Data Science and Innovation

Patrick Valduriez

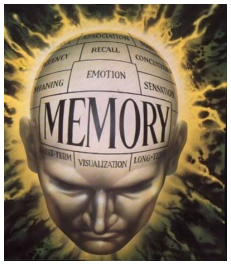
Outline

- Data science
- Technological innovation
- Some success stories
- Innovation at Inria-Brasil
- Hints to promote innovation

Data Science – a brief history

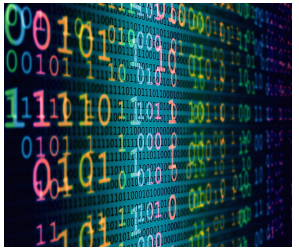
- The use of data started a long time ago
  - 2000 BCE: agricultural census in China
  - 300 years later: land tax collection in Egypt
  - The origins of statistics and accounting
- John Tukey (1962)
  - Describes the merging of statistics and computers as the future of **Data Analysis**
- Peter Naur (1974)
  - Defines **Data Science** as the usefulness of data and data processes derived from their application in building and handling models of reality
- Thomas Davenport and DJ Patil (2012)
  - **Data Scientist**: The Sexiest Job of the 21st Century

The Impact of the Data Medium



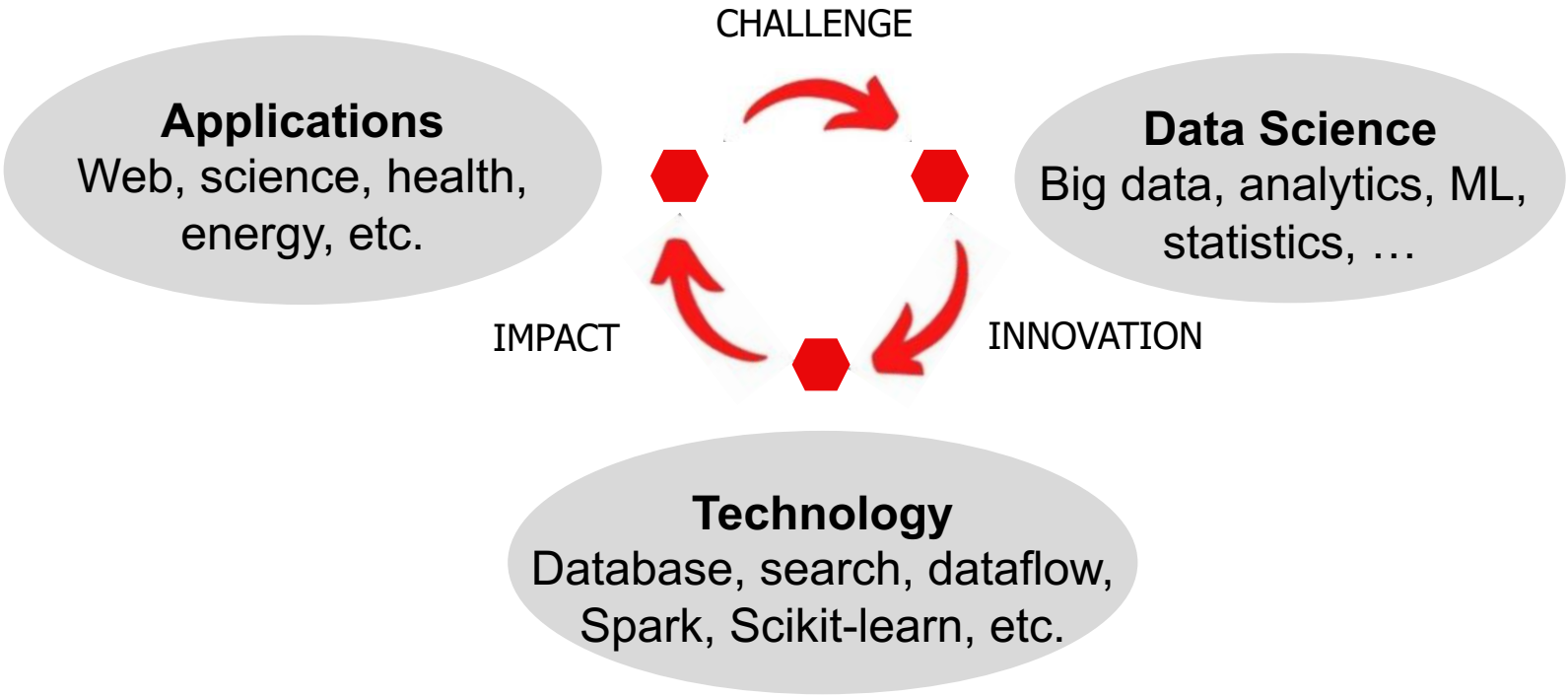
- Physical data
  1. Human memory: limited, unreliable
  2. Clay, stone: limited, reliable, heavy
  3. Papyrus: light, sequential scrolling
  4. Parchment: page organization
  5. Paper: like parchment, just better

