Data Science and Innovation

Patrick Valduriez

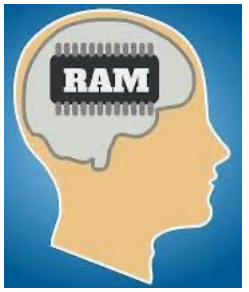
Outline

- Data science
- Technological innovation
- Some success stories in data science
- Hints to promote innovation

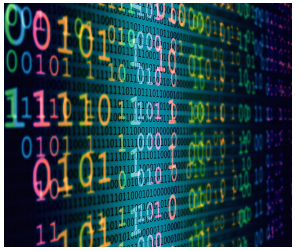
Data Management

- Definition
  - The collection, cleaning, organization, storage, updating and analysis of data to produce information
  - The basis for prediction and decision making
- 1975: ACM started SIGMOD
  - Special Interest Group on the Management of Data
- In the ancient times, data was used to provide information to help manage the state
  - 2238 BC: census of agricultural production by the Chinese
  - 1700 BC: cadastre (for land tax collection) by the Egyptians
- The origins of statistics, accounting and other scientific disciplines

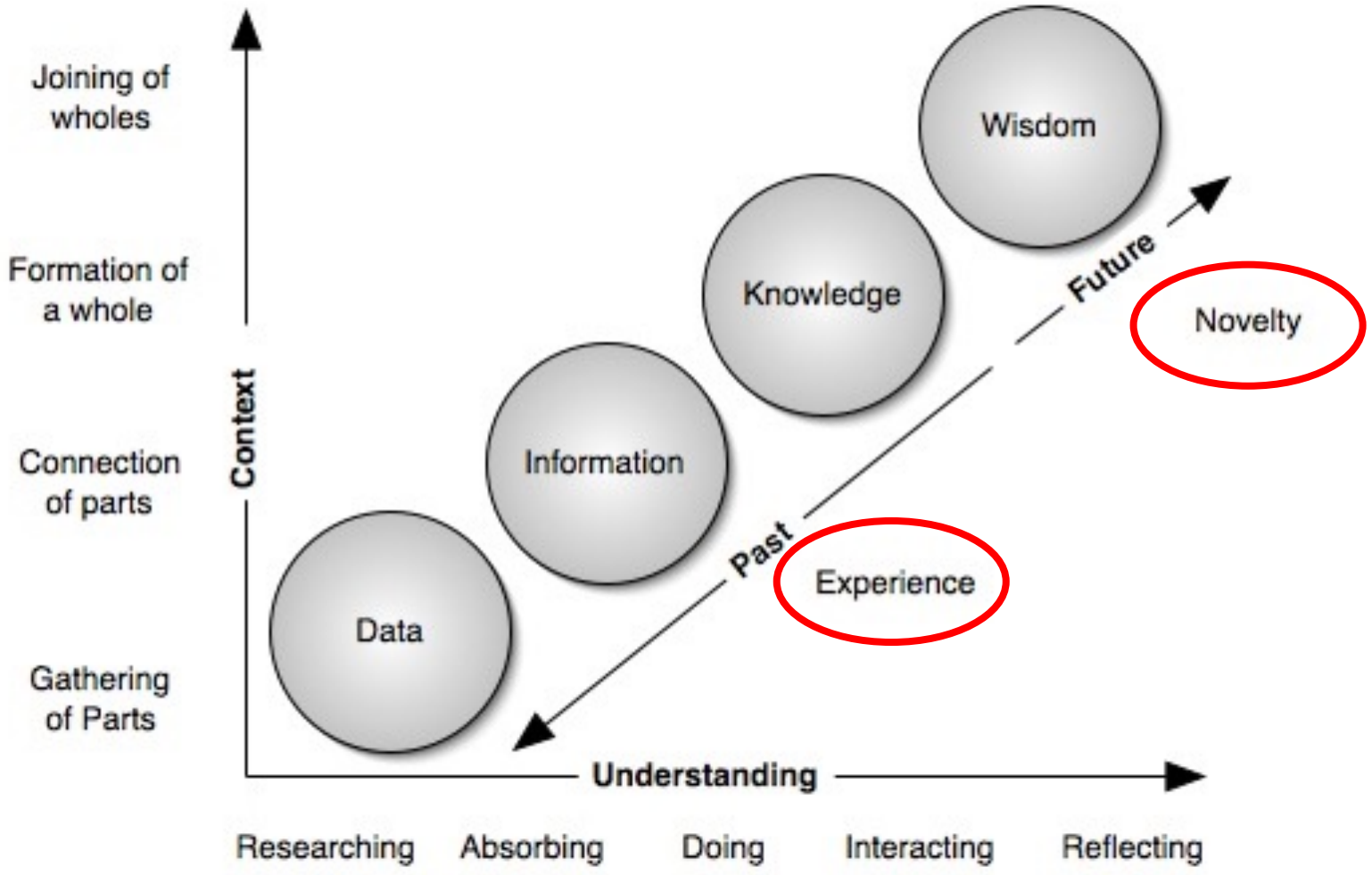
The Impact of the Data Medium



- To store, structure, share and manipulate data
- Physical medium
  1. Human memory: limited, unreliable
  2. Clay, stone: limited, reliable, heavy
  3. Papyrus: light, sequential scrolling
  4. Parchment: page organization
  5. Paper: like parchment, just better
- Today: digital data
  - Independence of data from physical medium
  - Data can be copied, transformed, manipulated as desired and communicated easily

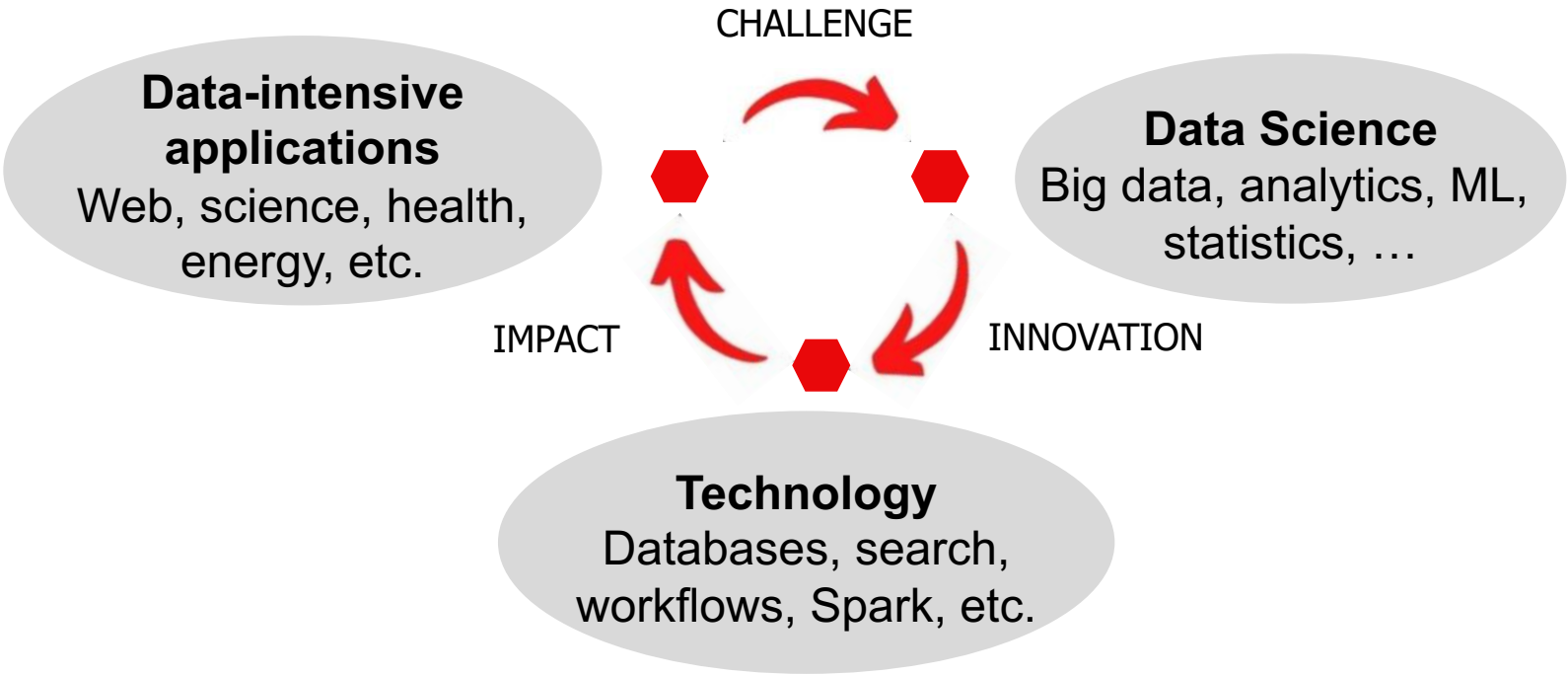


The Continuum of Understanding



Data Science

- Promises universal access to data
  - All human activities (within organizations, businesses experimental sciences, …) now depend on data
- With much innovation and impact



