From Databases to Data Science

Impact on Information Systems

Patrick Valduriez

Data versus Information

- Data
  - Elementary definition of a fact
    - E.g. temperature, exam grade, account balance, message, photo, transaction, etc.
  - Can be very simple, and taken in isolation, not very useful
  - But the integration with other data becomes useful

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

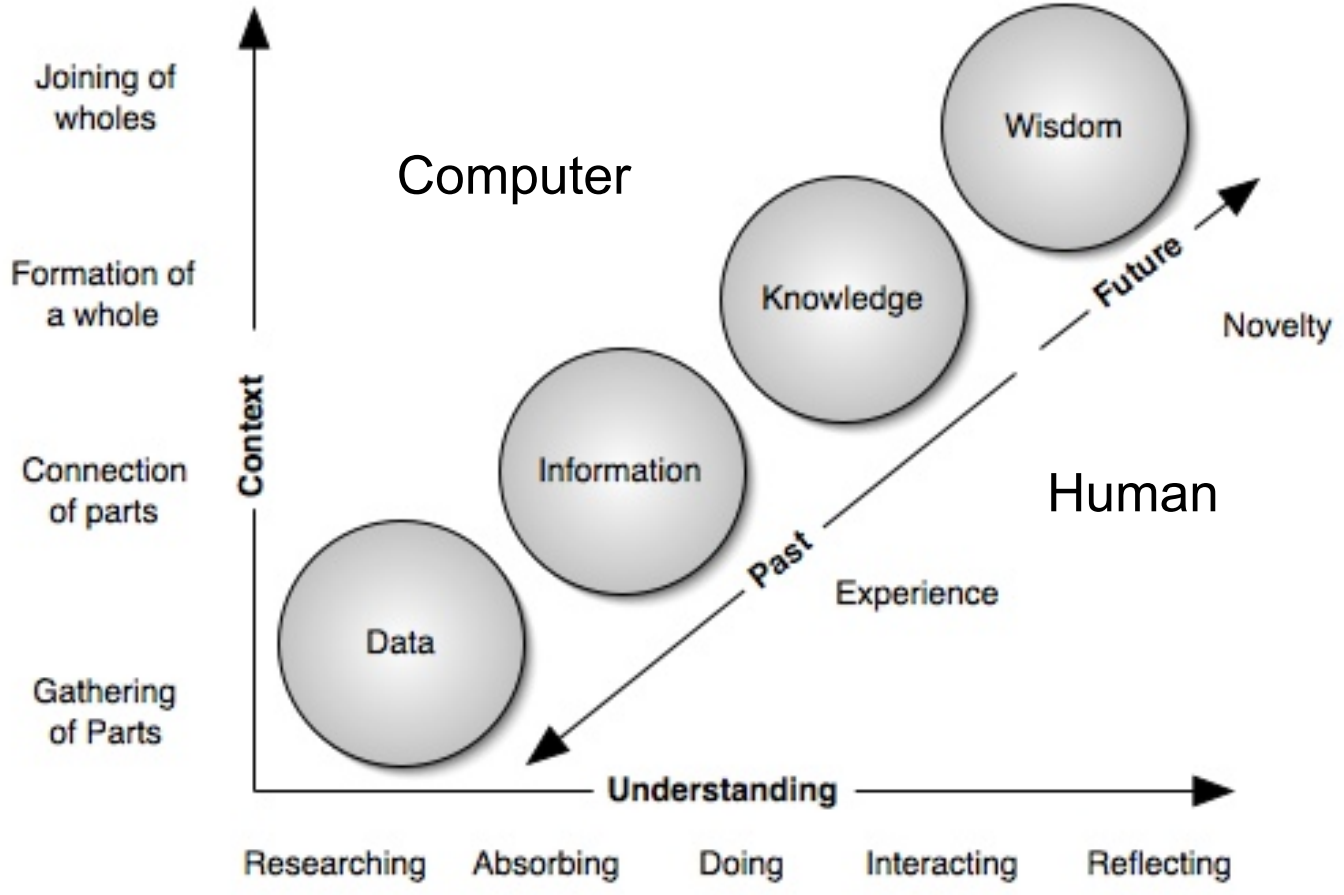
- Information
  - Obtained by interpretation and analysis of a collection of data to yield sense in a given context
  - Can be very useful to understand the world
    - E.g. climate evolution, ranking of a student, etc.

Data and Algorithm

> "Content without method leads to fantasy,
> method without content to empty sophistry."
>
> Johann Wolfgang von Goethe (Maxims and Reflections, 1892)

- The better the datasets, the better the machine learning algorithms
- Examples
  - 1997: IBM Deep Blue defeats Garry Gasparov
    - Negascout planning algorithm (1983)
    - Dataset of 700 thousand chess games (1991)
  - 2016: Google Alphago defeats Lee Sedol (4-1)
    - Monte Carlo method based algorithm (from the 1940's) and neural network
    - Dataset of 30 million of go moves

