EC09 – MSII

Big Data

Icam Strasbourg-Europe
Computer Engineering and Mathematics department

Lecturer : Cédric Bobenrieth (cedric.bobenrieth@icam.fr)

Organisation

3 CM

2 TD

4 Labsessions

1 Project

Objectives

Know how to plan the implementation of a Big Data analysis approach in an industrial environment

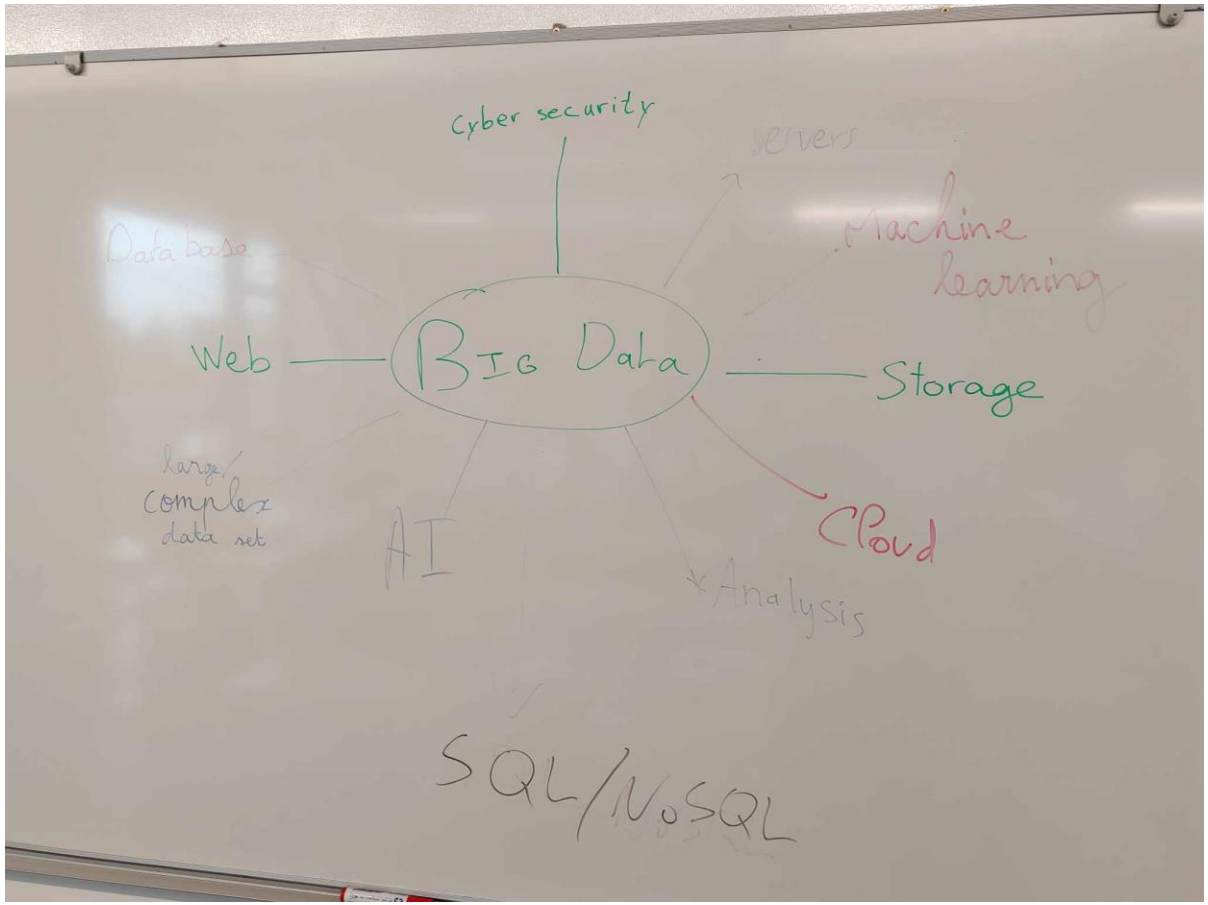
To be able to understand and improve the exploitation of data in an industrial company using Big data tools.

Know the tools, methods and sensors specific to industrial environments to obtain data.

