

Big data and analytics: framework and case studies

**Chiara Francalanci
Politecnico di Milano**

Outline

- Objectives
- Artificial Intelligence, Machine Learning and Deep Learning
- Big data and big data technologies
- Case studies
- Q&A

Objectives

- To understand the business opportunities offered by big data technologies and different approaches to analytics (including artificial intelligence, machine learning, deep learning and advanced statistics).
- To tie these opportunities to the evolution of technology.
- To provide a classification framework for existing case studies.
- To discuss key issues and related managerial decisions.
- Q&A

Artificial Intelligence vs. Machine Learning?

- **Artificial intelligence** (AI, also machine intelligence, MI) is intelligence demonstrated by machines, in contrast to the natural intelligence (NI) displayed by humans and other animals.
- **Machine learning** (ML) is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.

Source: Wikipedia.

Artificial Intelligence

- **Capabilities** generally classified as AI include successfully understanding human speech, competing at the highest level in strategic game systems (such as chess and Go), autonomous cars, intelligent routing in content delivery network and military simulations.
- The traditional **problems** (or goals) of AI research include reasoning, knowledge representation, planning, learning, natural language processing (NLP), perception and the ability to move and manipulate objects.

Source: Wikipedia.

Machine Learning

- **Unsupervised learning:** learning from data without a need for «ground truth», e.g. clustering or pattern recognition.
- **Supervised learning:** learning from data with «ground truth», e.g. predictive analytics.

«Ground truth» is data on the «true» behaviour or status of a system, typically obtained from direct measurement of real-world data.

Example – Earth observation for crop classification (agriculture)

- Satellites provide images of the earth, divided in pixels (e.g. 30mx30m pixel size), represented as reflectance values
- *Unsupervised approach*: from reflectance values, it is possible to identify larger areas (e.g. fields) with clustering techniques. Larger areas are then labelled based on the reflectance footprint of different types of crop.
- *Supervised approach*: from a classification of pixels (ground truth) it is possible to predict the crop for next year by training a random forest on reflectance values of previous time period (e.g. last year's satellite observations with corresponding ground truth).

Deep learning

Deep learning is a class of machine learning algorithms that:

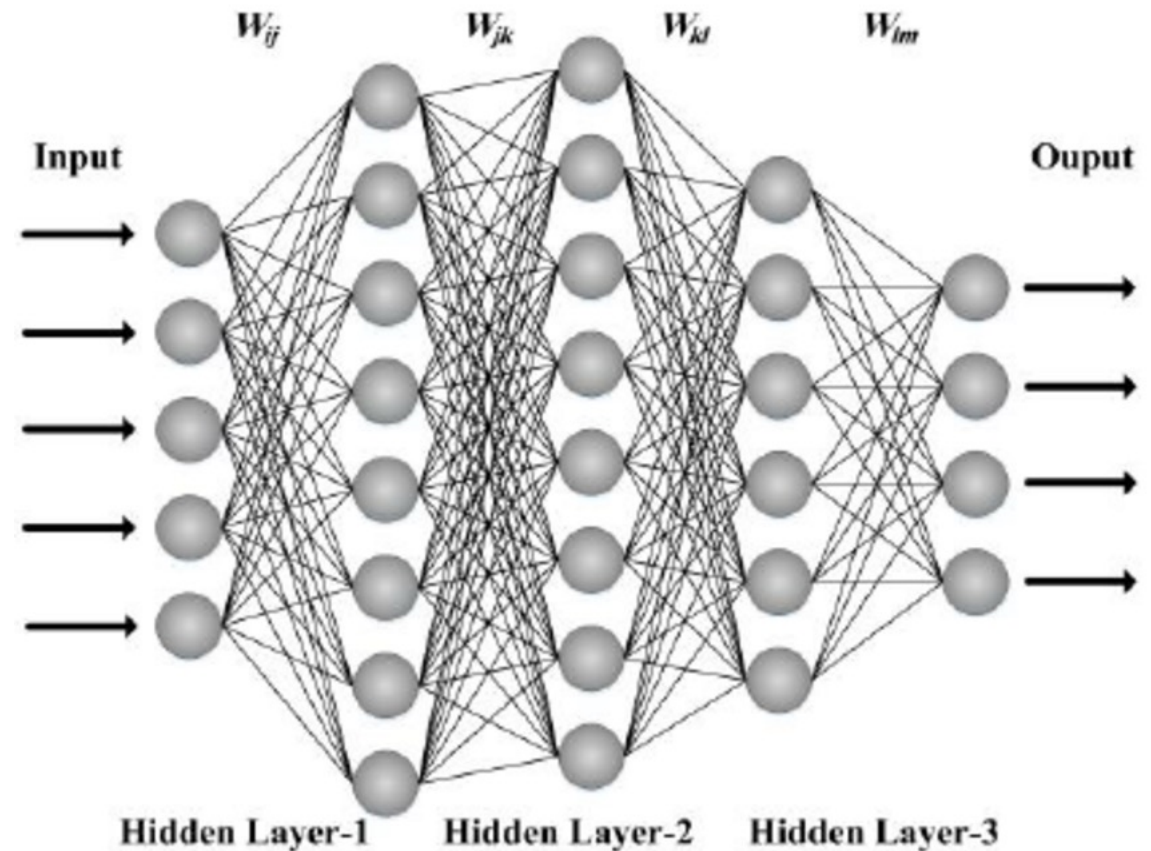
- Use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input.
- Learn in supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis) manners.
- Learn from multiple levels of representations that correspond to different levels of abstraction, e.g. the levels form a hierarchy of concepts.

Source: Wikipedia.

Example – Multi-layer neural network

Each layer performs a different transformation, e.g. image recognition:

- Layer 1 Segment image in pixels.
- Layer 2 Aggregate pixels in areas to obtain a simplified image recognition problem.
- Layer 3 Train a (single) network to recognize image.



Natural Language Processing (NLP)

- **Natural-language processing (NLP)** is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.
- **Challenges** in natural-language processing frequently involve speech recognition, natural language understanding, and natural language generation.

Source: Wikipedia.

Applications of NLP

- 1990 – Document management (document classification and retrieval, topic extraction)
- 2000 – Document and Web search
- 2005 – Speech to text (Voxforge 2006-, Sfynx 2011-)
- 2010 – Social media analytics, Web reputation, sentiment analysis
- 2015 – (chat)BOTs
- <https://www.youtube.com/watch?v=IXUQ-DdSDoE>

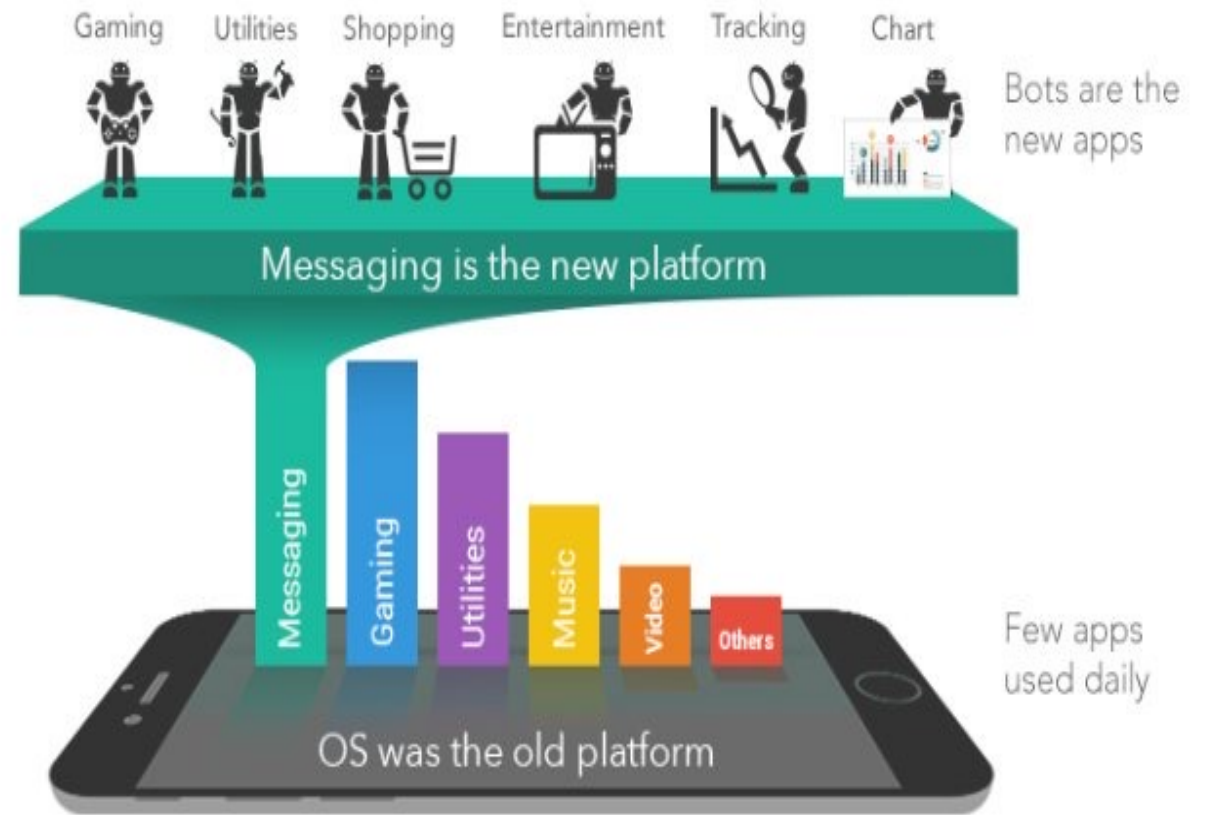
Source: Wikipedia.

(chat)BOTs

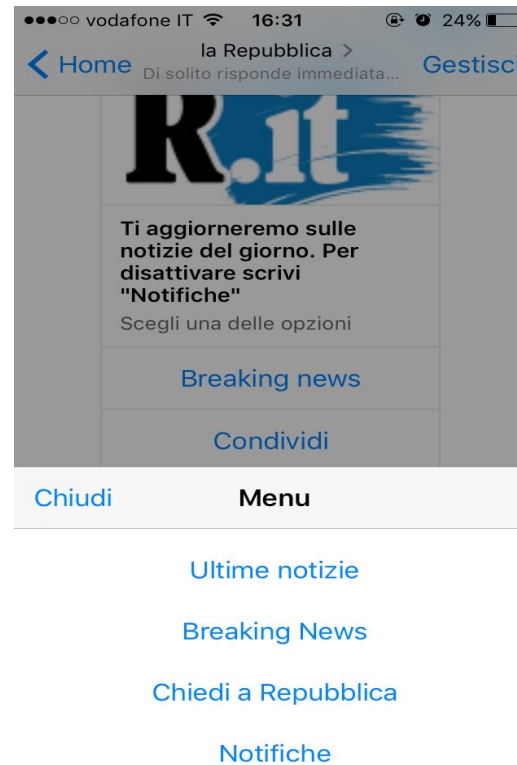
- A BOT works «for us», just like any other algorithm.
- The difference between a BOT and an algorithm is that a BOT does what *we* (humans) would do.
- For example, if we ask for the square root of a number, the BOT opens the calculator, digits the number, presses the square root button and then reads the result.
- The interaction with us is the best state-of-the-art interaction, initially based on text chats, now more often based on speech recognition.

