HOW DATA SCIENCE REALLY LOOKS LIKE

FAVIO VÁZQUEZ
DATA SCIENTIST

CLOSTER H_2O.ai

WHO AM I?

- Venezuelan
- Bachelor of Physics and Computer Engineer
- Master's at PCF-UNAM (Cosmology)
- Data Scientist at **H2O.ai**
- Faculty Member and **Emeritus** Instructor
- CEO and Chief Data Scientist at **Closter**
- Creator of **Ciencia y Datos**

- Very active on Linkedin;)
- Editor International Journal of Business Analytics and Intelligence
- Writer at Towards Data Science, Heartbeat, Becoming Human and Planeta Chatbot
- Medium Top Writer in Business, AI and Technology

CLOSTER H2O.ai

Contents

- Motivation
- Important concepts
- Data Science Timeline
- Creating a data product (theory)
- Learning data science
- **Data science problem and solution (Workshop)**

Motivation: Data science products are complex

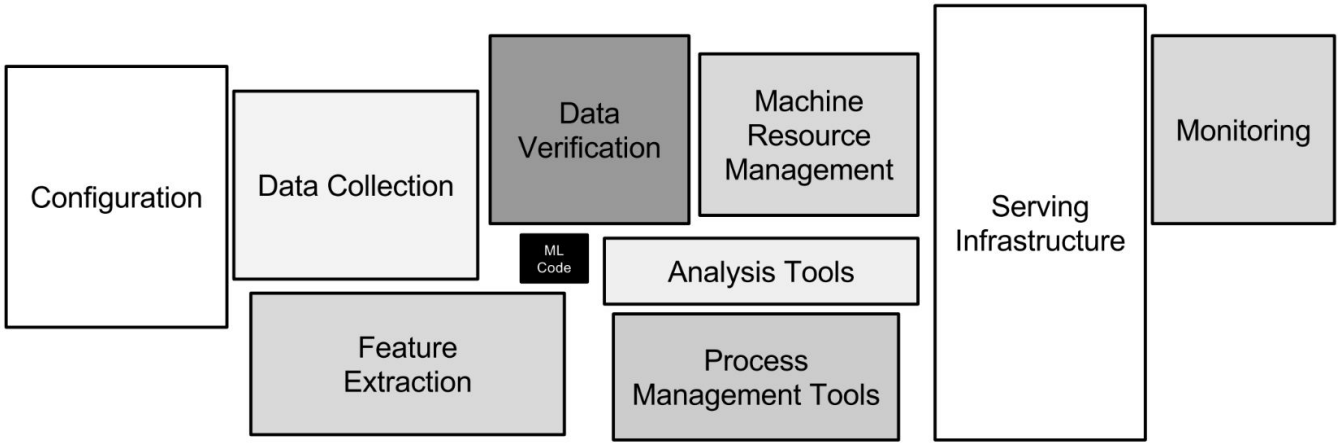
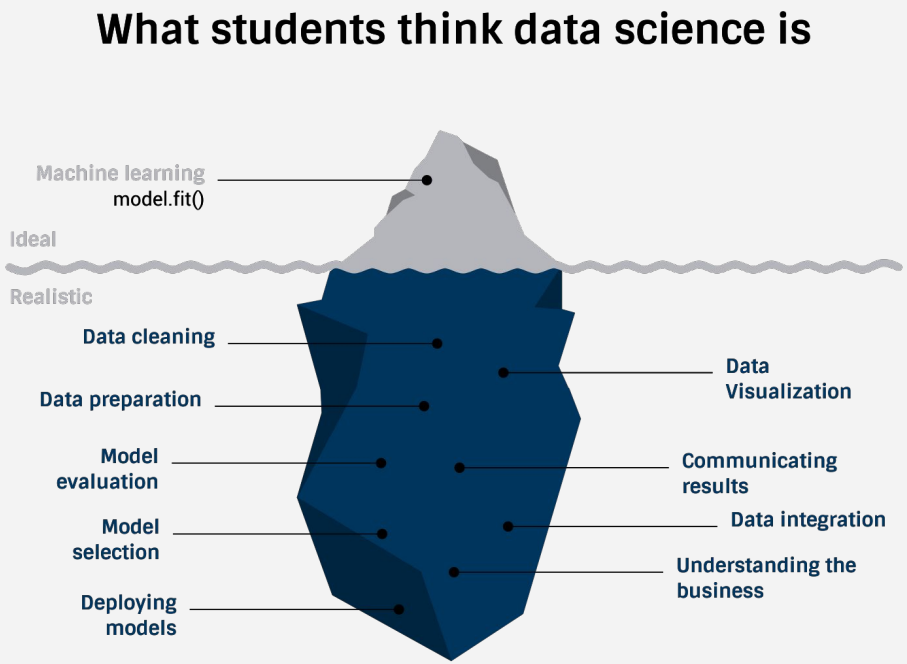


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Source: https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

CLOSTER H2O.ai

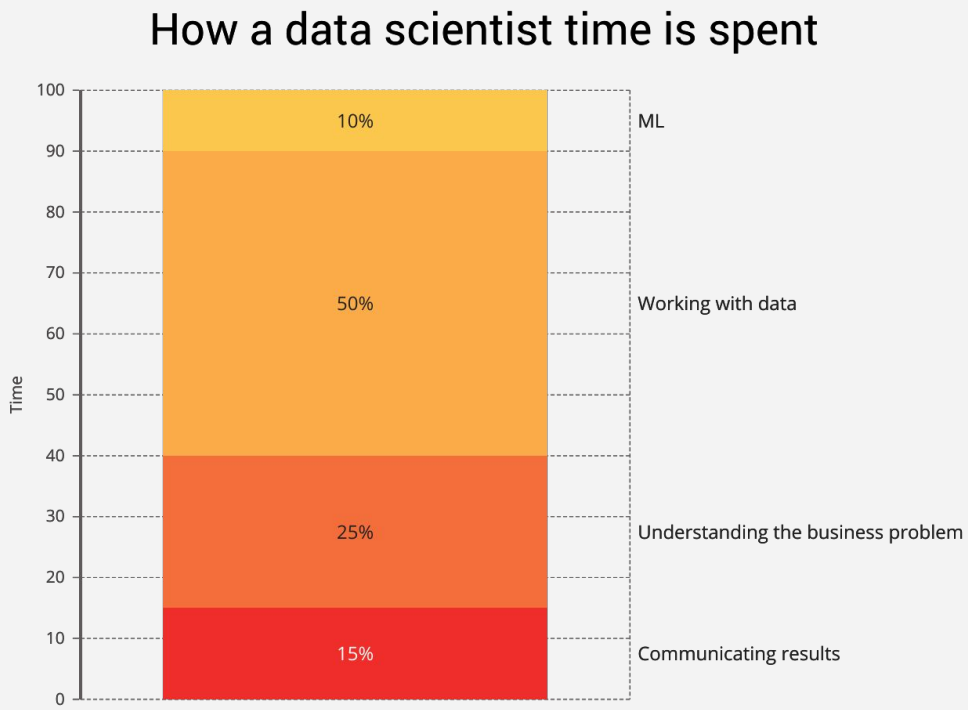
Motivation: Machine learning is not only model.fit()



Source: https://www.business-science.io/business/2019/06/27/data-science-workflow.html

CLOSTER H$_2$O.ai

Motivation: How data science really looks like

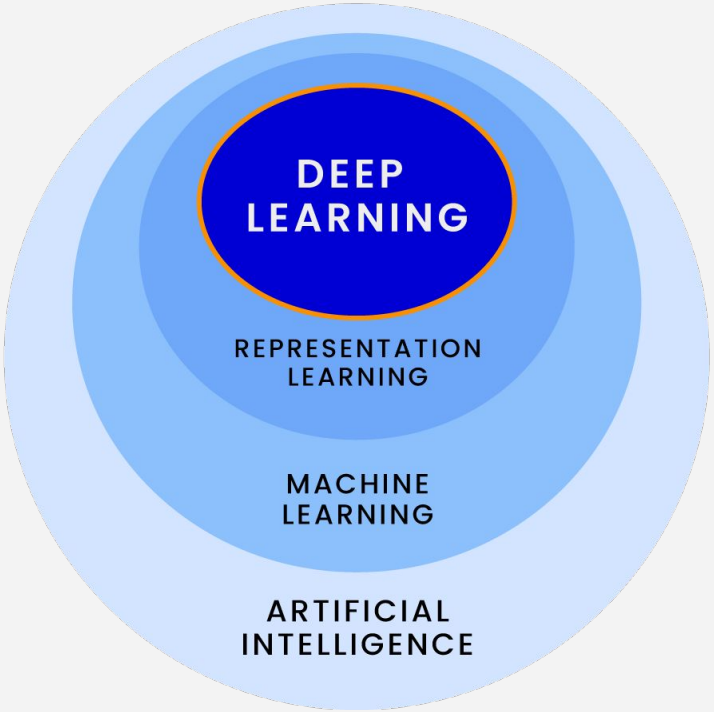


Source: https://www.business-science.io/business/2019/06/27/data-science-workflow.html