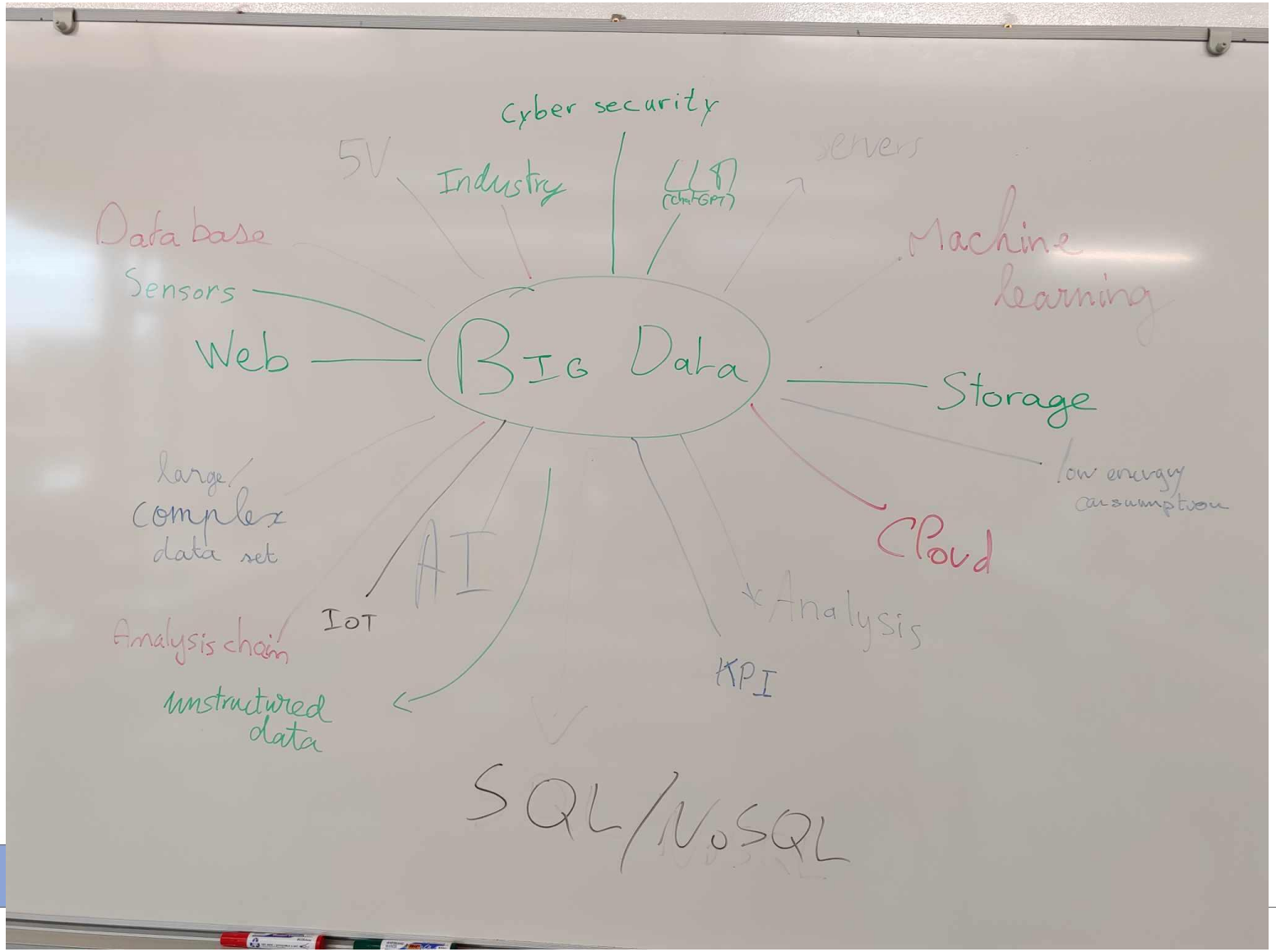
Use the main algorithms to process this data.

Mind Map

Mind Map


Cyber security
servers
Machine learning
Data base
Web
BIG Data
Storage
large/ complex data set
AI
Cloud
Analysis
SQL/NoSQL

Mind Map


Cyber security
5V
Industry
LLM
(ChatGPT)
servers
Data base
Sensors
Web
Big Data
Machine learning
Storage
low energy consumption
Cloud
large/complex data set
AI
IoT
Analysis chain
unstructured data
Analysis
KPI
SQL/NoSQL

Mind Map


Storage
Internet
Digital
Information
Analysis
Amount
Database
Business
Byte
Network
Size
Big Data
Cloud
Analytics
Cluster
File
Web
Volume
Server
Computing
Research
Software

The Lecture Outline

1. Concept of Big Data

- Introduction of Big Data
- The V's rules of Big Data
- Domains of Use
- The Data Analysis Chain
- Real example

2. Common tools of the Big Data

3. Process the data

Evaluation

coefficients	Written exam	Project presentation
	1	1

1 supervised work and 1 project

Introduction of Big Data

Definition

What's your definition of « big data » ?

Do you have examples ?