The Continuum of Understanding

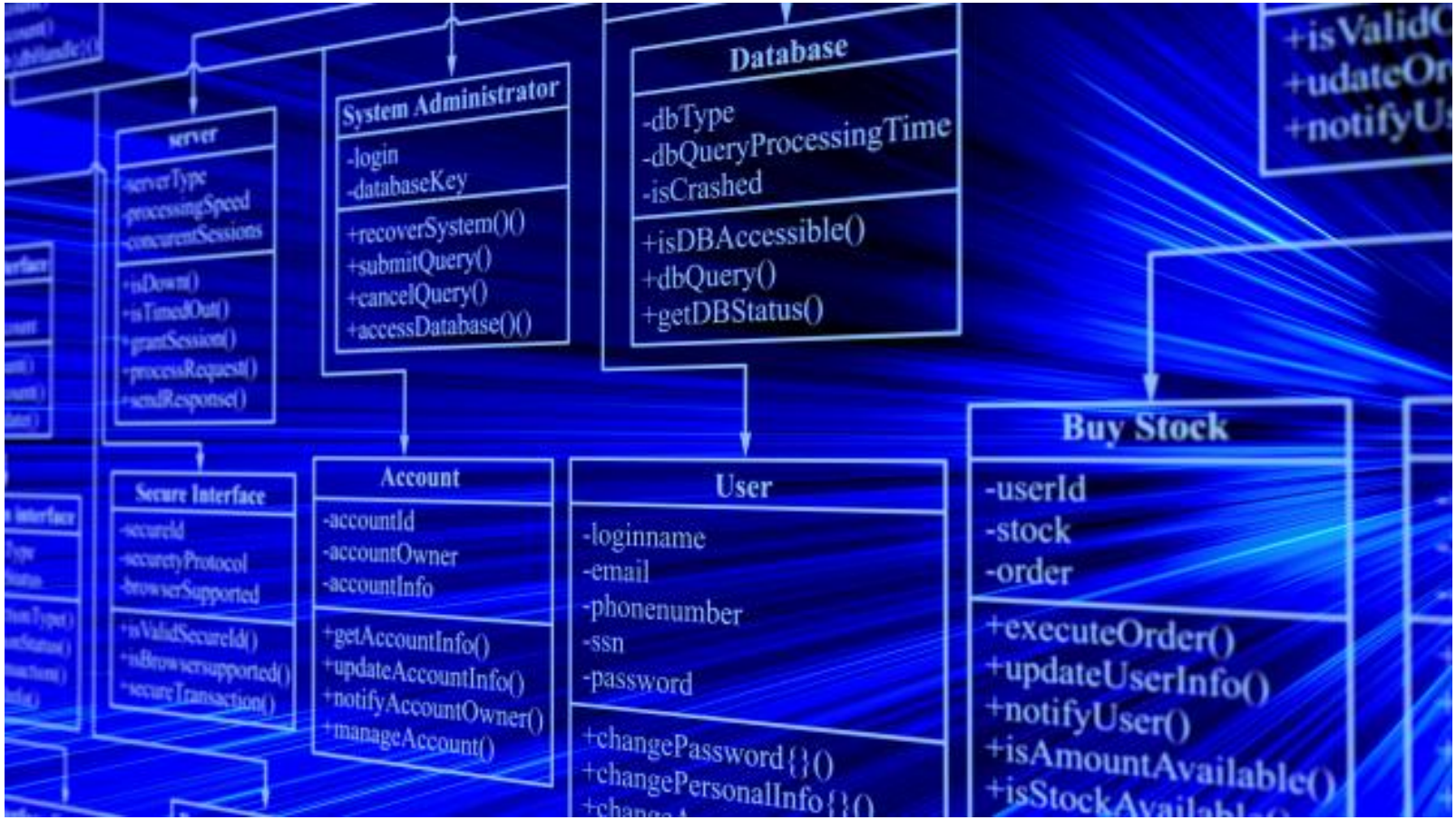


- The more the data, the better the understanding
  - If we (humans) do a good job

Outline

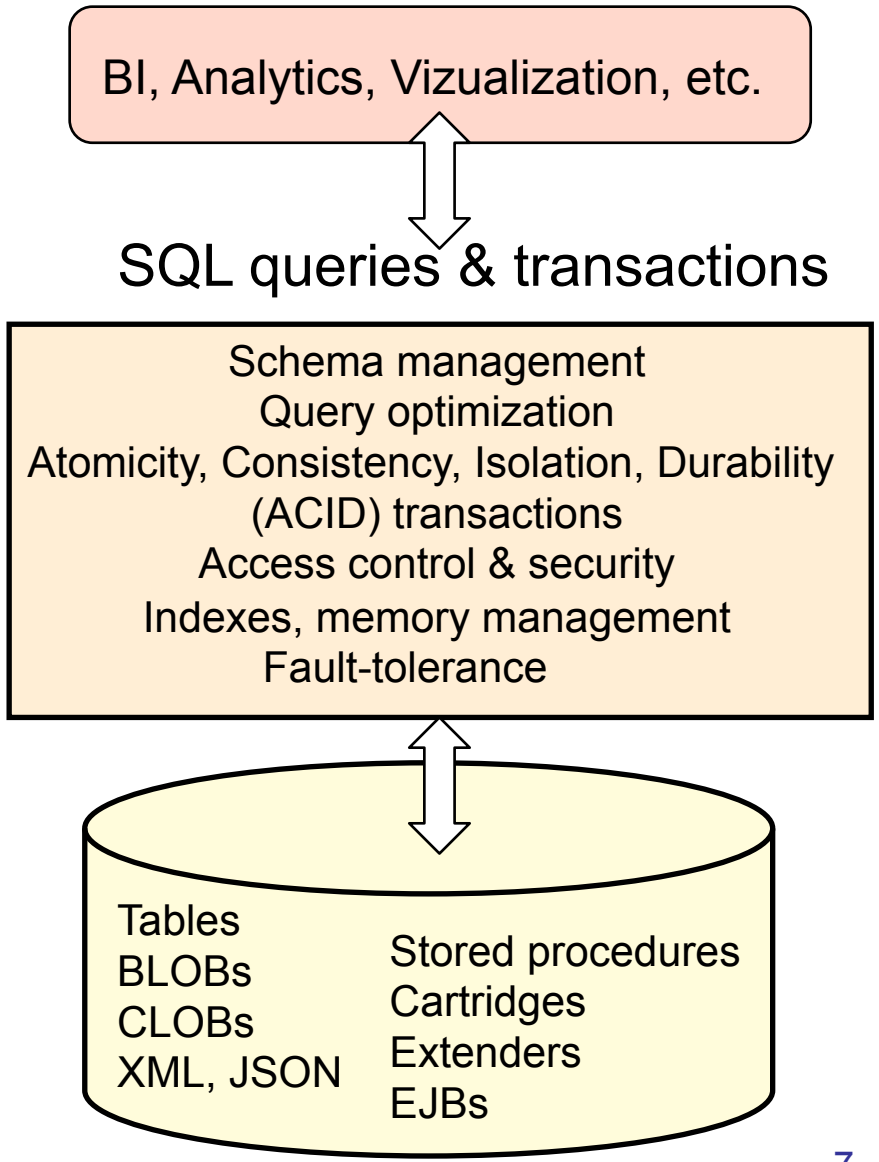
1. Databases
2. Data science
3. The good, the bad and the ugly
4. Impact on information systems
5. Opportunities and risks

