Data Science

Opportunities and Risks

Patrick Valduriez

Data versus Information

- Data
  - Elementary definition of a fact
    - E.g. temperature, exam grade, account balance, message, photo, transaction, etc.
  - Can be complex
    - E.g. a satellite image
  - Can also 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

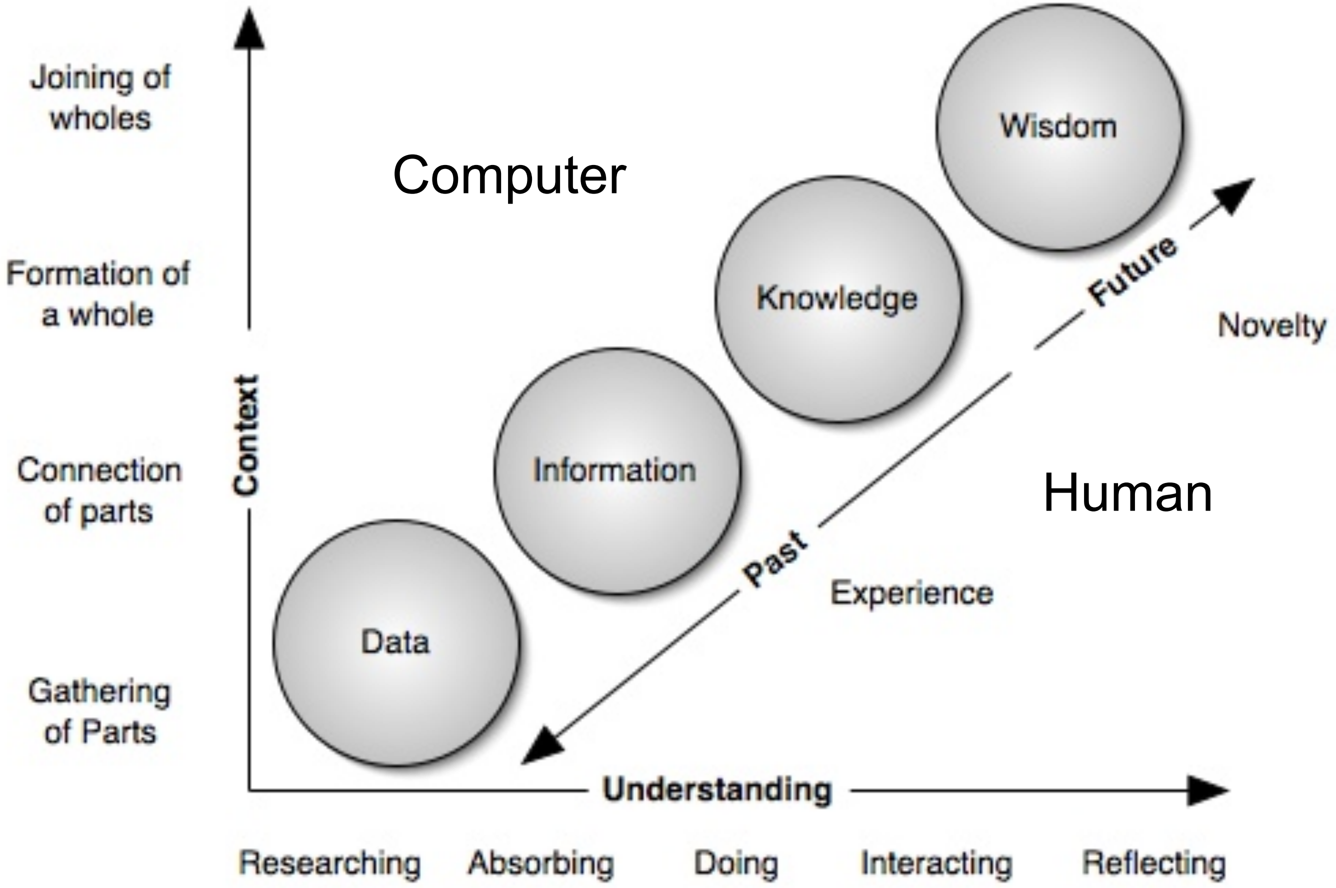
- Information
  - Obtained by interpretation and analysis of data to yield sense in a given *context*
  - Can be very useful to understand the world
    - E.g. climate evolution, ranking of a student, etc.

