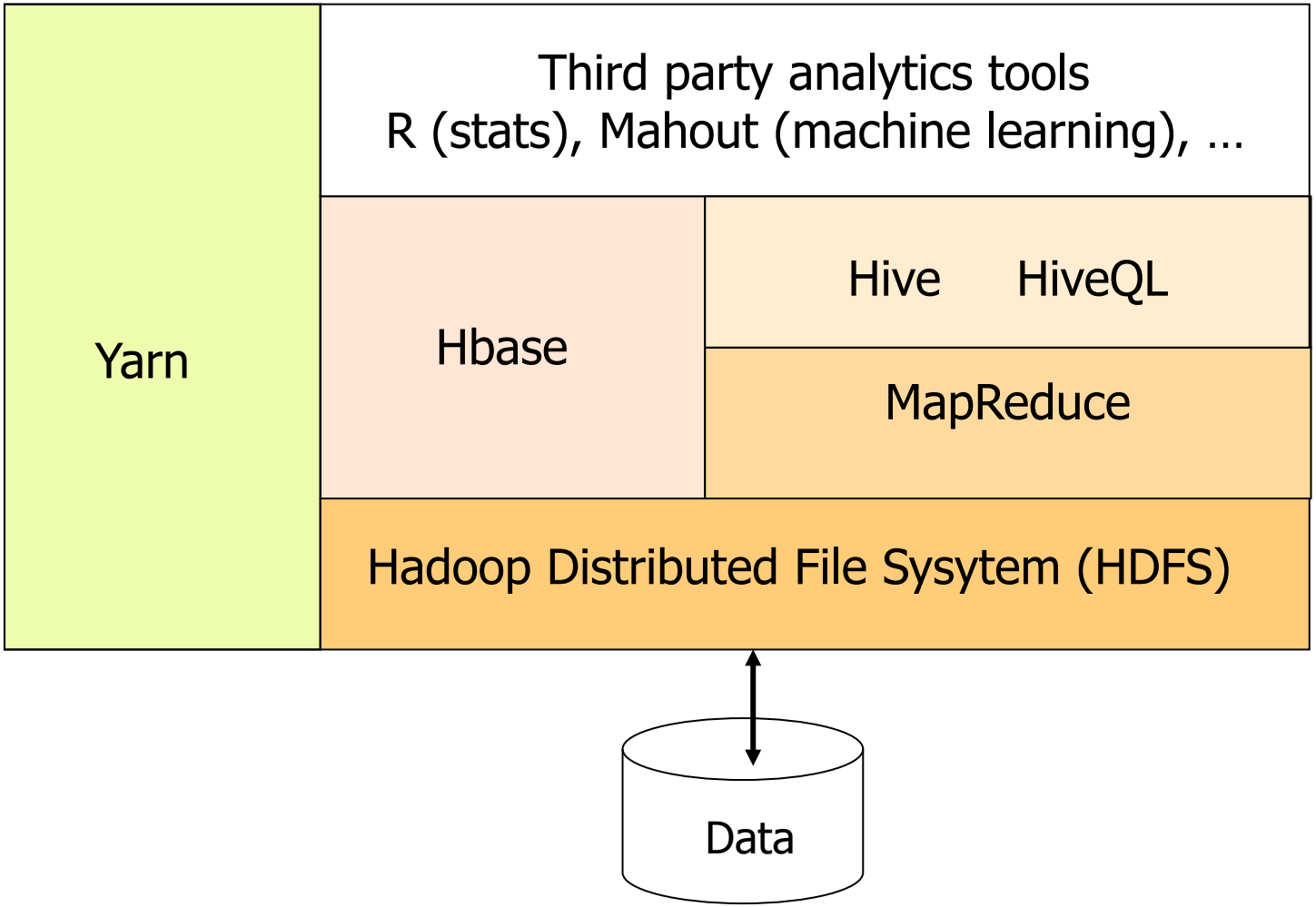
- Big data of interest
  - Categorical data, e.g. the customer bought the item or not
  - Continuous data, e.g. a temperature
  - Time series, e.g. temperature evolution during one year
  - Graphs, e.g. social network data
  - Uncertain data, e.g. temperature readings
  - Data streams, e.g. sensor data

- Methods for big data analytics
  - Pattern mining
  - Clustering
  - Outlier detection
  - Probabilistic data mining
  - Privacy preserving data mining
  - Visual analytics

