fondazione banfi

SANGUIS JOVIS

ALTA SCUOLA DEL SANGIOVESE

VI Edizione
SUMMER SCHOOL SANGUIS JOVIS

SANGIOVESE PHYGITAL:

L'impatto della tecnologia dalla vigna al Metaverso

fondazione banfi

SANGUIS JOVIS

ALTA SCUOLA DEL SANGIOVESE

VI Edizione

SUMMER SCHOOL SANGUIS JOVIS

DIETRO LE QUINTE DEL DIGITALE:
Il dato

DIETRO LE QUINTE DEL DIGITALE: il dato

- **AGENDA**

- Who am I ?
- Introduction
- The Dataverse
- A different view on Data

Introduction

01

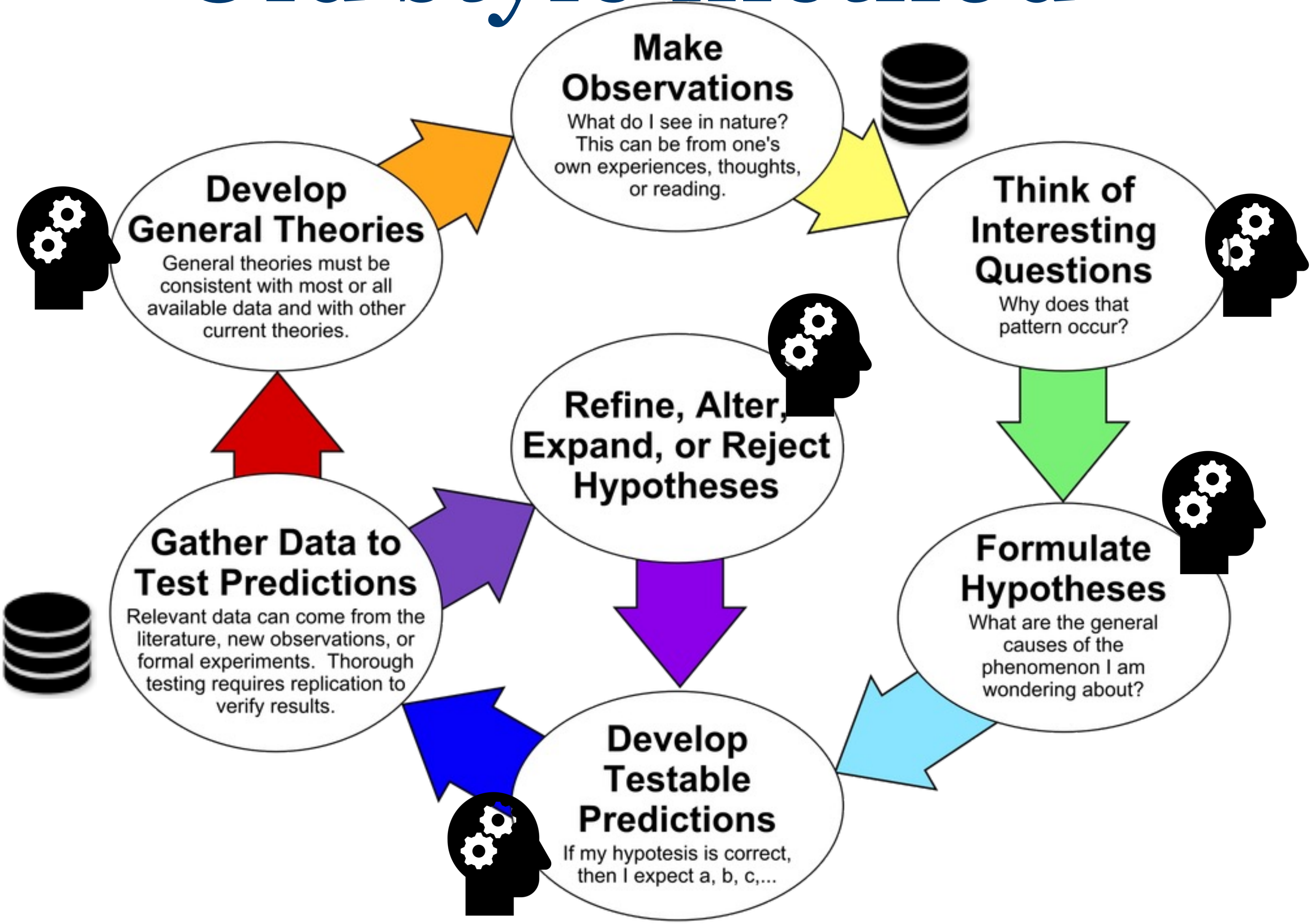
Introduction

BEHIND THE SCENES ???

It’s not **DIGITAL** transformation

It’s **THOUGHT** transformation !

Old style method



Make Observations

What do I see in nature? This can be from one's own experiences, thoughts, or reading.

Think of Interesting Questions

Why does that pattern occur?

Formulate Hypotheses

What are the general causes of the phenomenon I am wondering about?

Develop Testable Predictions

If my hypotesis is correct, then I expect a, b, c,...

Gather Data to Test Predictions

Relevant data can come from the literature, new observations, or formal experiments. Thorough testing requires replication to verify results.

Develop General Theories

General theories must be consistent with most or all available data and with other current theories.

Refine, Alter, Expand, or Reject Hypotheses

Old style method

The Scientific Method

The scientific method might be the single most powerful idea humans have ever had, and progress since the Enlightenment has been simply astonishing.

Old style method

The Scientific Method	Limits
The scientific method might be the single most powerful idea humans have ever had, and progress since the Enlightenment has been simply astonishing.	The problem is that these challenges are **so complex**.