Databases


server
-serverType
-processingSpeed
-concurentSessions
+isDown()
+isTimedOut()
+grantSession()
+processRequest()
+sendResponse()
System Administrator
-login
-databaseKey
+recoverSystem()()
+submitQuery()
+cancelQuery()
+accessDatabase()()
Database
-dbType
-dbQueryProcessingTime
-isCrashed
+isDBAccessible()
+dbQuery()
+getDBStatus()
Secure Interface
-secureId
-securityProtocol
-browserSupported
+isValidSecureId()
+isBrowsersupported()
+secureTransaction()
Account
-accountId
-accountOwner
-accountInfo
+getAccountInfo()
+updateAccountInfo()
+notifyAccountOwner()
+manageAccount()
User
-loginname
-email
-phonenumber
-ssn
-password
+changePassword{}()
+changePersonalInfo{}()
Buy Stock
-userId
-stock
-order
+executeOrder()
+updateUserInfo()
+notifyUser()
+isAmountAvailable()

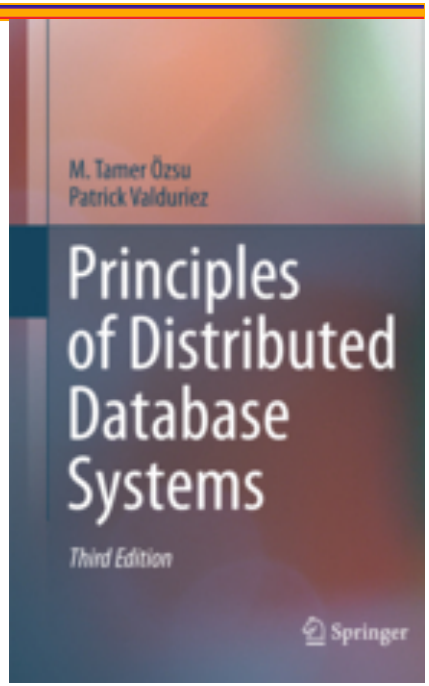
Relational DBMS (RDBMS)

- Appeared in the 1980's
  - After 10 years of research
- SQL: the universal database language
- ACID transactions
  - Strong database consistency
- Many services and tools
- At the heart of any information system



Distributed Databases

- RDBMS have succeeded in leveraging distributed architectures
  - Multi-tier: from 2-tier to n-tier client-server
  - SOA with Web services (SOAP, …)
    - Great for software vendors and integrators
    - Heavy and complex for IS
  - WOA with Web protocols (HTTP, REST, …)
    - Large adoption from web giants
    - On the rise
  - Massively parallel computing
    - Data warehousing
  - Integration in the cloud
    - Data as a service



The CAP Theorem

- Polemical topic
  - "A database can't provide consistency AND availability during a network partition"
  - Argument used by NoSQL to trade consistency for availability
- Two different points of view
  - Relational databases
    - Consistency is essential
      - ACID transactions
  - Distributed systems
    - Service availability is essential
      - Inconsistency tolerated by the user, e.g. web cache

What is the CAP Theorem?