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
- Milestones
  - 1997: IBM Deep Blue defeats Chess world champion Garry Kasparov
    - Negascout planning algorithm (1983)
    - Dataset of 700 thousands of chess games (1991)
  - 2016: Google Alphago defeats Go master Lee Sedol (4-1)
    - Monte Carlo method based algorithm (from the 1940's) and neural network
    - Dataset of 30 millions of go moves

The Continuum of Understanding



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

Outline

1. Data science
2. The good, the bad and the ugly
3. Technologies for data science
4. HPC & big data analysis
5. Opportunities and risks

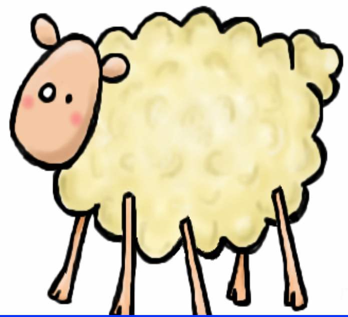
Data Science

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, process, 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

M3P Montpellier Plant Phenotyping Platforms

Observation

Experimentation

Collaboration

Processing
Integration
Analysis
Search

Data
Information
Knowledge

Zenith

Example: data-intensive science

The problem

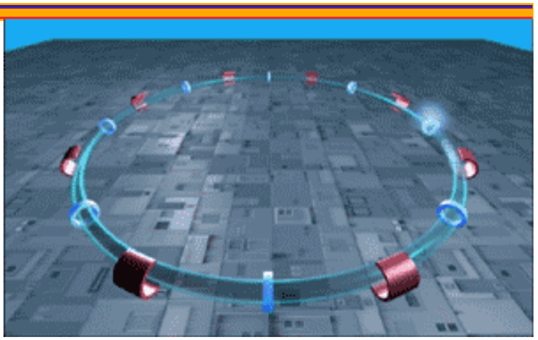
"*Scientists are spending most of their time manipulating, organizing, finding and moving data, instead of researching. And it's going to get worse*"

The Office Science Data Management Challenge (USA DoE 2004)

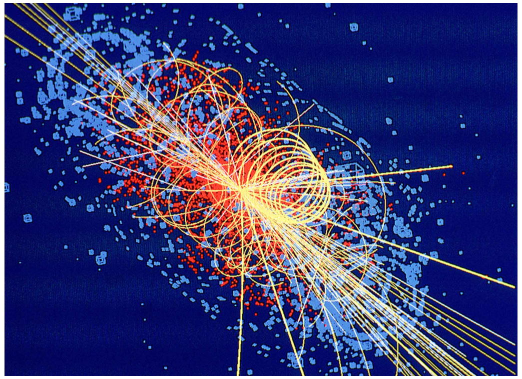
In bioinformatics, the time to deal with data can be well above 50% (IBC annual review 2017)