- Digital data
  - Independence from electronic medium
  - Data can be easily captured, copied, transformed, manipulated and communicated
  - The basis for data science today



Data Science

- Enables to uncover hidden insights from data
  - Knowledge discovery, actionable intelligence
- Much innovation and impact



Technological Innovation

Innovation

- Introduces something new
  - Economy: process, product, business model, …
  - Society: idea, belief, religion, political system, …
- May yield “progress”

Technological Innovation

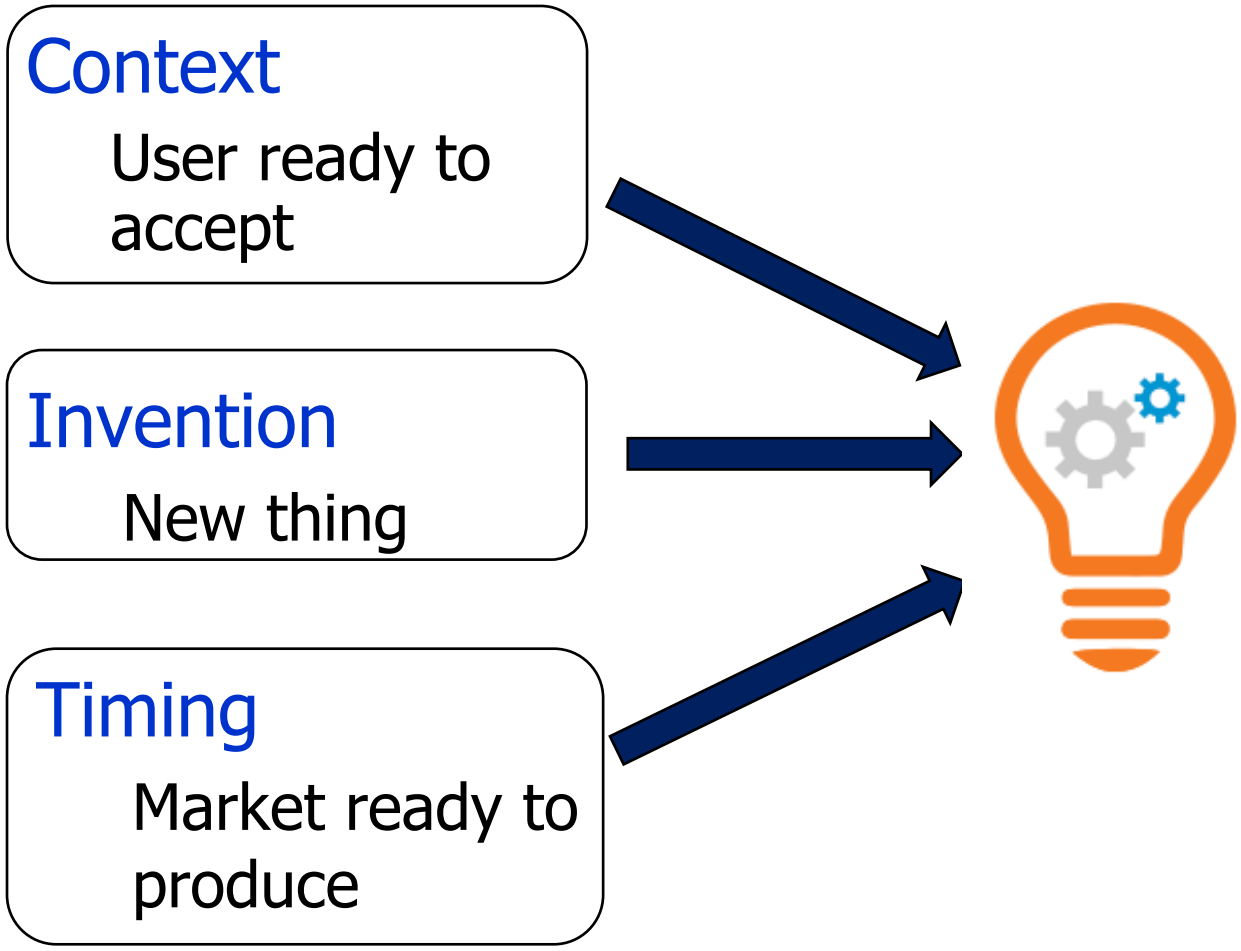
- New technology
  - As a result of research (or R&D), e.g. a new code library (implementing a new algorithm)
- Strategies to innovate
  - Within an organization, the market and customer base are well-know, hence, one can have a formal process, driven by *managers*
  - Within a startup, the context may 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 app environment, etc.
- An innovation is an invention that causes change in user behavior or business
  - Hard: only a few inventions lead to innovation
  - May be accidental
    - E.g. the pacemaker (Wilson Greatbatch)
  - Can take much time
    - E.g. the airplane (who's the inventor?)