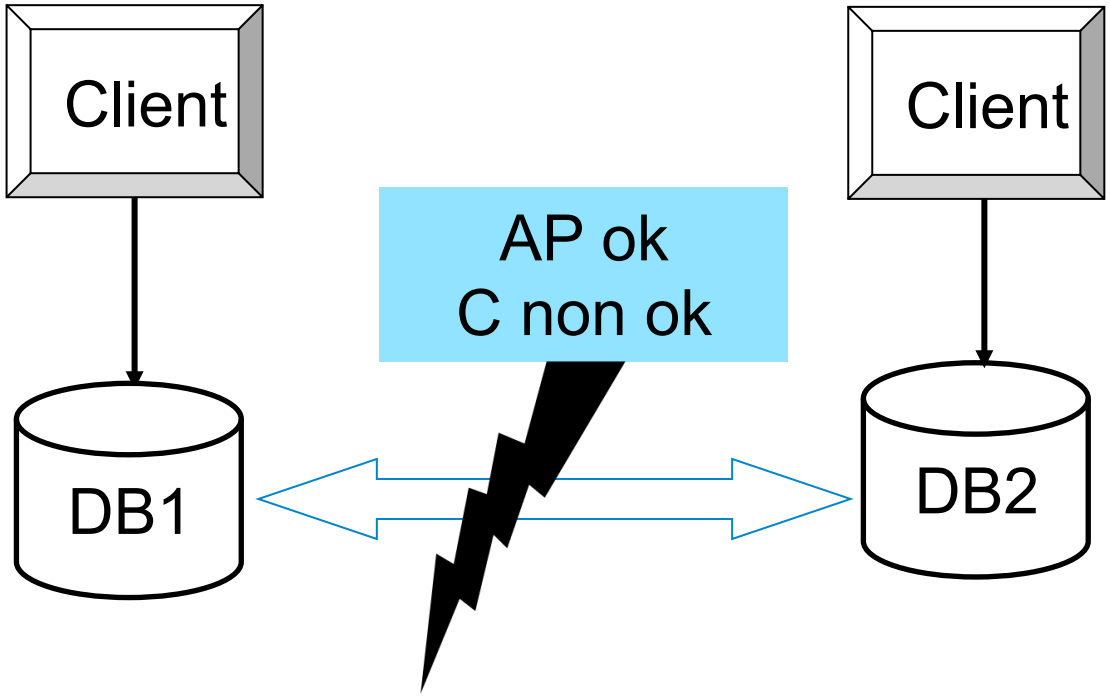
- The desirable properties of a distributed system
  - ***C****onsistency*: all nodes see the same data values at the same time
  - ***A****vailability*: all requests get an answer
  - ***P****artition tolerance*: the system keeps functioning in case of network failure
- History
  - At the PODC 2000 conference, Brewer (UC Berkeley) conjectures that one can have only two properties at the same time
  - In 2002, Gilbert and Lynch (MIT) prove the conjecture, which becomes a theorem

Strong vs Eventual Consistency

- Strong consistency (ACID)
  - All nodes see the same data values at the same time
- Eventual consistency
  - Some nodes may see different data values at the same time
  - But if we stop injecting updates, the system reaches strong consistency

- Illustration with symmetric, asynchronous replication in databases

Symmetric, Asynchronous Replication



But we have eventual consistency

- After reconnexion (and resolution of update conflicts), strong consistency can be obtained

Limits of RDBMS

- The "one size fits all" approach has reached the limits[1]
  - Loss of performance, simplicity and flexibility for applications with specific, tight requirements
  - New specialized DBMS engines better
- For many apps, RDBMS provide both
  - Too much: ACID transactions, complex query language, lots of tuning knobs
  - Too little: lacks specific optimizations for data analysis, flexible programming model, flexible schema, scalability

[1] "One Size Fits All": An Idea Whose Time Has Come and Gone. Michael Stonebraker, ACM Turing Award 2015.

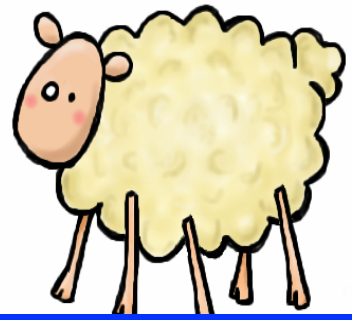
Data Science

Data Science: definition

- Data science
  - The science of making sense of data
  - The use of data management, 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

Hard to find data scientists !



New training programs all over the world

Should we all be teaching “Intro to Data Science” instead of “Intro to Databases”?

ACM SIGMOD panel 2014

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 and manage
- Velocity
  - Continuous data streams are being produced
  - Makes it hard to process online
- Variety
  - Different data formats, different semantics, uncertain data, multiscale data, etc.
  - Makes it hard to integrate
- 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?

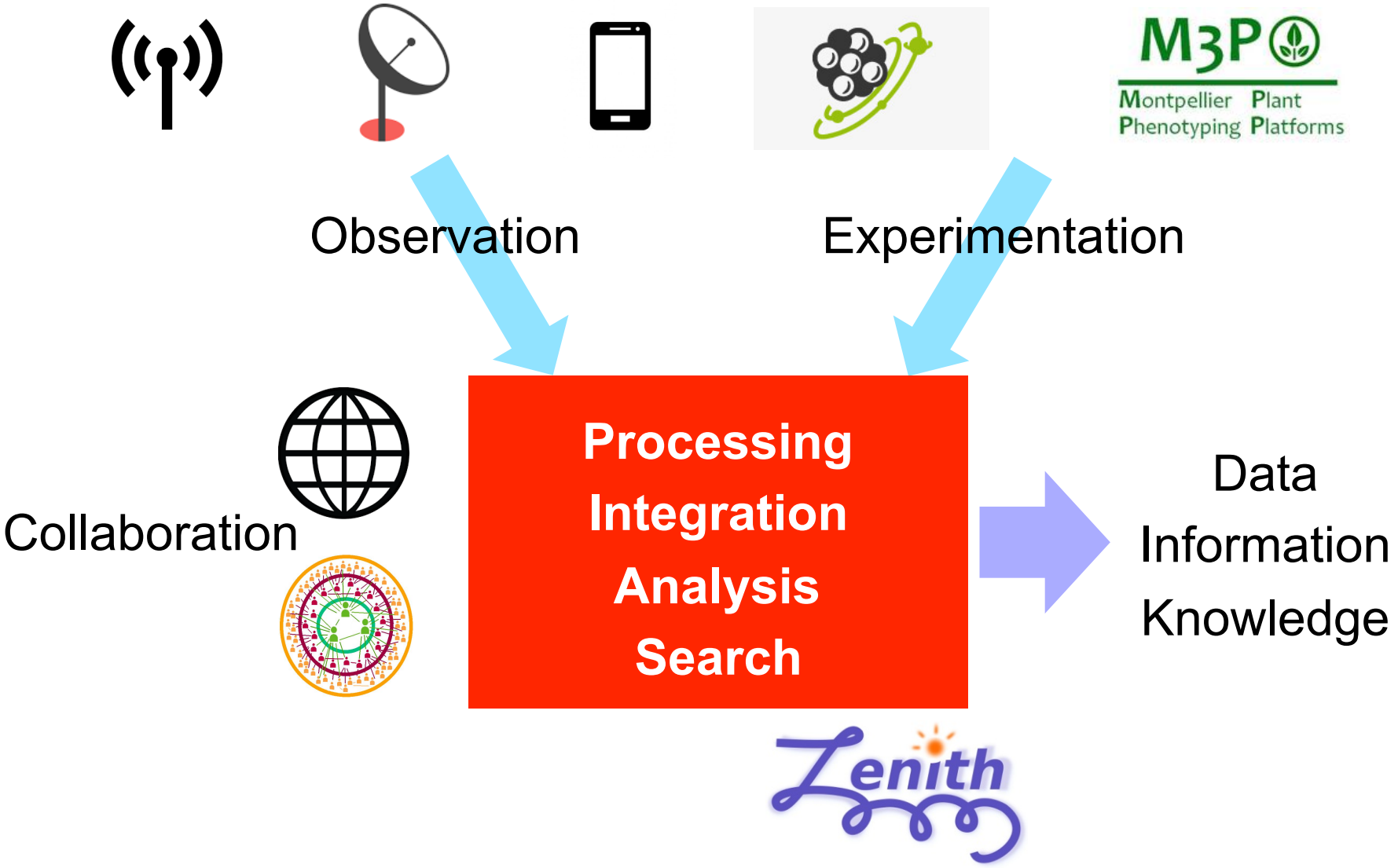
Big Data Analytics (BDA)