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

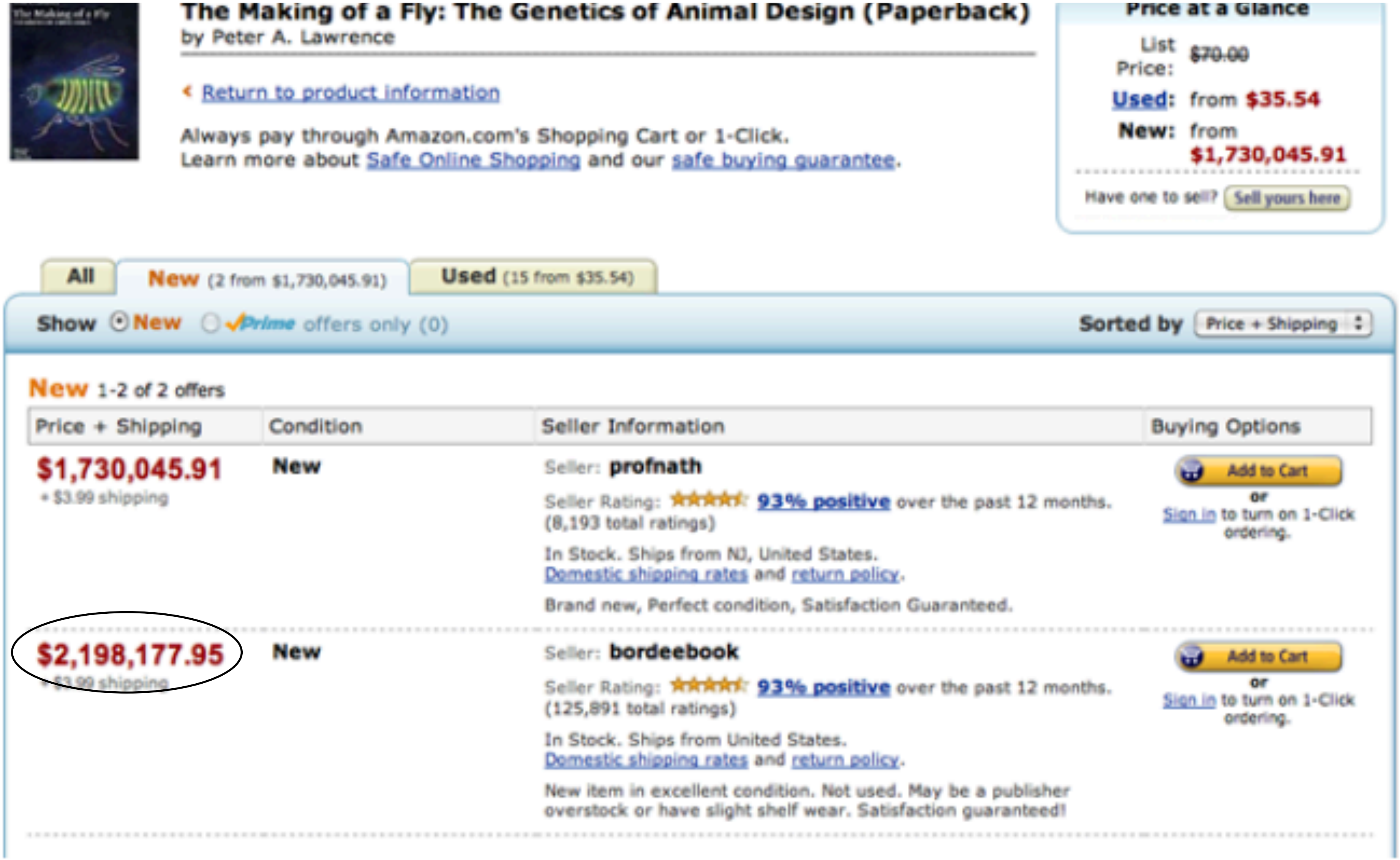
November 5, 2013: 1:00 PM ET

Recommend 32

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.

