11-Big Data Analytics- framework and case studies

Closely related to machine learning, artificial intelligence and deep learning...

Artificial Intelligence vs. Machine Learning

- Artificial intelligence is intelligence demonstrated by machines, in constrast to the natiral intelligence displayed by himand and other animals
- - Machine Learning 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
    So you cannot use them in cases where you cannot have less than 100% reliability and trust as LLMs will give an answer even if they do not know the answer, LLM and errors go hands in hands so we need to think if we can overcome these errors. You can combine different methods to reduce the error on data.

Artificial Intelligence

- Capabilities generally classified as AI include successfully understanding human speech, competing at the highest level in strategic game systems(such as 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
  Having data in your system is a key factor in using artificial intelligence

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 w.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. It can be created, also with algorithms, but need correctness in the construction as this will be the foundation of your work.
  Topic extraction needs a semantic network to be constructed. Ground truth is expensive to be built as it is manual and for this is also error prone.

Deep Learning

Class of machine learning algorithms where:

- Use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation, Each successive layer uses the

Natural Language Processing

NLP was performed with semantic(embeds structures of sentences to the words to perform more refined tasks) or syntactic(recognize a string, highly precisional) engines. Problems is that different languages have different structures so a semantic engine is specific for only one type of language.
Many challenges in NLP:

- Speech recognition
- Natural Language Understanding
- Natural Language Generation
  Still limited in the benefits of talking to a machine instead of typing. Still difficulties in length of the prompt and the costs that they required(energy, infrastructures,...)

ChatBots

It is a Bot that works for us. It can do work for us that we could do in some clicks. This involves an inversional paradigm of mobile technologies: instead of installing and opening an app we talked to a chatbot so if a company wants to provide us with services through the chatbot they would become invisible as the chatbot will not promote your brand but only give out the services.
Company are questioning if LLMs are enterprise ready. Nowadays we can integrate LLMs in enterprise world with APIs, but there are some barriers:

- Allucinations: LLMs **always** provide an answer, if they do not have the answer to the question they will invent it. Depending on the training set they could also know wrong answers. LLMs learn by whatever content that is provided to them, no filter to their learning process so they can deteriorate. Can be still used if the answer doesn't need to be always

correct/errors can be coped. Data science want to provide a consistent story and do not care too much of the correctness of the story(wrong by scientific means, but it is not a wrong process as can help to hide some errors/need to have someone that contradict you). Power BI has LLMs integrated in the backend, not exposed to the user yet.

- Need to extract data from large datasets. If you want to extract data from the original dataset you can upload it to the LLMs, but if these data are an asset of the company and do not want to let them leave your corporation you need to be protected by data theft from LLMs. Do not know how much errors on data analytics the LLMs do. Simple data acquisition is good to do, but too complex can be detrimental to the objective of learning from data(Example LLMs still cannot replace coding, still need good developers).
  As LLMs have replaced web searches these give them an edge to the distribution of information.

Machine Learning application priority(by industry)

Customer segmentation is still done by people, not using data analytics to confirm the experience as LLMs are not mature enough for the task.
Changing market segmentation to search for a better revenue, when you are in a good situation, is not really a good idea as it is difficult/expensive to change my market segmentation. Changes take a lot of time to be enacted.
LLMs are far away from doing data science, there are ML application to do data analysis but there is still need to be an expert that do the majority of work.
Can build an agent that do a lot but there is still time to have a correct agent that use new ML and not the old way of doing ML.

Machine learning (AI) business KPIs

SECTOR	PROBLEM	KPI
Finance	Risk exposure assessment	Loss reduction
Accomodation	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

Why should AI be related to business KPIs?

Problem with ML is that it is a blackbox. ML better than a linear model as can catch non linear relation in our data and relationship.
Explainable AI: tries to gain trust for ML even from non technical people, guide people through the exploration process for ML. Done through qualitative analytics, use simple qualitative stuff with elementary school Math. Make sense of the complex relationship between data without using ML. Then make a step further and introduce a correlation model. Not just a matter of model but also share methodology.
Also the adoption of technologies is tied to the monetary value gained(if not much still better to maintain status quo). Can also automate decision creating new jobs and not touching the work of existing people. Easier if you are creating knowledge and not substitute it.
You should have some KPIs tied with ML applications.