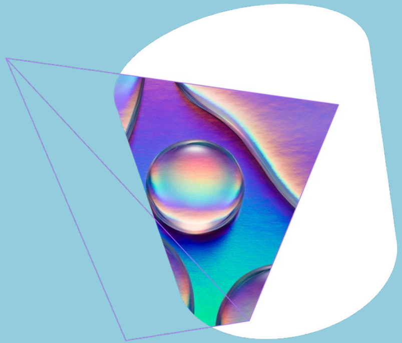
Complexity

Go has **10^{170}** legal positions
Observable universe contains **10^{82}** atoms*

*Scientific estimation are between 10^{78} to 10^{82} atoms in the known, observable universe. Example of source: https://www.universetoday.com/36302/atoms-in-the-universe/

Complexity

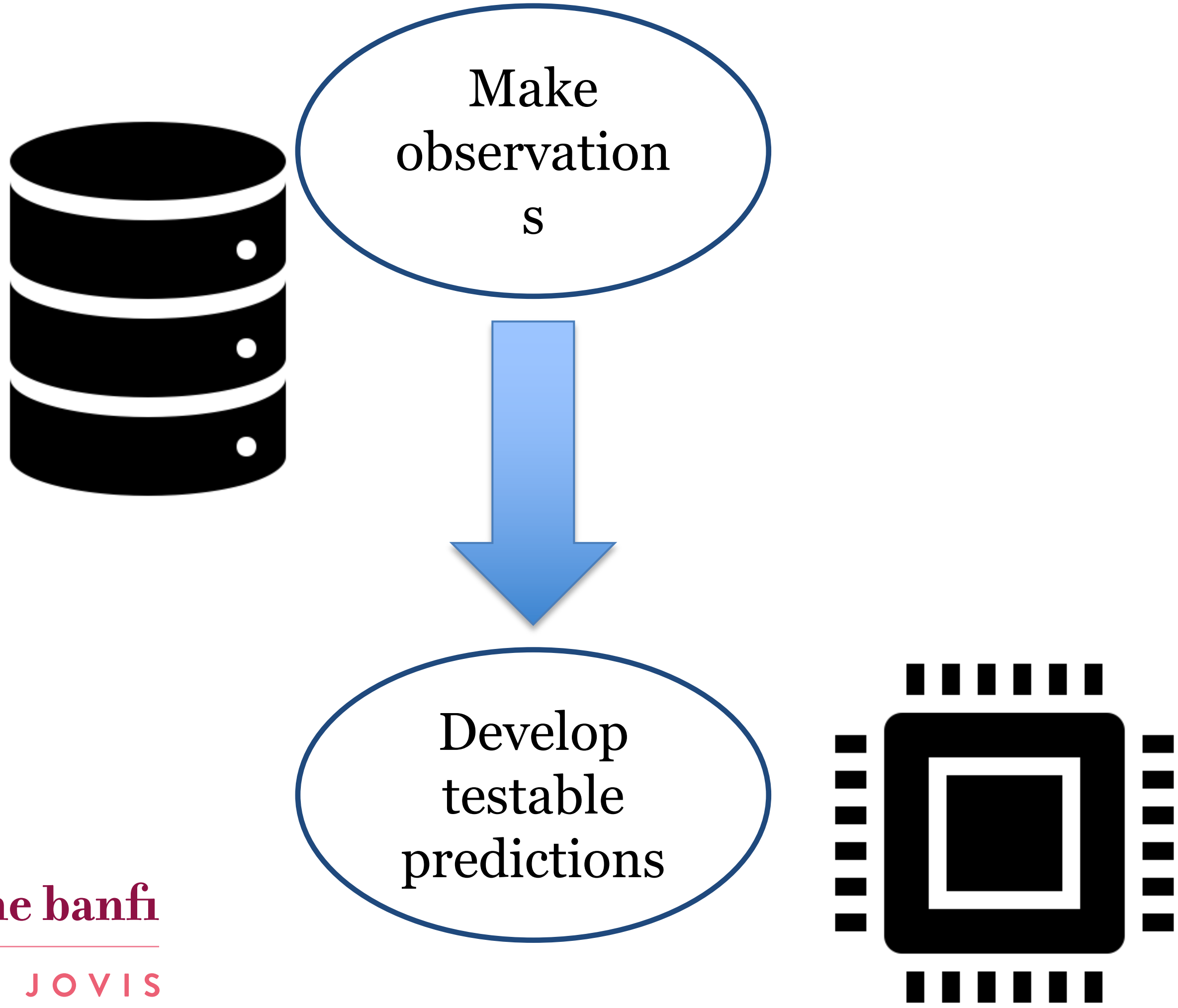
Go has **10^{170}** legal positions
Observable universe contains **10^{82}** atoms*



AlphaGo is the first computer program to defeat a professional human Go player, the first to defeat a Go world champion, and is arguably **the strongest Go player in history.**

*Scientific estimation are between 10^{78} to 10^{82} atoms in the known, observable universe. Example of source: https://www.universetoday.com/36302/atoms-in-the-universe/

How did this happen ?