BOTs and the inversion of paradigm

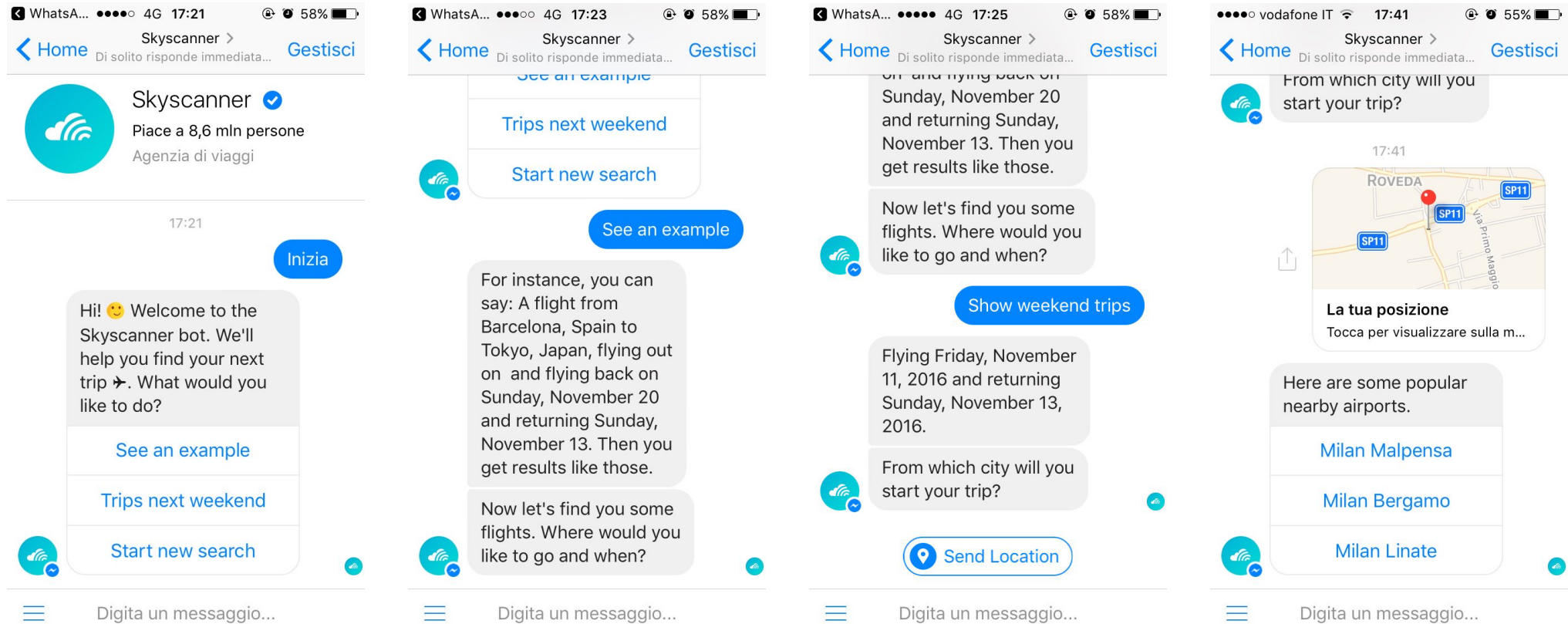
- A BOT is more than a mobile app, it is a new paradigm to integrate mobile apps.
- The current trend is to develop a mobile app equipped with APIs for BOTs.
- Users are reached on the apps that they use most frequently, typically messaging apps such as Messenger or Telegram



Example: La Repubblica



Example: Skyscanner



Example: PAMela, full set of functionalities

- Quali sono gli orari di apertura della filiale più vicina a casa mia?
- Vorrei ordinare una spesa in via Politecnico contenente pomodori e zucchine
- Siete aperti il 24 dicembre?
- Posso lavorare con voi inoltrando il mio cv?
- Ho smarrito la password dell'area clienti, posso reimpostarla?
- Posso trovare del pane carasau presso la filiale di via Sardegna?
- Quali sono i punti vendita più vicini a me?
- Quanti punti servono per poter ordinare la lampada in premio?
- Potrei avere il numero di telefono della filiale di via Bazzini?
- Mi mostri il volantino promozionale?
- Quanti punti ho sulla carta fedeltà?
- Hai delle ricette da consigliarmi per il pane carasau?
- Invio di messaggi di lamentela o di gradimento al bot: es. Che bello che adesso vendete anche il pane carasau!
- Accettate carte di credito?
- *Da parte del Bot randomly* : "Hei! Sono PAMela, la tua assistente, hai qualcosa da consigliarmi?"
- *Da parte del Bot randomly* : "Hei! Sono PAMela, la tua assistente, quali prodotti vorresti vedere in offerta?"
- *Da parte del Bot randomly* : "Hei! Sono PAMela, la tua assistente, buon natale!! Il panettone è in sconto da noi da oggi fino al 6 gennaio!"
- *Da parte del Bot randomly* : "Hei! Sono PAMela, partecipa anche tu al concorso di Natale!"

Amazon Alexa

- <https://www.youtube.com/watch?v=b4uG9dfFtE4> from 5 m 45 s
- Companies can add services to Alexa (es. online shopping) and integrate them with Alexa predefined services (e.g. shopping list).
- The design of the semantic capabilities and intelligence of Alexa is not entirely predefined and is largely part of the design of each business application of Alexa (e.g. commands, equivalent formulations of commands, suggestions, interaction paths,...). It's called *situational design*, it includes designing Alexa's conversational skills in *your context*.

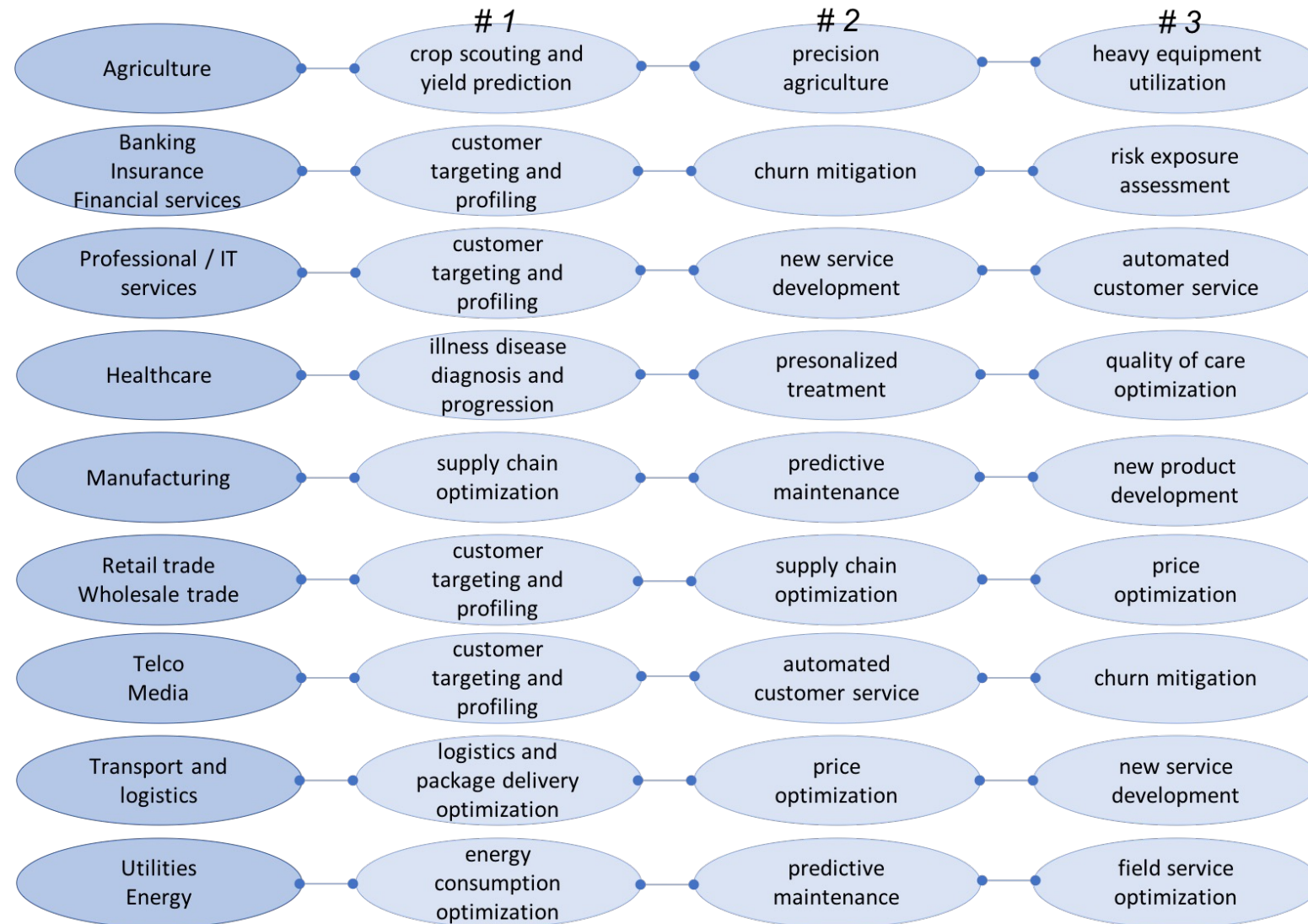


Google Assistant / Google Home



- <https://www.youtube.com/watch?v=FPfQMvf4vwQ> da 30 s
- Business risks of Web and social media: adversarial attacks on Google home....
<https://www.youtube.com/watch?v=t7Krn-DH3tw>
- General risks: privacy (Google home is always listening and private data stored in the cloud to be interpreted with AI/NLP), security, unwanted triggering of commands,...

Machine learning (AI) application priority (by industry)



Source: H2020 Databench,
Nov.. 2018

Machine learning (AI) business KPIs

SECTOR	PROBLEM	KPI
Finance	Risk exposure assessment	Loss reduction
Accommodation	Targeting	Increased sales/margins
Manufacturing	Predictive maintenance	MTBF, availability/productivity
Health	Compliance checks	Quality of care
Telecom	Network analytics	Quality of customer service
Media	Marketing optimization	Increased revenues/margins
Transport	Churn prediction for targeting promotions	Churn reduction
Utilities	Customer behavior analysis and custom pricing	Increased margins
Oil&Gas	Natural resources exploration	Increased ROI from plant investments
Retail/Wholes	Optimization of assortment choices, price optimization	Increased sales/margins
Professional Services	Customer profiling	Offer redemption
Government	Contract analytics	Reduced expenses/ service improvement
Education	Student data analysis	Workload balancing

Source: H2020 Databench report D2.1, March 2018

Why should AI be related to business KPIs?

