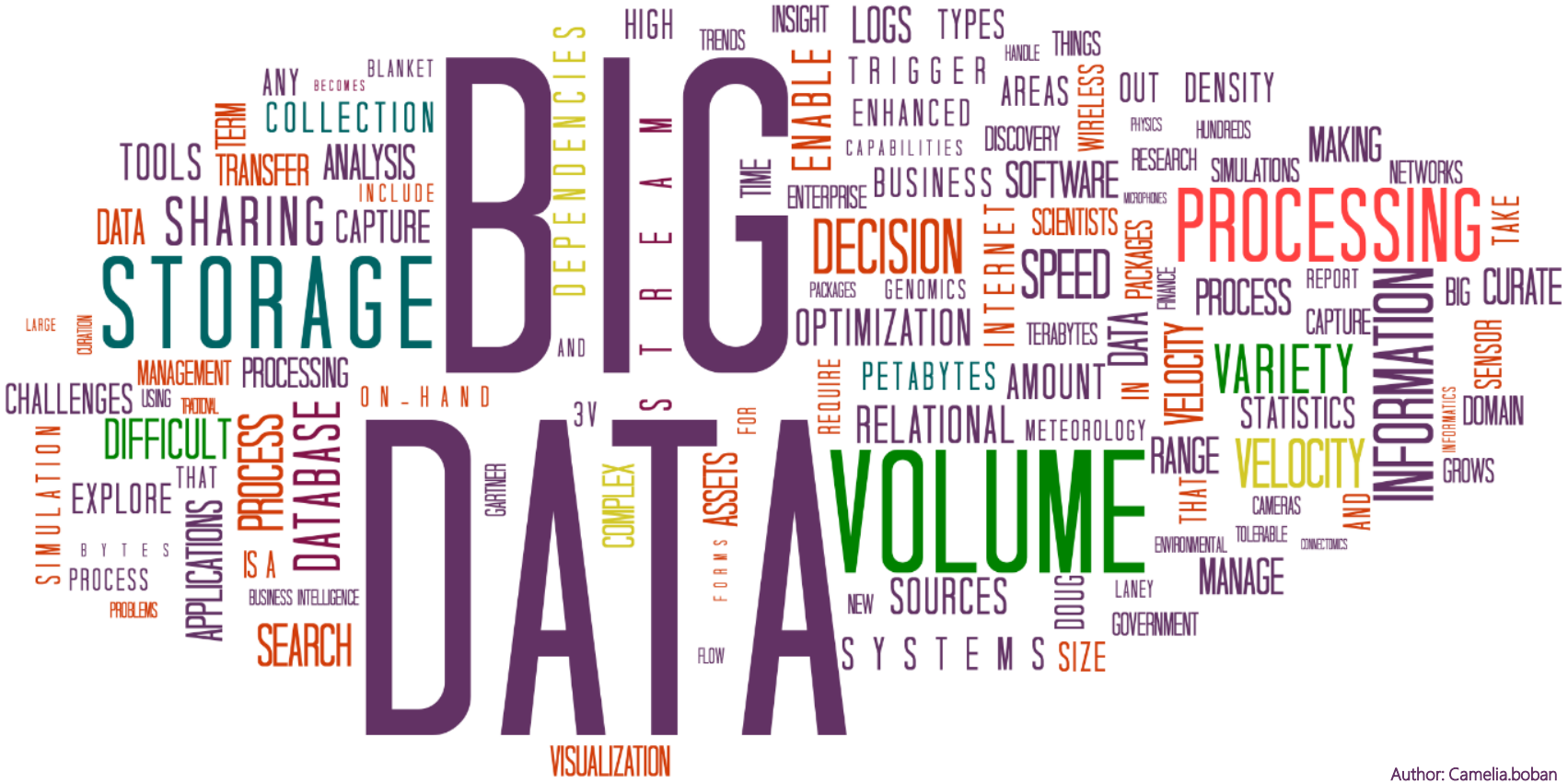
BIG DATA: A BIG CHALLENGE FOR THE INDUSTRY

ISAE-Supaéro, November 27th, 2018
Presented by Jong-Mo Allegraud, Mews Partners


WHAT IS BIG
DATA FOR YOU ?

Big Data has no clear definition



Author: Camelia.boban

Big Data scope & definition is unclear

It is used to designate sets of data that are too large for human intuitive understanding – and even beyond for traditional processing application tools

A few figures on big data

An exponential data growth

2012
2.8 exabytes (10^{18}) generated in one year

2017
16.3 zettabytes (10^{21}) generated in one year

5800x in 5 years

src: IDC data age 2025 white paper, 2017

A booming market

2013
$20bn revenues

2018
$42 bn revenues

2x in 5 years
(projection for 2027 : $103bn)