CLOSTER H2O.ai

Important concepts

Panorama of AI



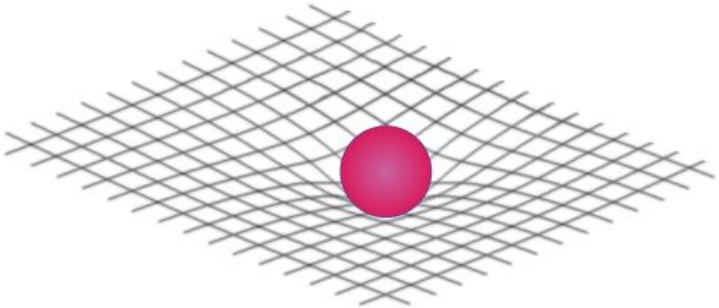
CLOSTER H2O.ai

Artificial Intelligence (AI)

Automated processes to emulate human skills and capabilities with computing.

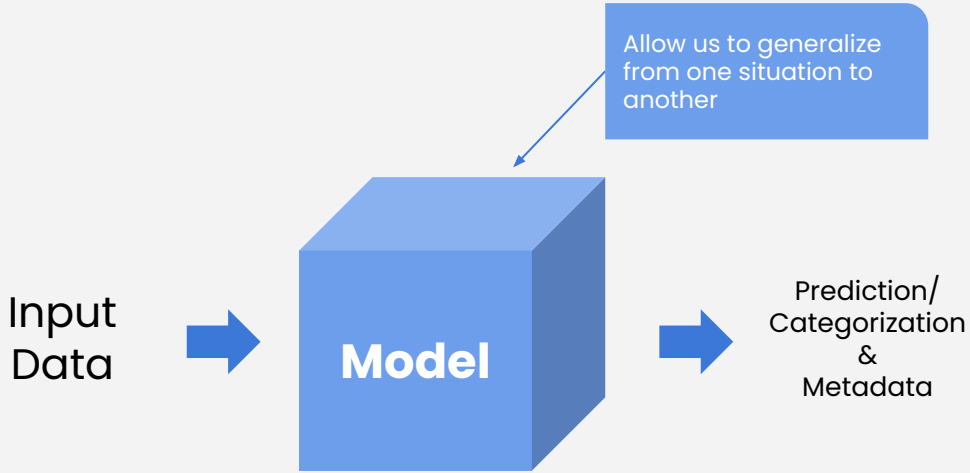
CLOSTER H$_2$O.ai

Model



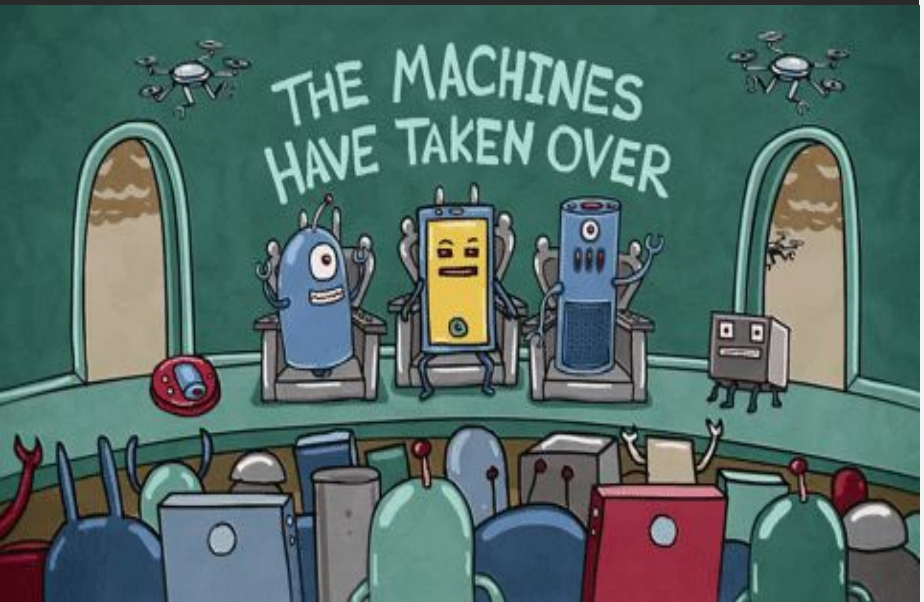
$$G_{\mu\nu}+\Lambda g_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu}$$

Abstraction of reality to understand a process or phenomenon using mathematical tools.



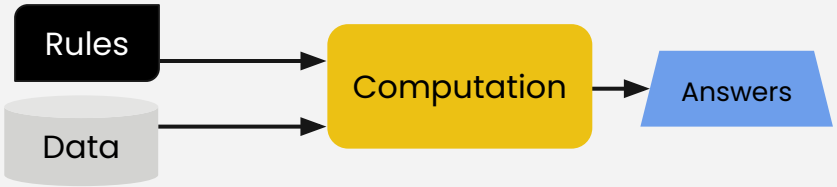
CLOSTER H2O.ai

Machine Learning

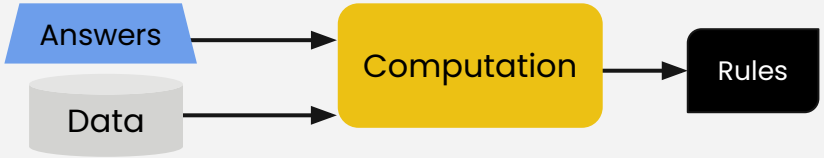


Discovery of patterns from data to predict the operation of a process or phenomenon using algorithms that are not explicitly programmed.

Traditional Programming



Machine Learning



CLOSTER H2O.ai


methodology
sciences
development
effort
areas
support
project
evaluation
center
engineering
campus
software
data
director
collaboration
space
impact
working
career
students
reproducibility
studio
science
groups
budget
evaluation
institute
faculty
work
tools
statistics
scientists
university
enviroment
activities
new
domain
research
education
graduate
program

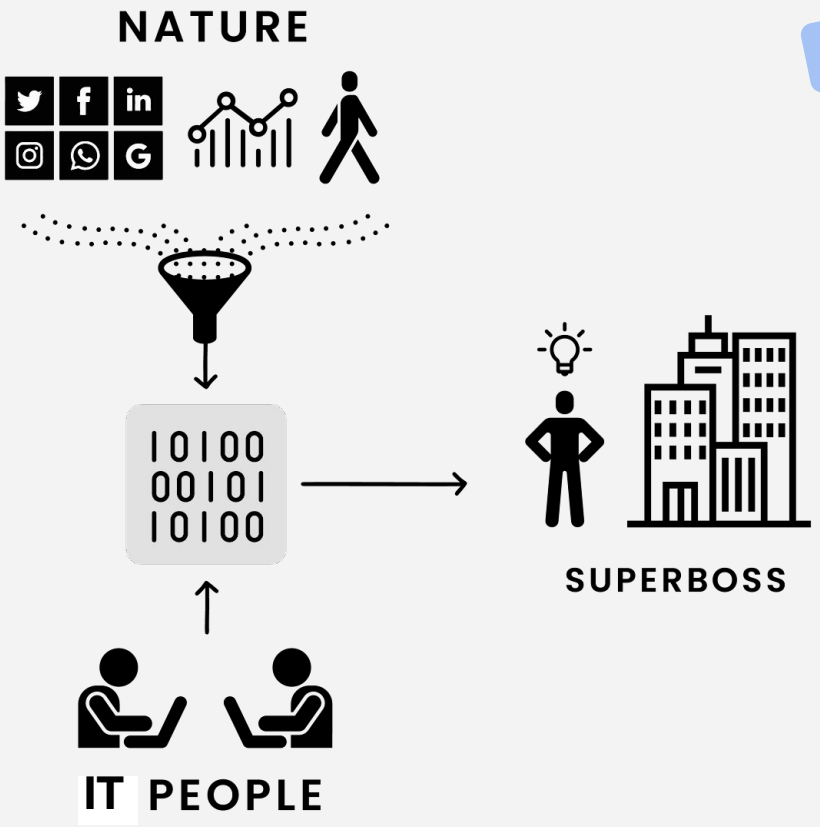
THE ROLE OF DATA SCIENCE

Data science is the mediator on the path of using AI to impact business.