Technological Innovation

Innovation

- Introduces something new to the world
  - Economy: process, product, business model, …
  - Society: idea, belief, religion, political system, …
- May yield "progress"
- But sometimes considered armful
  - England, 1546 (during the Protestant Reformation): Innovation Ban by King Edward 6 to protect the state from disorder and chaos

Innovation

- Introduces something new to the world
  - Economy: process, product, business model, …
  - Society: idea, belief, religion, political system, …
- May yield "progress"

Letter against AI: Elon Musk and experts call for pause in development

Jefferson Tafarel
March 30th, 2023

Letter against AI signed by more than 1000 experts warns about the risks of the race in the development of Artificial Intelligence (AI) models and asks for 6 months of suspension of activities

Technological Innovation

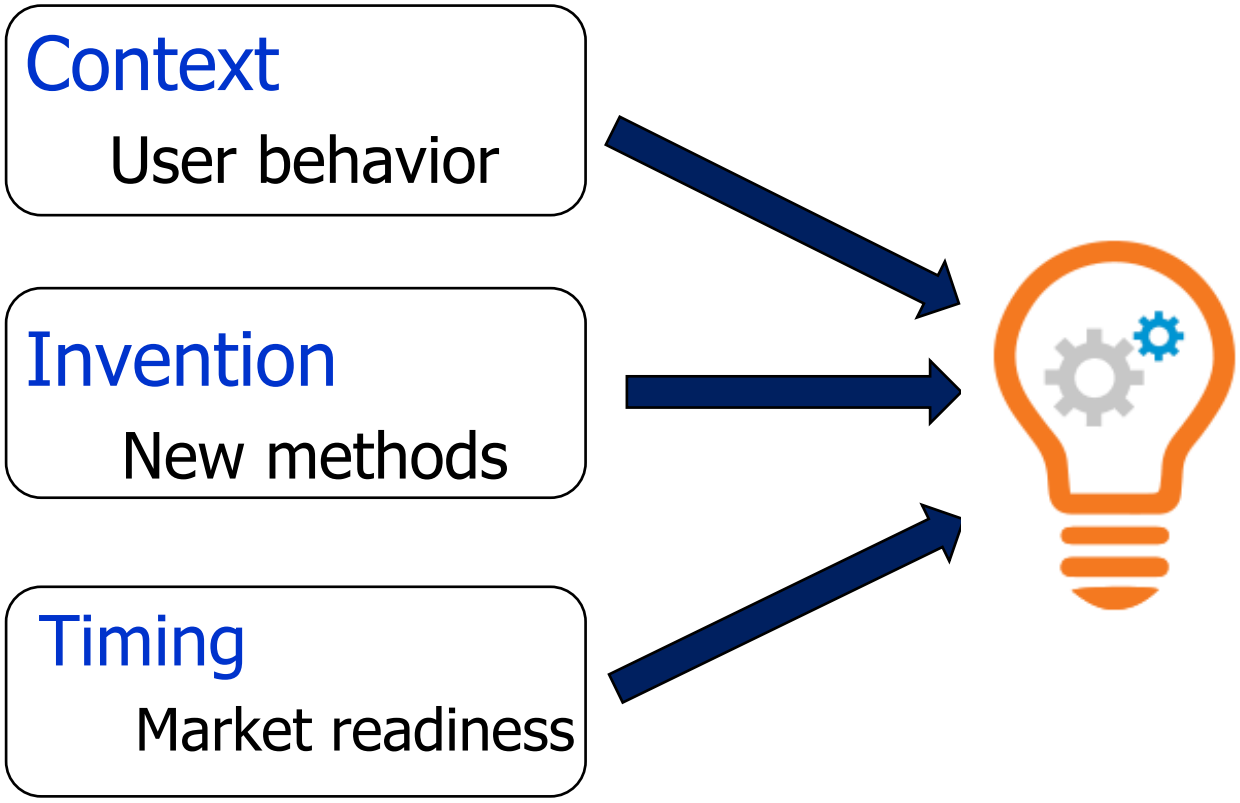
- New technology (as a result of research)
  - E.g. a new code library (implementing a new algorithm)
- Strategies to foster tech innovation
  - Within an organization, the market and customer base are well-know, hence, one can have a formal process, driven by *managers*
  - Within a startup, the context may be unknown or quickly changing, and hard to formalize (and manage), hence the need for *leaders*