Protecting Inventions

- 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 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, Naughton Moriarty & McNett

[57] **CLAIM**

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

DESCRIPTION

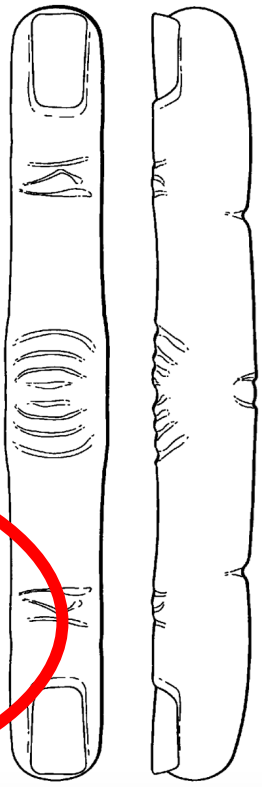
FIG. **1** is a plan view of a nose pick, showing my new design;

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

1 Claim, 1 Drawing Sheet



The Coincident-core Memory Patent (1956)

- U.S. Patent 2,736,880
  - 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

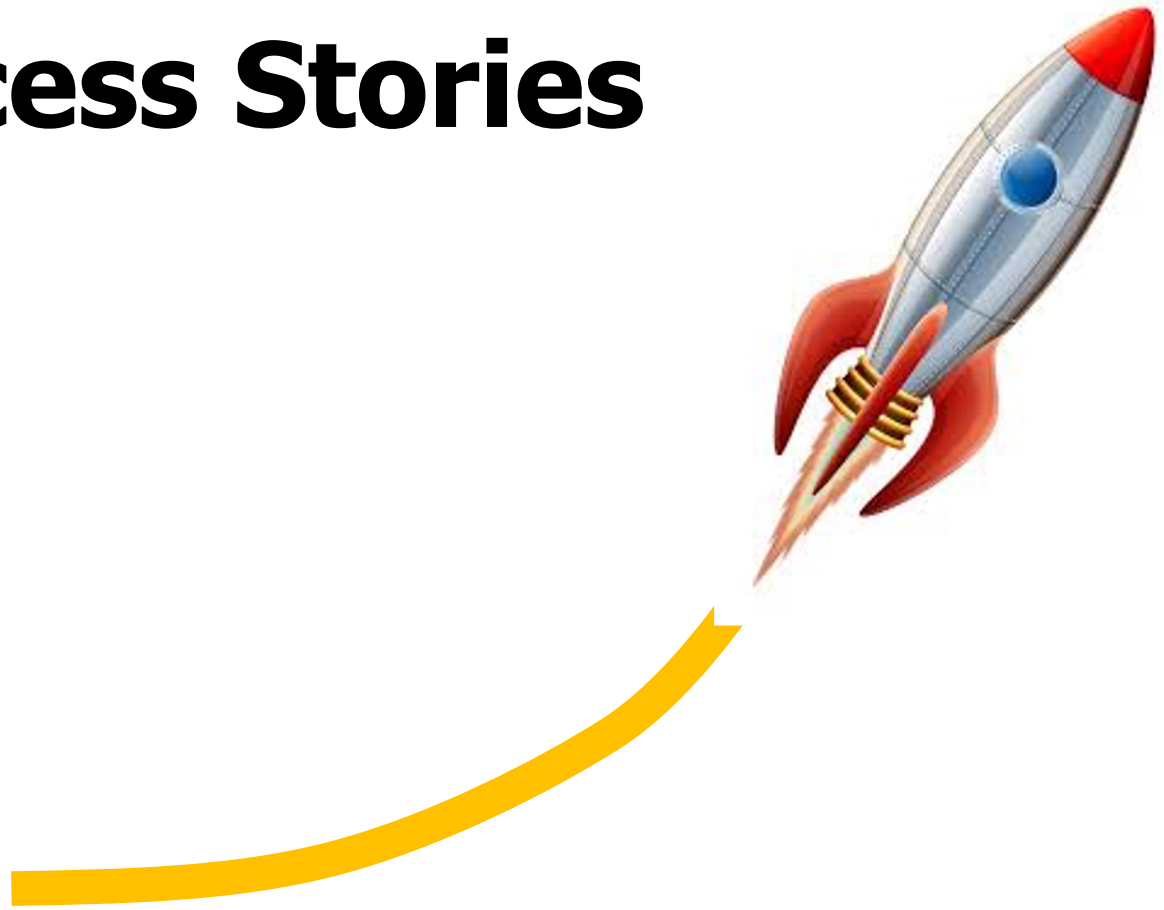
Critique of Patents

- 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
    - Academics suffer the Publish-or-Perish pressure
    - Patenting takes time and may conflict with the publication in a conference or journal (which must come next)

Public Licenses

- Protect and leverage the inventions' artifacts
  - Artifacts: software, hardware, data, …
  - The invention is described in research papers, white papers, …
  - The license specifies how the artifact can be used
  - Open source software licenses
    - Copyleft (GPL, CeCILL, EUPL): constraining, viral
    - Weak copyleft (LGPL, Mozilla): less constraining
    - Permissive (Apache, BSD, MIT)
- The basis for much innovation
  - Linux, Apache, PostgreSQL, Spark, TensorFlow, Scikit-learn, …
- Strong impact in the cloud service-based business
  - Some $10+ billion acquisitions
    - Redhat by IBM, GitHub by Microsoft

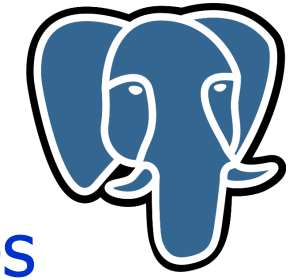
Success Stories



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)
- Only a few innovations in Oracle 2.0 (1980)
  - Implementation of the SQL language (from IBM)
    - With techniques published in others' research papers
  - Accidental incompatibility with IBM's SQL database
    - Thanks to IBM that kept its error codes secret
  - Support of UNIX and other operating systems
- But a lot more later on
  - E.g. Oracle Parallel Server (Benoit Dageville et al)