Perhaps the name "data science" is not very comfortable for everyone, but I will try to show that we can take advantage of it for now.

how does **data science** emerge in a company?

CLOSTER H2O.ai

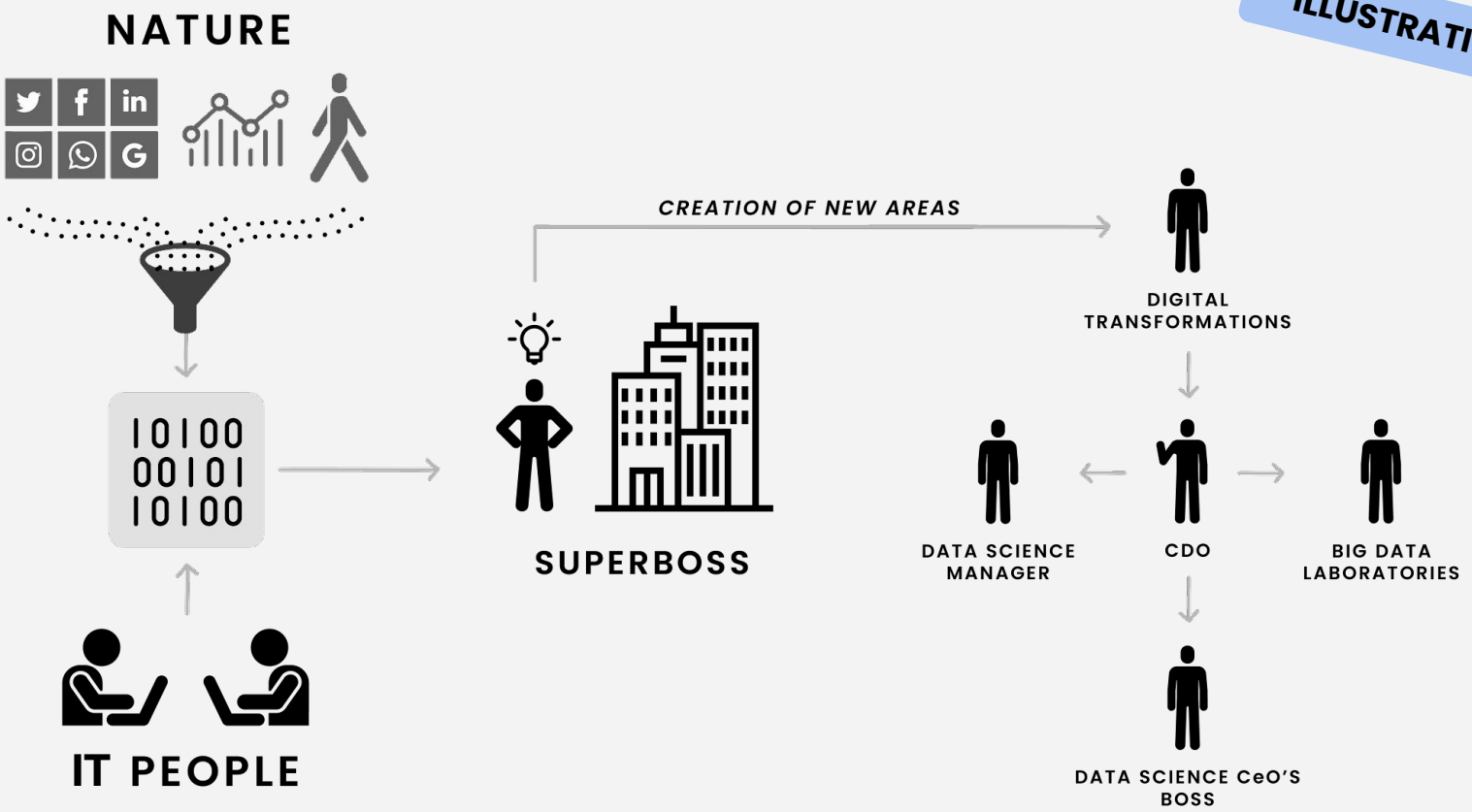


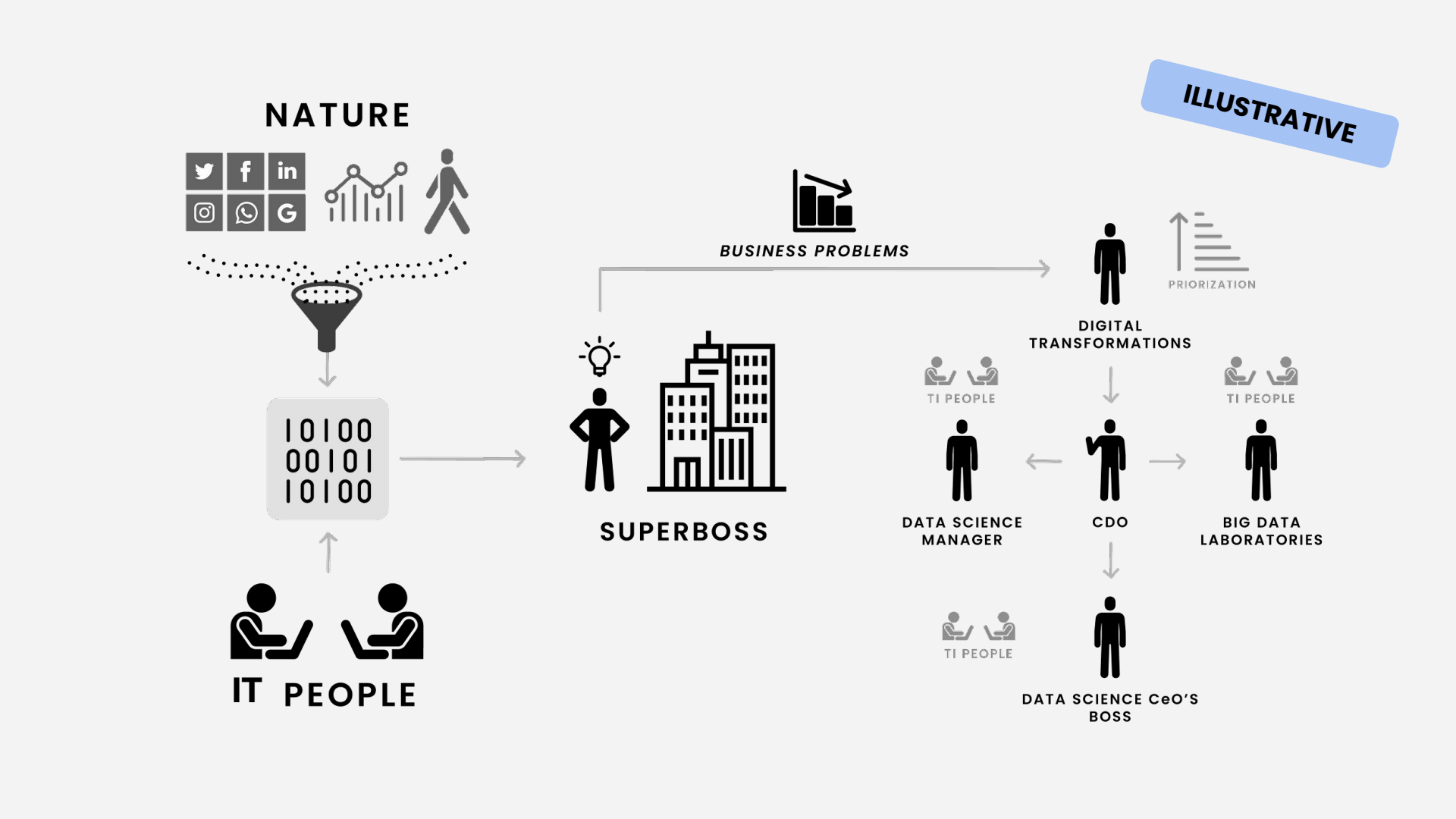
ILLUSTRATIVE

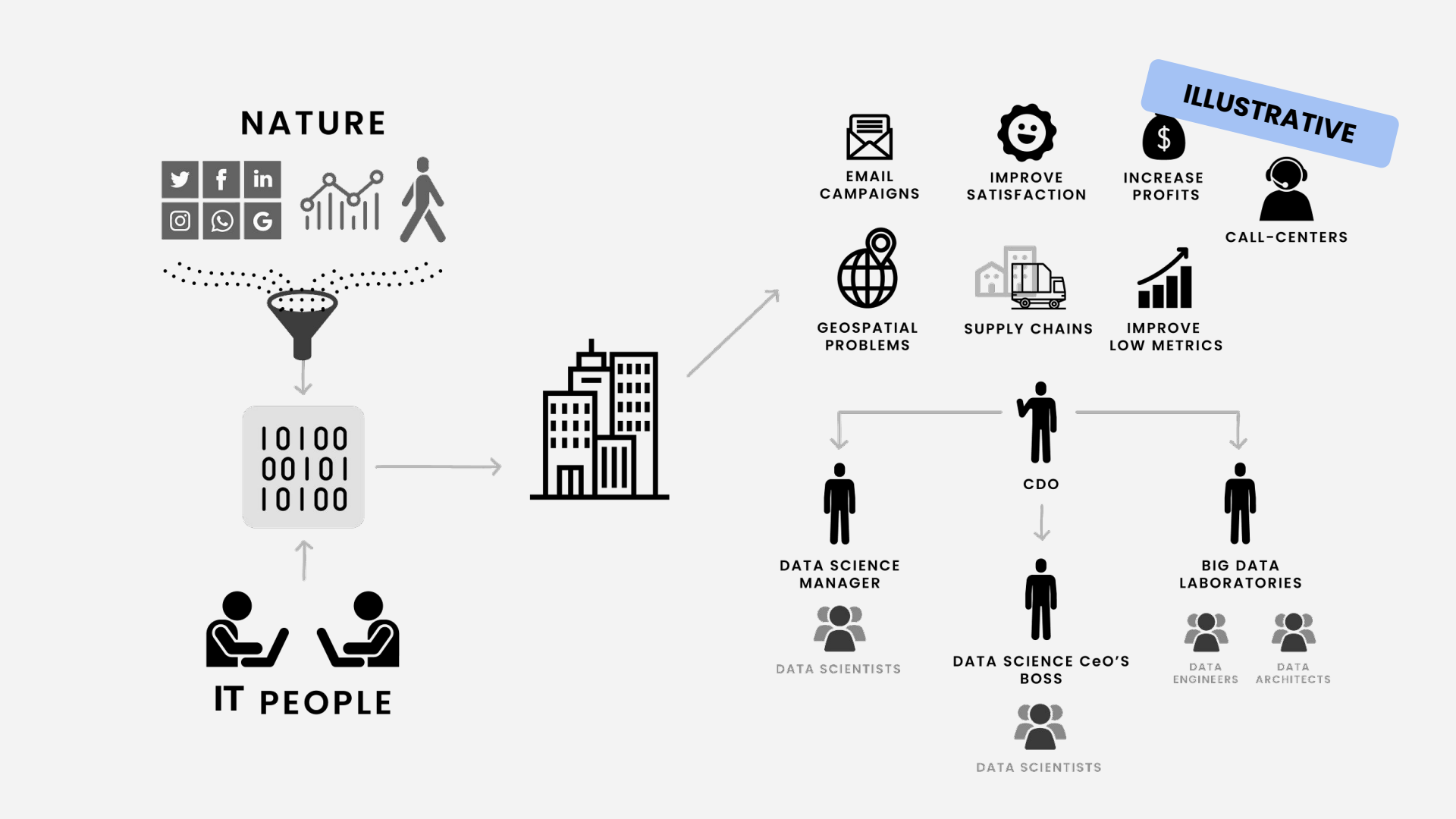
We have a lot of data, we need to do something about it