Evaluation of business KPIs

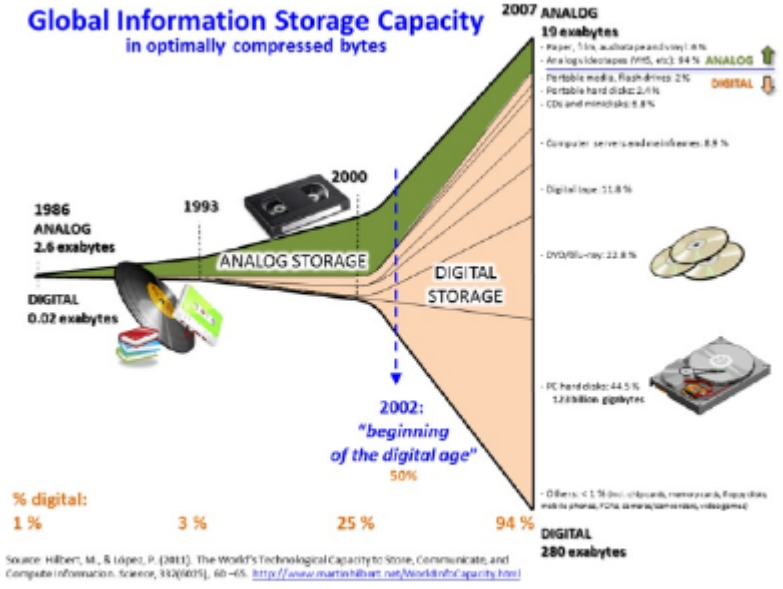
The benefits of AI/machine learning use cases are rarely quantified.
There’s a lack of business benchmarking initiatives. Quantitative evidence almost exclusively comes from suppliers of technology solutions.
For some use cases, economic benefits are difficult or impossible to quantify (KPIs are simultaneously affected by multiple initiatives).

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.
A lot of answers derive from technologies application and flops from computer engineers.
Usually innovation take in new skills so the people that could be replaced need to be able to pick up these new skills to be desirable for the company, introduce new tasks that are more complex and then simplify the old one.
ChatGPT let know passed the message that AI is complex but its usage is simple and anybody can use it(good approach to market, the product is complex but it is perceived as simple).
A pilot can have positive and negative results, but if there is some continuity there could be some beneficial aspects, but a company when it see a failure stop to invest in the technology without even looking inside the possible reasons and capabilities.
Data needs to be aligned in time and space for each feature, problem as the company have different databases, data warehouse and different views for the same object, no fully integrated data. If accuracy is key to the decision you need data quality, so getting high quality data is complex.

Why big data? Why now?

ML, LLMs need a lot of data to be trained.



26

Integration of data is needed for big data as they let the automation of gathering data, the web boosted big data analytics. It created a new application where the marketing people needed to set the parameters of the recommendation system and keep an eye on the recommendation system performances. Need to keep an eye on the ML algorithm as they could make mistakes.

How big is big data

> 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 accellerating.»

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.
  Few companies can develop technology for big data so they are limited from what is on the market. From complex analytics I can extract information. Big data and analytics take in consideration complex Math.
  We live under impression that few Terabytes are not big data but when we look at ML algorithm we see that even some Gigabytes are difficult to handle correctly, scalability issues. Need to hit barriers when the competitors hit barriers.
  Two types of scalability
- Technical: the technology can be applied to reach the scope
- Econiomical: the technology is profitable for the company

Types of Big data

Conversation Text Data

Amount of data depends on how big the brand is, small amount of data possible for local brands.
The problem reside in the semantic engines. You count the resources needed for each brand,

Image, audio and video data

Can reduce the data with a less resolution image, possible to have on demand analytics creation.

Sensor Data

Create data on the collection of the sensors analytics, usually thrown out after some days as they can also create privacy concern.