- Because with AI we embed (or support) «decisions» inside software
- Decisions should be driven by business KPIs
- Example:
 - In yield prediction, the precision of yield estimates for different types of crop has a direct impact on the ROI of financial investments.
 - In turn, automating (or supporting) trading decisions should be driven by ROI.

Evaluation of business KPIs

- The benefits of AI/machine learning use cases are rarely quantified.
- There's a lack of business benchmarking initiatives.
- Quantitative evidence almost exclusively comes from suppliers of technology solutions.
- For some use cases, economic benefits are difficult or impossible to quantify (KPIs are simultaneously affected by multiple initiatives).
- A typical managerial question (still) is: do I really need machine learning?

Evaluation of business KPIs - example

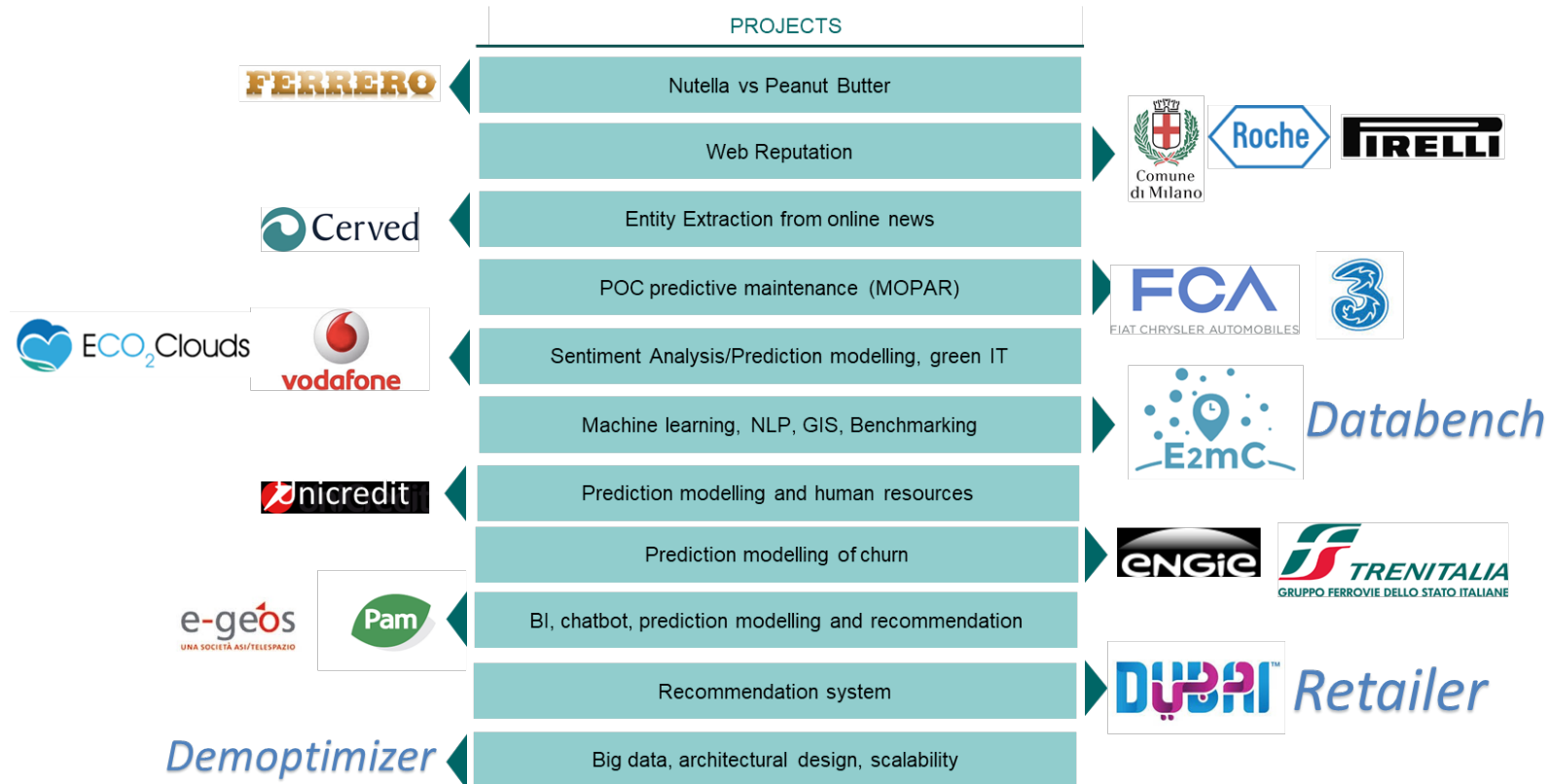
Linear vs. non linear prediction models in the retail industry

	Average % error exponential smoothing	Average % error Holt-Winters	Average % error machine learning (XGBoost)
Total daily revenue	13.18%	36.02%	4.9%
Daily revenue of individual shop	12.9%	19.54%	6.23%
Daily revenue of group of shops with similar seasonality	-	avg 26% opp 10% flat 20.02%	avg 3.44% opp 5.44% flat 5.92%
Daily revenue of individual product	24.11%	26.86%	16.7%
Daily revenue group of similar products	-	14 prod 11.53% 408 prod 16%	14 prod 6.03% 408 prod 5.88%

Test set: July 2016

Evaluation of business KPIs

We have observed the benefits of machine learning in a variety of contexts:

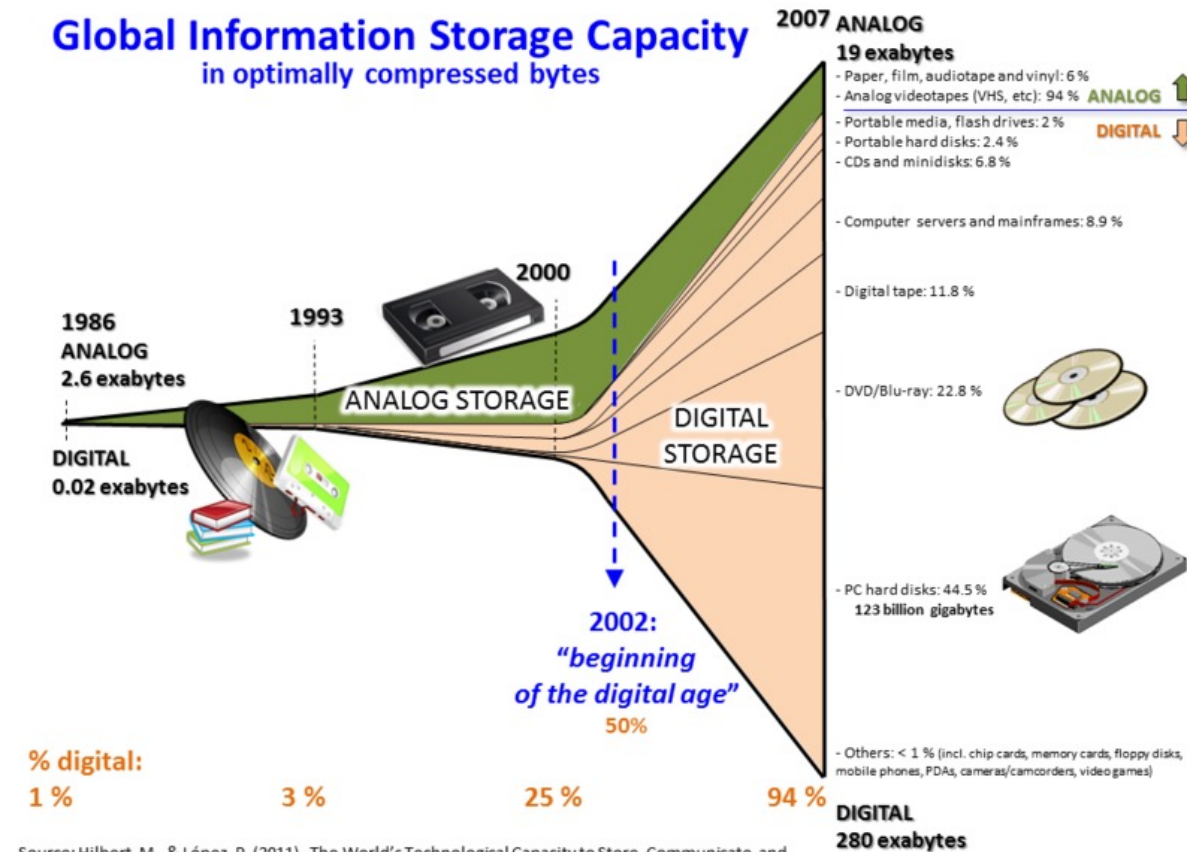


Why are managers (still) skeptical?

- Because AI and machine learning are (and are perceived as) complex.
- Because there is no off-the-shelf technical solution.
- Because technology is special-purpose and expensive.
- Because AI and machine learning are seen as a threat by decision makers (in fact, it may replace some of them).
- Because AI and machine learning are associated with the concept of «big data» which adds to their complexity.

Why big data? Why now?

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate (*Source: Wikipedia*).



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

How big is big data (1/2)?

Erik Schmidt (Executive Chairman Google): «From the dawn of civilization until 2003, humankind generated 5 Exabytes of data. Now, we are producing 5 exabytes every two days, and the pace is accelerating.»

Name	Symbol	Power
Kilobyte	KB	10^3
Megabyte	MB	10^6
Gigabyte	GB	10^9
Terabyte	TB	10^{12}
Petabyte	PB	10^{15}
Exabyte	EB	10^{18}
Zetabyte	ZB	10^{21}

How big is big data (2/2)?

....and global IP traffic is forecasted to reach 3 ZB/year, roughly 4 EB/day by 2021

Year	Global Internet Traffic
2001	1 EB/year
2004	1 EB/month
2007	1 EB/week
2013	1 EB/day
2021	4 EB/day

What is big data?

- Big data is «a collection of data from traditional and digital sources inside and outside your company that represents a source for ongoing discovery and analysis.» (Source: Forbes)
- Big data is any amount of data that raises technical scalability challenges for a given company due to the increasing growth rate of data and a need for continuous analysis.