CLOSTER H2O.ai

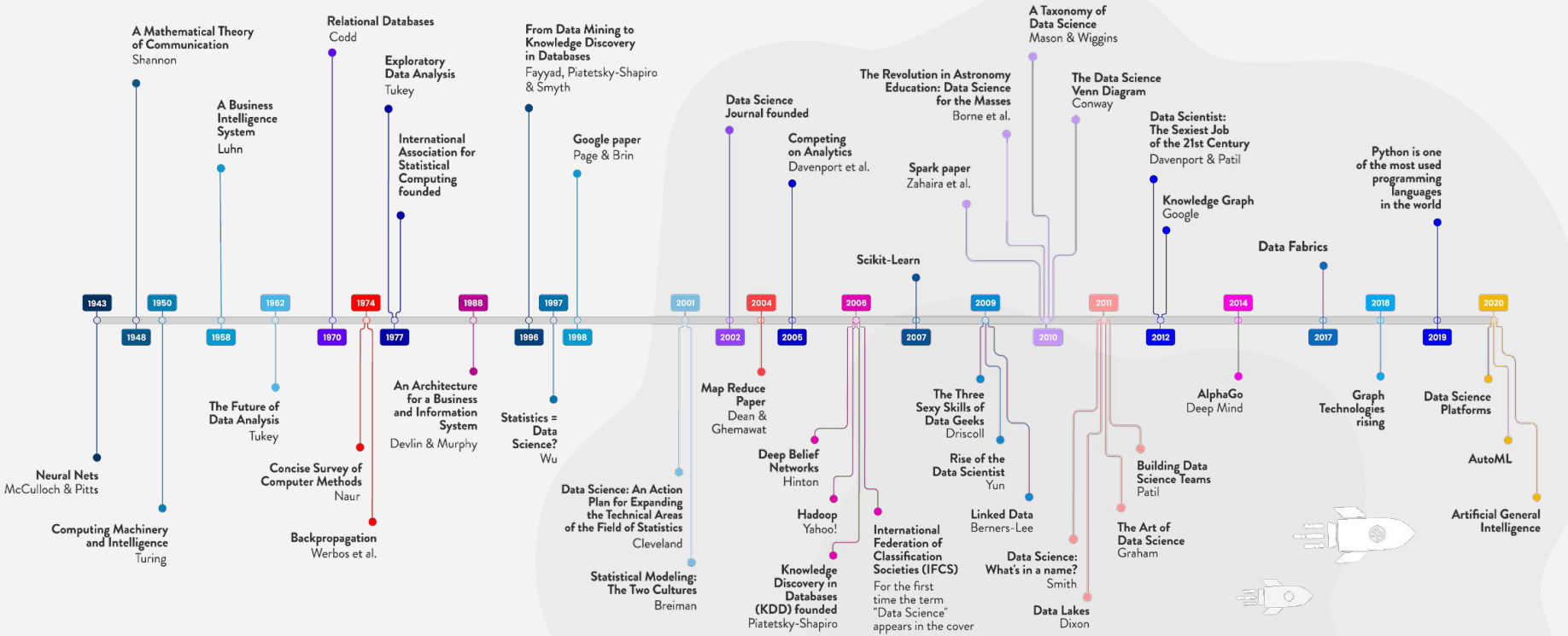
ILLUSTRATIVE

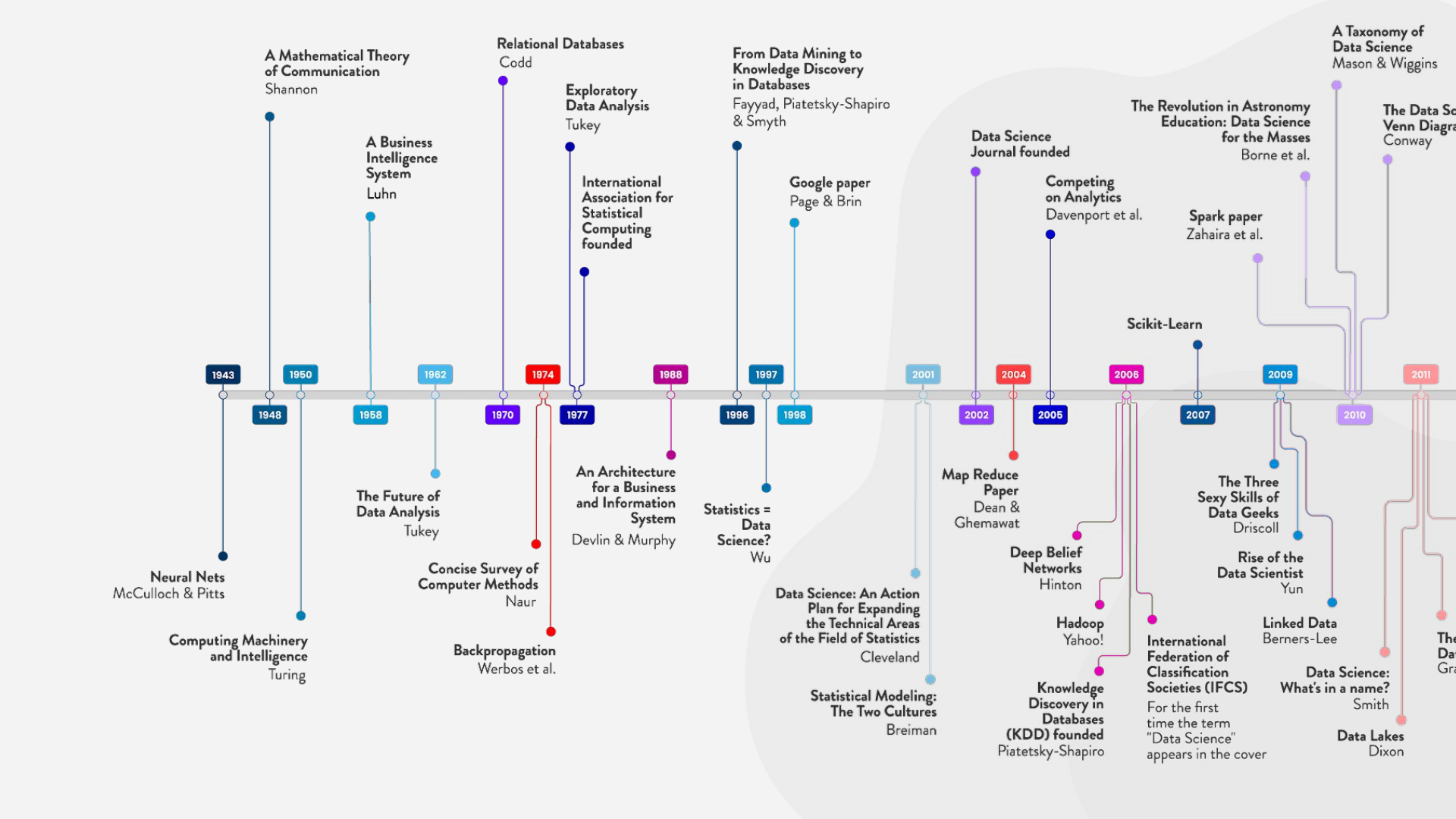


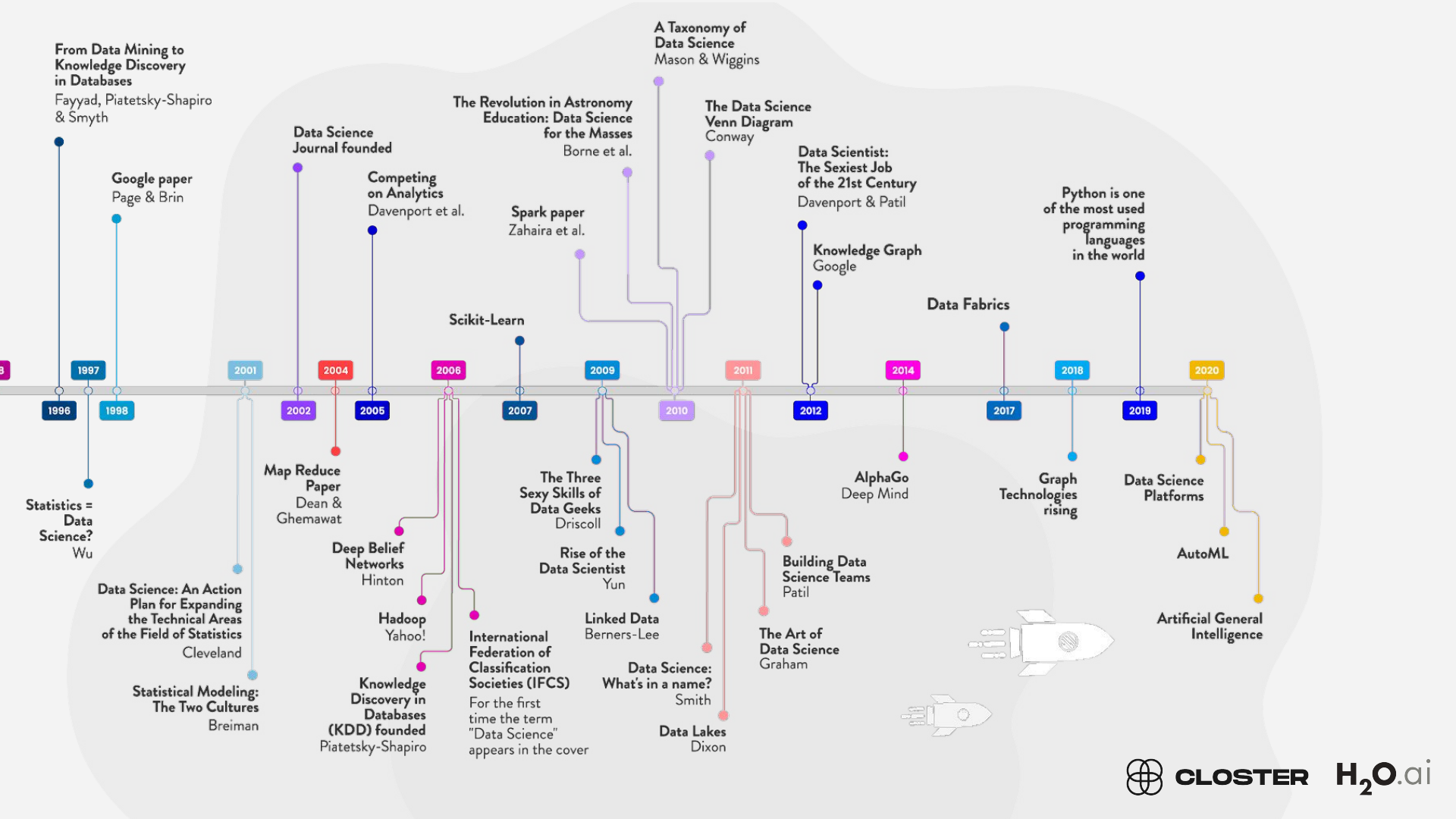

ILLUSTRATIVE
NATURE
BUSINESS PROBLEMS
PRIORIZATION
DIGITAL TRANSFORMATIONS
TI PEOPLE
TI PEOPLE
10100
00101
10100
SUPERBOSS
DATA SCIENCE MANAGER
CDO
BIG DATA LABORATORIES
TI PEOPLE
IT PEOPLE
DATA SCIENCE CeO'S BOSS


NATURE
IT PEOPLE
ILLUSTRATIVE
EMAIL CAMPAIGNS
IMPROVE SATISFACTION
INCREASE PROFITS
CALL-CENTERS
GEOSPATIAL PROBLEMS
SUPPLY CHAINS
IMPROVE LOW METRICS
CDO
DATA SCIENCE MANAGER
DATA SCIENTISTS
DATA SCIENCE CeO'S BOSS
DATA SCIENTISTS
BIG DATA LABORATORIES
DATA ENGINEERS
DATA ARCHITECTS