src: https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/#717e21d72926

An explosion of the computing power

1997
$2x10^5$ FLOPS per USD

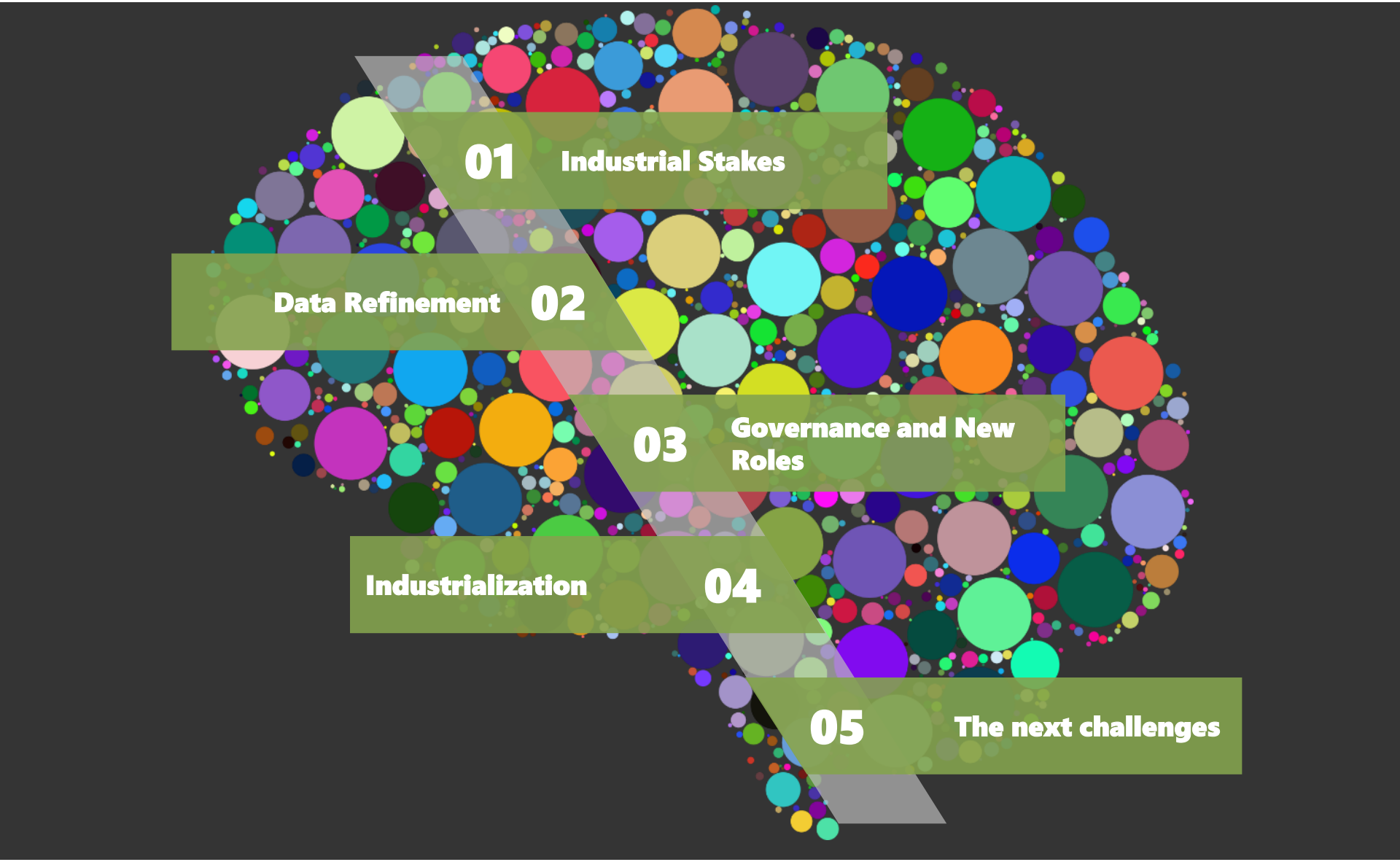
2017
$3x10^9$ FLOPS per USD

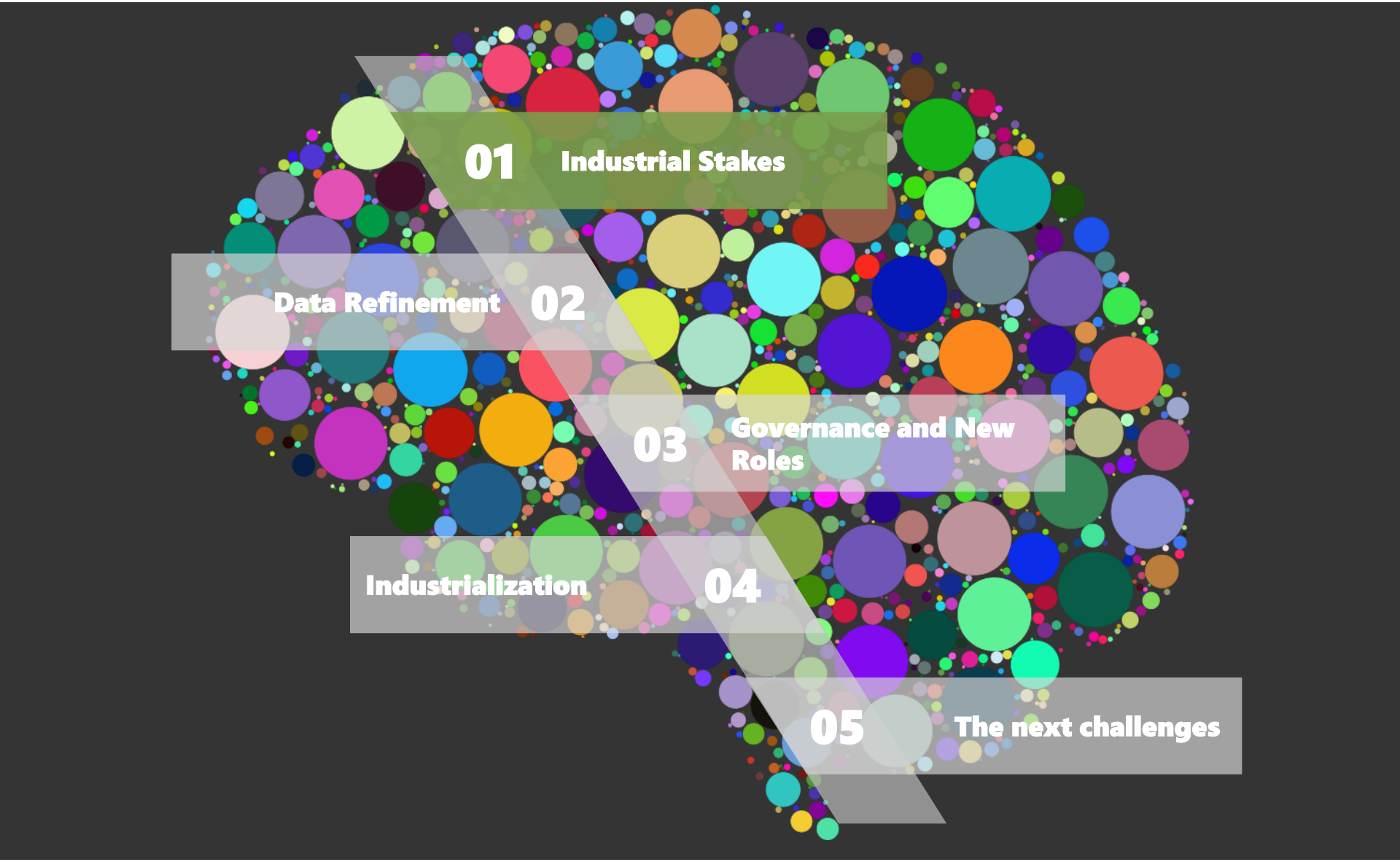
10,000x in 20 years

src: https://en.wikipedia.org/wiki/FLOPS

Today, data is one of the biggest challenges for the companies

Actually the real question is:


01 Industrial Stakes
Data Refinement 02
03 Governance and New Roles
Industrialization 04
05 The next challenges


01 Industrial Stakes
Data Refinement 02
03 Governance and New Roles
Industrialization 04
05 The next challenges

Big Data and Industry 4.0

Business stakes are evolving

- Real time
- Industrial work sharing more complex
- Mass customization
- Connectivity

Some new technologies now mature and affordable

- Artificial Intelligence
- Virtual Reality
- Data exploitation
- Industrial IoT
- Digital Twins
- Cobots
- 3D Prints
- Process evolutions
- Automation / AGV

Competitive
Performance, Quality, cost

Agile
Ability to reconfigure production, to better answer to changing customer needs

Digital
Digitalization of information flow, communication between product, machines, operators,

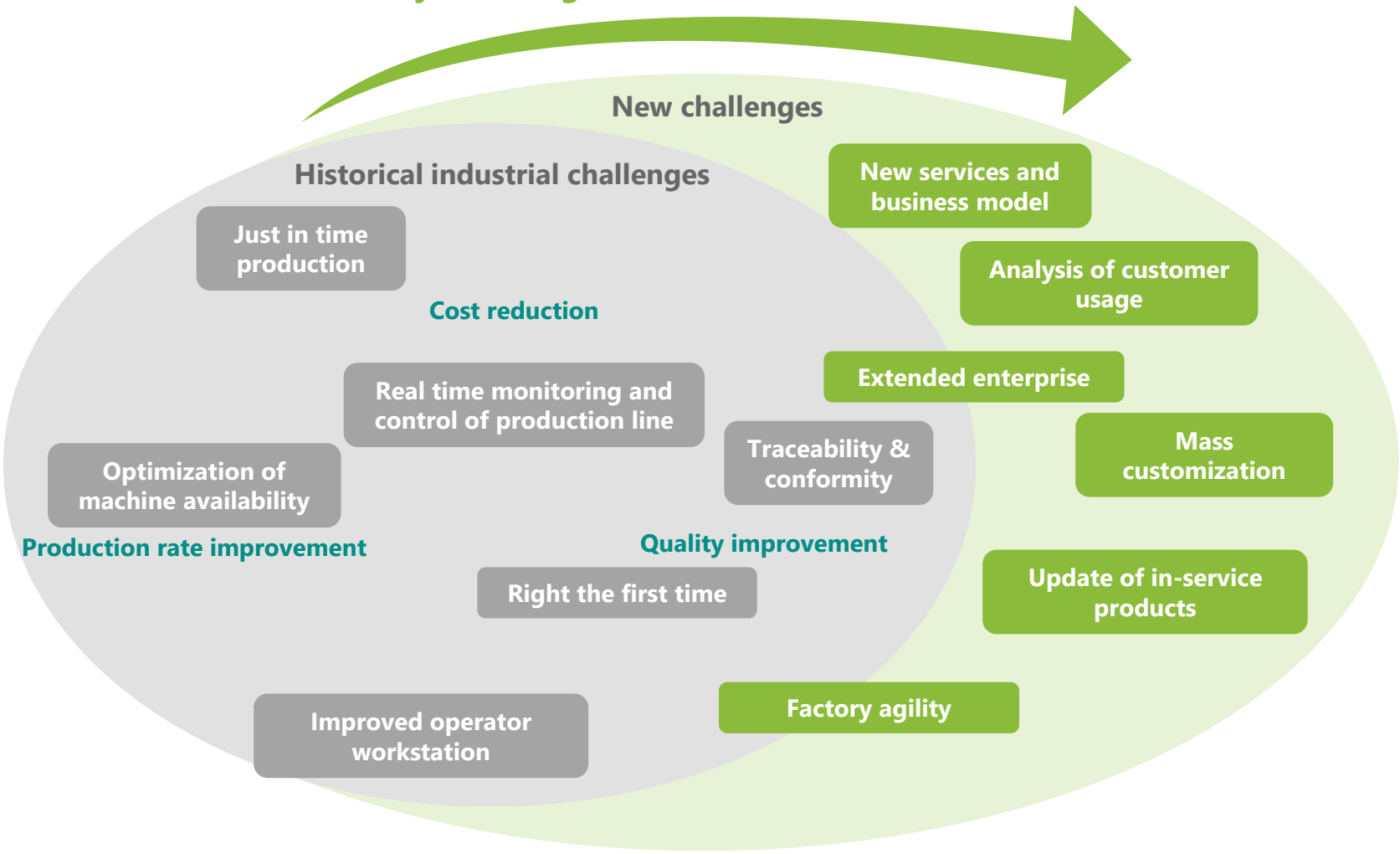
Extended
Integration with partners and suppliers