Types of big data

Conversation text data	<ul style="list-style-type: none">• e.g. Twitter, Facebook
Photo and video Image data	<ul style="list-style-type: none">• e.g. Youtube
Audio files	<ul style="list-style-type: none">• e.g. call centers
Sensor data	<ul style="list-style-type: none">• e.g. geo seismic data
The Internet of Things data	<ul style="list-style-type: none">• e.g. smart devices, smart phones
Web customer data	<ul style="list-style-type: none">• e.g. Web logs
Traditional customer data	<ul style="list-style-type: none">• e.g. receipts, loyalty programs, traffic data of telephone/Internet operators

Conversation Text Data (1/2)

- Social media conversation data can be «big data.»
- For example, every second, on average, around 6,000 tweets are tweeted on Twitter (visualize them [here](#)), which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year.
- In bytes, Twitter produces text for roughly 1 MB/s, that is 70 GB/day, 2 TB/month, 25 TB/year

Buzz volume

The volume of data on a specific brand is:

- Small if the brand is local. For example, Twitter buzz on Milan city (or Berlin or Madrid) is in the range of 3,000 tweets per day (60 MB/year)
- Larger if the brand is global. For example, Twitter buzz on «Nutella» is in the range of 15,000 tweets per day (1 GB/year)

Conversation Text Data (2/2)

- Every minute, on average, 293,000 statuses are updated on Facebook, which corresponds to roughly 25 TB/year (same as Twitter).
- On the Internet, there are 152 million blogs, a few million forums and almost 1 billion Web sites.
- Text data are relatively small compared to images and videos, but text analyses are computing intensive.
- Semantic engines can process roughly 1 MB of text in 1 day with 1 core.

Image, audio and video data

Image

The size of a JPEG picture file ranges between 1 and 6 MB, depending on the quality. On FB, 136,000 photos are uploaded every minute, corresponding to 400 TB/day and 150 PB/year.

Audio

The size of the audio recording of a 1 hour conversation is roughly 20 MB. A call center handling 1,000 calls/day with an average call duration of 4 minutes requires 1,3 GB/day and 0,5 TB/year for audio recordings

Video

The size of a good-quality video is (roughly) 100 MB/minute, with a high variability depending on resolution. On Youtube, 300 hour of videos are uploaded every minute, corresponding to 2.5 PB/day and 1 EB/year

Sensor data

- Sensor data can be:
 - Text
 - Photos
 - Videos
- Sensor data represent a «stream» of data, that is a time series of data points, and their value is in their timeline.
- Along a timeline, sensor data can easily reach the PB size (an enterprise tape library typically scales up to 75 PB).

Example

1. A video surveillance service can reach 1 PB in 5 days.
2. Real time data from car sensors in GM is 25 GB/hour, that is 1 TB every 2 days.

The Internet of Things data

- It has been estimated by Gartner that the Internet of Things (IoT) will include 26 billion units installed by 2020.
 - If each unit produces just 1 KB of text data per day, the IoT will produce 10 PB/year in text data only, 40 times Twitter/Facebook text data....
 -but IoT data includes photos, audio, and video.
- ➔ Transferring streaming data from IoT devices to a single location for processing is challenging both from a technology and economic perspective.

Web customer data (Web logs)

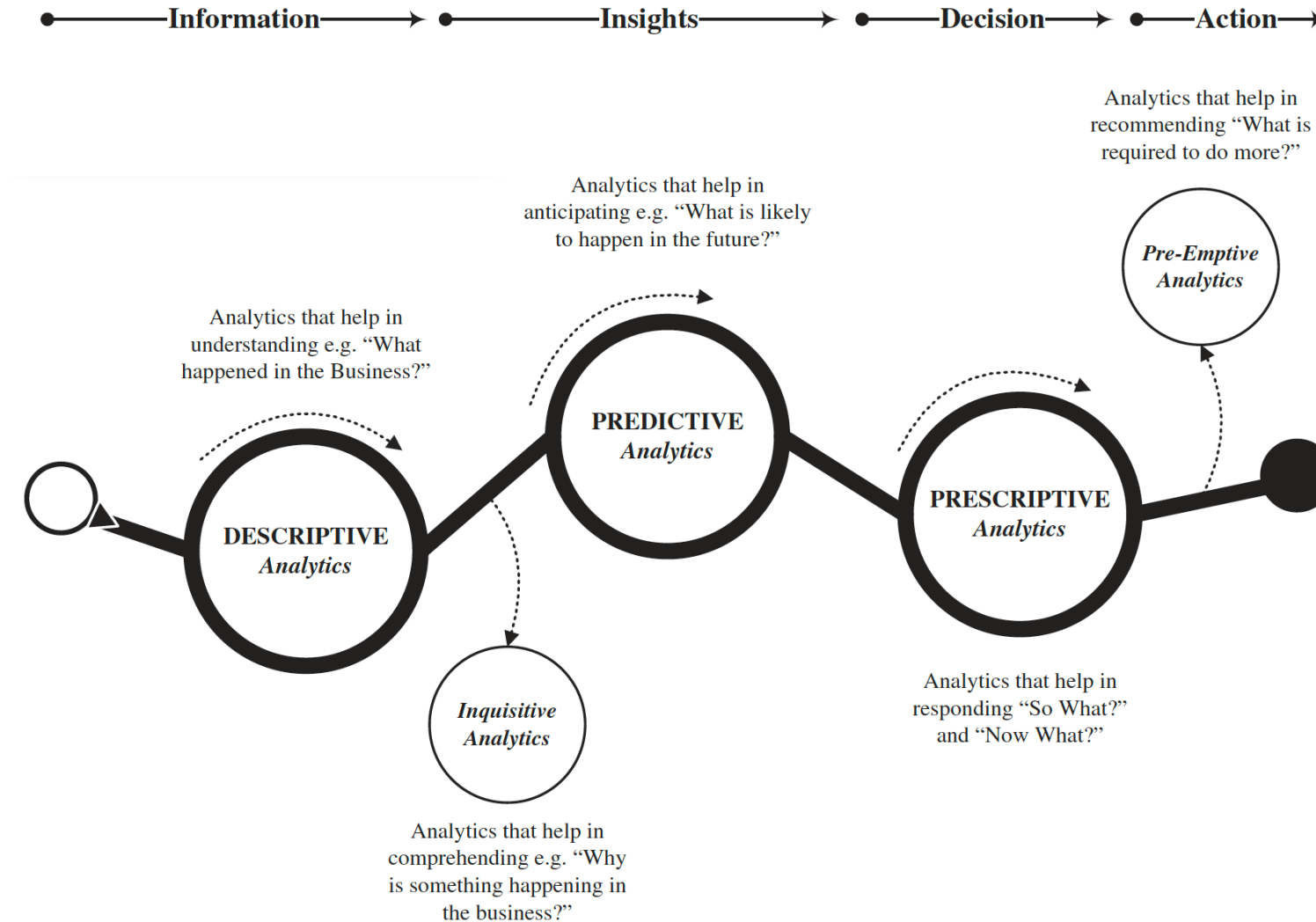
- In a log file, a typical hit might range roughly from 250 to 750 bytes. If a site experiences an average of 10,000 hits per day, the log file size can range between 2.5 MB to 7.5 MB.
- It is not unusual for enterprise-level sites to reach up to 5,000,000 hits per day and for the log file size to grow to several GB/day.
- For large organizations with extremely active web sites, generating a few TB of data in a year is common.
- Most organizations implement a log file rotation daily, weekly, or monthly.

Traditional customer data

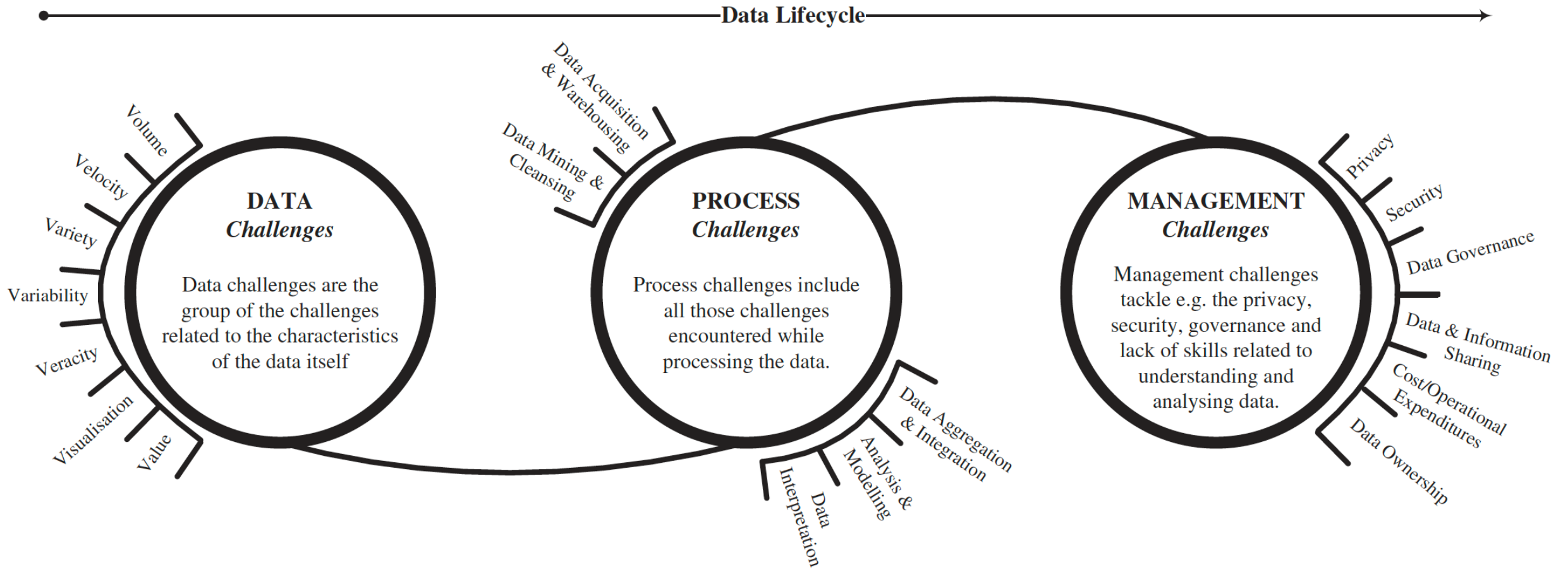
- Traditional customer data size ranges between 0,5 and 10 TB/year (compressed) depending on the type and size of the business.
- These data sets include both catalogue and transaction information.
- Complex processing operations have to be performed on these data, including customer segmentation and predictive analytics.
- While these data sets can still be managed with traditional relational technologies, queries can be very slow and non relational approaches can be advantageous.

Open source and free statistical tools may have a very low performance unless data are aggregated or short time periods are analysed (day, week, month).

Evolution of big data projects



Conceptual classification of BD challenges



Main issues with big data projects

1. Getting the technical skills needed to manage the new technologies for big data
2. Getting the data, which are very often stored in multiple databases, not integrated, not ready for analysis (e.g. not structured, not real time)
3. Getting the analytical skills to explore data and gather new and useful insights
4. Achieving business involvement

Once upon a time...

- Data were “small” and anybody could easily and inexpensively run analytics.
- Easy analytics can be processed with Excel.
- Excel can be used in a more sophisticated way by writing scripts and can accommodate a few complex analytics.

Example

Easy analytics: linear regression.

Complex analytics: combining easy built in analytics with ad hoc routines in Visual basic that embed a process with which the built-in analytics should be used and create specific outputs.