Make observations

Develop testable predictions

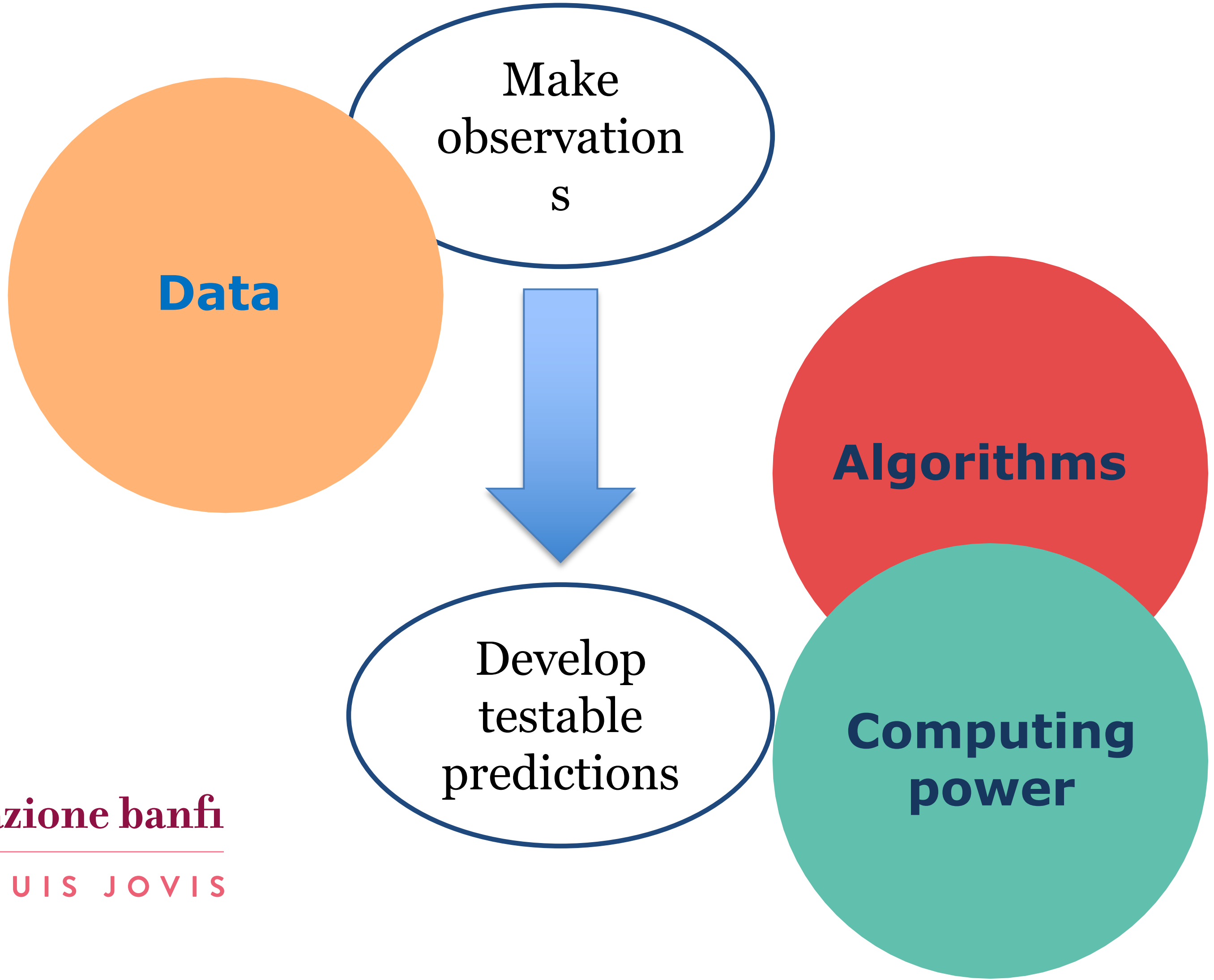
The ingredients

Data

Algorithms

Computing power

How did this happen ?



Data

Make observations

Develop testable predictions

Algorithms

Computing power

The Dataverse

Components, characteristics and challenges

02

Data

Structured

Unstructured

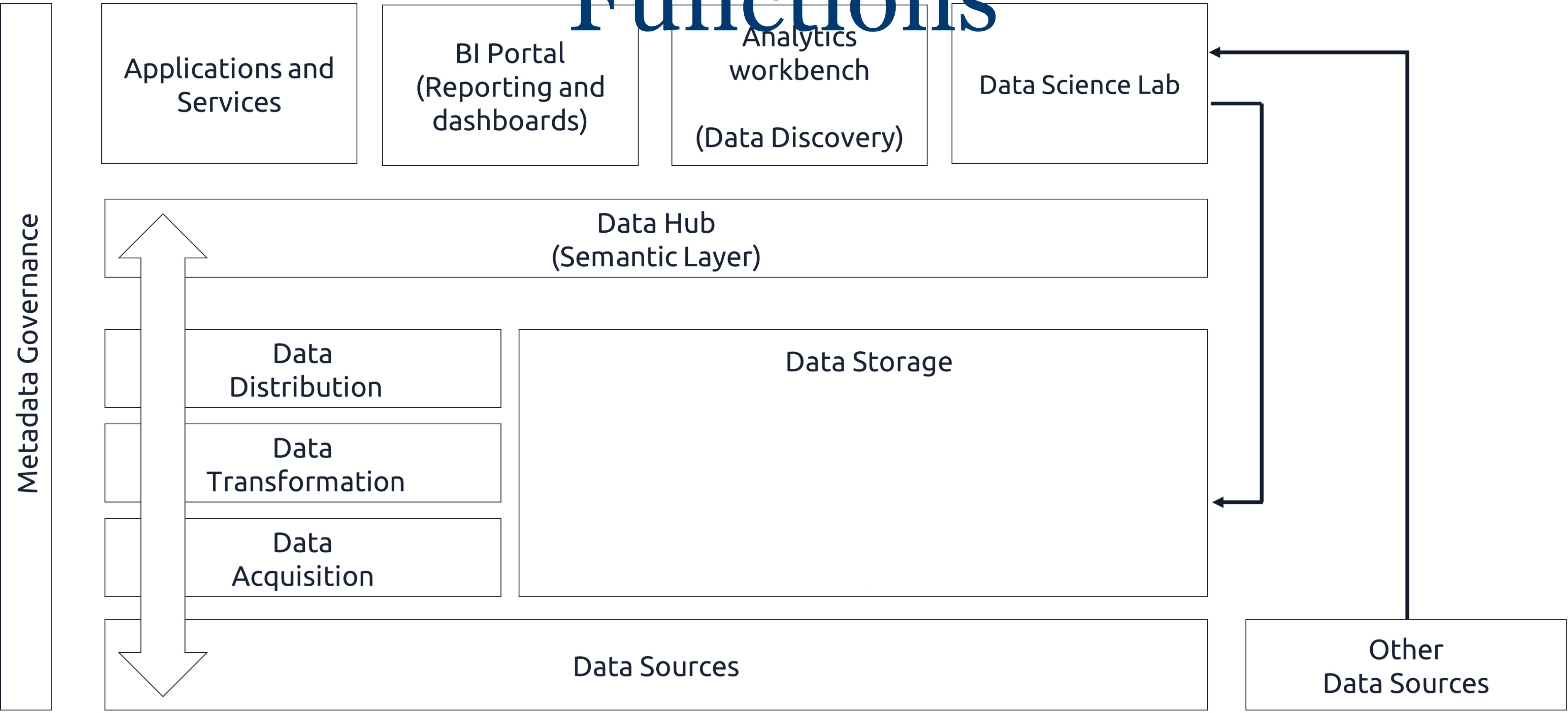
Which data has the highest growth rate ?