Human
Focus operators on added value tasks

Intelligent
Capability to self-assess and take corrective actions

Big data is an enabler for historic manufacturing stakes and new stakes

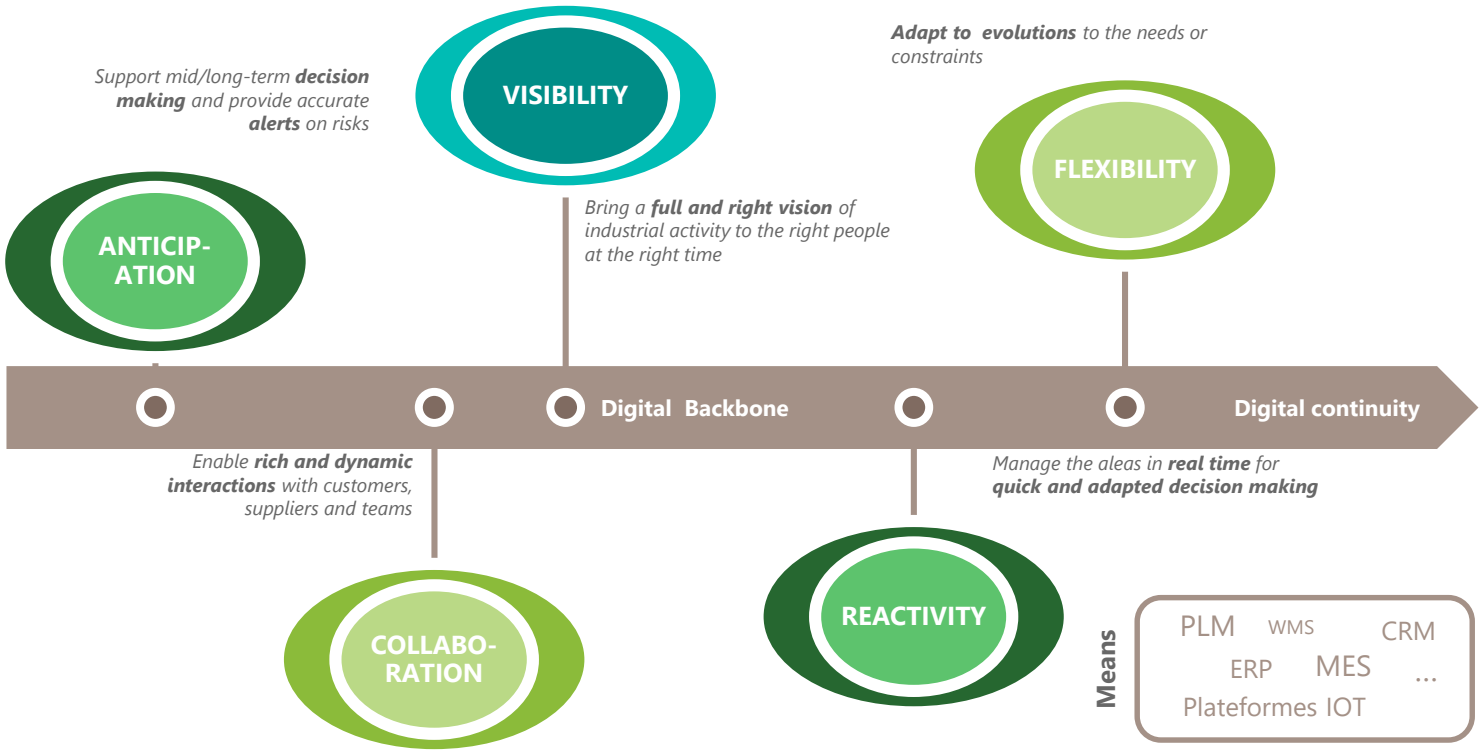
Not content to bring solutions to new industrial stakes, Industry 4.0 brings new answers to classic ones as well



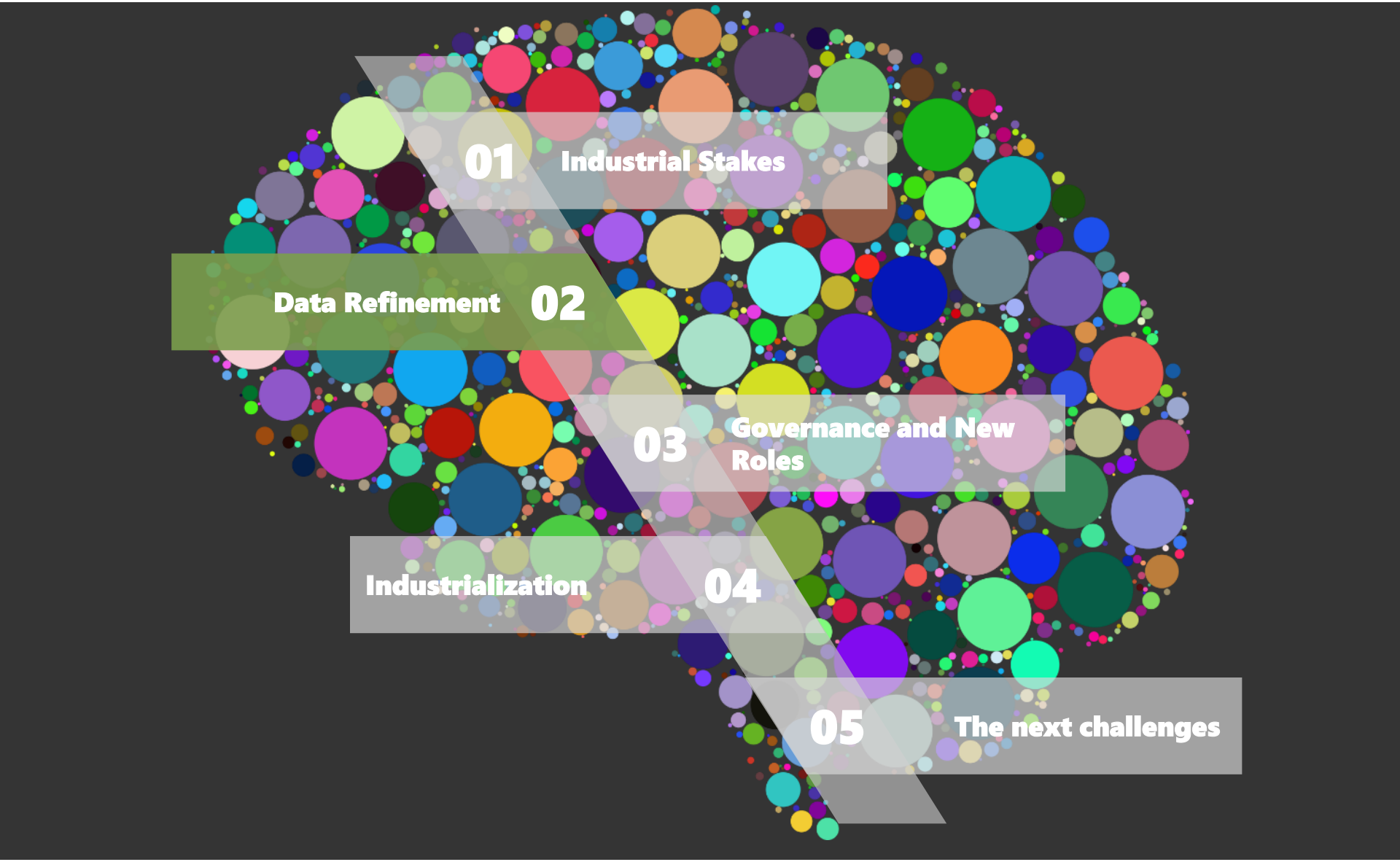
Big Data and Digital Continuity

Data and digital continuity at the heart of the industry transformation

Creating the digital backbone allowing to break silos between universes that are traditionally very distinct and leverage the mutualization of the various data