.....however Excel (2018) can grow to 1 M rows, only (and 16 K columns).

Easy and accessible open source: moving from Excel to MySQL

- Up to a few TBs, the issue is not the data storage: storage devices are relatively cheap (e.g. a 1 TB external drive for a PC is 50 euro at MediaMarkt)
- Up to a few TBs, the issue is not the DBMS per se: MySQL free can store 0,5 TB data by creating and uploading tables in a reasonable time frame on an entry-level server (32 GB RAM, 16 virtual cores).
- The issue is the end-to-end system that stores the data and provides the real-time analytics, starting with complex ETL (extraction, transformation and loading).

Example

Running a query to extract data prior to executing a statistical procedure that requires the join of two 100 MB MySQL tables runs out of resources on an entry-level server and does not terminate. Statistics on R quickly require days of processing on that size data on an entry-level server.

How large can a MySQL database become?

- Using a modern file system and OS, it is very likely that your database will perform very badly long before reaching theoretical limits.
- The useful size of your database is practically limited by the amount of RAM MySQL can use to cache information.
- Databases need memory to perform efficiently: memory accesses ($\sim 10^{-3}$ seconds) are one million times faster than disk seeks ($\sim 10^{-9}$ seconds)
- According to a rule of thumb, the maximum size of your database should be about 10 times the amount of caching memory

Memory is the limiting factor!

MySQL scalability

- MySQL would need 1TB of memory on a single machine to perform efficiently on a dataset of 10TB!
- To cope with Big Data, the best option is to run MySQL on a group of computers. The estimated limit of MySQL Cluster is 2 PB.
- To set up a cluster is expensive:
 - Hardware costs
 - Need for a large-bandwidth network (Gigabit ethernet)
 - Configuration is a time-consuming operation
 - Technical support only for commercial edition

Oracle scalability

- The maximum size of the Oracle database is 8 EB.
- This size can accommodate any type of data and is well beyond the typical multi-media big data requirements.
- Oracle is integrated with Hadoop (non relational approach, see next slides) and, therefore, can integrate large files without migrating them into relational tables.
- Oracle is integrated with R, enabling easy access to (open source) analytics and responding to end-to-end knowledge discovery requirements.
- However, schema is fixed, indexing is optimized for transactional systems, complex data manipulations are often difficult/impossible with sql and usually executed outside of Oracle, fast access to result tables in systems such as Neteeza and Teradata shows flexibility and performance limitations.

Hadoop scalability

- Framework that allows for the distributed processing of large data sets (terabytes or even petabytes) across clusters of computers.
- Used by some of the largest information technology and media companies in the world, such as Yahoo, Facebook, Twitter and Amazon.
- Designed to run on low-cost commodity hardware.
- Open source under the Apache license.
- The theoretical storage limit with Hadoop is estimated to be around 120 PB.

Hadoop – Strengths and weaknesses

Strengths

- Write-once-read-many model enables high-throughput access to data (works well for analytics, less so for transaction processing)
- Individual files are broken into blocks of a large size (default 64MB) and stored across the cluster.
- A typical block size of a common file system such as NTFS (Windows file system) is in the order of KBs.
- Impressive sequential read performance

Weaknesses

- Bad random read performance.
- Poor performance with small files.

Hadoop - Components

Hadoop has two main components:

- **Hadoop Distributed File System (HDFS):** a distributed, scalable and portable filesystem
- **MapReduce:** a distributed, fault-tolerant resource manager and scheduler for processing large data sets

Hadoop – MapReduce

MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster of servers.

A MapReduce program is composed of two procedures:

- Map(), which performs filtering and sorting (TaskTracker executed by each node separately, e.g. counting words in a text)
- Reduce(), which subsequently performs a summary operation (JobTracker typically executed by a dedicated node merging data from multiple nodes, possibly executing intelligent summary operations, e.g. counting the total number of words by adding up results from nodes)

Example – MapReduce

Word count example in Java: Map() function

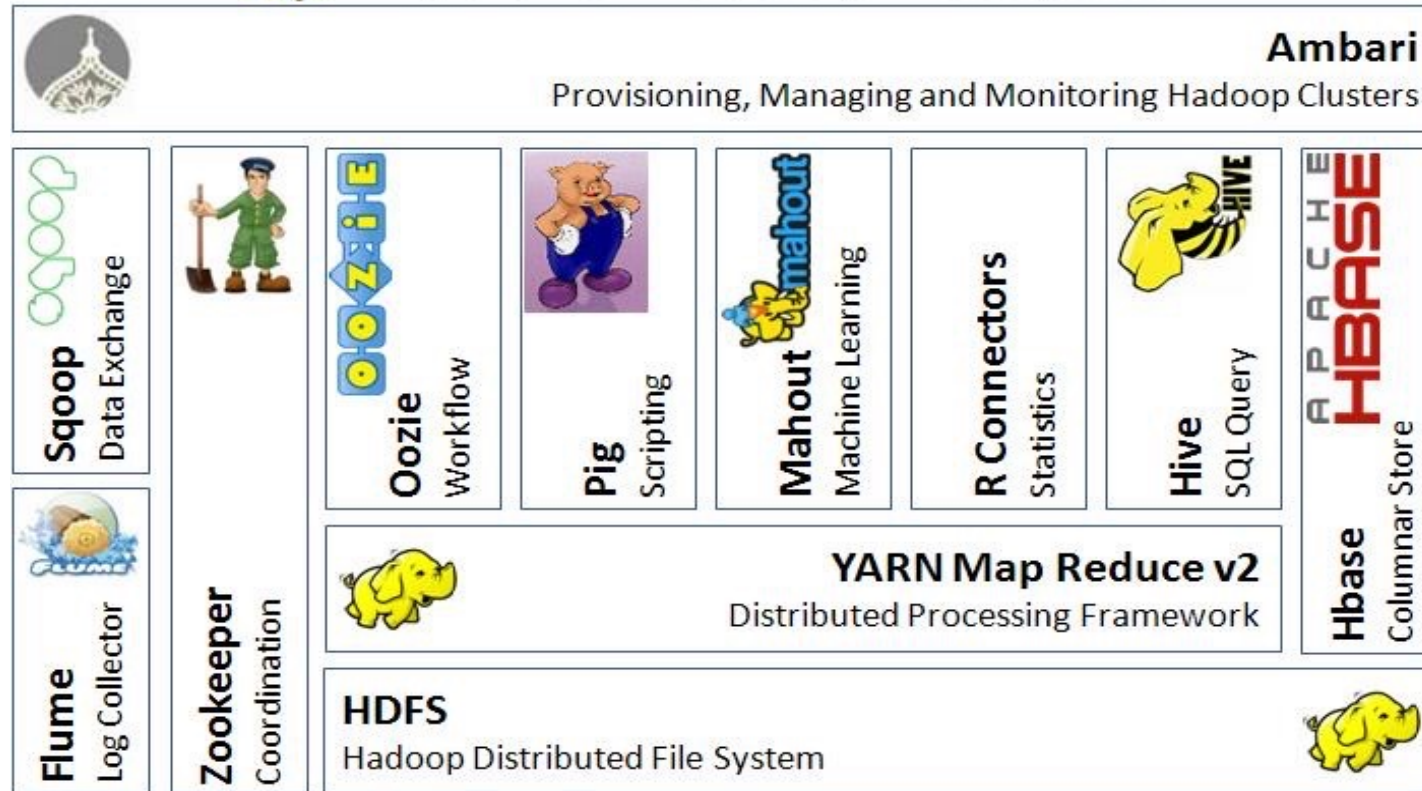
```
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>
{
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, Context context) throws
        IOException, InterruptedException
    {
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens())
        {
            word.set(tokenizer.nextToken());
            context.write(word, one);
        }
    }
}
```

Hadoop – Ecosystem



Apache Hadoop Ecosystem



Source: Apache Hadoop site www.hadoop.apache.org

Hadoop commercial distributions

Two commercial distributions

1. Cloudera + Hortonworks (merged in Oct. 2018)
2. MapR (partnership with IBM)
3. Pivotal (EMC)

They provide suites integrating a selection of Hadoop components and adds-on as well as related global support services.

They can be installed on any existing hardware or virtualized/cloud environment.

Getting the data can be a challenge

Companies may or may not have integrated data supporting their requirements for machine learning and advanced analytics.

Manufacturing

For example, in a manufacturing company based on custom design (e.g. oil & gas industry) may have different groups of engineers in different countries using different CAD tools. Procurement is based on estimates that could be improved with predictive analytics running on integrated CAD data that may not be readily available.

Telecom

Similarly, some telecom operators have vertical data silos (traffic, data, billing...) and have multiple non-integrated legacy systems. Integrating these data would allow them to improve their customer analytics and provide strategic insights on the effectiveness of their value proposition with different customer segments.

Data Lifecycle Management

- Data should be complete with respect to the data analytics requirements: this requires knowledge of organizational processes and practices.
- Data should be integrated: with big data, working on samples that are not statistically significant is a tangible risk.
- There is an objective difficulty in enforcing consistent quality as the data scales along any and all of the three Vs (volume, velocity, and variability).
- Quick to generate, quick to evaporate: data need to be deleted/cleaned frequently.
- Data are heterogeneous: data have different formats and sources. Integration may be coped with at different levels of abstraction (physical, logical, or only conceptual).

Organizational issues: getting the right skills

- The main organizational issue with big data is **human resource**. Talent management is critical to have good **data scientists** who can extract value from data. Along with the data scientists, a new generation of **computer scientists** are designing techniques for processing very large data sets.
- **Leadership** is another issue, since companies need clear vision and goals to enable coherent and target-oriented data analyses.
- **Company culture** should become data driven. This requires to move away from acting on instinct and HiPPO decisions (decisions based on the highest-paid person's opinion).

Organizational issues: achieving business involvement

- Even with the right skills and a strong leadership, building a data-oriented company culture can be a challenge.
- On one hand, culture can be changed when data are available. On the other hand, making the data available requires a considerable degree of business involvement and commitment.
- Evangelization and use cases are key to obtain the initial commitment to embark in a big data project.

Example

Utility companies have millions of smart meters installed providing quasi real-time data on customers. However, a few of them use this information to gather insights and improve their service. What are possible key analytics on data from smart meters that could demonstrate the potential of big data analytics? Simulation of more customized customer offers could constitute an interesting PoC.