Introduction of Big Data

Definition

Big data => Very large amounts of data

Including :

Data produced on **the Web**, or by **sensor** and **machine systems**

Systems and tools used **to integrate** and **analytically** explore this data.

Different points of view :

Technical : Framework, Technologies

Business : How to use it and to value it inside of the organization

Introduction of Big Data

Stats

Why « the Big Data » ?

Every day we create **2.5 quintillion** bytes of data

Equivalent to 10 million of blu-ray discs

Stacked it's the same height as 4 Eiffel Towers

Data is growing exponentially

1992 : 100 GB/DAY
1997 : 100 GB/HOUR
2002 : 100 GB/SECOND
2013 : 28,875 GB/SECOND
2018 : 50,000 GB/SECOND

3 Billion of people have acces to internet

=

Earth population in 1960

Introduction of Big Data

Pragmatic Definition

"You know you have big data when you possess **diverse** datasets from **multiple sources** that are **too large** to **cost-effectively** manage and analyze within a **reasonable timeframe** when using your **traditional** IT infrastructures.

This data can include **structured data** as found in relational databases **as well as unstructured data** such as documents, audio, and video."

The V's rules of Big Data

The first three V

First reference in a 2001 report « 3-D Data Management : Controlling Data **Volume, Velocity and Variety** » by Doug Laney

Volume : Refers to the size of the datasets.

Velocity : Refers to the increasing speed at which data is created, as well as the speed at which it can be processed, stored and analysed.

<u>Every minute there are :</u>

216,000 instagram posts

204,000,000 email sent

72 hour of footage uploaded on youtube

277,000 tweets

The V's rules of Big Data

The first three V

First reference in a 2001 report « 3-D Data Management : Controlling Data **Volume, Velocity and Variety** » by Doug Laney

Volume : Refers to the size of the datasets.

Velocity : Refers to the increasing speed at which data is created, as well as the speed at which it can be processed, stored and analysed.

Variety : Refers to the different types of data that are available to collect and analyze in addition to the structure data found in a typical database.

90% of data generated is « unstructured » : tweets, photos,…

The V's rules of Big Data

The 5 V of Big Data

IBM then added the fourth V :

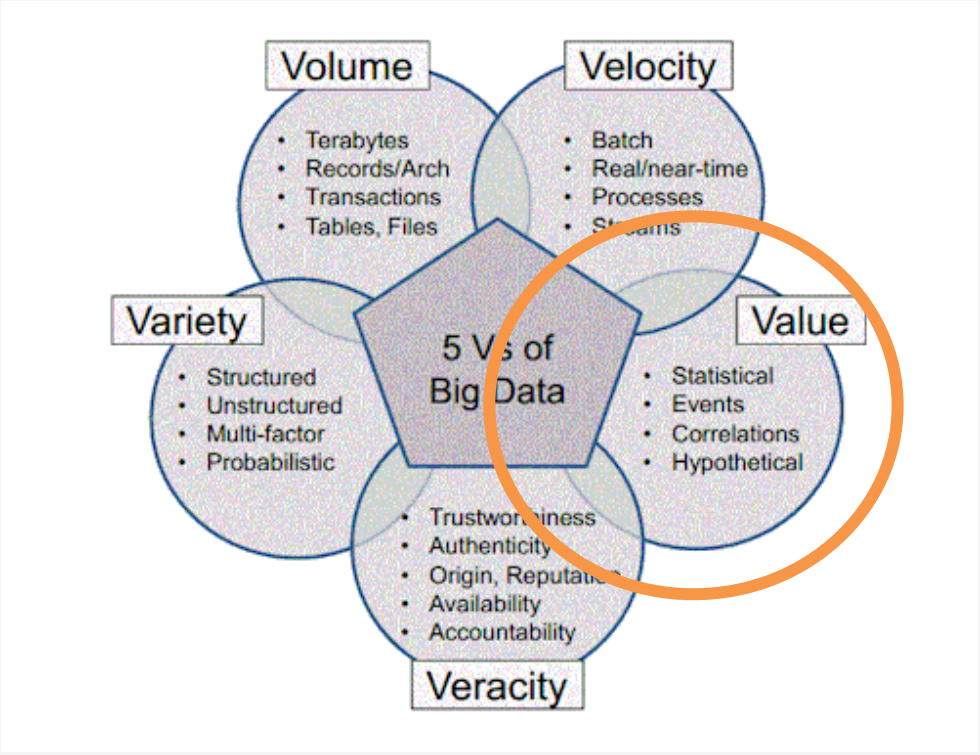
Veracity : Data quickly becomes obsolete and a lot of information that is shared over the Internet and social networks is not necessarily correct

IDC analyst Benjam Woo added one more V :

Value : Because big data is about supporting decisions, you need the ability to act on the data and derive value.