Big data is a major enabler for the digital continuity


01
Industrial Stakes
Data Refinement
02
03
Governance and New Roles
Industrialization
04
05
The next challenges

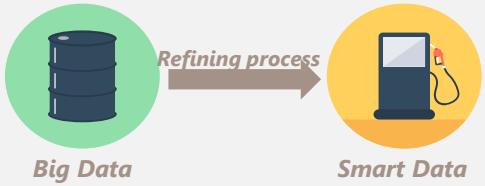
All that business needs is Smart Data

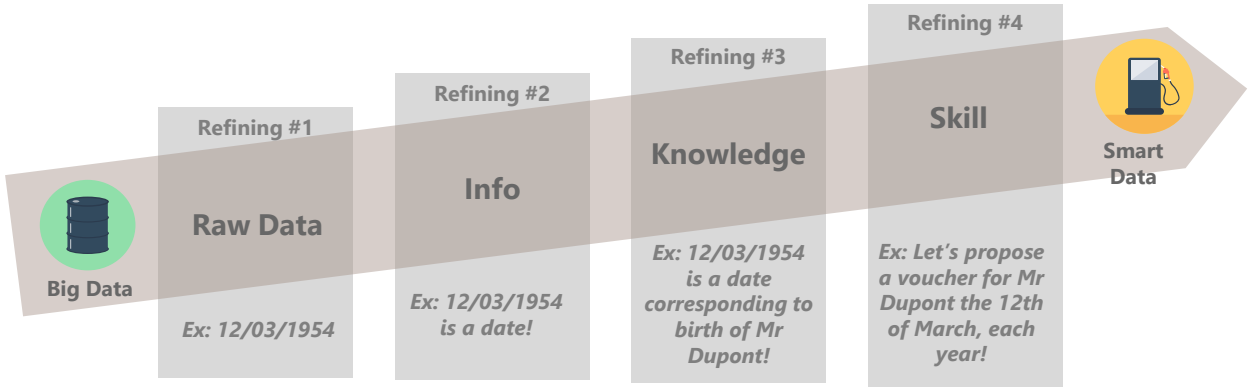
Smart Data vs Big Data

Principles

There is a similarity between Data and petroleum:

- **Data is now considered as the new Gold or new Petroleum**
- **Big Data is the crude oil** *(data in its raw format, even in emails)*
- **Smart Data is the refined oil** *(maximizing the value of data asset to enhance industrial companies performance)*

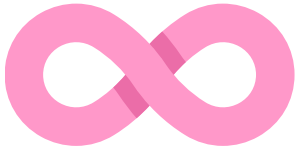




Source: François Cazals

... and Smart Data needs Big Data

Opportunities of smart data



Big data provides the opportunity to create digital continuity

By capturing data in its raw format (even in emails), without disturbing operations nor modifying the existing information system.

Improving decision making and forecast (real time cockpit)

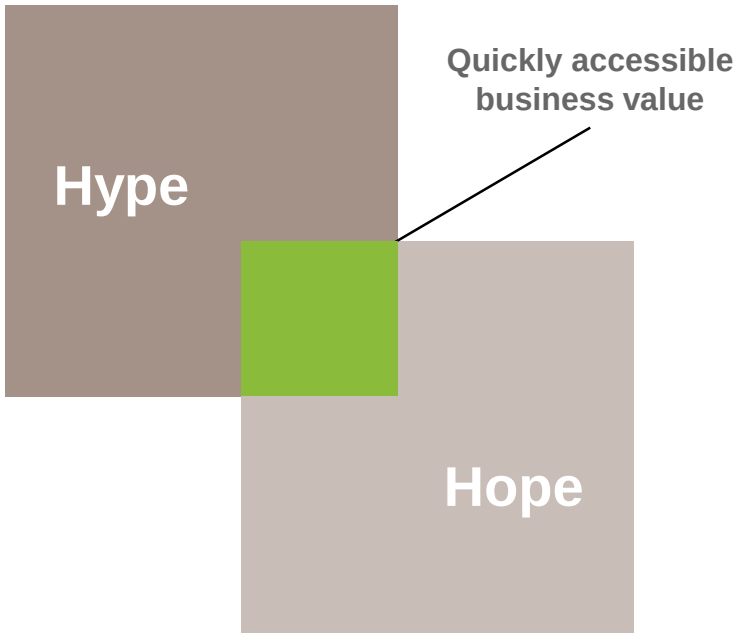
By setting-up 360° dynamic business views

Opening rooms for operational performance improvement (e.g. data quality, lead-time)