BI and analytics platforms

The Gartner Magic Quadrant shows **43 vendors**:

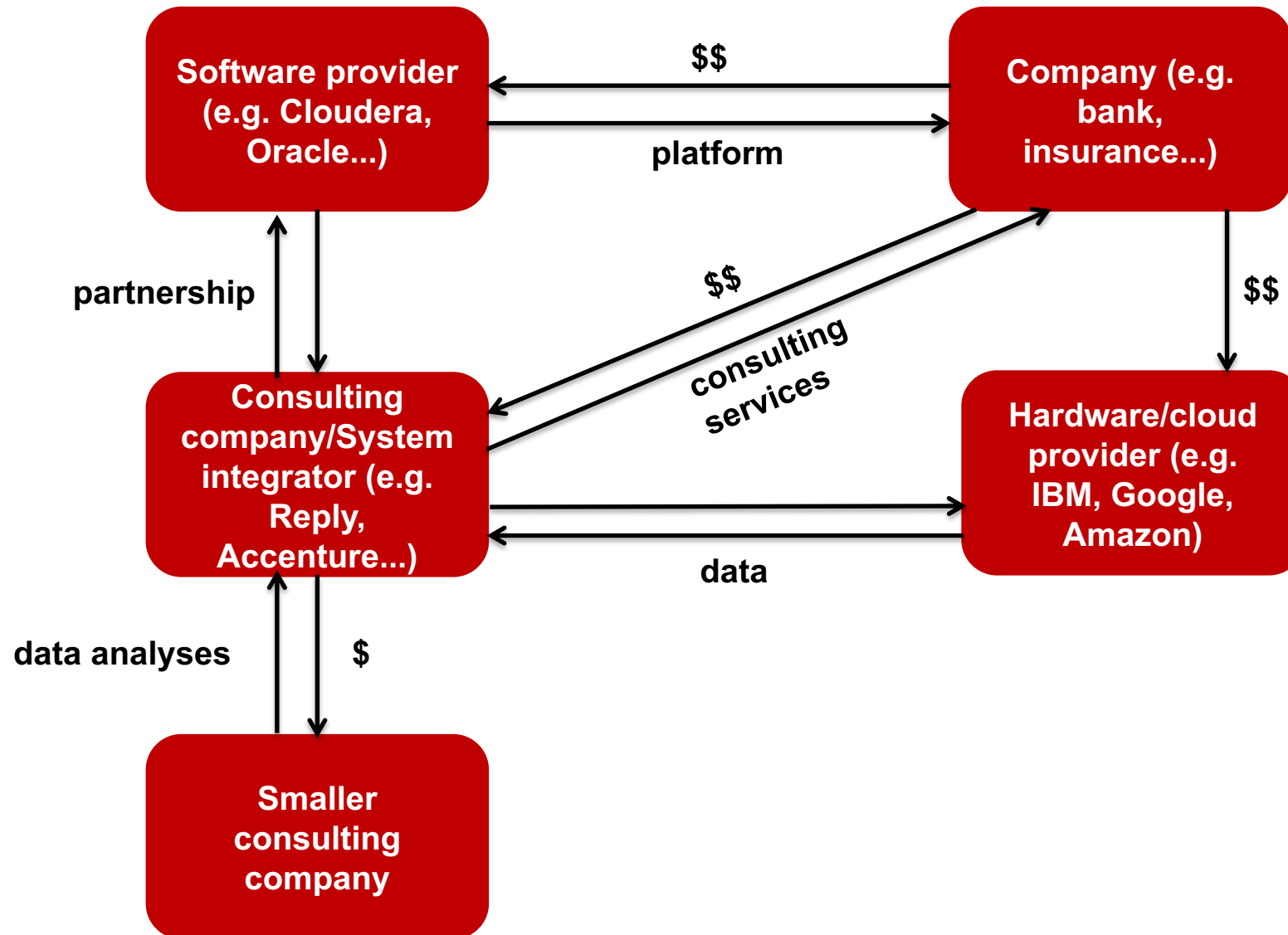
- Four vendors are called “Megavendors”: IBM, Microsoft, Oracle and SAP.
- Only two smaller vendors operate with a cloud-based approach.
- Only seven tools provide streaming BI functionalities.
- There exists a variety of smaller vendors.
- Megavendors show the lowest market growth rate, while data discovery leaders show the highest market growth rate.

Gartner's key findings (verbatim)

- Reporting continues to be the most widely used BI capability, but vendors that give a broader range of users access to more difficult types of analysis have the highest customer satisfaction and deliver the strongest business benefits.
- Data discovery vendors lead versus other vendor types in three key measures: aggregate product score, ease of use, and the complexity of analysis conducted by users. They also deliver the highest business benefits. This likely explains their market momentum.

While the megavendors are good at managing large data sizes, their ability to extract knowledge from data seems a weakness compared to other players.

The long knowledge extraction process



Big data analytics and consulting services

- Over **40 global competitors**, including Accenture, Atos, Avanade, Bain & Company, Booz Allen Hamilton, Boston Consulting Group, Canon, Capgemini, Capita, Cisco, Cognizant, CSC, Dell, Deloitte, EMC, EY, Fujitsu, HCL, Hitachi, HP, IBM Global Services, Indra, Infosys, KPMG, Lockheed Martin, McKinsey, Microsoft, Northrop Grumman, NTT DATA, Oracle, PwC, SAIC, salesforce.com, SAP, Tata Consultancy Services, Tech Mahindra, Unisys, VCE, VMware, Wipro, and Workday....
- A **large number of local**, smaller, yet growing and profitable **consulting companies** (e.g. the ecosystem of SAS developers&consultants).

Role played by smaller consulting companies

- they gather data from external sources: for example, they develop crawlers or crawl/classify/score manually to fill in a DB for Web reputation analyses
- they perform small and relatively simple system integration projects
- they develop complex analytics ad hoc (for example, they develop customers segmentation or sales prediction models tailored to their customers' sales data)
- they provide industry dependent knowledge (for example, they are specialized in mass retailing and provide marketing insights based on their customers' sales data)
- they are partners of larger technology providers and perform software customization, analytics, and reporting on behalf of the larger provider

Is anything changing?

- Open source technologies are growing out of their children shoes, granting big data capabilities to **smaller consulting companies** (without the intermediation of large technology vendors).
- **Success stories** from companies applying innovative approaches to BI are increasing in number and are no longer limited to smaller organizations.
- Companies are under pressure and have an increasing need for extracting value from data to improve their competitiveness in the **short term**, as opposed to starting long and risky projects involving major organizational changes.

Big data and business innovation

Two types of innovation:

1. «**Quick fix**» innovation: companies look for insights that can provide business value quickly without any major organizational change or integration with operations
2. «**Data-driven business process reengineering**»: companies embark in long-term organizational change to transform into a data-driven organizations basing decisions on evidence

Big data and business innovation – quick fix innovation examples (1)

No.	Industry	Project
1	Oil & gas	Saipem, Implementing a prediction model to improve effort estimates in complex engineering projects, without changing the organizational practices of engineers, but by integrating data from different CAD applications.
2	Retail	PAM, Improving general pricing strategy by applying discounts according to customer perceptions and related success of past offerings.
3	Banking (HR)	Ubis, Predicting candidate behaviour and improving employee performance (getting the data is a challenge, often unstructured).

Big data and business innovation – quick fix innovation examples (2)

No.	Type	Project
4	Manufacturing	Whirlpool gathers new data by embedding sensors in products to track actual product usage and mine social media for customer sentiment for product innovation.
5	Financial services	SEC (US Securities and Exchange Commission) needed insights to highlight hedge funds that required further investigation and used data analytics to identify outliers based on data inconsistencies (under a program called Aberration Performance Inquiry).
6	Media	NBC Universal makes changes to television programming in response to real-time customer sentiment, e.g. quasi real-time decisions on time slots (hours)
7	Grocery	7-Eleven Japan is Japan's most profitable retailer and heavily invested in data analytics by providing store clerks with a dashboard to make decisions on fresh food (stores order and receive deliveries three times a day). Each year 70% of the products sold are new products to the chain as a whole

Data driven business process reengineering examples – Amazon

- Epitomy of data-driven company, always.
- Predictive analytics are applied to cross-selling and advertising.
- Today's recommendations are based on each customer's wish list, items they have reviewed and what similar people have purchased – this creates a rounded profile of a customer used for predictive analytics.
- Are the only company that have a patent that allows them to ship goods before an order has even been placed.

Data driven business process reengineering examples – e-Quest

- One of the most recognized brands in HR industry.
- Has the majority of the Fortune 500 companies among its customers.
- Highly consulting-oriented approach: they help their customers treat talent acquisition as a strategy and help create efficient and effective sourcing strategies.
- eQuest's big data for HR division collects approximately 500 million job board performance statistics weekly (analyzed with QlikView).
- They have the ability to analyze how well a company's postings are performing against those of competitors to determine whether hiring goals are met.
- Help customers allocate funds by advertising on the most appropriate job posting sites.
- Help customers put more dollars behind critical jobs.

Data driven business process reengineering examples – Nissan motor company

- Nissan have localised websites designed to help consumers determine which Nissan is ideal for them (car types, models and colors).
- Nissan has made available a “request form” for the potential customers to fill out.
- Nissan aggregates these data points from individual customers to draw a picture as to the vehicles which were in demand throughout a particular region.
- Advertising campaigns and production are tailored to suit the needs of a region instead of a country as a whole.

Data driven business process reengineering examples – German World Cup Win

- FIFA World Cup won by Germany in 2014 (Brazil).
- Their victory is partly credited to Germany's use of data and analytics during training.
- The German football federation partnered with SAP to analyse video data and both individual and team performance.
- Using analytics, they cut down average possession time from 3.4 s to 1.1 s. This made the difference when they defeated Argentina, (supposedly, Mario Gotze's goal in extra-time).