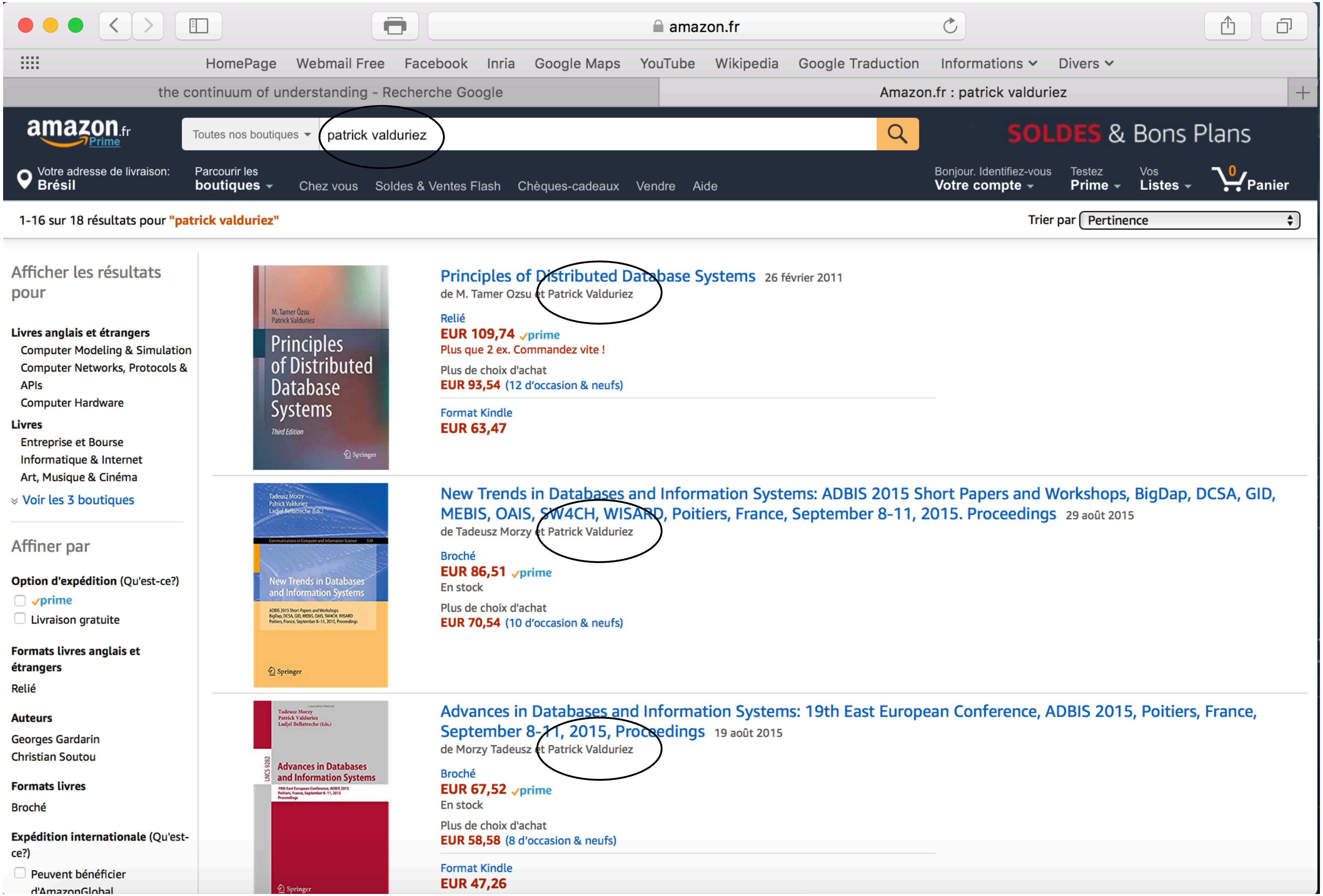
The Bad

- Excerpts:

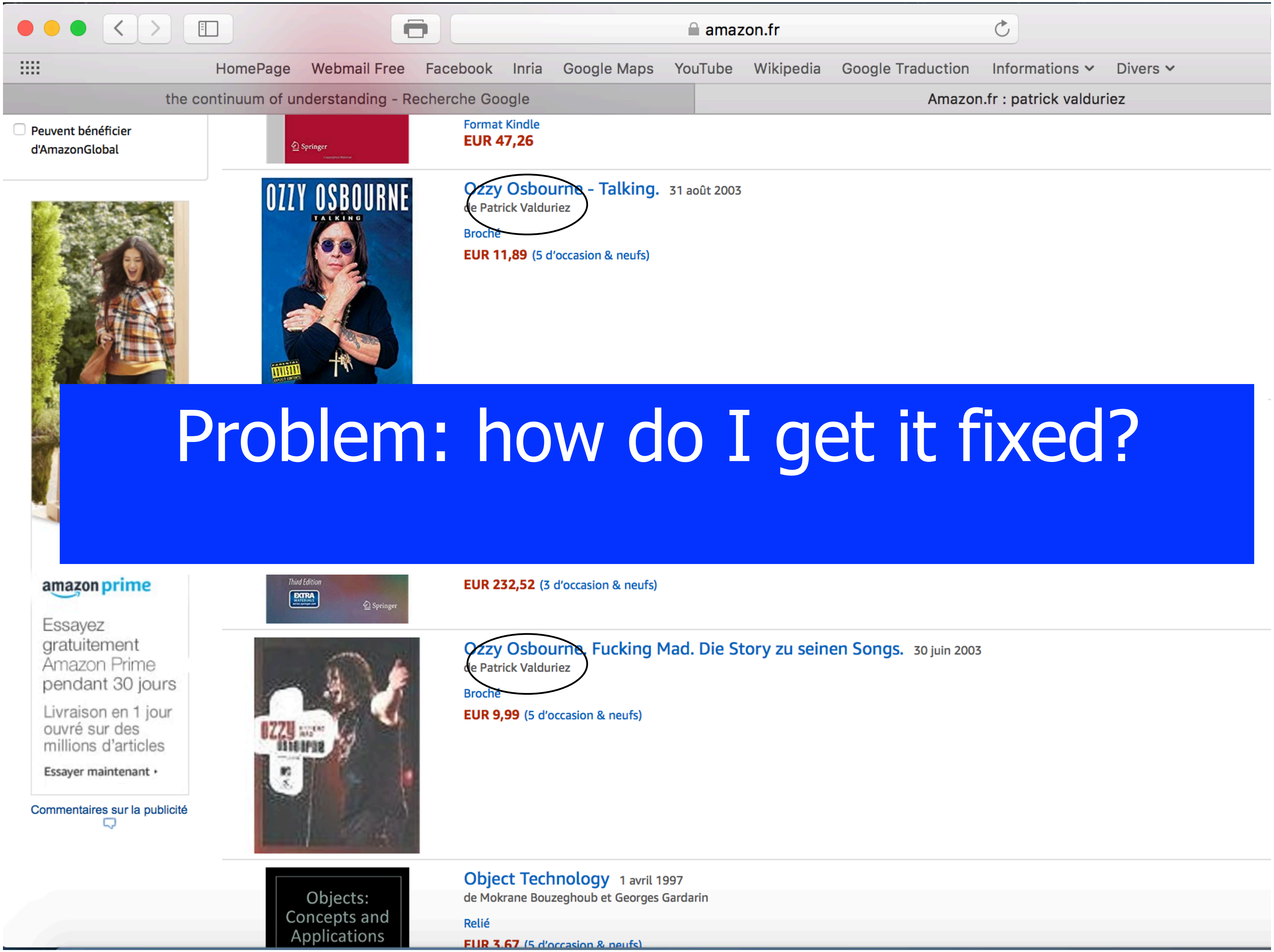
  What had happened was that two automated programs, one run by seller "bordeebook" and one by