DATA SCIENCE TIMELINE

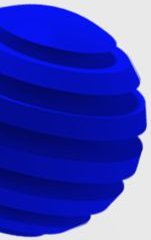



1943
Neural Nets
McCulloch & Pitts
1948
A Mathematical Theory of Communication
Shannon
1950
Computing Machinery and Intelligence
Turing
1958
A Business Intelligence System
Luhn
1962
The Future of Data Analysis
Tukey
1970
Relational Databases
Codd
1974
Concise Survey of Computer Methods
Naur
Backpropagation
Werbos et al.
1977
Exploratory Data Analysis
Tukey
International Association for Statistical Computing founded
1988
An Architecture for a Business and Information System
Devlin & Murphy
1996
From Data Mining to Knowledge Discovery in Databases
Fayyad, Piatetsky-Shapiro & Smyth
1997
Statistics = Data Science?
Wu
1998
Google paper
Page & Brin
2001
Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics
Cleveland
Statistical Modeling: The Two Cultures
Breiman
2002
Data Science Journal founded
2004
Map Reduce Paper
Dean & Ghemawat
2005
Competing on Analytics
Davenport et al.
2006
Deep Belief Networks
Hinton
Hadoop
Yahoo!
Knowledge Discovery in Databases (KDD) founded
Piatetsky-Shapiro
2007
Scikit-Learn
International Federation of Classification Societies (IFCS)
For the first time the term "Data Science" appears in the cover
2009
Spark paper
Zahaira et al.
The Three Sexy Skills of Data Geeks
Driscoll
Rise of the Data Scientist
Yun
Linked Data
Berners-Lee
2010
The Revolution in Astronomy Education: Data Science for the Masses
Borne et al.
A Taxonomy of Data Science
Mason & Wiggins
Data Science: What's in a name?
Smith
Data Lakes
Dixon
2011
The Data Sc
Venn Diagra
Conway



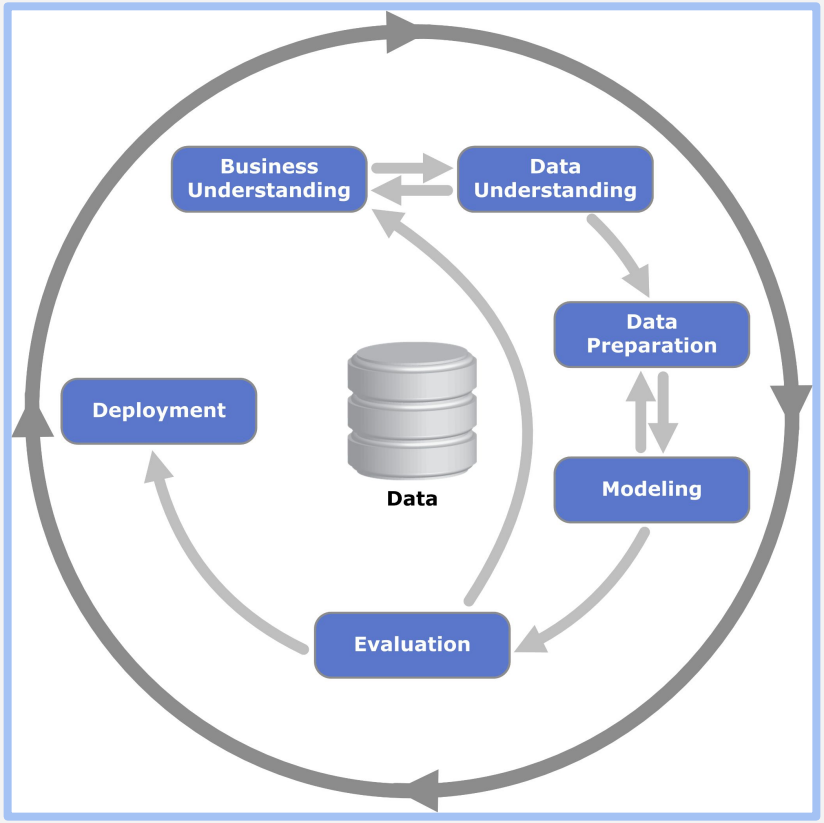
CLOSTER H2O.ai

Data Science Products



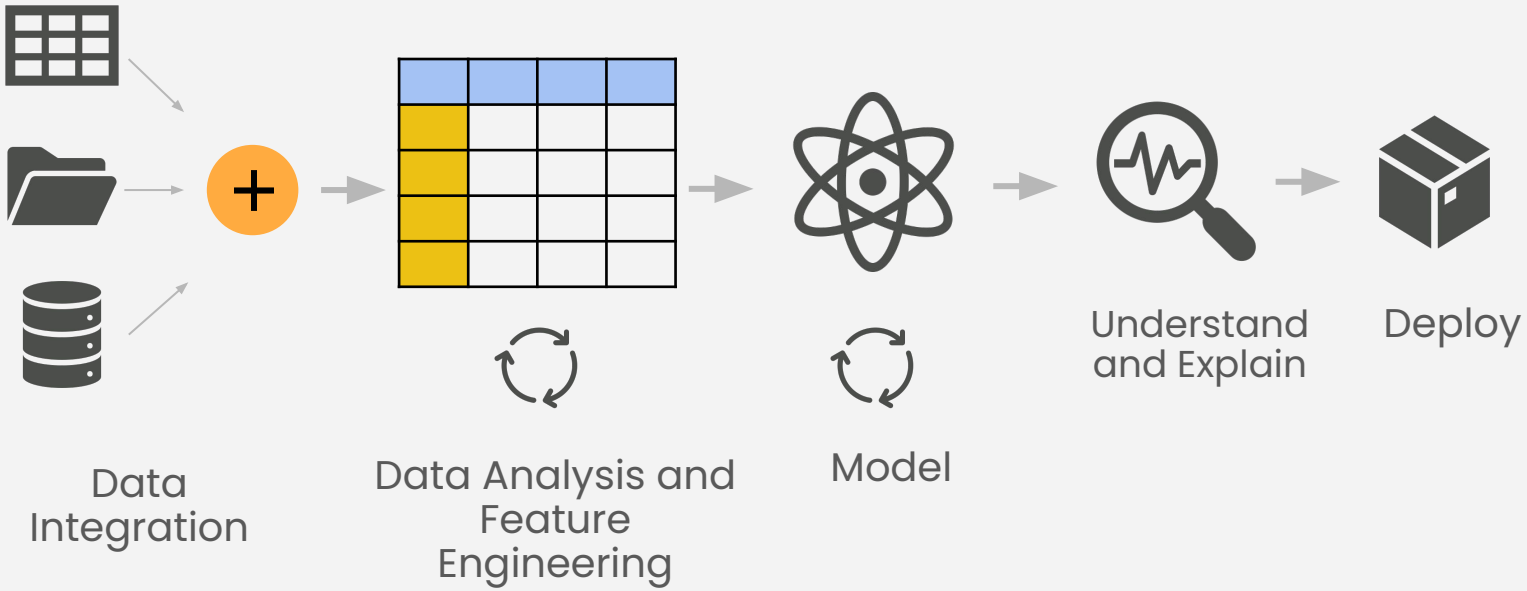
CLOSTER H2O.ai

CRISP-DM (One of the fathers of DS)