By tracking accurately key end-to-end data flows

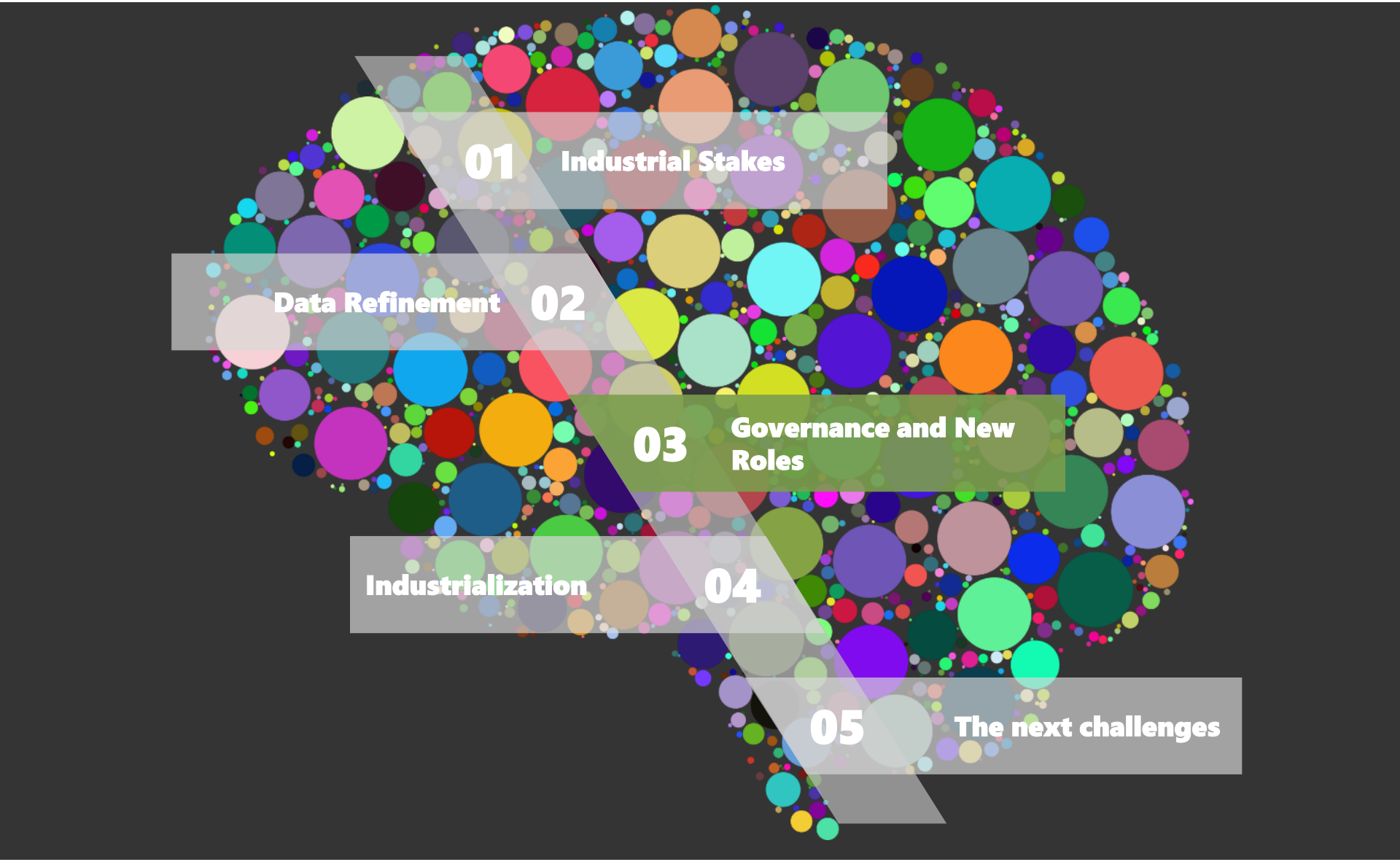
How can we generate smart data ?

Data Analytics and AI: Which approach to create Smart Data



Capturing the business value requires a combination of :

- **In-depth understanding** of your business processes
- A strong **data-science knowledge and experience**
- A successful and **proven methodology**


01 Industrial Stakes
Data Refinement 02
03 Governance and New Roles
Industrialization 04
05 The next challenges

What is data governance ?

Data Governance: a MUST in a big data context

ORGA

Data governance is moving from IT to Corporate level

Due to evolving context:

Data volumes and diversity, **within and out of the company** (big/open data)

Competitors, Customers and Suppliers **relationships**

Regulatory impacts (export control, GPDR, II901…)

MISSIONS

4 missions regarding data

Integrity
Availability
Usability
Confidentiality

2 ways of supporting company business

Secure running business
and
Enable new opportunities

TASKS

#1 – identify the right level of data

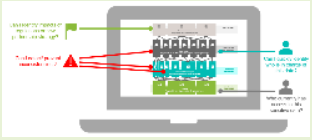
Detailed enough to comply with constraints, wide enough to be manageable

#2 – identify and animate levels of responsibility

From business to IT, data owners have to be identified at business and technical levels

#3 – create and maintain the data cartography

It supports the global risk analysis from business process to IT user rights



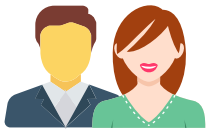
All of this, considering the work done in the past or still in progress at IT or business sides…

Change management is key to accompany the transformation towards this corporate Data Governance

The must company rethink its organization

Data implies new roles and jobs

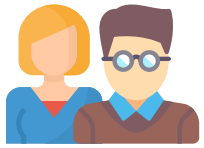
CDO/DTO
Leads digital/data initiatives at corporate level



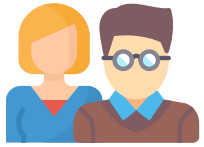
Data Governance Manager
Manages all data aspects (business value, data protection)

Data Custodian
Ensures data integrity and usability

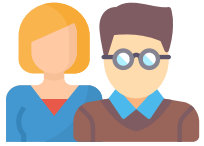
Data Steward
Ensures that regulation is respected



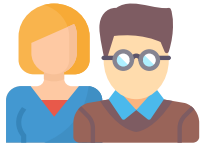
Data Architect
Designs de data architecture of the company



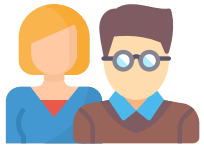
Data Engineer
Designs a data-based solutions