  Problem: over simplified models, but reality is complex!

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


amazon.fr
HomePage
Webmail Free
Facebook
Inria
Google Maps
YouTube
Wikipedia
Google Traduction
Informations
Divers
the continuum of understanding - Recherche Google
Amazon.fr : patrick valduriez
amazon.fr Prime
Toutes nos boutiques
patrick valduriez
SOLDES & Bons Plans
Votre adresse de livraison: Brésil
Parcourir les boutiques
Chez vous
Soldes & Ventes Flash
Chèques-cadeaux
Vendre
Aide
Bonjour. Identifiez-vous Votre compte
Testez Prime
Vos Listes
0 Panier
1-16 sur 18 résultats pour "patrick valduriez"
Trier par Pertinence
Afficher les résultats pour
Livres anglais et étrangers
Computer Modeling & Simulation
Computer Networks, Protocols & APIs
Computer Hardware
Livres
Entreprise et Bourse
Informatique & Internet
Art, Musique & Cinéma
Voir les 3 boutiques
Affiner par
Option d'expédition (Qu'est-ce?)
prime
Livraison gratuite
Formats livres anglais et étrangers
Relié
Auteurs
Georges Gardarin
Christian Soutou
Formats livres
Broché
Expédition internationale (Qu'est-ce?)
Peuvent bénéficier d'AmazonGlobal
M. Tamer Özsu
Patrick Valduriez
Principles of Distributed Database Systems
Third Edition
Springer
Principles of Distributed Database Systems 26 février 2011
de M. Tamer Ozsu et Patrick Valduriez
Relié
EUR 109,74 prime
Plus que 2 ex. Commandez vite !
Plus de choix d'achat
EUR 93,54 (12 d'occasion & neufs)
Format Kindle
EUR 63,47
New Trends in Databases and Information Systems
New Trends in Databases and Information Systems: ADBIS 2015 Short Papers and Workshops, BigDap, DCSA, GID, MEBIS, OAIS, SW4CH, WISARD, Poitiers, France, September 8-11, 2015. Proceedings 29 août 2015
de Tadeusz Morzy et Patrick Valduriez
Broché
EUR 86,51 prime
En stock
Plus de choix d'achat
EUR 70,54 (10 d'occasion & neufs)
Advances in Databases and Information Systems
Advances in Databases and Information Systems: 19th East European Conference, ADBIS 2015, Poitiers, France, September 8-11, 2015, Proceedings 19 août 2015
de Morzy Tadeusz et Patrick Valduriez
Broché
EUR 67,52 prime
En stock
Plus de choix d'achat
EUR 58,58 (8 d'occasion & neufs)
Format Kindle
EUR 47,26

Problem: how do I get it fixed?



The Ugly

amazon
Search
Keep Calm and Rape a Lot (Black) Fine Jersey T-Shirt
Solid Gold Bomb
£14.99 - £16.99
Select
Sizing info
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!

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.