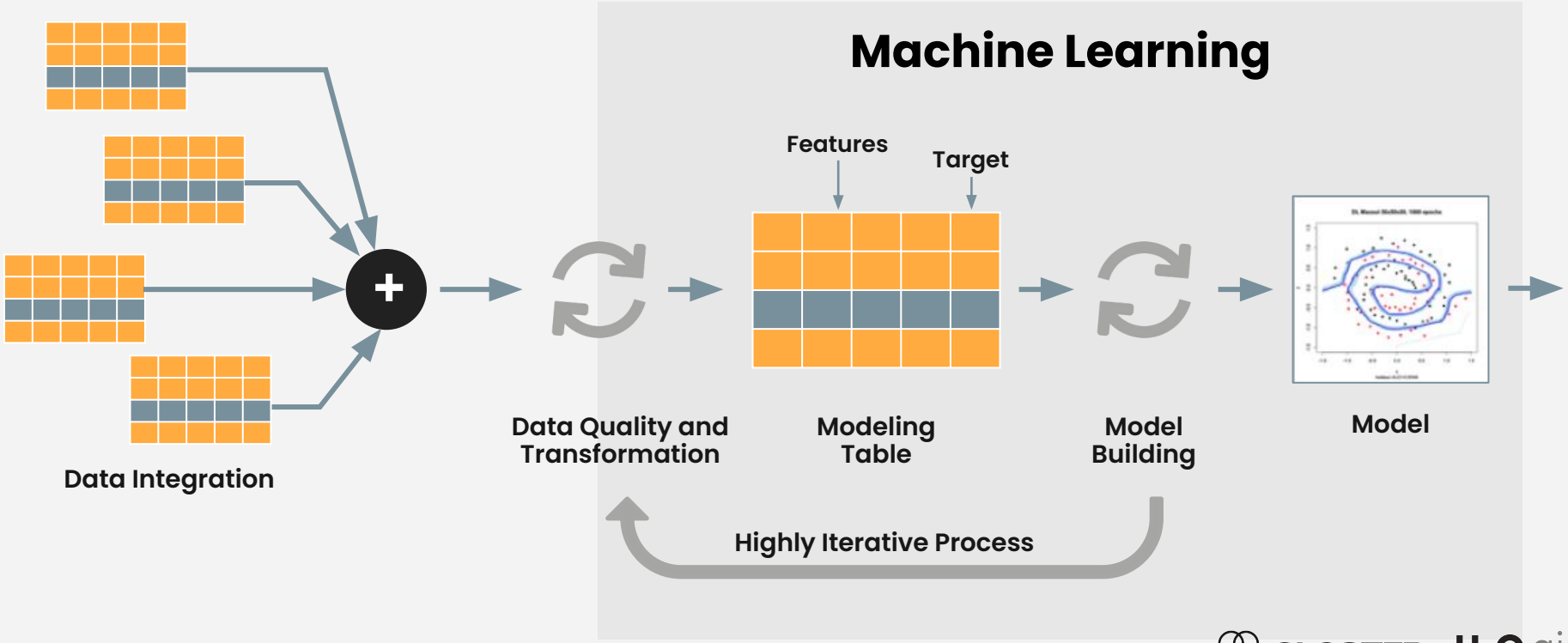
CLOSTER H2O.ai

Data Science Workflow



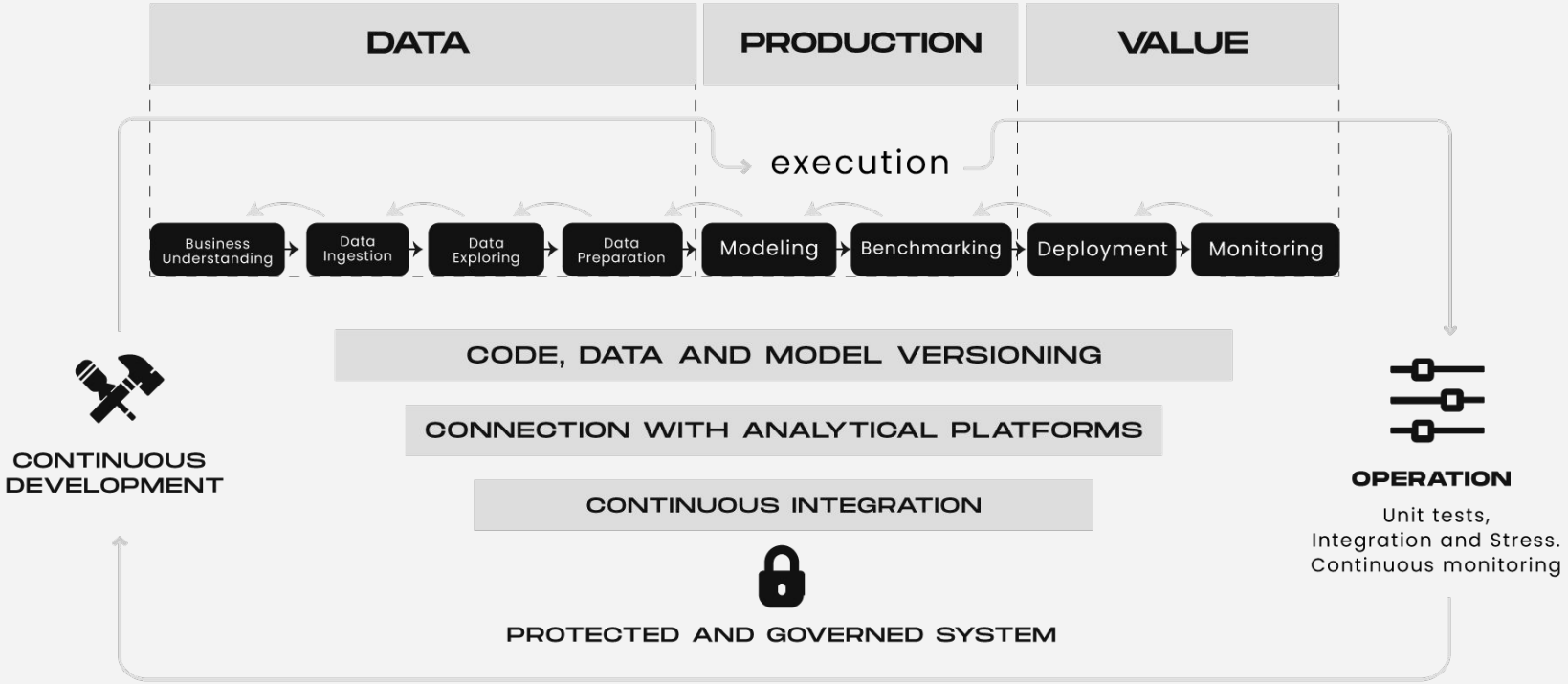
CLOSTER H2O.ai

Data Science Workflow



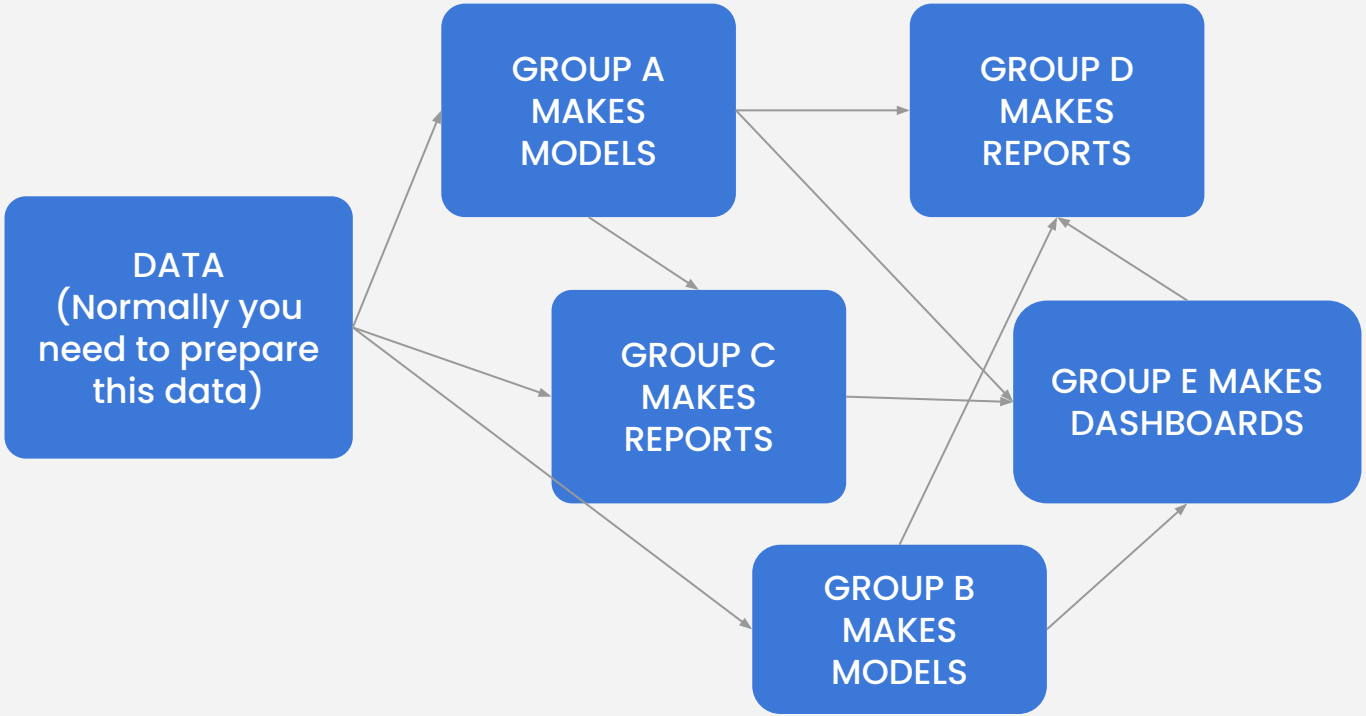
CLOSTER H2O.ai

DATA SCIENCE METHODOLOGY



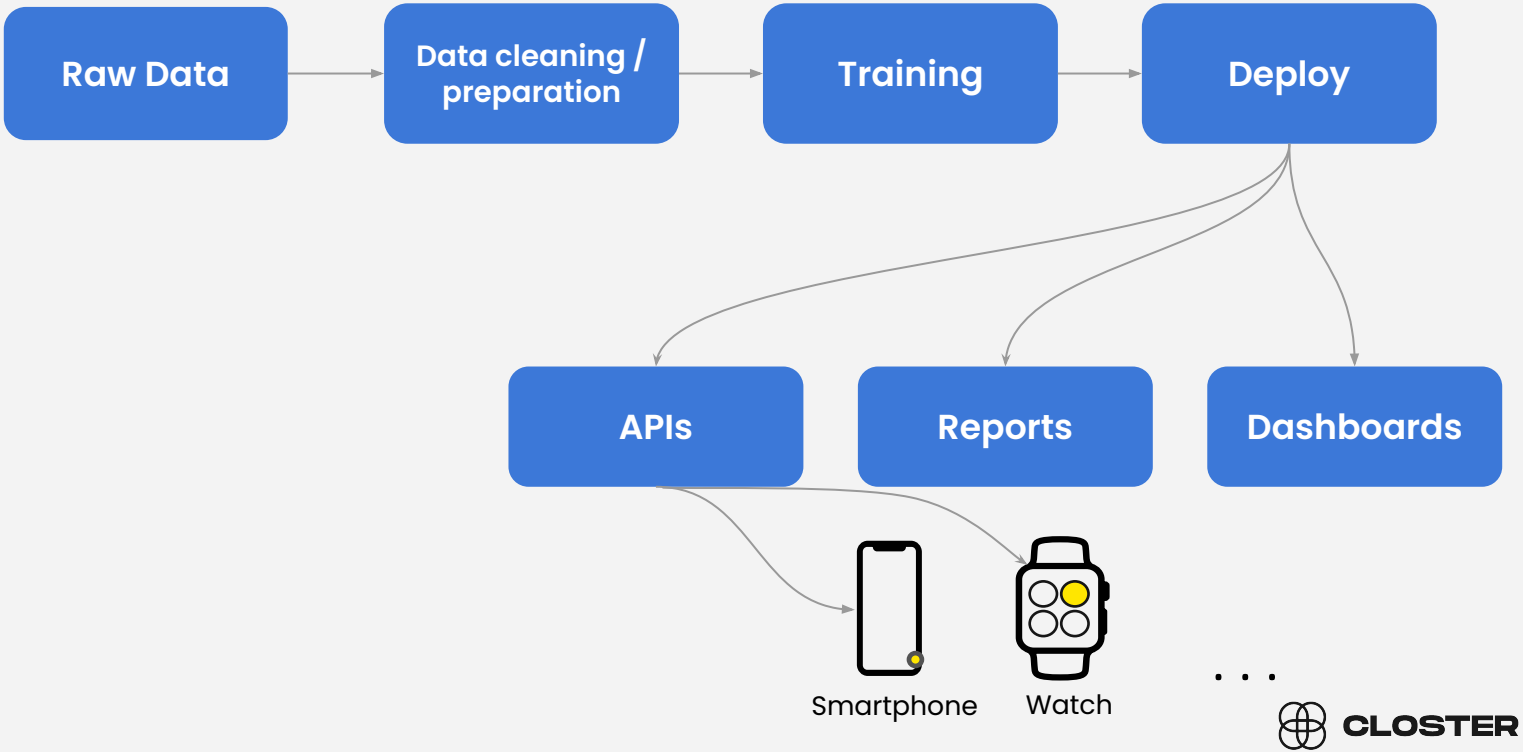
CLOSTER H2O.ai