PostgreSQL

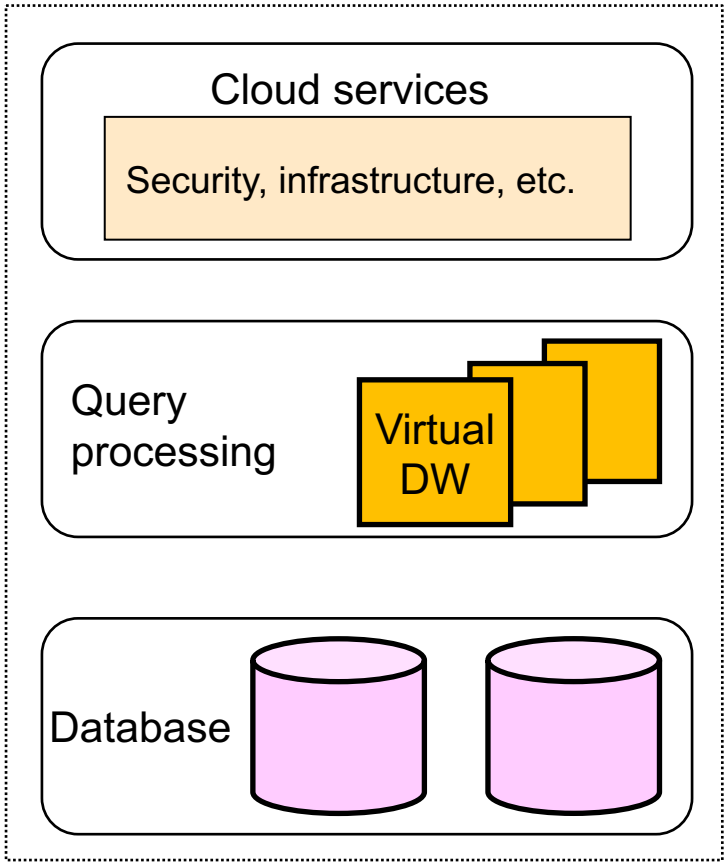


- The next generation of relational databases
  - Postgres (Post-Ingres) project at UCB (1985-1995)
    - The Postgres Next Generation Database Management System. Michael Stonebraker and Greg Kemnitz, Commun. ACM, 1991
- The first open source database
  - Abstract data types
    - Makes the DBMS extensible 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, …

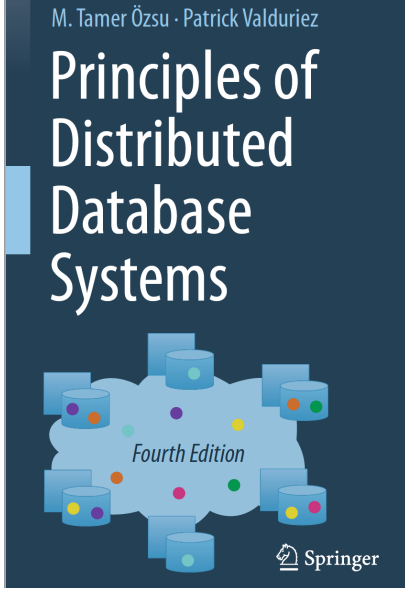
snowflake

- 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

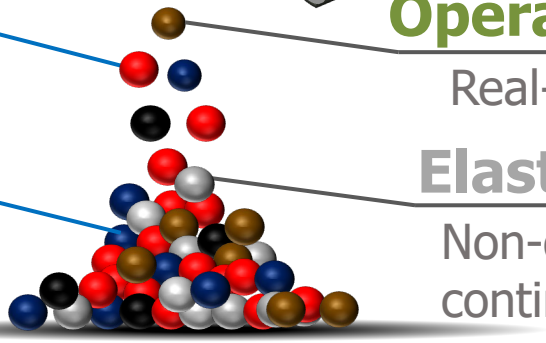
Queries across SQL, HBase, Neo4J, MongoDB, & Hadoop data lakes Integration with Data Streaming

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 citizen science platform that uses ML to help people identify plants with their mobile phones