Structured

Unstructured

Data
Management
Systems

High Level Data Platform Functions



Architectures

Data Warehouses & Data Marts & ODSs

Data Lakes

Hybrid models

Logical DWH & Data Virtualisation

Architectures

Data Warehouses & Data Marts & ODSs

Data Lakes

Hybrid models

Logical DWH & Data Virtualisation

Architectures

Data Warehouses & Data Marts & ODSs

Data Lakes

Hybrid models

Logical DWH & Data Virtualisation

Architectures

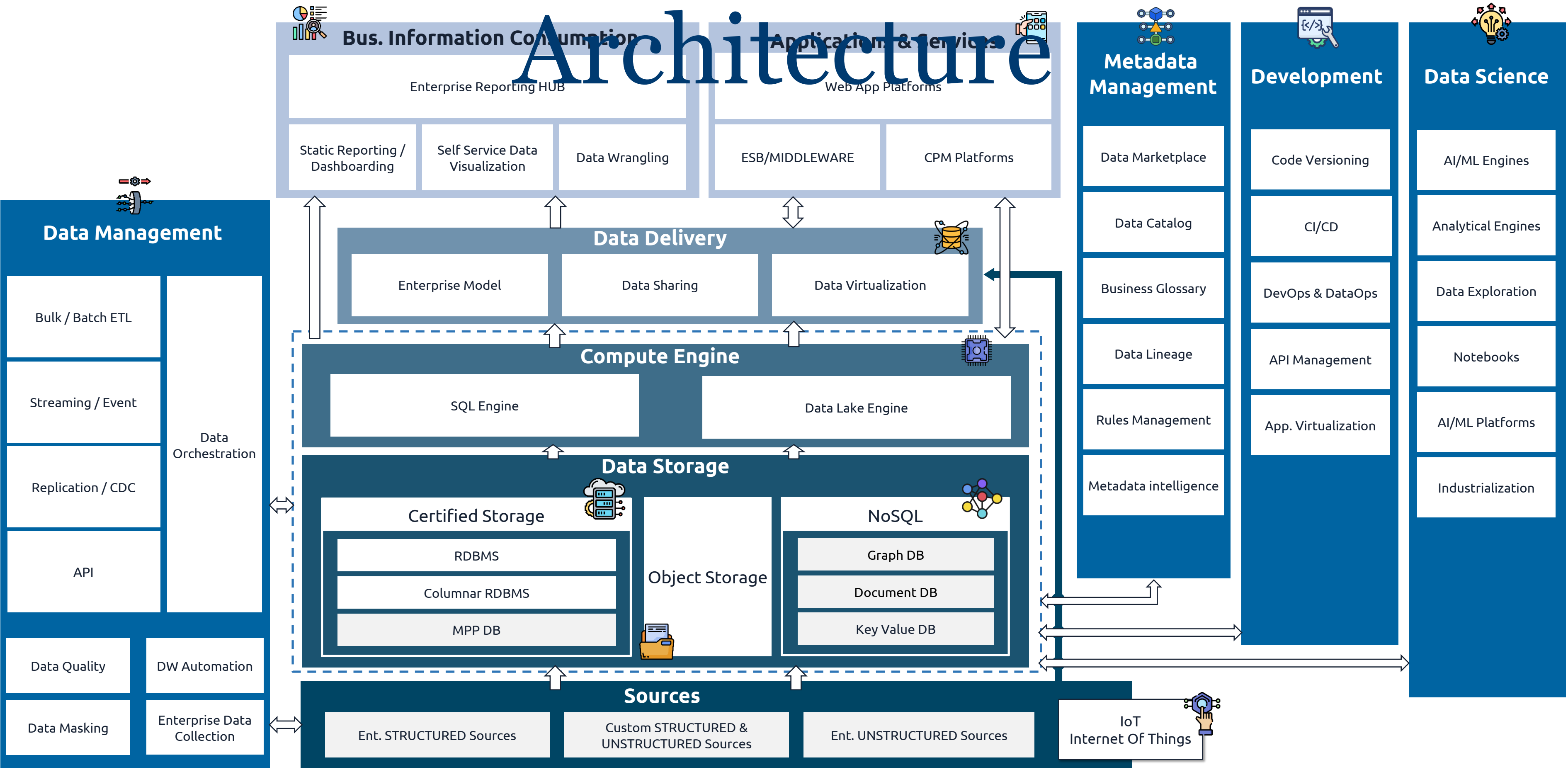
Data Warehouses & Data Marts & ODSs

Data Lakes

Hybrid models

Logical DWH & Data Virtualisation

Detailed Data Platform Architecture



What is the key difference between a data lake and a data warehouse ?