Manager versus Leader

- Both should have common skills
  - Knowledge, experience, dynamism, charisma, communication, benevolence, organization, ...
- Manager
  - In charge of implementing the company strategy
  - May lack technical skills
    - Makes communication with techies difficult
- Leader*
  - Able to create an inspiring vision, and guide and motivate a team towards a common goal
  - Strong technical skills
    - Helps getting respect from techies

*P. Valduriez. Making the Right Move to Senior Researcher
ACM SIGMOD Record, 50(2), 2021

Technological Innovation Process



Invention versus Innovation

- An invention is a *new "thing"*
  - Method, process, machine
    - E.g. algebra, printing, smartphone
  - Can combine several inventions, e.g. the smartphone is a computer, a mobile phone, an appdev, etc.
- An innovation is an invention that causes change in user behavior or business
  - Hard: only a few inventions lead to innovation
  - Can be accidental
    - E.g. the pacemaker
  - Can take much time
    - E.g. the airplane

Invention and Innovation

- Documenting, protecting, and leveraging inventions is critical for innovation
- Two main solutions
  - Patents
  - Public licenses
- Choosing a solution should depend on the particular situation
  - But often is a polemical topic (proprietary versus open)

Patents

- Patents are evidence of inventions with
  - Legal protection of intellectual property
  - Documentation of the invention (unlike trade secret), so that others can improve on
- Some (heavily cited) patents yield innovations while many do not

M. Campbell-Kelly, P. Valduriez: A Technical Critique of Fifty Software Patents. Marquette Intellectual Property Law Review, 249, 2005

The Nose Pick Patent (2000)

US00D430934S

United States Patent [19]

Willard

[11] **Patent Number: Des. 430,934**

[45] **Date of Patent: ** Sep. 12, 2000**

[54] **NOSE PICK**

[76] Inventor: **Charles E. Willard**, 453 W. Mechanic St., Shelbyville, Ind. 46176

[**] Term: **14 Years**

[21] Appl. No.: **29/097,842**

[22] Filed: **Dec. 15, 1998**

[51] **LOC (7) Cl.** **24-02**

[52] **U.S. Cl.** **D24/147**; D11/157; D11/160; D24/133

[58] **Field of Search** D24/133, 146, D24/147, 155; 606/162, 161; D11/157, 160; D21/811, 812; D1/109; D32/40, 43, 46

[56] **References Cited**

U.S. PATENT DOCUMENTS

D. 260,866	9/1981	Richards	D11/160
D. 353,239	12/1994	Briscoe	D32/43
D. 360,720	7/1995	Drevo et al.	D32/4
D. 400,326	10/1998	Fisher	D32/4
5,895,408	4/1999	Pagan	606/16

Primary Examiner—Ian Simmons

Attorney, Agent, or Firm—Woodard, Emhardt, Naughtor Moriarty & McNett

[57] **CLAIM**

The ornamental design for a nose pick, as shown an described.

DESCRIPTION

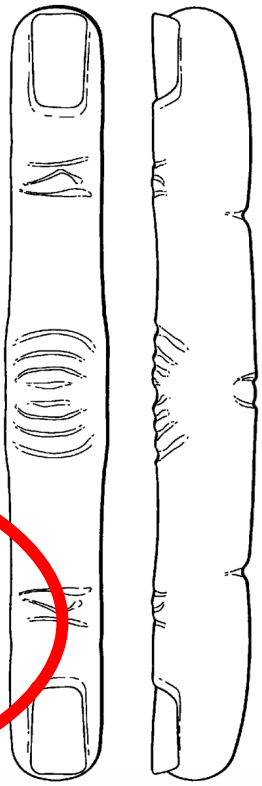
FIG. **1** is a plan view of a nose pick, showing my nev design;

FIG. **2** is a side view thereof with the opposite side viev being a mirror image thereof.