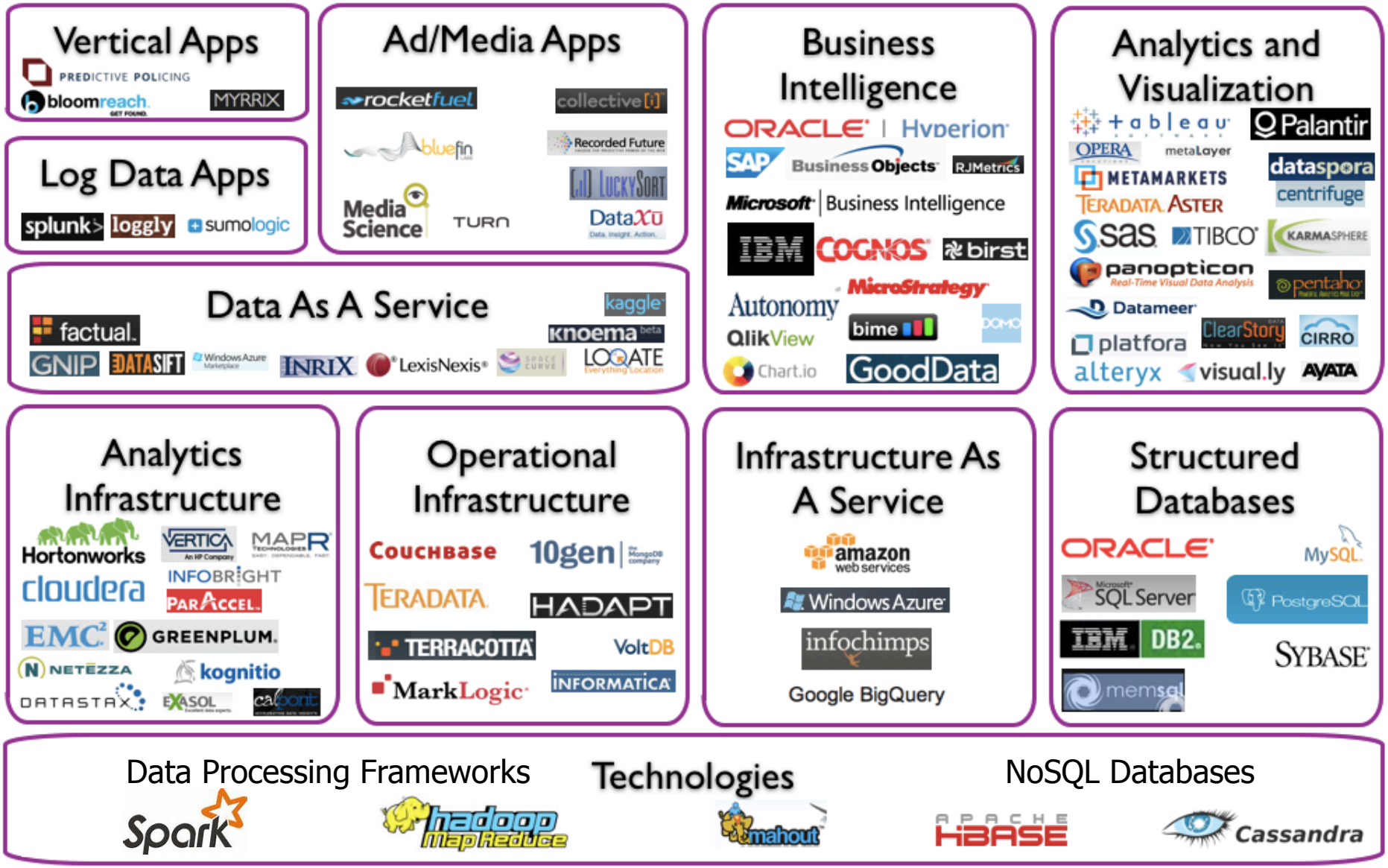
A New Software Stack



Hadoop Architecture



Cloud & Big Data Landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib

Cloud & Big Data Landscape

dave@vcdave.com

blogs.forbes.com/davefeinleib

A New Breed of Systems: multistores

- Also called *Polystores**
- Provide integrated access to multiple, heterogeneous cloud data stores such as NoSQL, HDFS and RDBMS
- Great for integrating structured (relational) data and big data
- Much more difficult than federated database systems
- A major area of research & development

*Michael Stonebraker (ACM Turing award 2015). The Case for Polystores. July 2015

CloudMdsQL Multistore

1. Carlyna Bondiombouy, Boyan Kolev, Oleksandra Levchenko, Patrick Valduriez. Integrating Big Data and Relational Data with a Functional SQL-like Query Language. *DEXA 2015* (extended version to appear in Springer *TLDKS* journal).

2. Carlyna Bondiombouy, Patrick Valduriez. Query Processing in Cloud Multistore Systems: an overview. *Int. Journal of Cloud Computing*, 38 pages, to appear, 2016.

3. Boyan Kolev, Patrick Valduriez, Carlyna Bondiombouy, Ricardo Jiménez-Peris, Raquel Pau, José Pereira. CloudMdsQL: Querying Heterogeneous Cloud Data Stores with a Common Language. *Distributed and Parallel Databases*, online, 43 pages, 2015.

4. Boyan Kolev, Carlyna Bondiombouy, Oleksandra Levchenko, Patrick Valduriez, Ricardo Jiménez-Peris, Raquel Pau, José Pereira. Design and Implementation of the CloudMdsQL Multistore System. *CLOSER 2016.*

5. Boyan Kolev, Carlyna Bondiombouy, Patrick Valduriez, Ricardo Jiménez-Peris, Raquel Pau, José Pereira. The CloudMdsQL Multistore System. *SIGMOD 2016.*

Opportunities and Risks

Opportunities

- Cost reduction (vs. traditional data warehousing)
  - New open source technologies (Hadoop, Spark, etc.)
  - Cloud services
- Faster, better decision making
  - Realtime data processing (e.g. online fraud detection)
  - Data crowdsourcing (e.g. Ebola example) to produce timely, precise data
- Better knowledge discovery
  - Virtuous circle between machine learning and big data
- New data products and services
  - Two-sided markets
  - Digital medicine, e-agriculture, etc.

Risks

- Data security
  - The bigger your data, the bigger the target it presents to attackers
- Data privacy
  - Personal data can be misused by people who have responsibility for analytics, and may violate data protection laws
- Cost
  - Data collection, aggregation, storage, analysis, and reporting
  - Data security and privacy
- Bad analytics
  - Oversimplified or wrong models (see "when big data goes bad")
  - Misinterpreting the patterns shown by the data and drawing wrong conclusions
- Bad data
  - Many projects start off wrong by collecting irrelevant, out of date, or erroneous data

Thanks