A data lake is for all data, a dwh is only for structured data

A dwh only containes quality data, a data lake does not

A dwh is only for reporting, a data lake is for artificial intelligence

A data lake is faster than a dwh

Deployment

On Premise

Cloud

Hybrid Cloud

Multi Cloud

Cloud

Infrastructure As A Service

Platform As A Service

Software As A Service

Benefits (??) of cloud architecture

- It avoids vendor lock-in
- You can more easily migrate to new technologies
- It is more secure
- It is more reliable
- It can scale up better

Development

Programming

Training

Analytics

Descriptive

Predictive

Prescriptive

Analytics

Supervised (Classification, Regression, …)

Unsupervised (Clustering, Dimensionality reduction,..)

Deep Learning

Usage

Infusing analytics into Apps

customer-facing, employee facing

Creating new Apps

Which is the industry where there is, at the moment, the highest # of use cases ?

Transportation and logistics

Telco

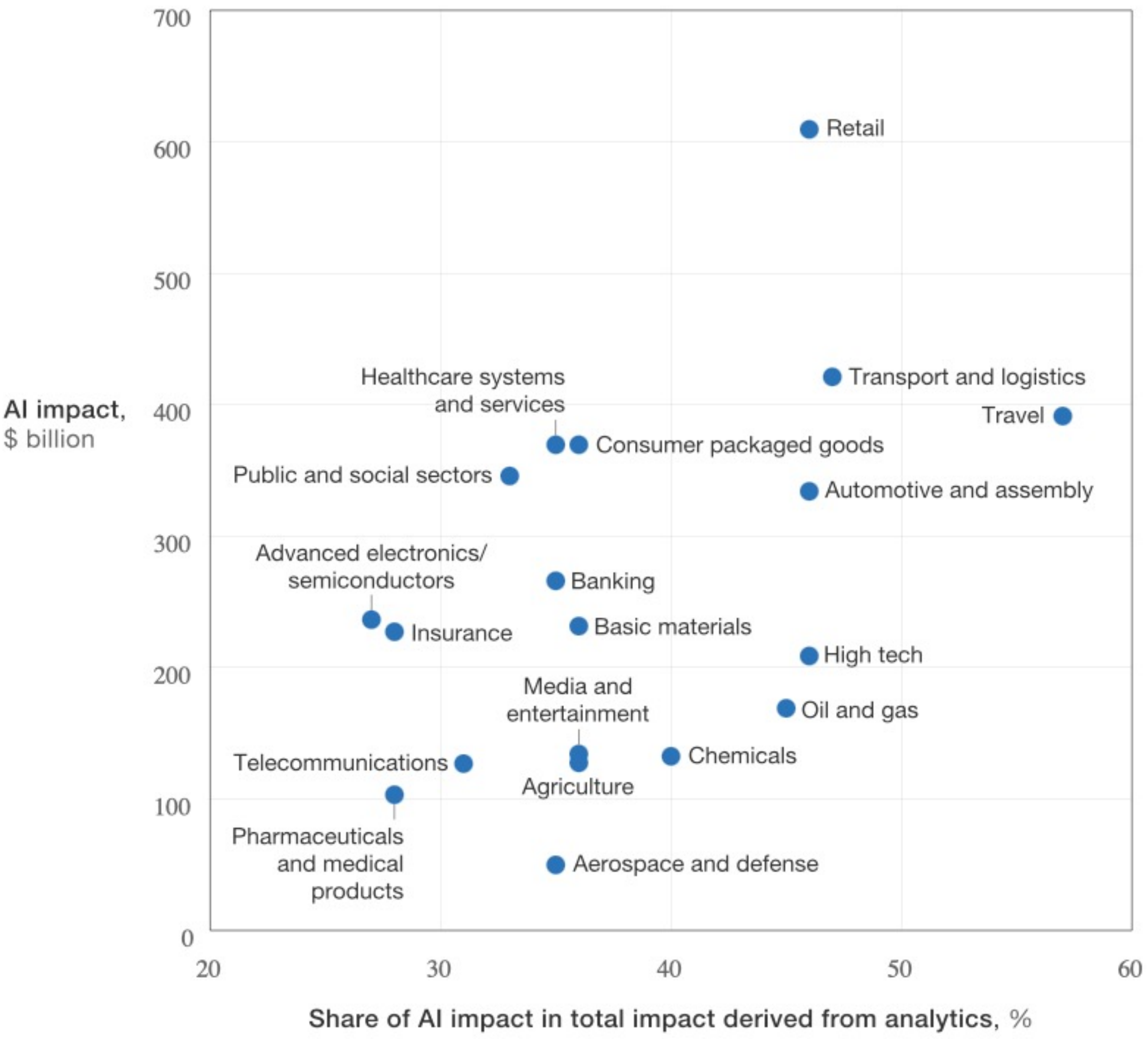
Retail

Advertising

Travel

Healthcare

Public sector



McKinsey&Company | **Source:** McKinsey Global Institute analysis

Which is function with the most AI use cases at the moment ?