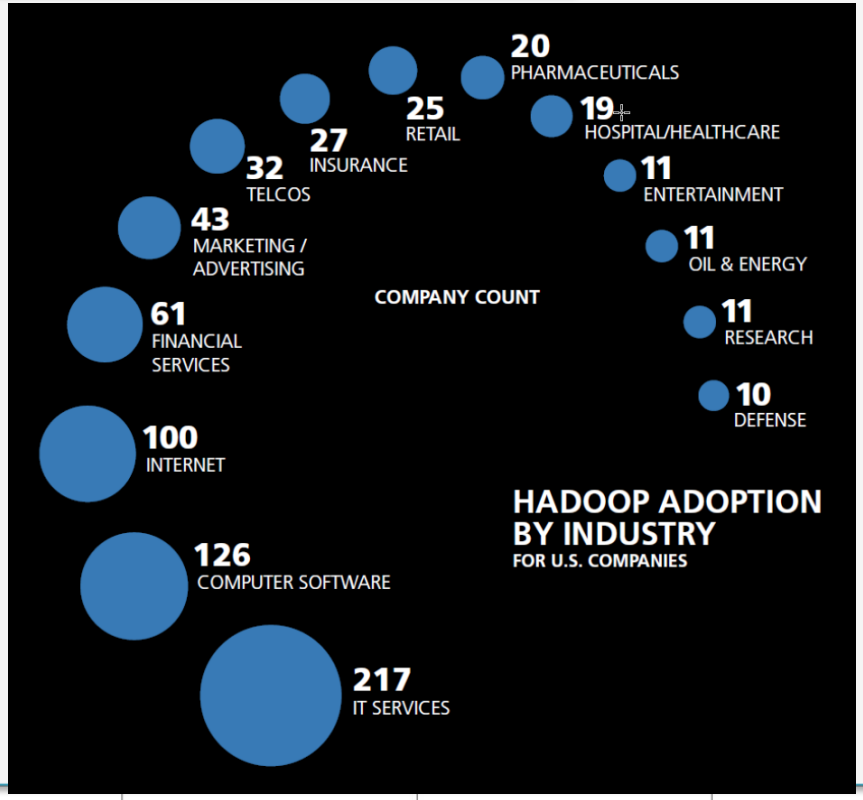
The V's rules of Big Data

The 5 V of Big Data



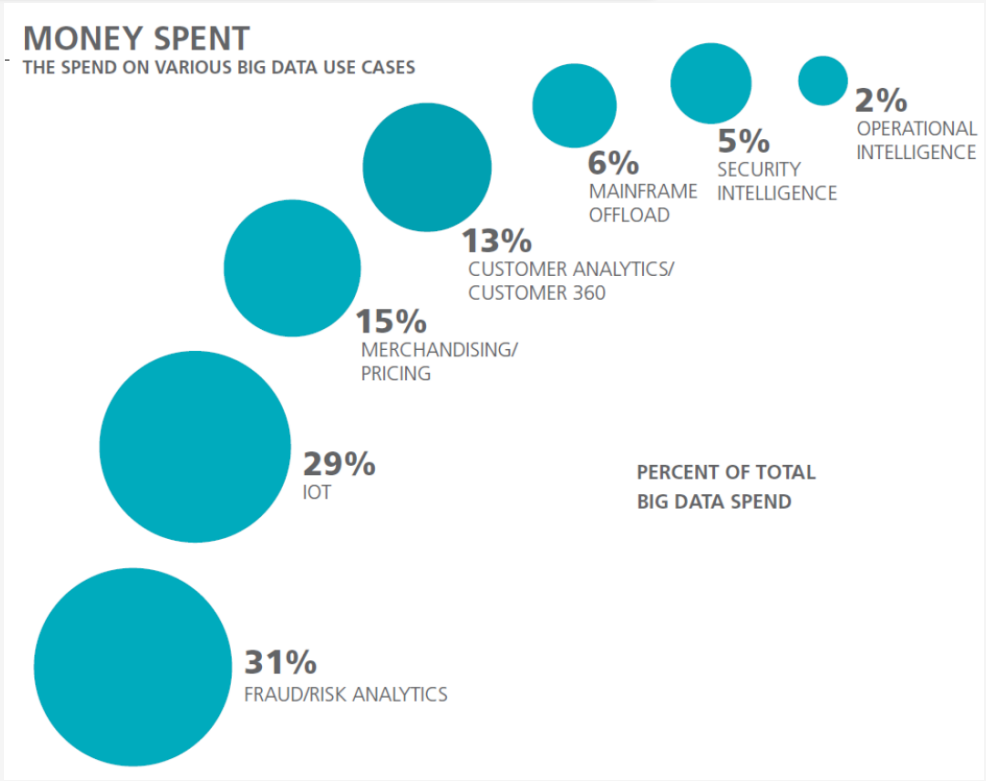
Domains of Use

By industry



Domains of Use

Money spent

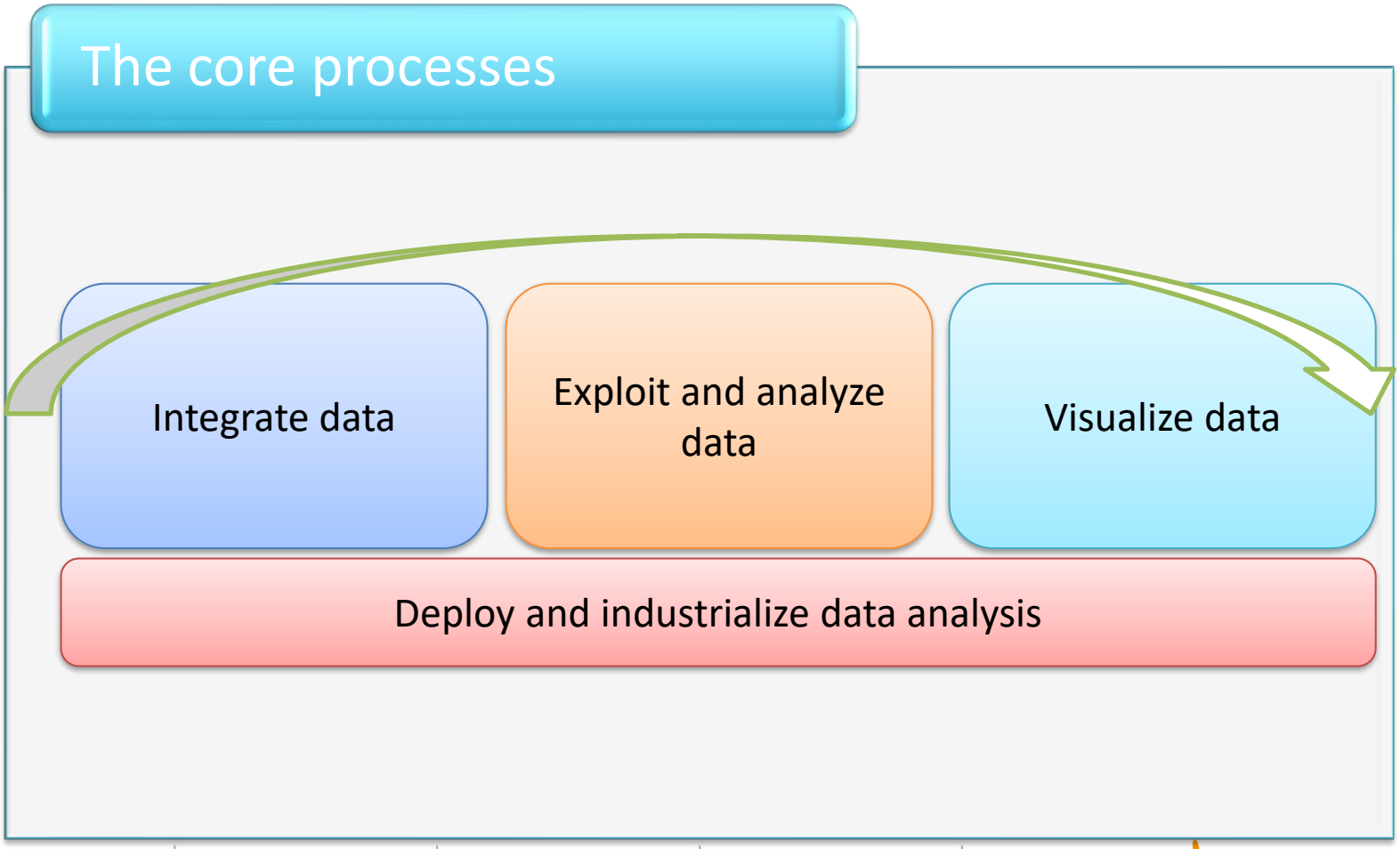


Domains of Use

Concrete use cases

Industry	Big data use cases
Automotive	Auto sensors reporting vehicle location problems
Financial services	Risk, fraud detection, portfolio analysis, new product development
Manufacturing / Production	Quality assurance, warranty analyses Digital factory (simulation), sensor-driven operations (reduce waste)
Healthcare	Patient sensors, monitoring, electronic health records, quality of care
Oil and gas	Drilling exploration sensor analyses
Retail	Consumer sentiment analyses, optimized marketing, personalized targeting, market basket analysis, intelligent forecasting, inventory management
Utilities	Smart meter analyses for network capacity, smart grid
Law enforcement	Threat analysis, social media monitoring, photo analysis, traffic optimization
Advertising	Customer targeting, location-based advertising, personalized retargeting, churn detection/prevention