Technologies

Data Science Landscape

Vertical Apps
Predictive Policing, bloomreach, Myrrix

Log Data Apps
splunk, loggly, sumologic

Ad/Media Apps
rocketfuel, collective[i], bluefin, Recorded Future, LuckySort, Media Science, Turn, DataXu

Business Intelligence
Oracle | Hyperion, SAP Business Objects, RJMetrics, Microsoft | Business Intelligence, IBM Cognos, birst, Autonomy, MicroStrategy, QlikView, bime, Domo, Chart.io, GoodData

Analytics and Visualization
tableau, Palantir, Opera, metaLayer, Metamarkets, dataspora, centrifuge, Teradata Aster, sas, TIBCO, Karmasphere, panopticon, pentaho, Datameer, platfora, ClearStory, Cirro, alteryx, visual.ly, Ayata

Data As A Service
factual, kaggle, knoema, GNIP, DataSift, Windows Azure Marketplace, INRIX, LexisNexis, Space Curve, Locate

Analytics Infrastructure
Hortonworks, Vertica, MapR, cloudera, Infobright, ParAccel, EMC², Greenplum, Netezza, kognitio, DataStax, Exasol, calpont

Operational Infrastructure
Couchbase, 10gen | MongoDB company, Teradata, Hadapt, Terracotta, VoltDB, MarkLogic, Informatica

Infrastructure As A Service
amazon web services, Windows Azure, infochimps, Google BigQuery

Structured Databases
Oracle, MySQL, Microsoft SQL Server, PostgreSQL, IBM DB2, Sybase, memsql