FIG. **3** is a bottom view thereof; and,

FIG. **4** is an end view with the opposite end view being mirror image thereof.

1 Claim, 1 Drawing Sheet



The Magnetic-core Memory Patent (1956)

- U.S. Patent 2,736,880
  - Multicoordinate digital information storage device (coincident-core memory)
  - Jay Forrester (MIT): filed May 1951, issued Feb. 1956
  - 10 pages, highly technical
- Context: Whirlwind computer project at MIT in 1950
  - Required a fast memory for real-time aircraft tracking
  - MIT computer scientist Jay Forrester invents the coincident-core memory that enables the 3D storage of information
- Impact
  - 9 other patents from other inventors
  - Used by all mainframe computers from 1955 to 1975
    - Big $ in patent royalties for MIT

Critique of Patents

- By protecting inventors' rights, they encourage inventions, investment and ROI
- But they may hurt
  - Innovation
    - Patent term is often considered too long (e.g. 20 years) and may hurt competition (monopoly situation)
  - Collaboration with academia
    - (Most) academics suffer the Publish-or-Perish pressure
    - Patenting takes time and may conflict with the publication of research results (which must come next)

Public Licenses

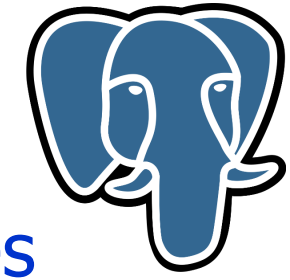
- Protect and leverage artifacts
  - Artefacts: open source software, open source hardware, open data, …
  - The invention is described in research papers, white papers, …
  - The license specifies how the artifact can be used
  - Many different licenses with different constraints for the users
    - Copyleft (GPL, CeCILL, EUPL): viral
    - Weak copyleft (LGPL, Mozilla): for code libraries
    - Permissive (Apache, BSD, MIT)
- The basis for many successful projects
  - Linux, Apache, PostgreSQL, Spark, TensorFlow, Scikit-learn, …
- Strong impact in the cloud service-based business
  - Some $10+ billion acquisitions: Redhat-IBM, GitHub-Microsoft

Success Stories

ORACLE®

- The beginning of relational databases
  - Invention of the relational model by E. F. Codd, 1970
  - Ingres project at UC Berkeley (1975-1980)
  - System R project at IBM Research (1975-1980)
    - Invention of the SQL language
- A few innovations in Oracle 2.0 (1980)
  - Implementation of the SQL language
    - With techniques published in others' research papers
  - Accidental incompatibility with IBM System R
    - Thanks to IBM that kept its error codes secret
  - Support of UNIX and other operating systems
- But many more later on
  - E.g. Oracle Parallel Server (Benoit Dageville et al)