- 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
- Online fraud detection across massive databases
  - Applicable in many domains (e-commerce, banking, telephony, etc.)
- National security
  - Signal intelligence, cyber analytics
- Real-time processing and analysis of raw data from high-throughput scientific instruments
  - E.g. to detect changing external conditions
- Health care/medical science
  - Drug design, personalized medicine

Example: data-intensive science



Example: data-intensive science



The problem

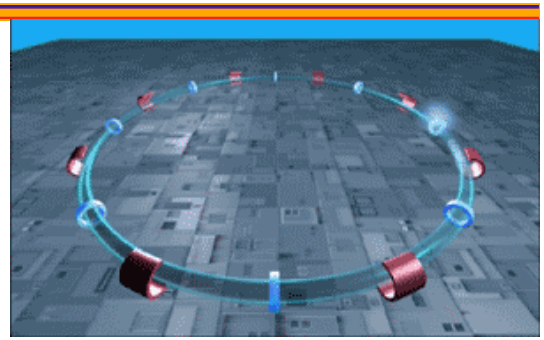
"*Scientists are spending most of their time manipulating, organizing, finding and moving data, instead of researching. And it's going to get worse*"

The Office Science Data Management Challenge
USA DoE 2004

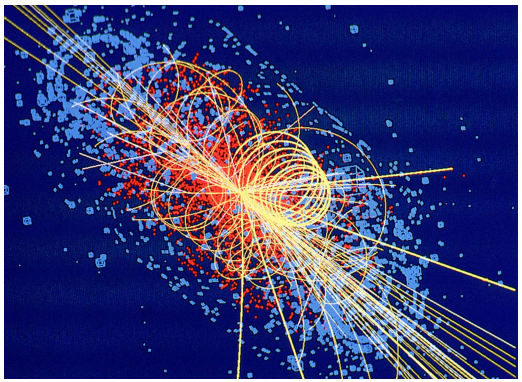
Zenith

Data Science
the good, the bad and the ugly

The good: Higgs Boson @ 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



- High Boson discovery
  - 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

The good: Google Sponsored Search Links

- Google Adwords and Adsense programs
  - Revenue around $50 billion/year from marketing
  - The user defines its maximum cost-per-click bid (max. CPC bid), the most she's willing to pay for a click on her ad
- 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