Technologies
Data Processing Frameworks: Spark, hadoop MapReduce, mahout

NoSQL Databases: Apache HBase, Cassandra

dave@vcdave.com

blogs.forbes.com/davefeinleib

Data Science Landscape

Easy to get lost

Many diverse solutions

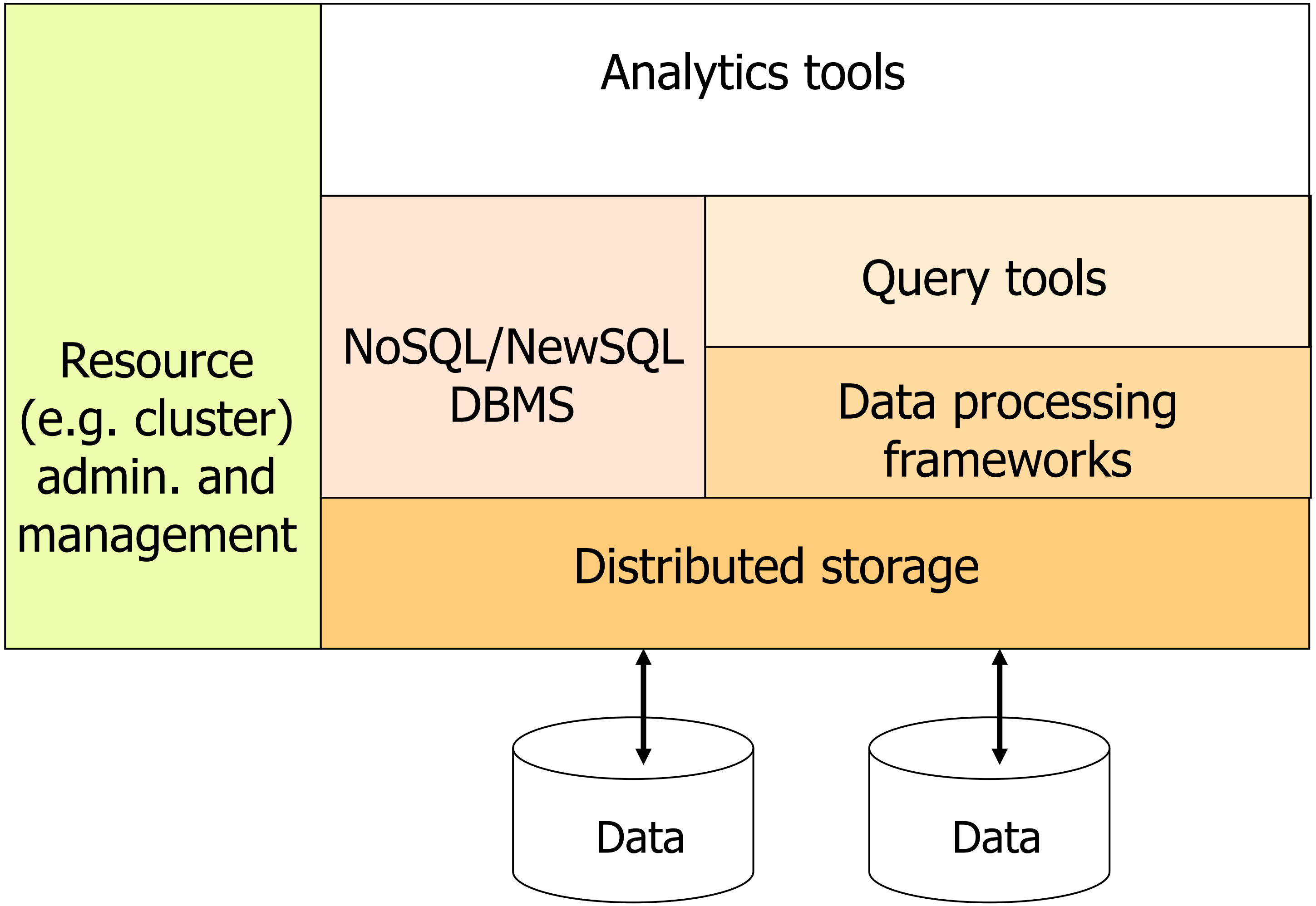
No standards

Keeps evolving

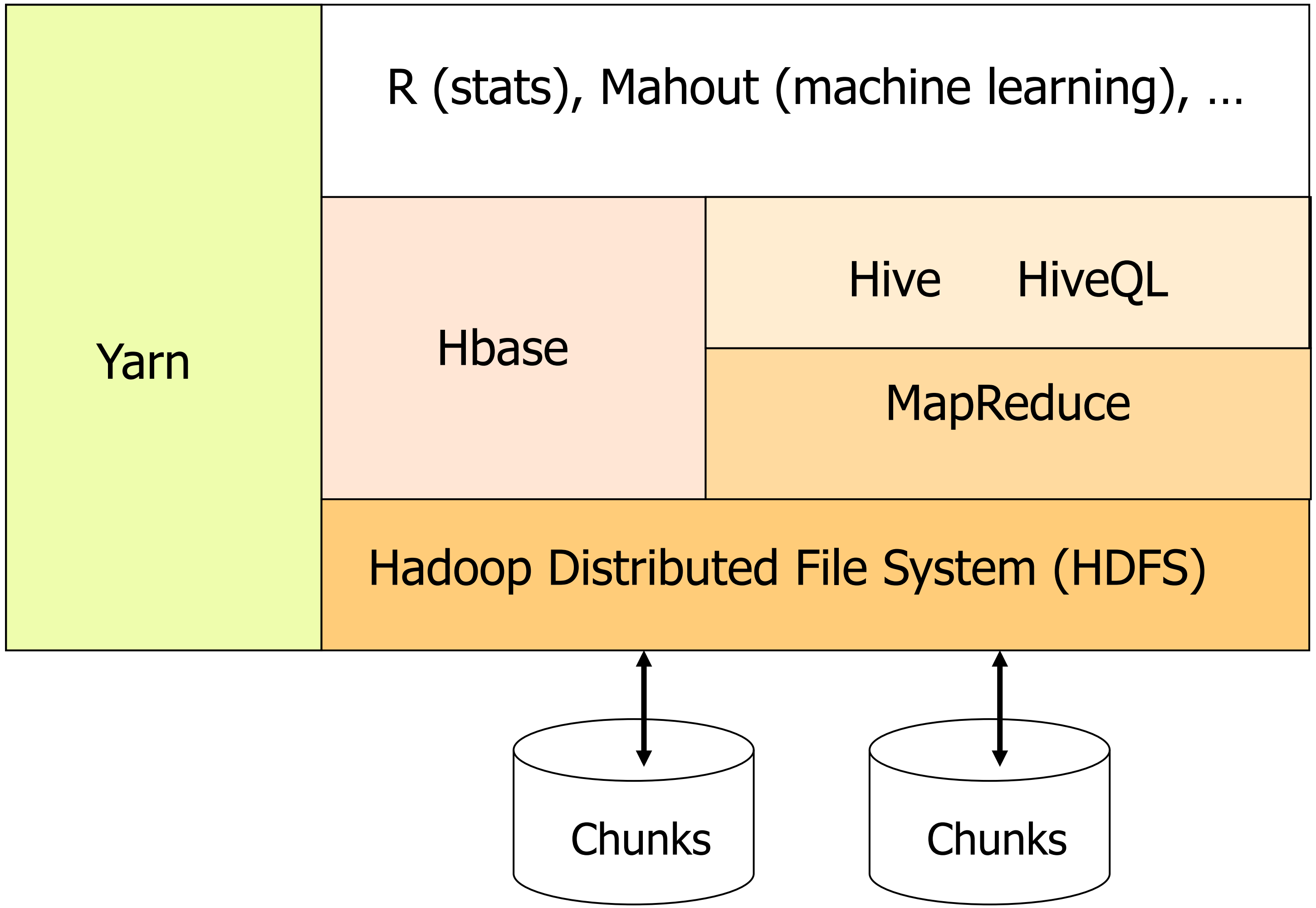
dave@vcdave.com

blogs.forbes.com/davefeinleib

A New Software Stack

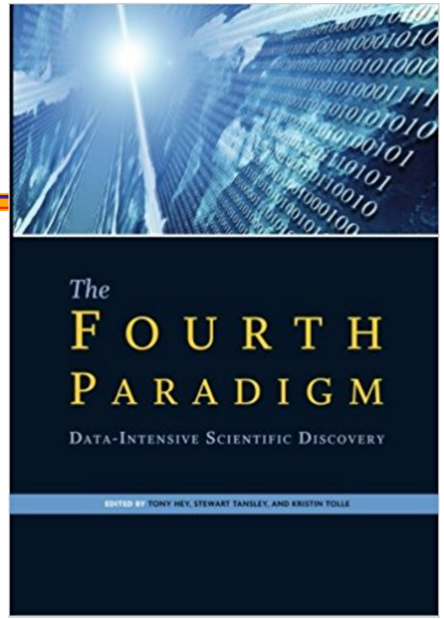


Hadoop Architecture



HPC & Big Data Analysis

Context: data-intensive science



- Modern science such as astronomy, biology and computational engineering must deal with overwhelming amounts of data
  - Generated by sensors, scientific instruments or simulation
- Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore these massive datasets
- Requires the integration of two paradigms
  - High-performance computing (HPC)
    - From high-end supercomputers to compute clusters
  - Data-intensive scalable computing (DISC)
    - Hadoop, Spark, Pregel, Giraph, NoSQL, NewSQL

HPC versus DISC

Dimensions	**HPC**	**DISC**
Focus	Compute-centric	Data-centric
Applications	Science, engineering	Web, business
Target	Simulation	Data management, data analysis
Objectives	High-performance	Scalability, fault-tolerance, availability, cost-performance
Programming models	Low-level (MPI, OpenMP) Operator libraries	High-level operators (Map, Reduce, Filter, …)
Programming languages	C, C++	Java, Python, Scala
Parallel architectures	Shared-disk and specific hardware	Shared-nothing clusters of commodity hardware

Approaches