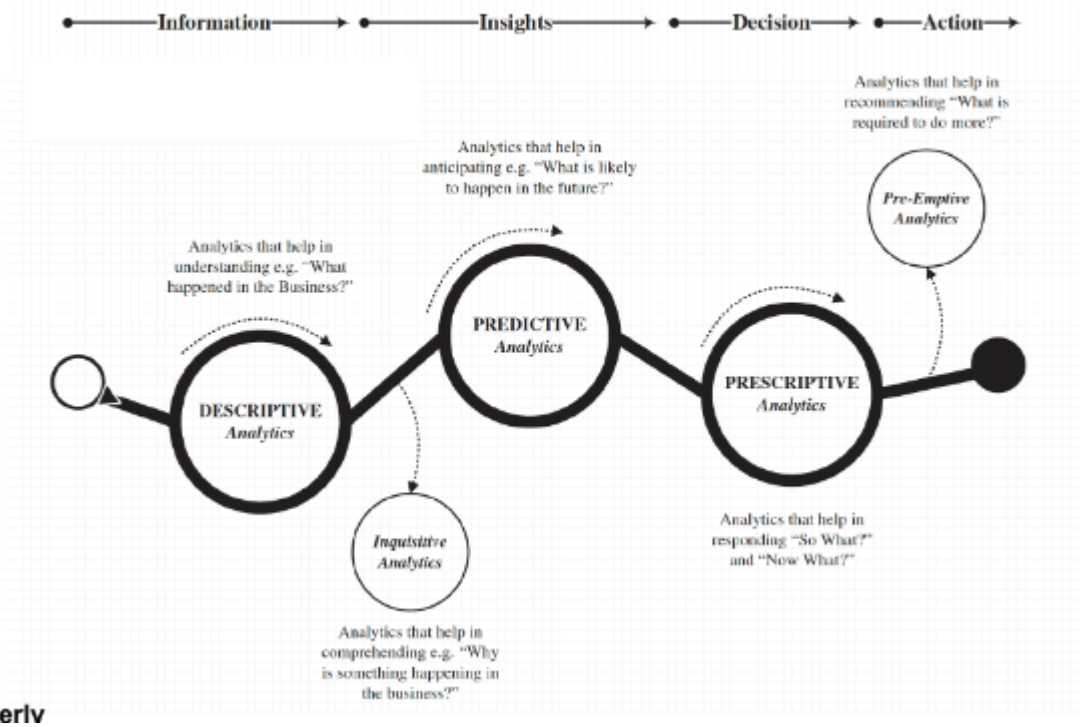
The Internet of Things data

IoT produce a vast amount of data.

Web customer data (Web logs)

Click stream analysis(pre-sales behavioural data). Useful to understand the behaviour of the clients before they buy something(High end of the funnel, they generate big data amount).

Evolution of big data projects



arterly

At the start no idea of how the data are disposed, if there is seasonality, etc... so you do Descriptive Analytics. DO these analytics to make sense of data. Then you do predictive analytics helps in predicting trends/do better predictions(managers like it more). Then there is Prescriptive analytics if there is a system that can process well the outputs of your models. For example satellites can see the movement of the earth and transmit the possible changes that will happen in the world, can set alarms to see if there are some catastrophes happening. Innovation is though, a lot of challenges as innovation can be misinterpreted or not trust.

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
   Consultant company will put labels on their employees even before the employee has the skill, need to be updated on the new things(only impression, maybe there is a small budget

on the technology using pilot). Even if a pilot goes wrong the next one will be more successful as the employees have more experience. Usually companies are not in a hurry and are sceptical of your work. Things do not happens quickly, probably need more rounds. Difficulties to decoupling data analytics and economics as they are correlated but, usually, the economic employees use only excels and can see only a fraction of data making wrong decisions.

Easy and accessible open source: moving from Excel to MySQL

Problem with MySQL is RAM bottleneck, maximum data size is 10 times the amount of data in the cache you can have for the machine. Usually divide by ten from the row of Excel to have good performance.