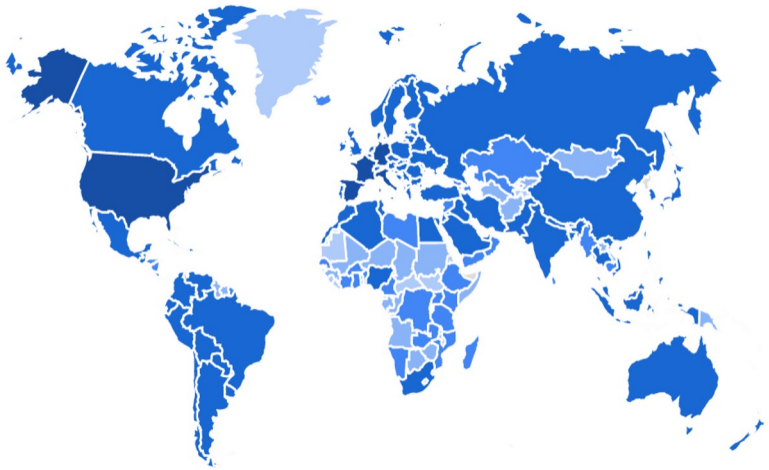
Download on the App Store

GET IT ON Google Play

Pl@ntNet Online Version

Pl@ntNet

20M users per year
200+ countries
2M identifications per day



Personal use

Nature

Gardening

Phytotherapy

Professional use

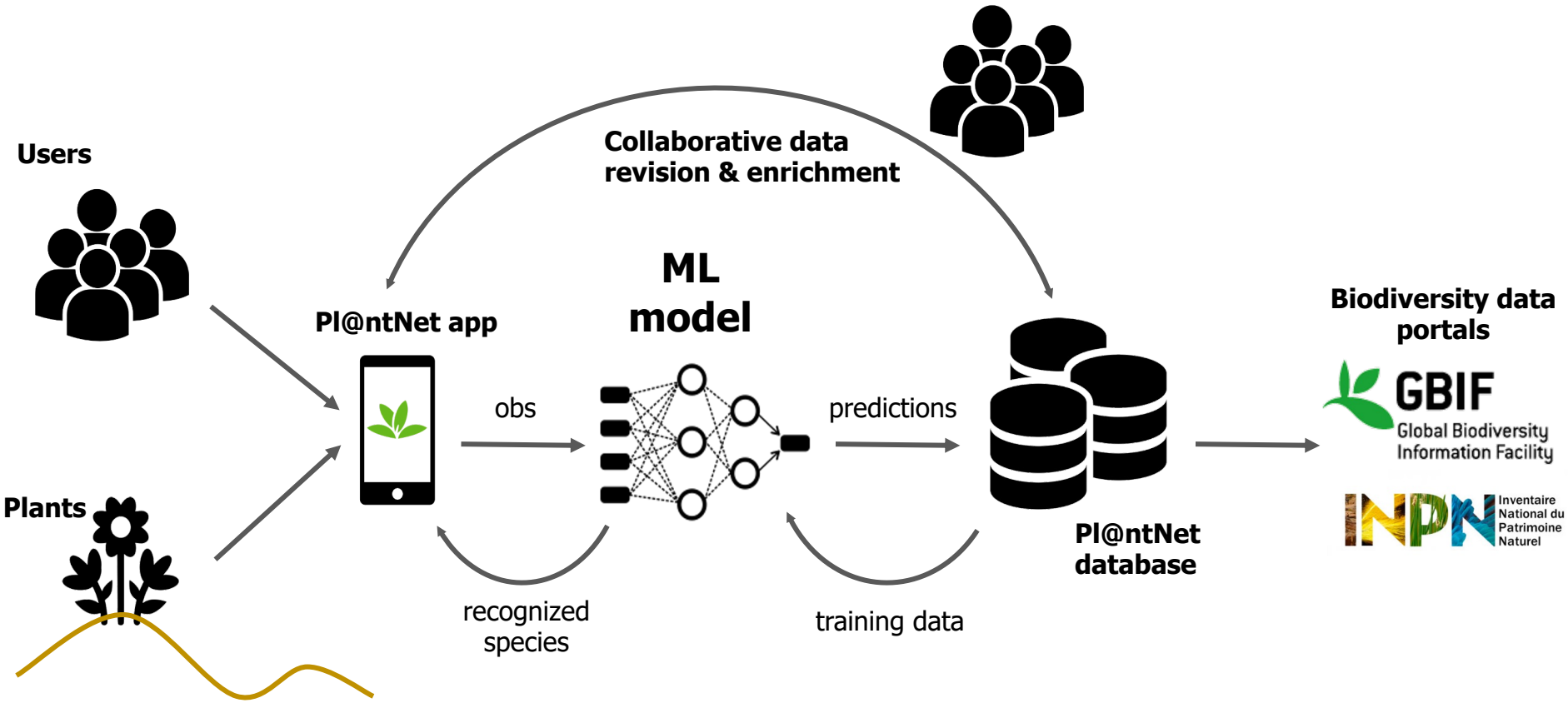
Agro-ecology

Management of natural resources

Education, entertainment

Commerce

Collaborative Machine Learning



Innovation at

Inria Brasil

A PARTNERSHIP WITH LNCC AND BEYOND

Together, we do high-quality research
in HPC, AI, Scientific Computing and Data Science
to create innovative technology and address the challenges
of scientific and industrial applications