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

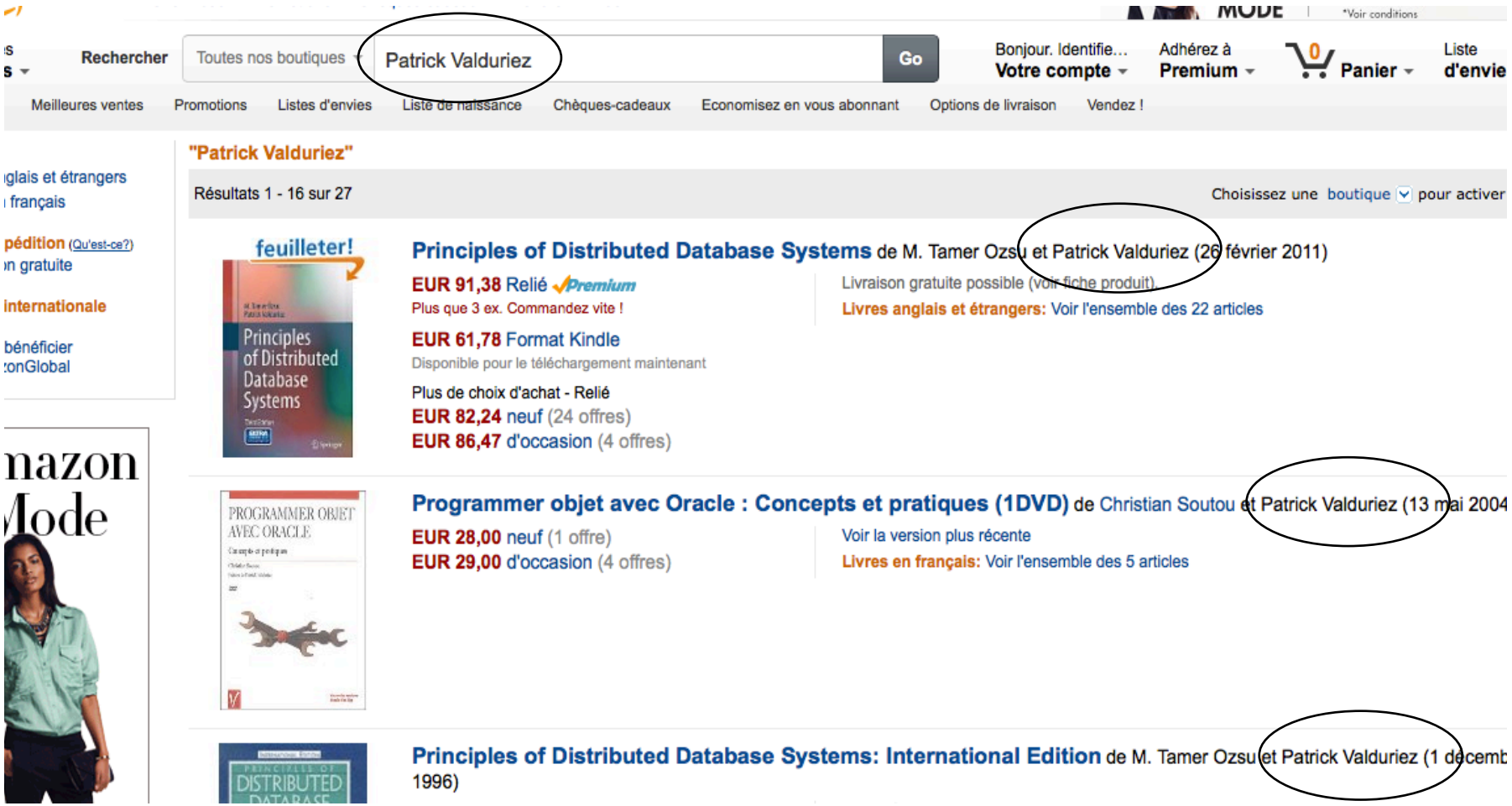
- Excerpts:

  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.

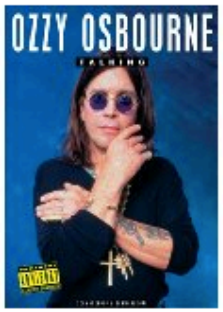
Problem: over simplified models, but reality is complex!

The Bad (for me)



The Bad (for me)

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



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 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

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

template of a T-shirt and automatically get posted as an Amazon item for sale.

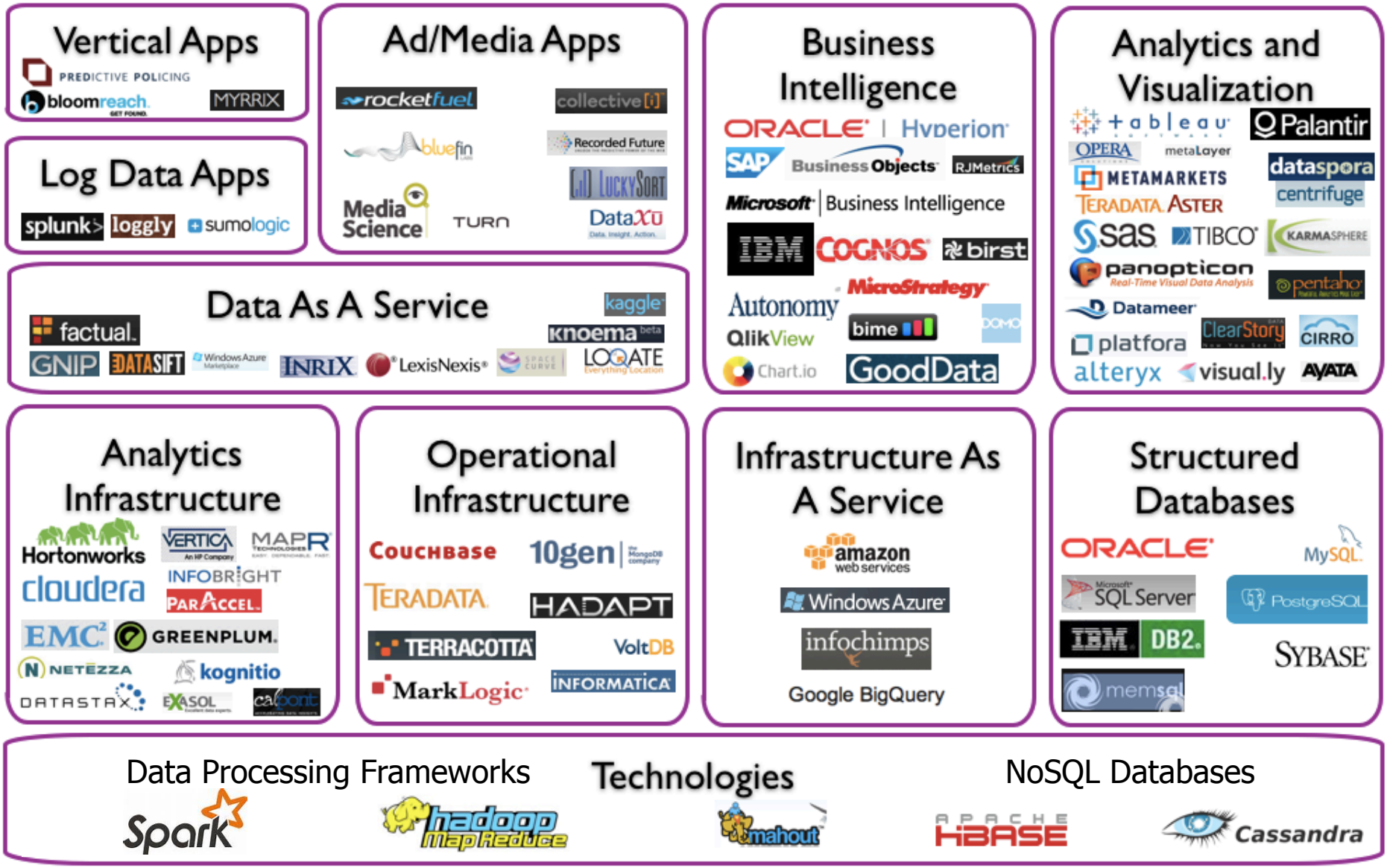
Their mistake was overlooking a single word in a list of 4,000 or so others.