Finance

HR

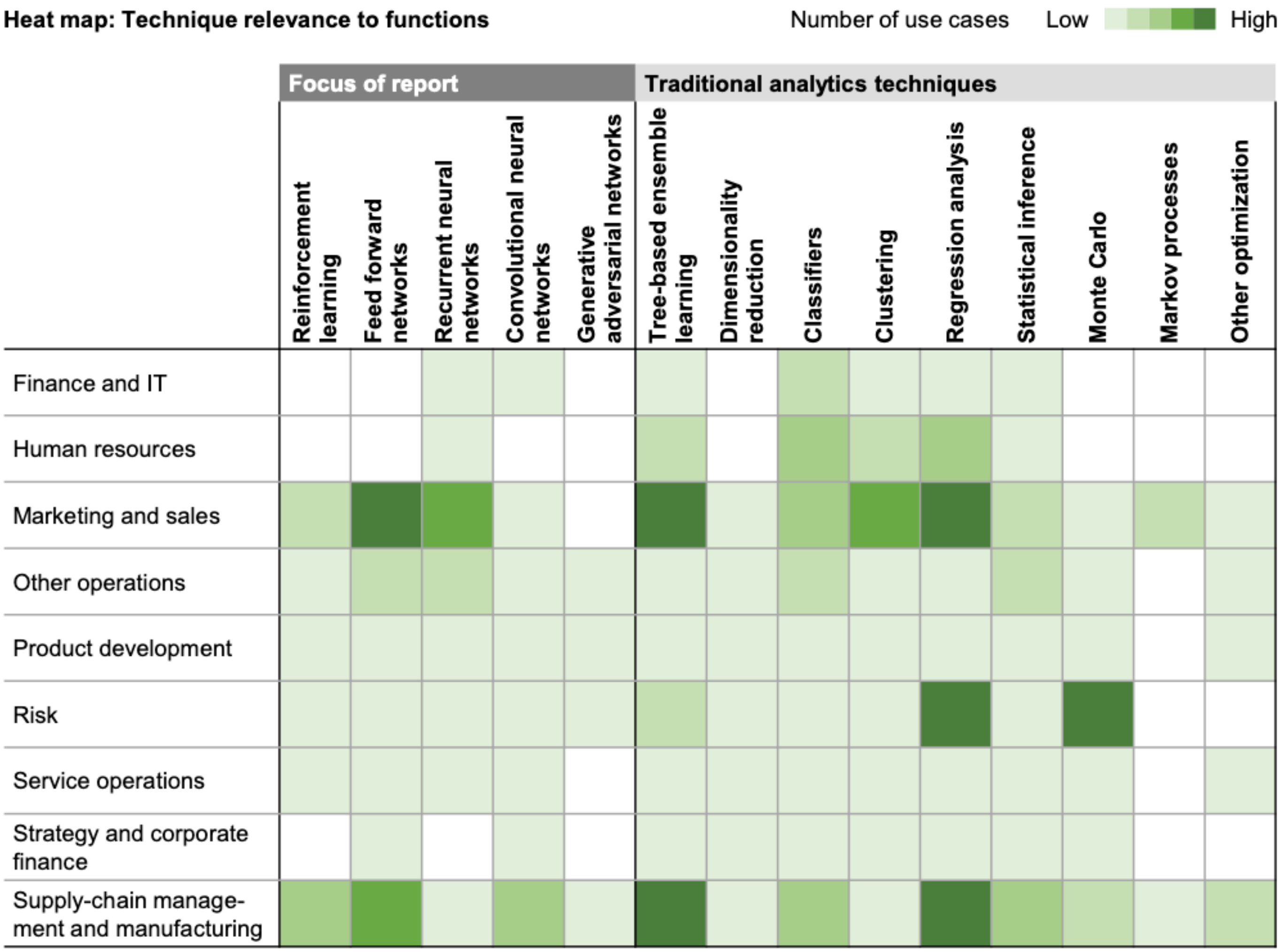
Marketing and sales

Product development

Risk management

Supply chain management

Cybersecurity



Heat map: Technique relevance to functions

SOURCE: McKinsey Global Institute analysis

Which is the key characteristics of the Dataverse ?

Polymorphism

Complexity

Consistency

Reliability

Accuracy

How to tackle complexity ?

Step #1

Data Knowledge

Knowledge

Structured

Semantic

Consolidated

Structured

Modeling

Mapping

Semantic

Glossary

Consolidated

Integration
Lineage

A different perspective on data

02

- Is everyone producing data ?
- What does data want ?
- Is data virtual or real ?
- Is data born to live alone ?
- Does data flow smoothly ?
- Does data exist beyond countries ?
- Is data good or not ?
- Does data come with responsibilities ?
- Is data tidy ?
- Is data eternal ?

Is everyone producing data ?