Data science products are complex



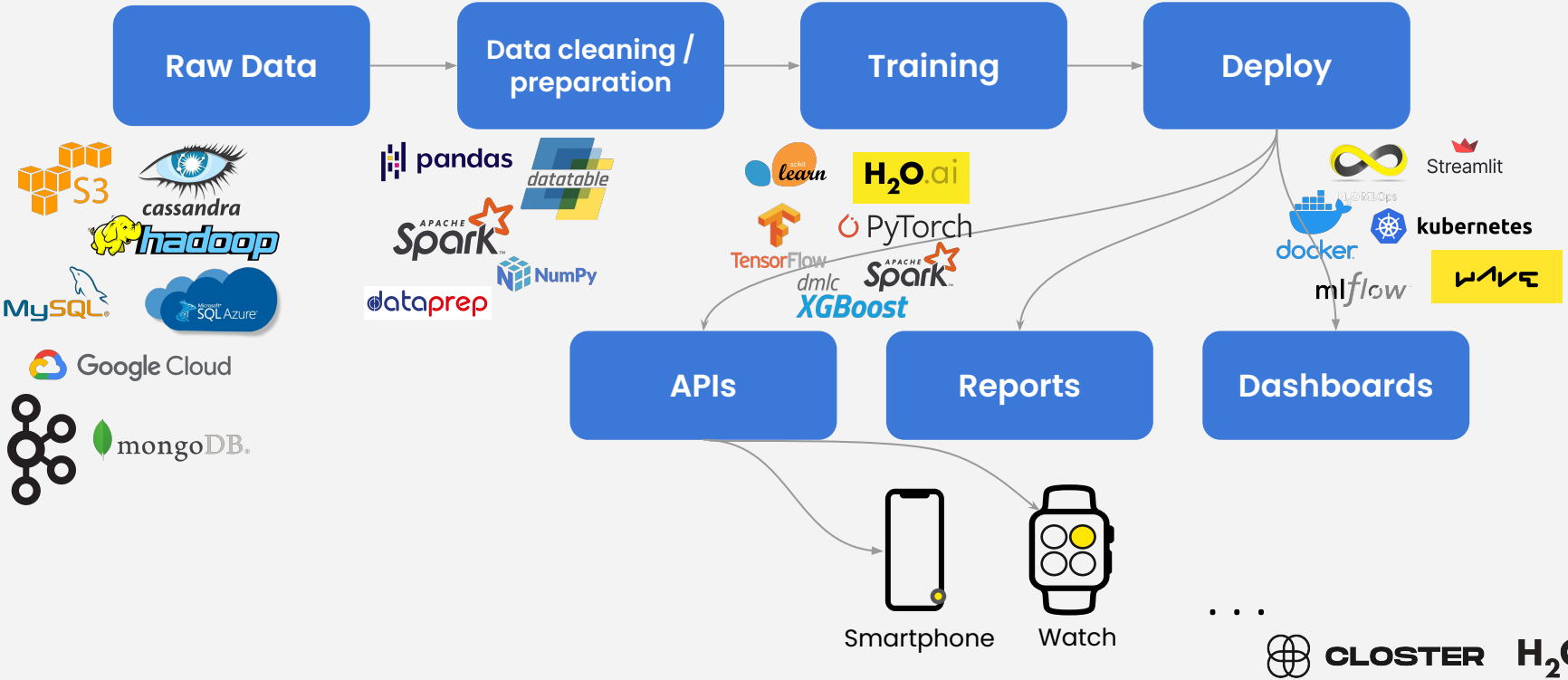
CLOSTER H2O.ai

Data science products are complex



CLOSTER H2O.ai

Data science products are complex



Data science products are complex

Demo links

https://github.com/FavioVazquez/dsgo-workshop-2021

https://aquarium.h2o.ai/