- Postprocessing analysis
  - Performs analysis after simulation, e.g. by loosely coupling a supercomputer and a DISC cluster (possibly in the cloud)
  - Simple, non intrusive but is restricted to batch analysis
- In-situ analysis
  - Runs on the same compute resources as the simulation, e.g. a supercomputer
  - Intrusive, but makes it easy to perform interactive analysis
- In-transit analysis
  - Offloads analysis to a separate partition of compute resources, e.g. using a single cluster with both compute nodes and data nodes
  - Less intrusive than in-situ, but requires careful synchronization of simulation and analysis

SciDISC

Scientific data analysis using Data-Intensive Scalable Computing

Project coordinators: Marta Mattoso & Patrick Valduriez

Inria – Brazil Associated Team

2017 - 2019

Opportunities and Risks

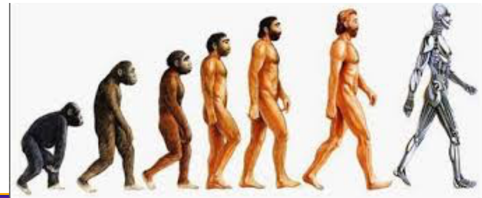
Opportunities

- 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 (Uber, Airbnb, Leboncoin, etc.)
  - Digital health, digital 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

Impact on Homo Sapiens



- More and more intelligent tools
  - Self-driving cars, autonomous robots, digital assistants, drones, terminators, ...
  - Questions
    - Responsibility in case of problem (failure, collateral damage, …)
    - Towards a job-less society
    - Freedom and privacy

Big Brother

- Transhumans (augmented humans)
  - Human enhancement through natural or artificial means
  - Questions
    - The end of natural selection
    - A new human species
    - Immortality, e.g. replacing a dead person by an AI

Thanks