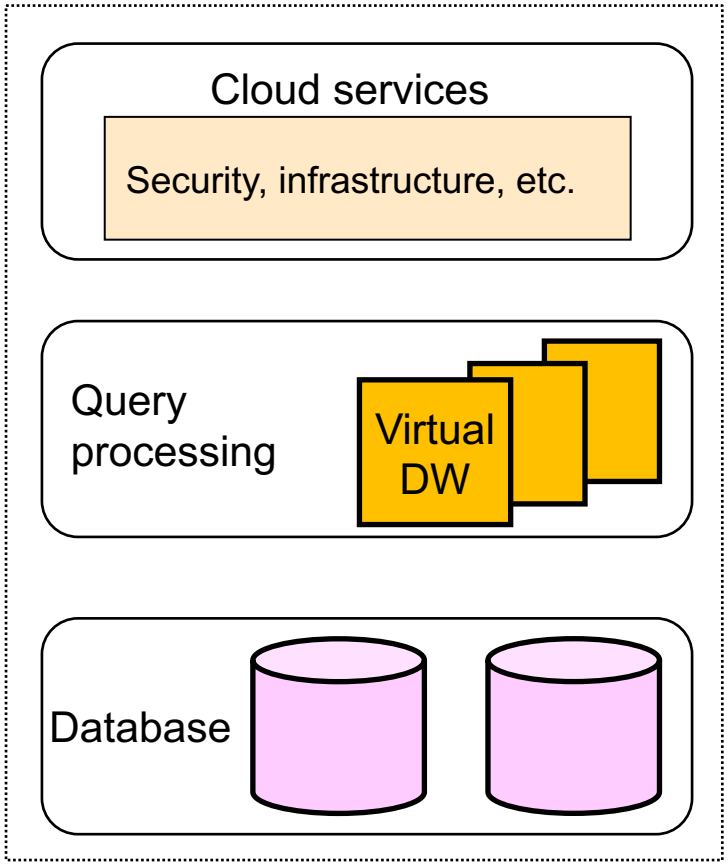
PostgreSQL



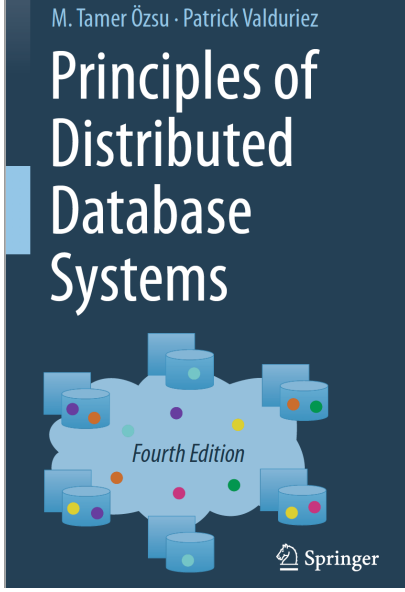
- The next generation of relational databases
  - Postgres (Post-Ingres) project at UCB (1985-1995)
    - The Postgres Next Generation Database Management System. Michael Stonebraker and Greg Kemnitz, Commun. ACM, 1991
- The first open source database
  - Abstract data types
    - Makes the DBMS extensible with user-defined code
  - Rule-based programming
    - Makes the DBMS intelligent
- Impact
  - 4th most popular (db-engines.com/en/ranking)
  - Many successful commercial variations
    - Aster Data, CitusDB, EnterpriseDB, Netezza, ParAccel, …

- Created in 2012
  - Founder: Benoit Dageville
  - 2020: largest IPO at Nasdaq ever ($3.4 billion)
- Cloud agnostic
  - AWS, Azure, Google, …
- Innovations
  - Ease of use
  - Storage disaggregation
    - Independent levels of cloud services
    - Separate provisioning and invoicing

Cloud data warehouse



LEANXCALE



- Delivers a next generation NewSQL database
  - Created in Madrid in 2015 by R. Jimenez-Peris
  - Many innovations in distributed databases

Ultra-Scalable OLTP*

SQL Fully ACID DB

Polyglot

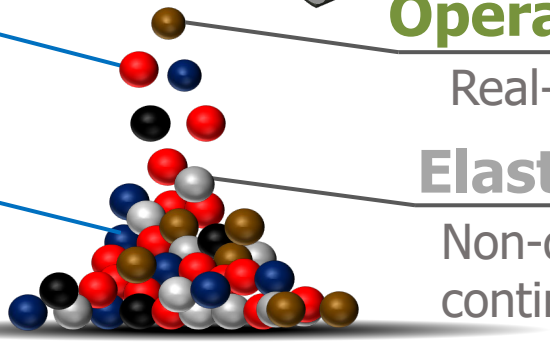
Queries across SQL, HBase, Neo4J, MongoDB, & Hadoop data lakes Integration with Data Streaming