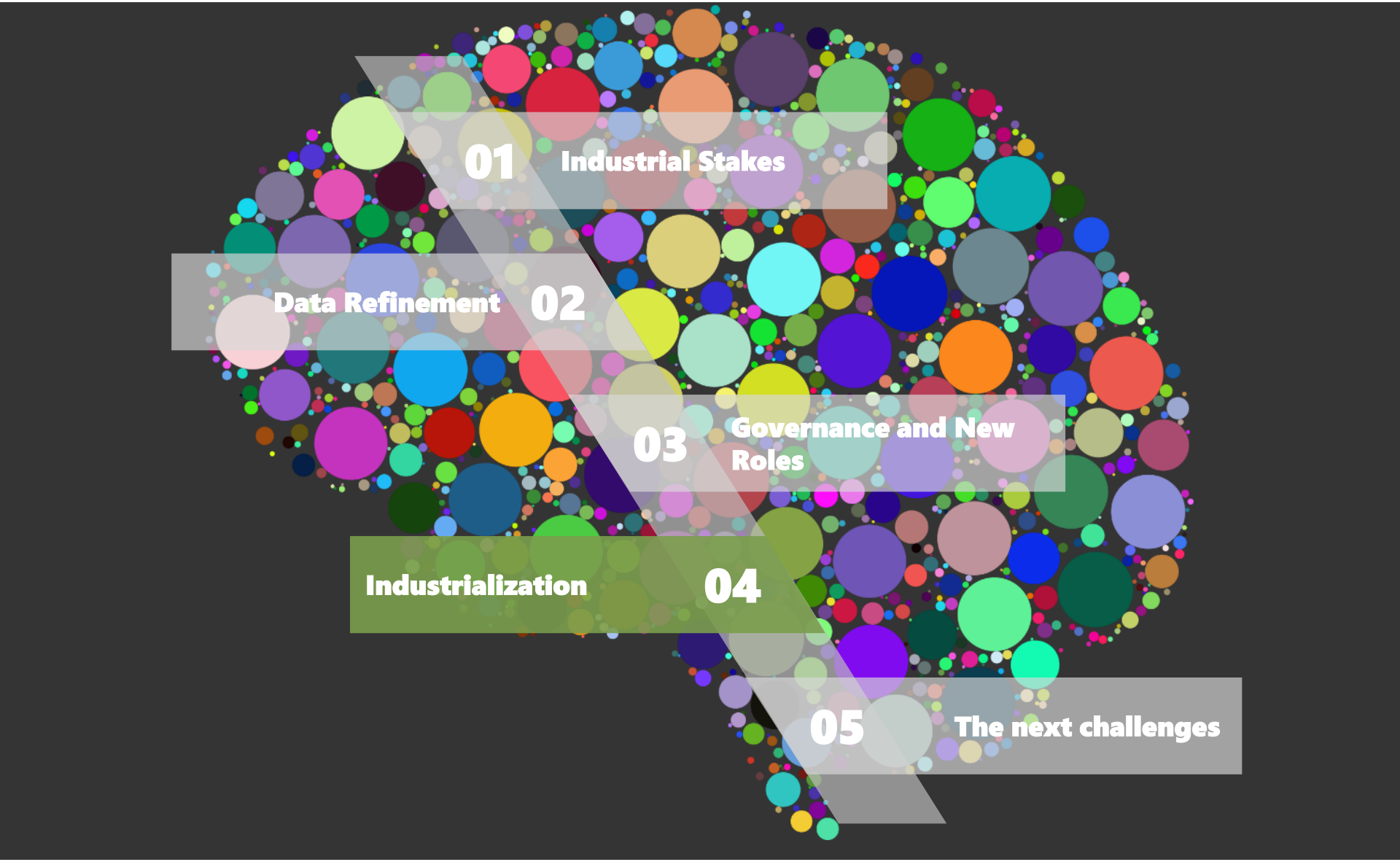
Data Analyst
Designs the data models of a solution



Data Scientist
Designs the algorithms to process the data



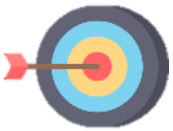
Data Visualisation Expert
Uses fancy data viz' tools to make clients understand what the team does


01
Industrial Stakes
Data Refinement
02
03
Governance and New Roles
Industrialization
04
05
The next challenges

Data Warehouse vs Data Lake

Data Analyst independence
News Needs

Data Warehouse

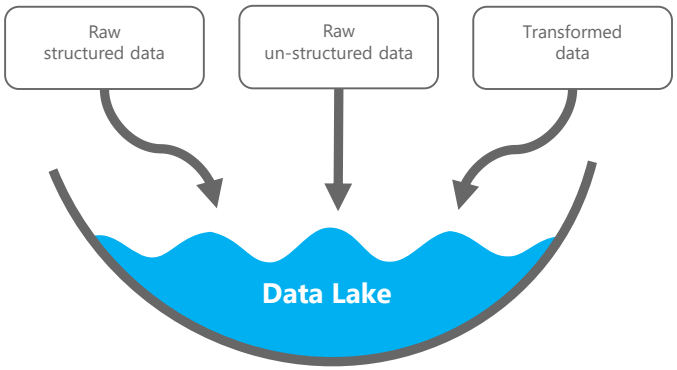


1. 1 place and 1 language for data from several IT systems
2. Recordability of data
3. Integrity and accuracy of the data

3 technical component

Staging Area	Data Warehouse	Data Mart
Prepare data before storage	Storage place	Sub-place of the warehouse providing transform data

Data Lake



- More analysis possible (Real-Time, Predictive, Data Discovery, ...)
- More flexibility for new Data Analysts, Data Scientists, Dev ➔ Schema on Read
- Best computing power based on new technologies

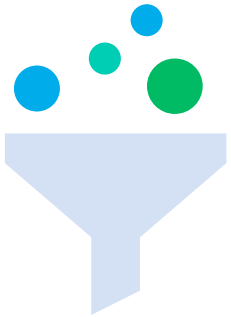
Source : http://www.decideo.fr/Le-Data-Lake-est-il-le-nouveau-Data-Warehouse_a9336.html

Data Lake is more flexible but can be painful to setup

Data ingestion

Identify data relevant for the use case

Select which data are needed



Gather the required data in a usable data set

Ensure the data integrity and usability

Make sure data flows continuously in production mode

The data stream must be automated

Key Success Factors

Ensure the project has a clear business value

Never forget a technical solution is used to solve a business problem

In the case of a PoC, make sure the PoC will resolve all the technical issues

One risk of a PoC approach is to simplify the problem too much

Ensure all the data needed are available

There can be a lot of reasons why the data are not available

Ensure all the data available provide a right description of the problem

A biased data set leads to a biased model

Scaling in production mode

Assess the workload of human actions in the process

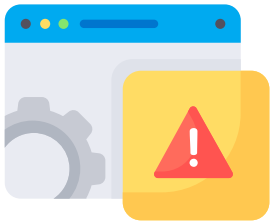
If some actions cannot be automated, make sure the human workload is sustainable

Incoming data must be clean

Data must be clean or cleanable automatically

Incoming data must be monitored

The availability of the data must be monitored and alerts must be triggered upon anomalies



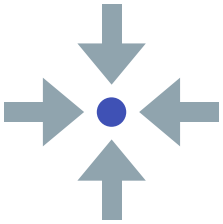
Output data must be monitored

Any undesired behaviour that can be detected automatically must raise an alert

Learning @ scale

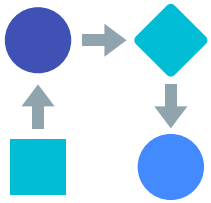
Assess labelling workload

Supervised AI algorithms require labelled data. Have a good view of the impact in terms of workload and delay



Define the process to collect new data to learn in production mode

How are they collected ? Do they need to be cleaned ? Labelled ?