Data driven business process reengineering examples – procurement

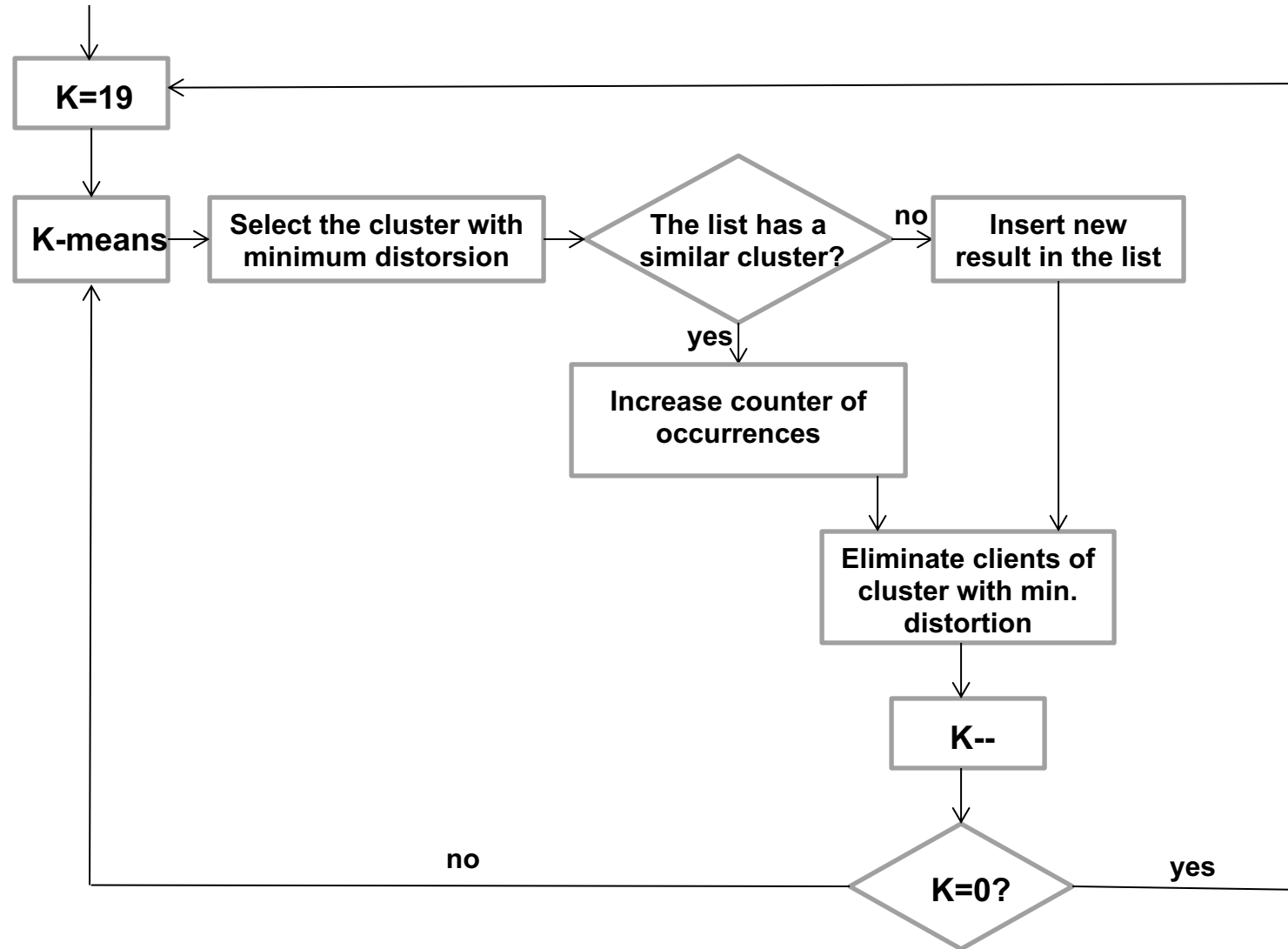
No.	Comp.	Software
1	AMD	Uses Zycus software to improve procurement efficiency with a reduced staff
2	ZF Friedrichshafen	Used IBM Emptoris software to reduce supplier risk
3	Orbitz Travel	Used Coupa to drive savings on travel
4	Primark	Used SciQuest software to improve global transportation procurement
5	Cox Enterprises	Used Ariba to consolidate information from all systems and centralize indirect spend
6	VF Corporation	Used Ista software to leverage the company's indirect spend information to identify key suppliers and analyze spending

Data driven business process reengineering examples – PAM

- Implementation and testing of a a customer segmentation model along the following dimensions:
 - Purchasing habits (spending and visits)
 - Price sensitivity
 - Lifestyles
- Clustering
- Strategic comparison between HI-LO and EDLP
- Sales prediction models
- Assortment optimization models

Analyses have been performed on PAM – Panorama receipts in 2013-2018 time frame.

Ad hoc algorithm to make k-means scalable



- The complexity of k-means grows exponentially with k
- With k=10 it converges in 1 hour on an entry-level server, with k=11 in one day, with k=12 in one year...
- However, the “best clusters” with minimum distortion are found quickly and an analyst can easily get a sense for a “good” value of distortion with a given data set

Life styles – results of k-means

Life Style	Cluster	%clients	%expense	%Low	%Medium	%High
Cherry pickers	Promotions	5.85	3.48	2.18	51.46	46.36
Price sensitive	Low-cost	7.57	8	33.8	61.57	4.63
	Low-cost and private-label	7.06	6.65	50.56	44.84	4.61
Loyal to brand	Private-label	9.17	13.46	8.04	78.29	13.67
Fresh food	Meat and fish	9.22	10.05	1.88	57.35	40.77
	Fruit and vegetables	6.98	8.91	16.75	71.57	11.68
Social	Entertainment and high calories	3.02	2.18	9.43	61.87	28.7
	High calories	3.81	3.23	7.06	57.08	35.86
Health enthusiast	Health and diet	3.92	3.72	4.27	64.13	31.6
Fast cuisine	Quick to cook	6.77	7.44	8.46	71.39	20.15
	Ready to eat	4.34	4.27	9.1	69.03	21.86
Young families	Kids	3.57	2.03	1.05	43.45	55.5
Traditional	Basic ingredients	3.7	1.93	5.01	52.63	42.35
	Preparations	4.33	3.63	20.03	66.44	13.54
	Seasonal	3.05	1.36	1.47	45.95	52.58
Creative cuisine	Gourmet	2.43	2.31	2.94	51.31	45.75
	Regional	2.16	1.94	4.23	65.74	30.03
	Exotic	2.36	1.63	9.43	67.29	23.28
Average customer	Average customer	10.69	13.78	3.63	78.1	18.27

The insourcing trend

The complexity of big-data related choices and analysis activities drives companies towards insourcing, with the following goals:

- Work with IT to identify the technical solution that best fits organizational requirements: this may also involve insourcing and consolidation of data centers (e.g. GM has recently consolidated 20 outsourced data centers into 1 internal data center to «save money»)
- Identify the competences needed to set up an internal group that:
 - performs the analyses of data and works with reference people from multiple organizational functions to tie the analytics to the business needs (requirements management)
 - works with management consultants to use specific analytics to facilitate organizational change (risk and change management)
 - possibly coordinates with smaller consulting companies to integrate ad hoc services with mainstream business needs (contract management)

Savings from insourcing and consolidating data centers

Savings from consolidation:

- economies of scale on technical staff
- standardization of technology platform and economies of scale on structure costs and various overheads
- discounts on larger contracts

Savings from insourcing:

- better control of all resources
- improved access to qualified human resources
- lower response times to organizational needs

Overall savings can be easily greater than 50% and in the billion \$ range for large corporations

Most common approach to make-or-buy decisions

“Buy” the infrastructure, since:

- It significantly reduces the time required to complete a big data project
- The loss of technical competences is limited

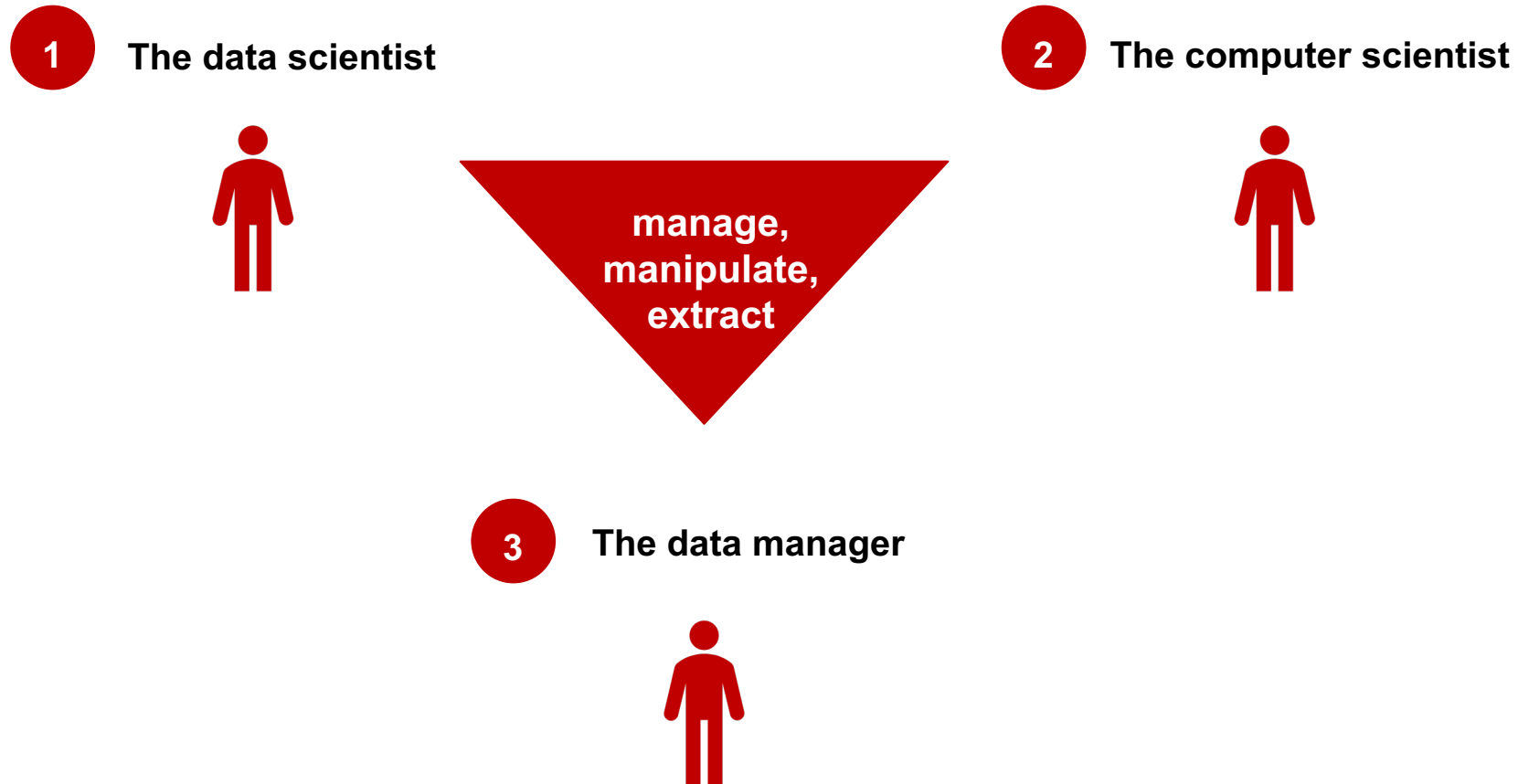
“Make” data management and analyses, since:

- Data analyses are recognized to be strategic
- Analytics are not a commodity yet and the tailor-made approach often provides an edge

Benefits from running the analytics internally

- Consultants often tend to standardize their reports and provide analytics according to a template: running the analytics in house allows for greater customization and flexibility.
- The level of trust in external information can be low, especially if analyses are performed by the same companies providing the «solutions» to the issues identified: for example, Web reputation analyses are very often performed by the same agencies in charge of the company's Web presence.
- Real-time business intelligence may require a tighter relationship with the business in order for knowledge discovered from the analyses to be quickly translated into action.

The «data science group»



A data scientist

The data scientist has expertise in statistical modelling and analysis, as well as business knowledge.

Example

General characteristics: experience in the analysis of economic and financial data with a variety of statistical techniques and in ad hoc algorithm design.