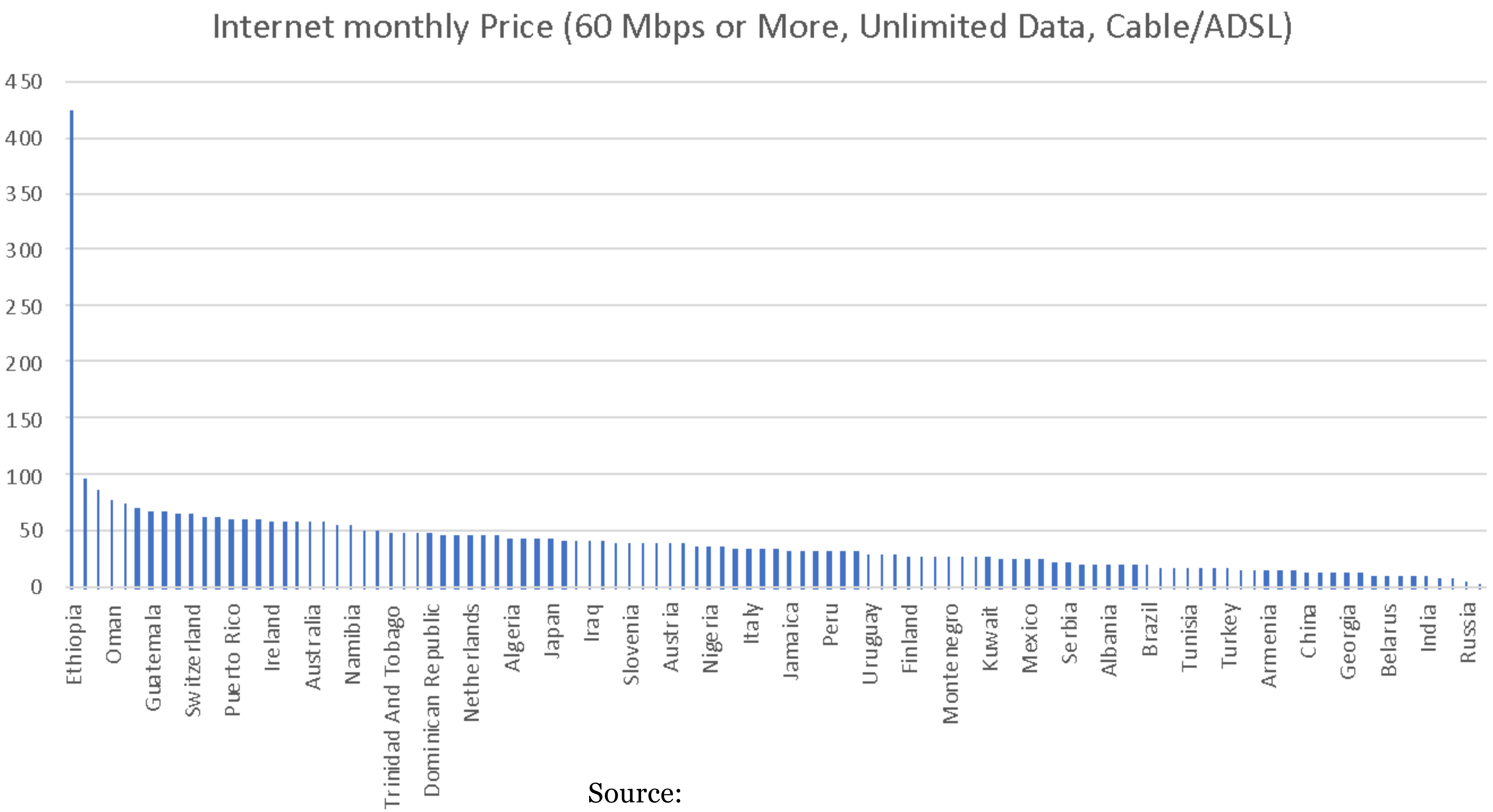
There are hot spots of data production and vast, empty gaps

Some people are producing more data than others

Some data appears to be from one person, but can be many behind

ATTENTION POINTS:

- Data is unevenly distributed
- Pricing structure of the internet access to it is different in different places
- Data access regulations are different from country to country
- Data connections (undersea cables) have different transfer rates
- Cultural lenses drive how data is produced, transformed, analyzed



Internet monthly Price (60 Mbps or More, Unlimited Data, Cable/ADSL)

Source:
Numbeo.com

What does data want ?

BIG DATA WANTS ACCUMULATION OF MORE OF ITSELF

Data wants more data

More data wants an algorithm to make sense of it

More analytics will require more algorithms

Why are we all now happy with the need of more data ?

What is that pushes all of us so strongly towards an empirical approach ?

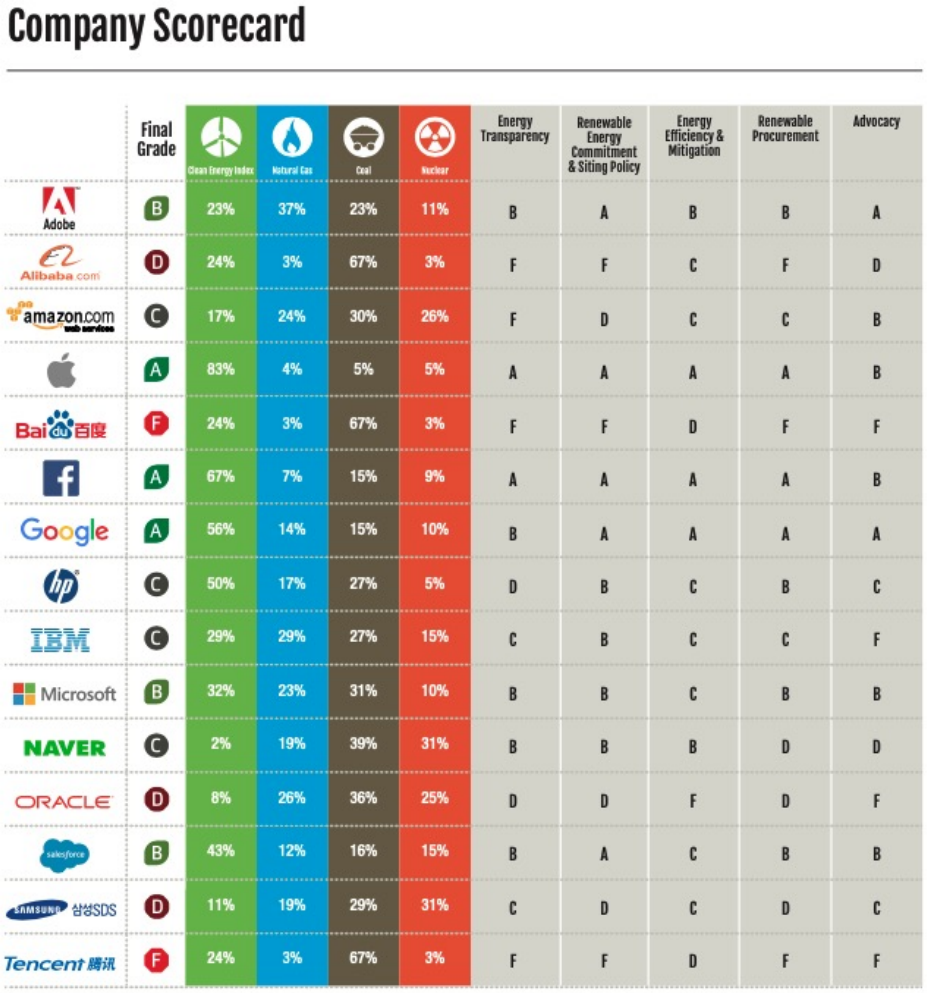
Is data virtual or real ?