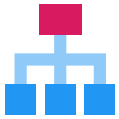
Define the algorithms update process

The process must guarantee there is no performance loss after an update

Some examples of missions @ Mews

Define and priorize the Big Data business cases for a digital roadmap

Identify business cases from data sets available. Assess technical feasibility and business value.



Support Data Governance Setup

Help to organize the data governance setup and facilitate the cooperation between the services.

Organization of a challenge to recognize ships in satellite images

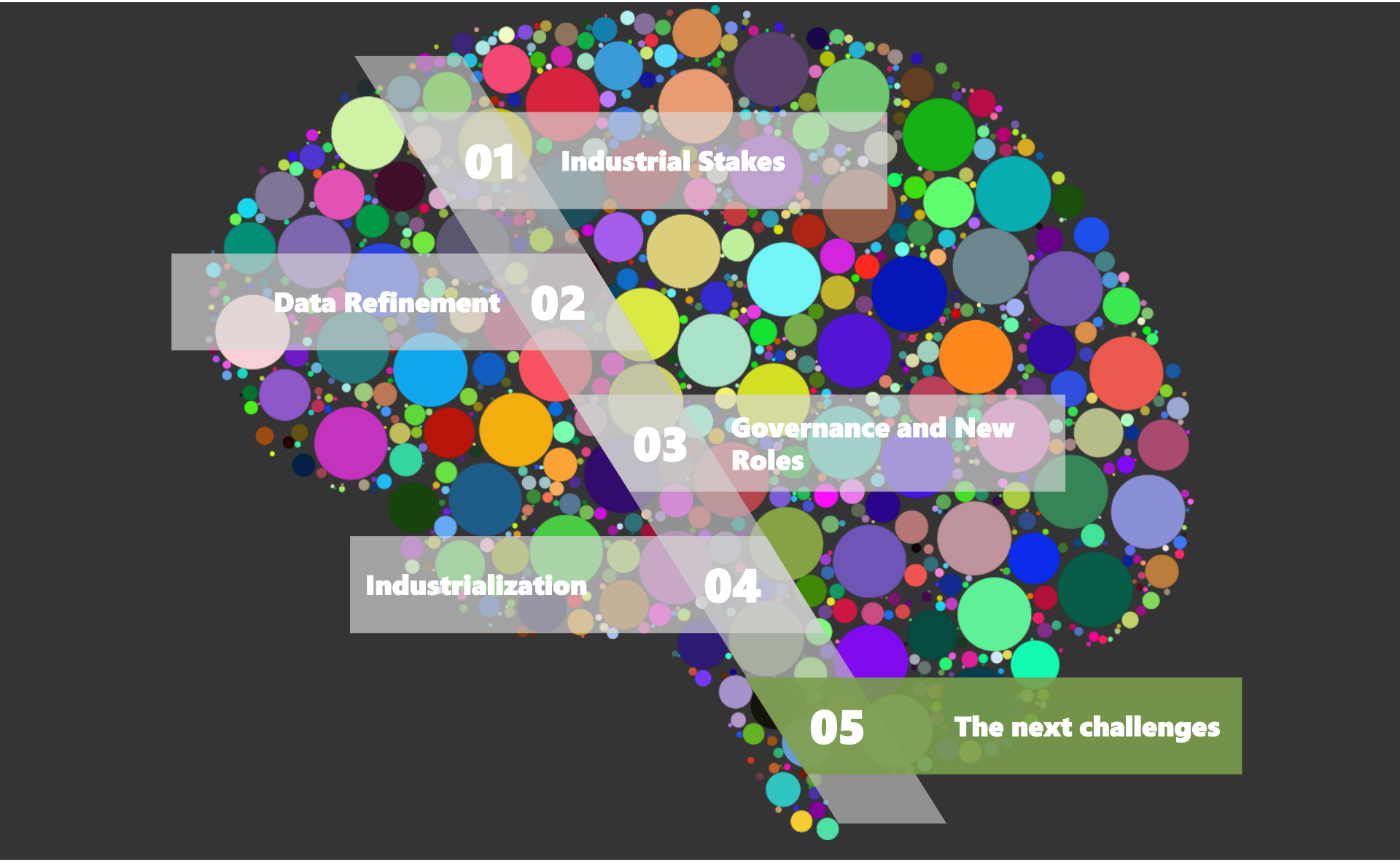
Collect and label the data, setup the evaluation procedures, post the challenge on the platform, communicate, follow-up

Establish the data model for a Transport Operating System

Clarify the data model, list the actions to undertake to ensure data usability, identify the data at risk (entered manually)

Simulate the impact of a ramp-up on the information system

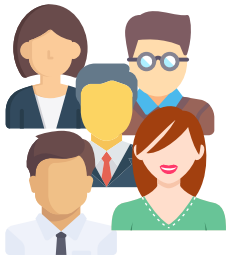
Model the network of applications involved in the business process and the applications health in a ramp-up scenario


01
Industrial Stakes
Data Refinement
02
03
Governance and New Roles
Industrialization
04
05
The next challenges

Biggest challenges in progress

Finalize data workflows

In a lot of industries data organizations is still in progress.



Onboard people in the transformation

Big data can raise a lot of fears towards employees. Any data initiative must keep people in the loop.

Extract value from company's legacy

A lot of knowledge is stored as unstructured data or people experience not present in any data set.

Handle the cost of data

The amount data generated has been multiplied by 5800 in 5 years but the storage cost has been divided by only 1.6 in the meantime.

Big data is used more by individuals than by companies

THANKS FOR YOUR ATTENTION

Jong-Mo Allegraud, Senior Data Scientist
jong.allegraud@mews-partners.com

4 bis rue Brindejonc des Moulinais, 31500 Toulouse – France
Tel. +33 5 62 88 78 00
www.mews-partners.com