The Data Analysis Chain



The Data Analysis Chain

Goal of the data chain

Analyse raw **data sets**

To address **one specific problem or question**

The Data Analysis Chain

Integrate data

Get the data needed to solve the problem

Gather data from **different sources**

Clean the data

The Data Analysis Chain

Exploit and analyze data

Explore the data

Identify significant features

Apply statistics and **Machine Learning algorithms**

The Data Analysis Chain

Exploit and analyze data

- Data set observation
- Analysis environment set up
- Exploratory data analysis
- Data cleaning
- Data Visualisation
- Statistical Analysis (ex: PCA; correlation)
- Machine learning Analysis (supervised; unsupervised)
- Reporting

The Data Analysis Chain

Visualize data

Show the original data

Show the output of analysis

Highlight the answer to the problem

The Data Analysis Chain

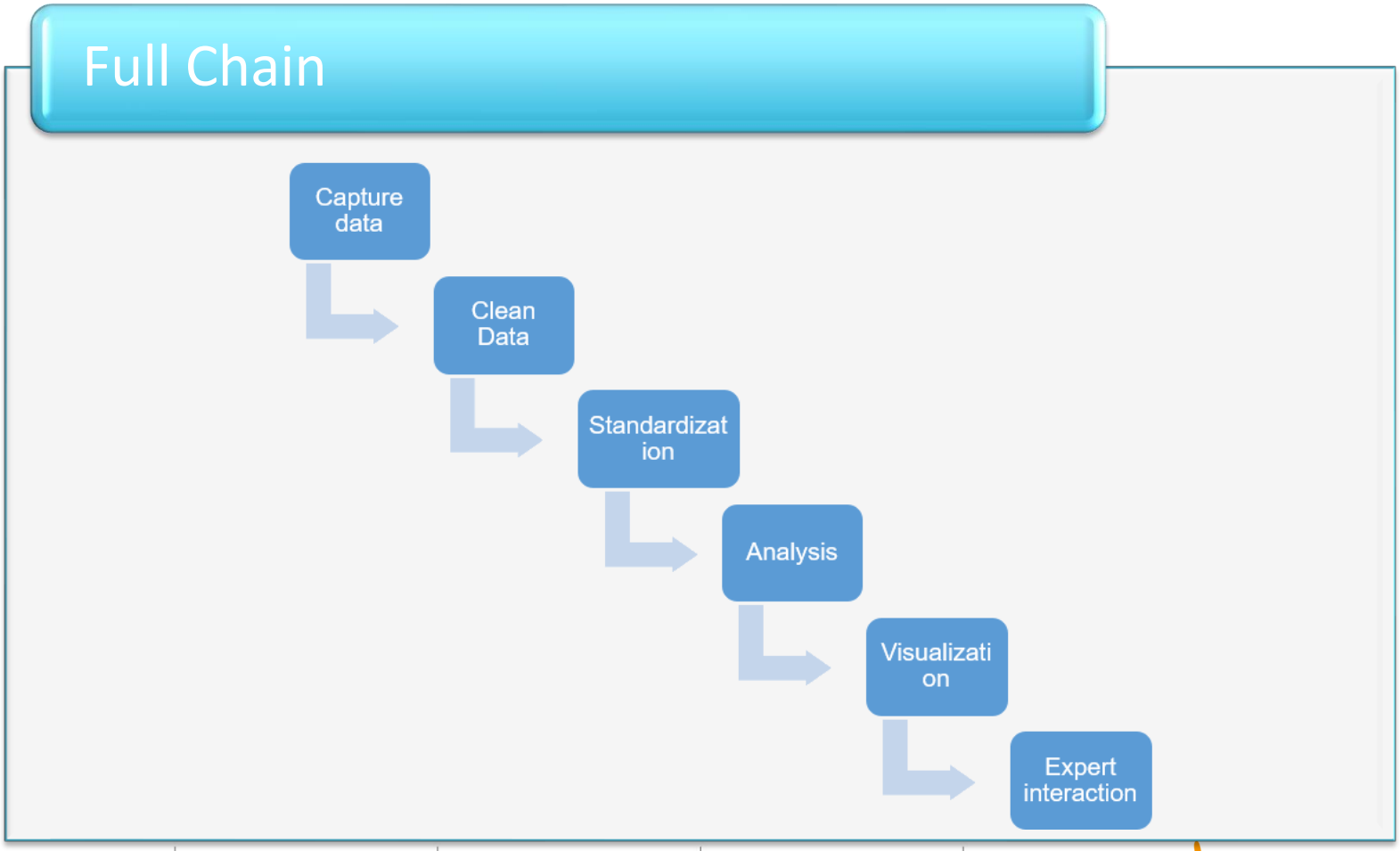
Deploy and industrialize data analysis

Ensure durability

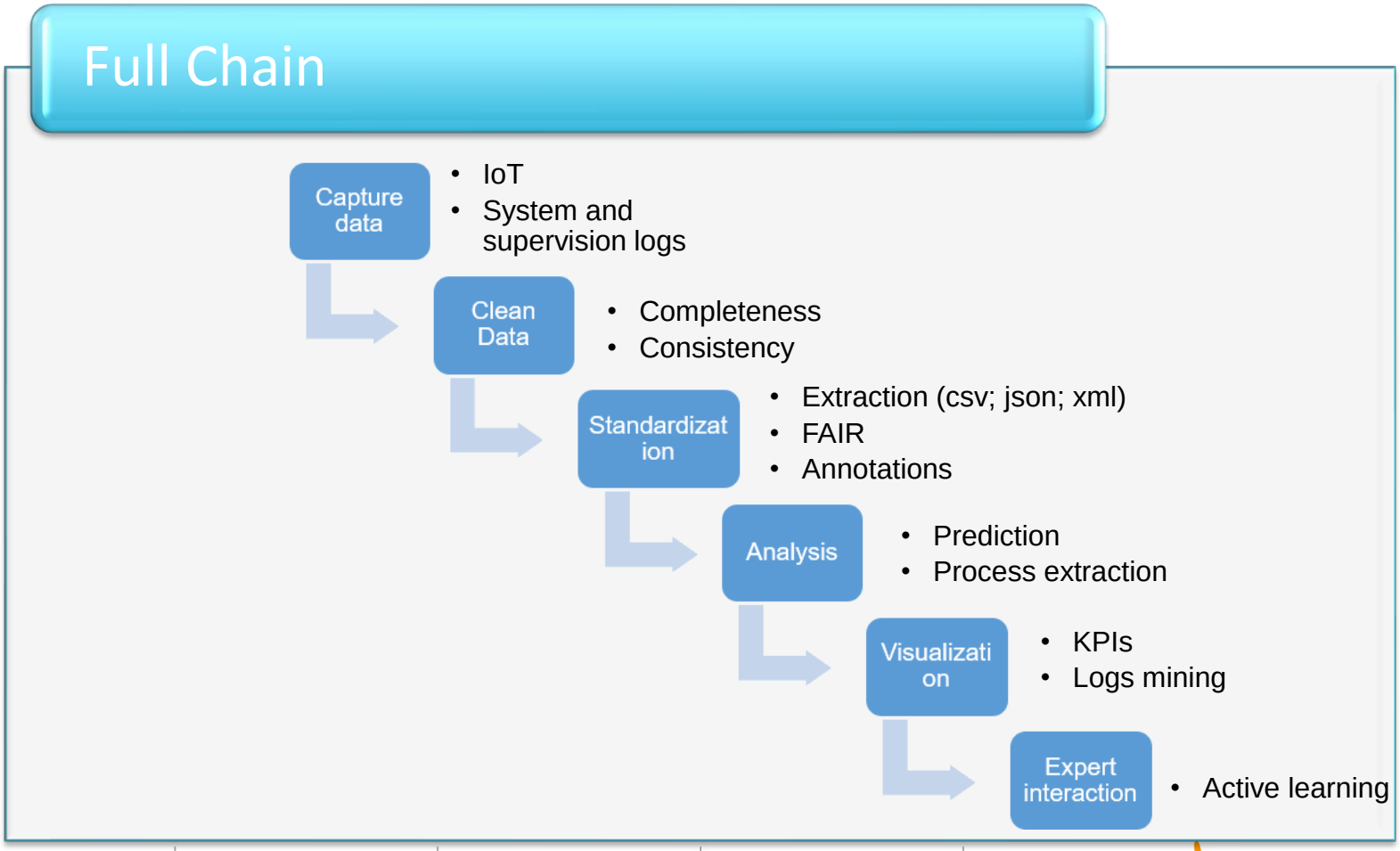
Ensure performance

Ensure compliance with legal requirements

The Data Analysis Chain



The Data Analysis Chain



Real examples

Capacity planning

Optimal utilization of resources is a key competitive advantage for logistics providers. Excess capacities lower profitability, while capacity shortages impact service quality and put customer satisfaction at risk.