Impact on Information Systems

Data Requirements of a Modern IS

- Automate data collection
  - Including from external data streams
- Ensure that data is understandable, trusted, visible, accessible, optimized for use, and interoperable
  - Processes for strategy, planning, modeling, security, access control and quality
- Realtime data analytics and automated reaction to external events
- High data quality and assurance
  - Enabling information sharing, and fostering data reuse by minimizing data redundancy

Cloud & Big Data Landscape



dave@vcdave.com

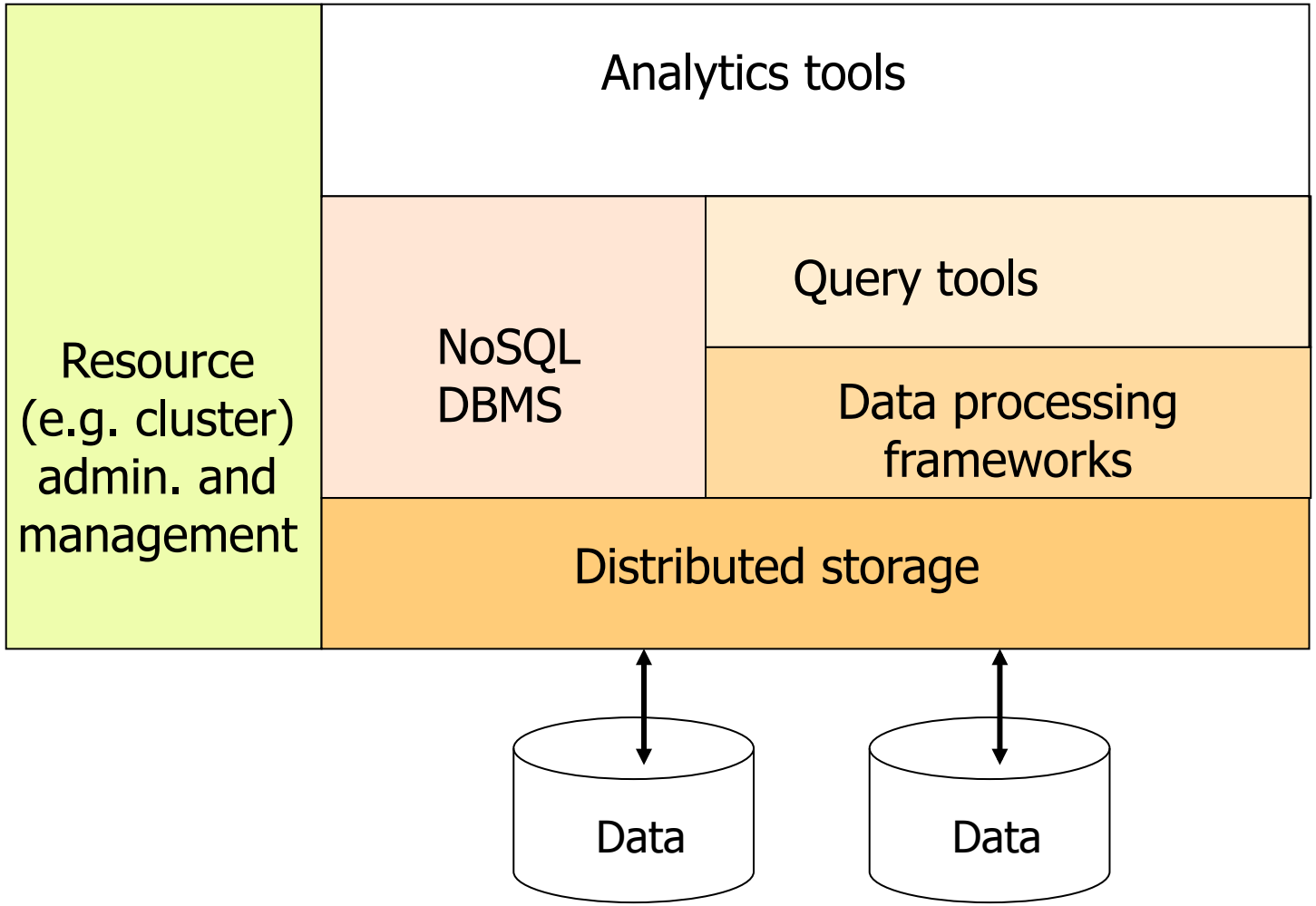
blogs.forbes.com/davefeinleib

Cloud & Big Data Landscape

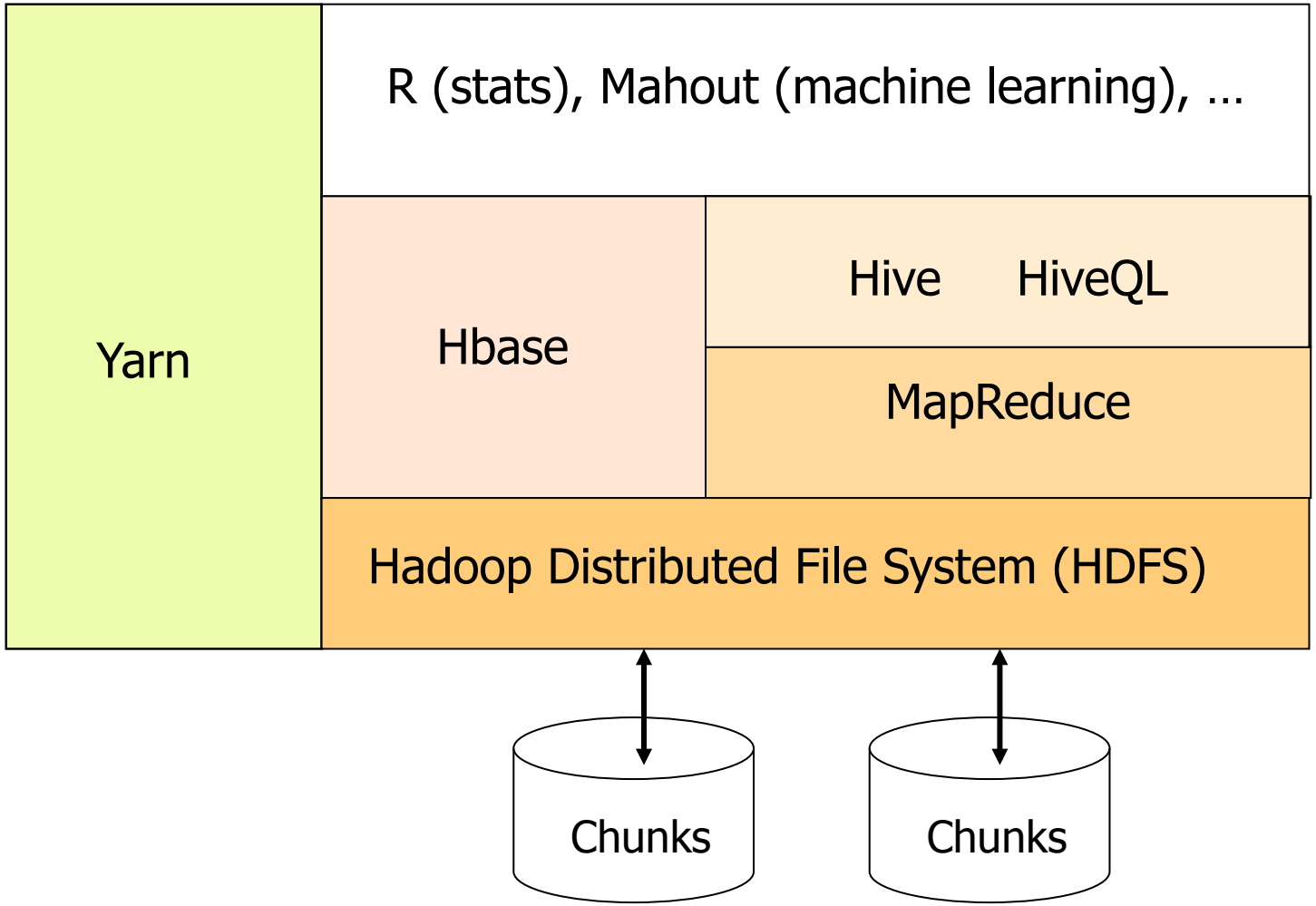
dave@vcdave.com

blogs.forbes.com/davefeinleib

A New Software Stack



Hadoop Architecture



Big Data Frameworks

- Many popular frameworks
  - MapReduce, Spark, Flink, Pregel, Giraph, etc.
- Alternative to traditional BI tools
  - And integration in big data offers of RDBMS vendors
    - IBM InfoSphere BigInsights, Microsoft HDInsight, Oracle
- Takeover of data analytics by the programmer
- Keeps evolving, strong impact from research
  - MapReduce abandonned by Google, replaced by Cloud Dataflow
  - Hadoop MapReduce challenged by Apache Spark

NoSQL DBMS

- Specific data model
  - Key-value: Amazon DynamoDB, Apache Cassandra
  - Document: MongoDB, CouchDB, Expresso
  - Tabular: Google Bigtable, Hbase
  - Graph: Neo4J, Sparcity
  - Multi-model: OrientDB
- Relaxing of RDBMS properties
  - Complete SQL, ACID transactions, data independence
- For more
  - Simplicity (schema, language)
  - Flexibility (integration within programming language)
  - Scalability
  - Performance

NewSQL DBMS

- Hybrids relational/NoSQL
  - Google F1, CockroachDB, SAP HANA, MemSQL, LeanXcale, VoltDB
- Google F1
  - F1 inspired by the term *hybrid filial 1* in genetics
  - For the AdWords killer app
    - + de 100 Terabytes, 100K requests per sec
    - Scalability problem with the MySQL / cluster solution
  - Objective: "F1 combines high availability, the scalability of NoSQL systems like Bigtable, and the consistency and usability of traditional SQL databases."
  - Geographic distribution of the data centers
    - Synchronous replication between data centers for high availability
    - Sharding and parallel processing within data center

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 to produce timely, precise data
- Better knowledge discovery
  - Virtuous circle between machine learning and big data
- New data products and services
  - Two-sided markets
  - E-health, 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