OLAP over Operational Data

Real-Time Big Data

Elastic & Ultra-Efficient*

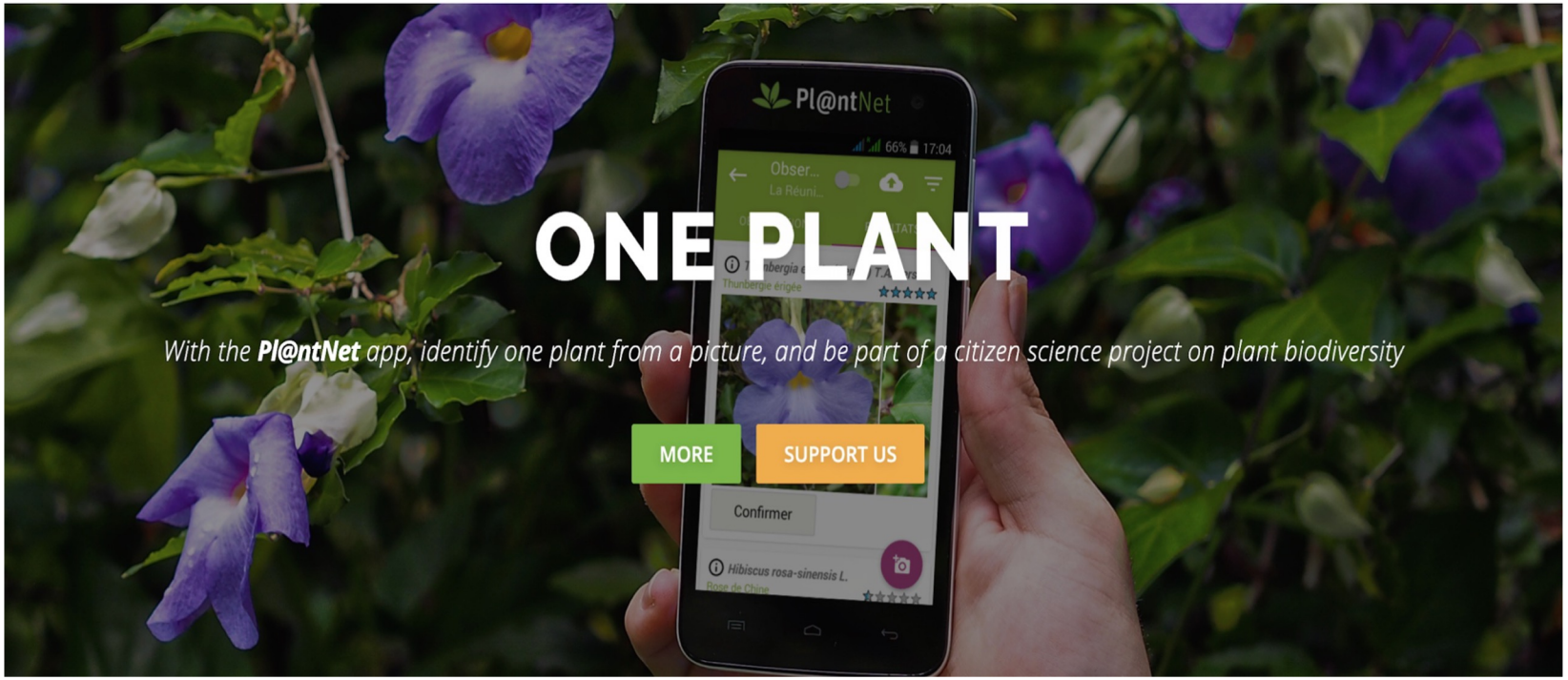
Non-disruptive data migration, continuous load balancing



*R. Jimenez-Peris, D. Burgos-Sancho, F. Ballesteros, Marta Patiño-Martinez, P. Valduriez. Elastic Scalable Transaction Processing in LeanXcale. Information Systems, 2022