The topology and capacity of the distribution network are adapted according to anticipated future demand

Transit points and transportation routes must be managed efficiently on a day-to-day basis.

Real examples

Customer value management

Big Data techniques, enriched by public Internet mining, can be used to minimize customer attrition and **understand** customer demand.

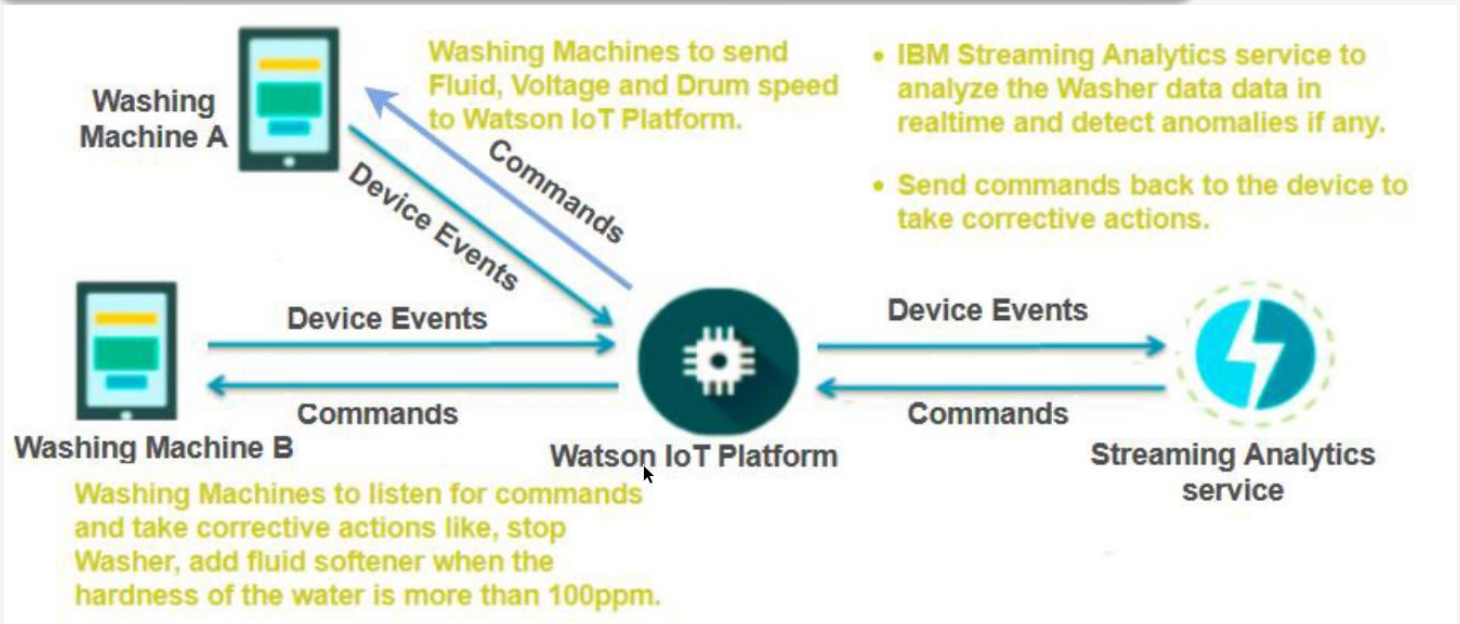
Smart use of data enables the identification of valuable customers

Produce an integrated view of customer interactions and operational performance, and ensure sender and recipient satisfaction.

Source : csi_studie_big_data

Real examples

IoT : Washing Machines



Source : *https://developer.ibm.com/recipes/tutorials/integrate-ibm-streaming-analytics-service-with-watson-iot-platform*

Real examples

Big Data save the Earth

DataCenter are **physical sites** that bring together all kinds **of servers** for storing, sending and receiving data.

They use a lot of electricity : 48 billion kWh

→ Produce a lot of heat

→ Need a ventilation system

→ Also use a lot of energy (40% of the total energy used by a data center)

Google use an **artificial intelligence system created by Deepmind** to automate cooling systems which result in 40% energy savings

And the AI ?

Big Data in AI

AI learn on data

To learn effectively…they need a lot of data !

Ex : ChatGPT

On the « whole » internet (Common Crawl

- → 499 billion of words
- → 45 TB of data

And the AI ?

Big Data in AI

AI learn on data

Ex : MidJourney 2,8 million of images + descriptions

Dall-E, 250 million of images + descriptions

LET'S FIGHT

Take your phone !

Questions ?