Specific technical competences:

- IT background and MBA.
- Experience in management consulting.
- Experience in business intelligence projects.
- Proficient in the use of different statistical tools (STATA, R, WEKA...).
- Knowledge on different types of statistical analyses, including mining and knowledge discovery.
- Knowledge of techniques for natural language processing.
- Expertise with different programming languages (C, Java).
- Good analytical skills and ability to design ad hoc algorithms.

A computer scientist

The computer scientist has expertise in the design of highly efficient algorithmic software.

Example

General characteristics: experience in the development of applications that require the management of large data volumes, both batch and real-time.

Specific technical competences:

- Knowledge of C and C++ programming languages
- Low-level I/O, in serial, parallel and asynchronous modes. I/O by memory mapping.
- Multi-process parallel applications.
- Multi-thread parallel applications.
- InterProcess Communication (IPC): socket, FIFO, Pipe, memory mapping, IPC via signals.
- Signal management.
- Experience in the development of crawling Web applications (in C and Java).
- Experience in the development of Natural Language Processing applications.
- Operating systems: Unix, Linux.

A developer and data manager

The computer scientist has expertise in the design of highly efficient algorithmic software.

Example

General characteristics: experience in application development with system and data management skills.

Specific technical competences:

- Knowledge of C, Java, Objective C programming languages.
- Experience in application development with Javascript, HTML, CSS, JSP and Java EE.
- Knowledge of relational databases (Access, MySQL).
- Knowledge of frameworks for distributed big data management (Hadoop, HDFS, MapReduce, Pig, Hive).
- Experience in mobile application development (iOS, Android).
- Experience in systems management.
- Experience in the development of Web crawling applications (in C and Java).
- Operating systems: Unix, Linux, Mac OS X, Android, iOS.

Skill shortage

- It is difficult to find and retain human resources with good technical skills
- Continuous and fast innovation makes technology more and more complex and difficult to manage
- Outsourcing is often inevitable, but:


**Outsourcing
the data center
(housing)**

**Outsourcing
processing
capacity
(hosting)**

**Outsourcing
software
development**

**Outsourcing
system
integration**

**Outsourcing
testing&validation**



Loss of
infrastructure
mgmt
competences

Loss of
infrastructure
procurement
competences

Loss of
software
development
competences

Loss of
software
management
competences

Loss of
data/process
competences

The risks of cloud

Google PAS and Amazon AWS address every imaginable need:

- Computing capacity
- Storage and databases
- Networking
- Big data
- Data transfer
- API platform and ecosystem
- Internet of things
- Cloud AI
- Management tools
- Developer tools
- ...



Cloud PAS becomes a **super convenient one-stop shopping** for both:

- Hardware and
- Software,

with a consequent conflict of interest.

In addition to a loss of competences, it represents an oligopoly (2 players is almost a monopoly), which in the long run typically causes lower quality and higher costs.

Diversification (as opposed to a one-stop shopping) is the only way to mitigate risks.

How to manage a big data project

1. Define Your Use Cases

Companies realize the most significant benefits from Big Data projects when they start with an inventory of business challenges and goals and narrow them down to those expected to provide the highest return. The following questions help:

- Who are the key stakeholders?
- What data do they own?
- What data do they need?
- How do they evaluate the flexibility of current IT services on data?
- What type of insights would be enabled by more integrated and more readily available data?

How to manage a big data project

1. Define Your Use Cases

- Select 1 or 2 use cases with clear KPIs to provide a proof of concept
- The stakeholders often raise a number of requirements in addition to analytics, such as:
 - Greater flexibility in managing data (e.g. be able to modify the data schema)
 - Lower lead time from data creation to data analysis and action (e.g. real-time BI)

Example

For example, a telecom company has opted for a multi-tenant Hadoop option, where they have defined a data schema making a distinction between attributes that can be modified by the business units and attributes that may be modified by a centralized coordination unit.

How to manage a big data project

2. Determine the Project Team

- Identify the project “sponsor” to remove obstacles, find the budget, provide organizational support, and champion the cause.
- Establish the project manager and the team. Define the roles and responsibilities of each team member.
- Understand the team’s availability and resource constraints for the project.

Example

A large bank has identified the CFO as the sponsor and the CFO has taken a chance to renegotiate the data management response times of IT, in addition to providing key use cases and sponsoring large infrastructural expenses (800 TB).

How to manage a big data project

3. Plan your project

- Specify expected outcomes in measurable business terms (e.g. increase revenues by implementing data-oriented pricing criteria)
- Determine any other quantifiable business requirement (e.g. improve data trust).
- Define what a successful Big Data implementation would look like (e.g. number of actual users).

Example

The HR dept. of a large IT company would like to equip employees with a tool that allows them to select training based on their lack of skills vis a vis desired job position. This initiative will ultimately be successful if it increases the percentage of employees updating their CVs in the HR application.

How to manage a big data project

4. Define your technical requirements

- Inventory all tools used today.
- Sketch the current architecture.
- Identify your data sources (internal, external, additional).
- Define your data schema.
- Design your infrastructure.
- Identify your suppliers.

A case study: targeting

1. Online recommendation engine
2. Assortment optimization engine (personalized shelf)
3. Personalized online search engine
4. Personalized shopping list
5. Couponing engine (with personalized pricing)
6. Personalized communication engine (mailing, flier, pop-ups, ...)
7. In-store proximity recommendation engine
8. Online and in-store instant promotion engine (with personalized pricing)
9. Real-time access to big data

The need

- Current targeting starts from the idea of showing personalized content, assuming that this is in and of itself good. But what if this shifts customer behaviour towards making less money, indiscriminantly, for too many products?
- We start from the idea that targeting should be based on business objectives and use personalization in different ways to reach business objectives.



Business objective:

- revenue
- percent margin
- cumulative margin
- quantity
- a *balanced* mix depending on the customer
- ...

Example: upsell, margin



Example: crosssell, revenue



How we use personalization

The system can learn how each individual customer reacts to targeting by answering questions, such as:

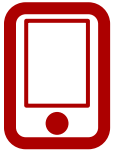
- Which items are most suitable for upselling?
- Which item pairs are most suitable for crossselling?
- What are the price elasticity limits for different items in upselling?
- What are the price elasticity limits for different item pairs in crossselling?
- Which is the comfort zone of each customer? To what extent should recommendations show habitual products?
- To what extent can we show interesting promotions without exceeding budget limits?

In-store simplicity: proximity services



1. Simple localization based on smartphone/mobile app (or smart device owned by the store)
2. Proximity services:
 - Instant recommendations: «you may be interested in pasta X, which is discounted and is right behind you»
 - Instant promos: «if you buy a second bottle of sauce, there is a 50% discount on pasta X, right behind you»
 - Instant memos: «you do not have pasta in your cart, sure you do not need it?»
 - Instant ads: «we have a new type of pasta right next to you»
 - Instant recipes: «if you buy mascarpone you can cook tiramisu»
 - ...

In-store simplicity: couponing



PERSONALIZATION enables a variety of options:

1. Coupons are generated batch and sent via email (to be printed/shown on smartphone/activated on Web site) BEFORE a customer's next shopping experience.
2. Customers can CHOOSE to which item to apply a coupon within a set of preselected items.
3. Customers are shown INSTANT COUPONS based on cart content.
4. Coupons are automatically applied and NOTIFIED to customers.
5. Products can be BUNDLED in different ways on one coupon (e.g. up-sell, cross-sell, agreements with suppliers etc.).
6. Special coupons can be generated for CLICK&COLLECT.
7. ...

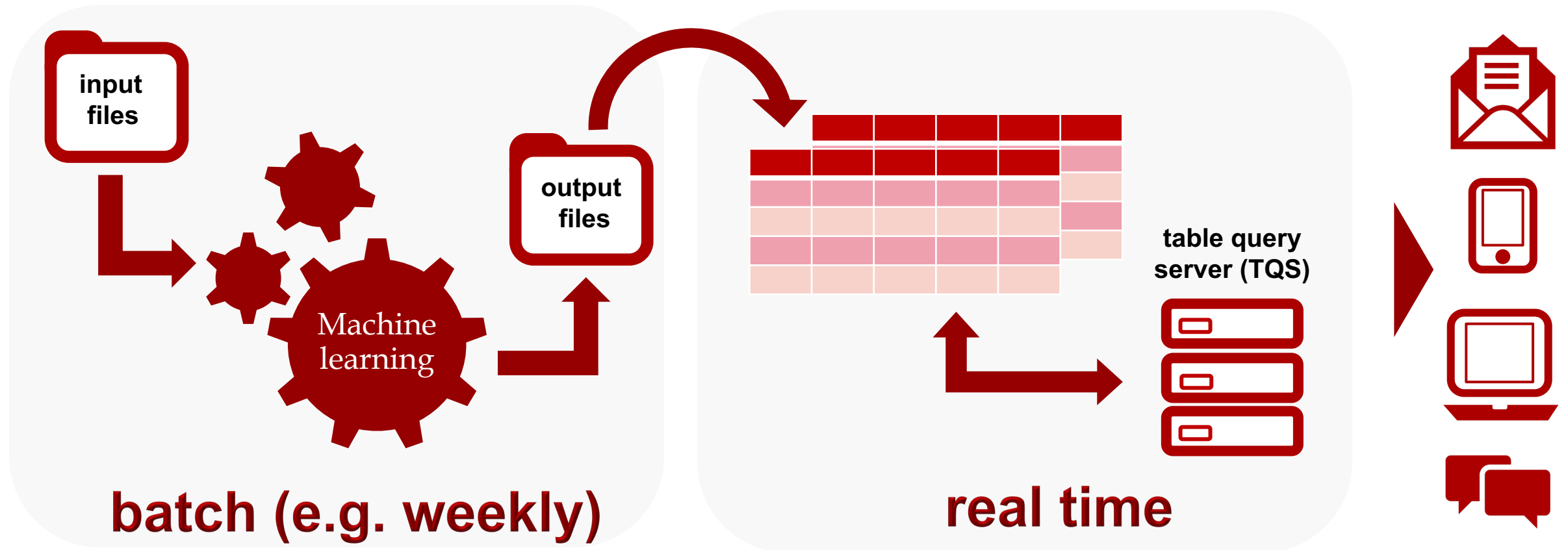
Benefits for customers

- Personalization of recommendations: items that are selected for recommendation have a strong connection with customers' habits.
- Product discovery mechanism: items that are selected for recommendation span across the entire product catalog.
- Serendipity: recommended items are often discounted or have a price that is lower than the habitual product's price (although they have a greater margin) or are involved in loyalty initiatives...

Architecture

production of tables – batch server

access to tables – real-time server



Deployment

- The batch component can be deployed on premises or in cloud.
- The real-time component should be deployed on premises to minimize latency.
- What is the TCO of the architecture and what the ROI from targeting?

Conclusions

- Big data is getting bigger: need for scalable and inexpensive solutions
- Start with basic analytics looking for quick fixes
- Experiment with advanced analytics looking for quick fixes
- Start on a longer term change program when big data projects have reached consensus with many stakeholders