The idea that data is virtual let us think it can be indefinitely accumulated.

Actually, data is physical. It requires physical objects to live in. Objects that change our ecosystem.

Furthermore, some data wants to be “objectified”.

Tanks containing coolant for servers at a Google Data center in Saint Ghislain, Belgium.YVES HERMAN / REUTERS



Company Scorecard

	Final Grade	Clean Energy Index	Natural Gas	Coal	Nuclear	Energy Transparency	Renewable Energy Commitment & Siting Policy	Energy Efficiency & Mitigation	Renewable Procurement	Advocacy
Adobe	B	23%	37%	23%	11%	B	A	B	B	A
Alibaba.com	D	24%	3%	67%	3%	F	F	C	F	D
amazon.com web services	C	17%	24%	30%	26%	F	D	C	C	B
	A	83%	4%	5%	5%	A	A	A	A	B
Bai du 百度	F	24%	3%	67%	3%	F	F	D	F	F
f	A	67%	7%	15%	9%	A	A	A	A	B
Google	A	56%	14%	15%	10%	B	A	A	A	A
hp	C	50%	17%	27%	5%	D	B	C	B	C
IBM	C	29%	29%	27%	15%	C	B	C	C	F
Microsoft	B	32%	23%	31%	10%	B	B	C	B	B
NAVER	C	2%	19%	39%	31%	B	B	B	D	D
ORACLE	D	8%	26%	36%	25%	D	D	F	D	F
salesforce	B	43%	12%	16%	15%	B	A	C	B	B
SAMSUNG 삼성SDS	D	11%	19%	29%	31%	C	D	C	D	C
Tencent 腾讯	F	24%	3%	67%	3%	F	F	D	F	F

Source: CLICKING CLEAN: WHO IS WINNING THE RACE TO BUILD A GREEN INTERNET?, GreenPeace, 2017

Is data born to live alone ?

Most data doesn’t exist in isolation

Sometimes the relationship is given by the object producing the data, others by its location, others by the individual producing it, others by the data itself referring to previous existing data

Algorithms themselves create new relationships

Are all this relationships real ?

Are all of them qualified for making judgements ?

Does data flow smoothly ?

Not all data is equal:

- Different speed of transmissions
- Different access capability and processing thru physical gateways
- Data is created different to flow on the network (voice and videos move differently than text)
- Data flows also depend on the physical environment

Imagining that all data will move freely everywhere in a kind of universal moment of splendor may not be the case

Does data exist beyond countries ?

Data is all produced under specific policy regimes

Data is not a denatured object, it comes with the inheritance of the culture producing it

Is data good or not ?

Data is feral

DATA WANTS TO GO WILD,

DATA WANTS TO GET OVER THE FENCE AND GET GOING

Data can defy the expectations of its originators:

- Algorithms can create new forms of it
- Can appear in unexpected places
- Can change format and get a usage completely different
- Can end up in the hands of unintended people

People think that data will do exactly what is told. This is not the case:

- In the hands of governments, there can be fear of totalitarism
- In the hands of corporation, as a minimum coupons and aggressive marketing

Does data come with responsibilities ?

Users of data must be educated on the responsibility of doing the right thing with data

It is not only about being the “custodians” of data, of knowing about Data Governance

It is also about “opening it up”, creating the conditions for data to be integrated with other data and tell new stories

Is data tidy ?

Data will resist being tidied up.

Data is often incomplete.

Data is often telling lies even when looks good.

Complete data quality is a chimera.

Is data eternal ?

Not all data wants to last forever

Not all data is meaningful forever

Data often changes its value over time

Data retention and data deletion policies

Conclusion: what to do next ?

- Learn how to read an algorithm, learn how to program an algorithm
- Study (and regulate) the new alchemists
- Challenge the new empiricism

Bibliography

1. "Data Science for Business" by Foster Provost and Tom Fawcett
2. "Competing on Analytics: The New Science of Winning" by Thomas H. Davenport and Jeanne G. Harris
3. "Data-Driven: Creating a Data Culture" by Hilary Mason and DJ Patil
4. "Data Management for Researchers" by Kristin Briney
5. "Big Data: A Revolution That Will Transform How We Live, Work, and Think" by Viktor Mayer-Schönberger and Kenneth Cukier
6. "Data Driven: Profiting from Your Most Important Business Asset" by Thomas C. Redman
7. "Data-First Marketing: How to Compete & Win in the Age of Analytics" by Janet Driscoll Miller
8. "The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling"
9. "Data Smart: Using Data Science to Transform Information into Insight"
10. "The Big Data-Driven Business: How to Use Big Data to Win Customers, Beat Competitors, and Boost Profits" by Russell Glass and Sean Callahan
11. Genevieve Bell. The secret life of big data, Prickly Paradigm Press

Thanks for your attention !

fondazione banfi

SANGUIS JOVIS

ALTA SCUOLA DEL SANGIOVESE

fondazionebanfi.it