Pl@ntNet

A platform based on citizen science and data science to study biodiversity



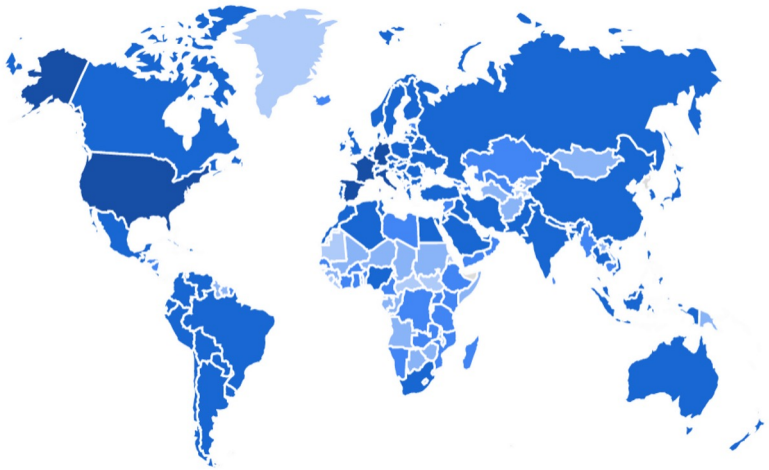
Download on the App Store

GET IT ON Google Play

Pl@ntNet Online Version

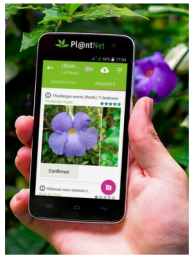
Pl@ntNet

25M users
200+ countries
2M identifications per day

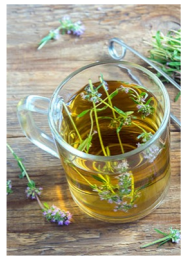


Personal use

Nature



Gardening



Phytotherapy

Professional use

Agro-ecology

Management of natural resources

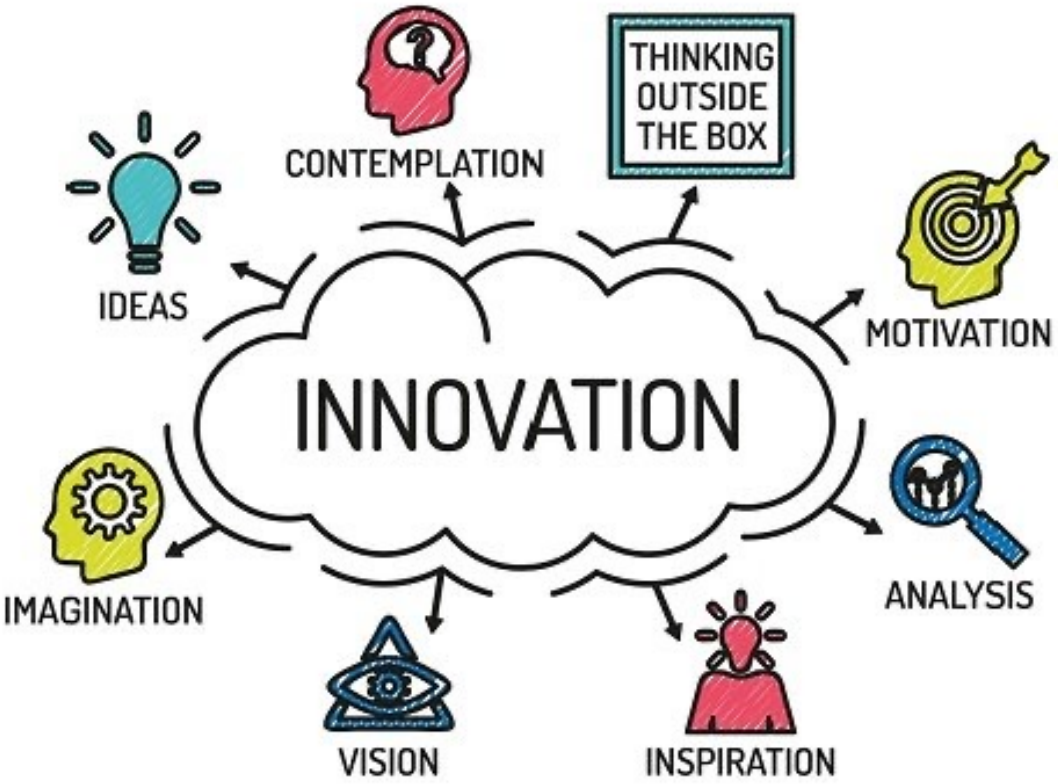
Education, entertainment

Commerce

Pl@ntNet


Training
Validation
BIODIV. DATA
MACHINE LEARNING
Observation
Identification
NATURE
OBSERVERS

Promoting Innovation



Some Hints

- Work with creative people
  - Universities, research labs, startups, partners, etc.
- Promote cerebration by creating a general sense of permissiveness
  - Avoid simplified PKIs and easy metrics
- Encourage creative employees to share their ideas, even preliminary
  - Avoid self-censorship
- Leverage leaders' years of experience and knowledge
  - Educate employees and help push ideas
- Work with key customers to select ideas
  - Turn them into innovations

Open Innovation

- Distributed innovation process across organizational boundaries
  - Competitions, hackathons, start-up incubators...
- Pros
  - For large groups, it promotes the creativity of employees, allows collective innovation and create special relationships with start-ups and research labs
  - For small companies, it allows to benefit from the infrastructures set up by large groups (incubator, hosting, etc.), financing, etc.
- Cons
  - For large groups, it is sometimes difficult to trust start-ups
  - For start-ups, the slow decision-making process is a major obstacle
  - On both sides, staff turnover can threaten the collaborative relationship