HPC4e

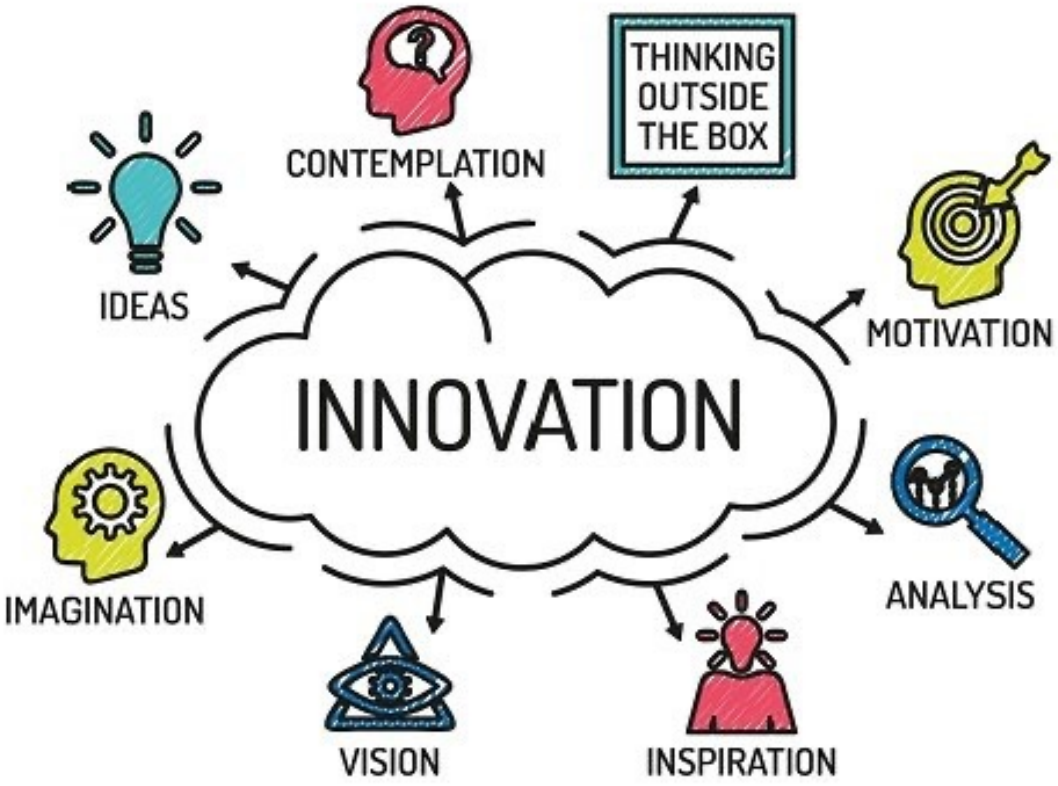
- BR-EU project 2015-2017
  - BR: ITA, LNCC, UFRGS, UFRJ, UFP, Petrobras, IBM Brazil
  - EU: BSC, Inria, Lancaster Univ, CIEMAT, Repsol, Total
- Objectives
  - Develop high performance simulation tools that can help the energy industry to respond to future demands and carbon-related environmental issues
  - Foster the cooperation between companies from Brazil and Europe
- Achievements
  - Strong scientific results in Exascale computing, geophysics and data science
  - Organization of international events
    - Int. Workshop: HPC and Data Science meet Scientific Computing, CARLA 2022 Conference, Porto Alegre, Brazil
  - Knowledge and technology transfer to industrial partners

Provlake (Inria, UFRJ, IBM Brasil)

- Computational Science & Engineering: challenge
  - Analyzing the data processed by workflows developed by multiple teams globally is critical for reproducibility, understanding and explaining
  - Hard to integrate heterogeneous data and workflows and make runtime data analysis efficient
- Result
  - Provlake: a data management system capable of efficiently querying data across distributed services, databases and workflows by leveraging provenance data
    - Open source available on github
    - Renan Souza, SBBD 2021 PhD award (advisors: M. Mattoso, P. Valduriez)
    - Many strong papers
- Impact
  - ProvLake software used a lot by other teams*

    *Management of Machine Learning Lifecycle Artifacts: A Survey. M. Schlegel, S. Kai-Uwe Sattler. ACM SIGMOD Record 51.4, 2023
  - https://ibm.biz/provlake: used by IBM for their clients

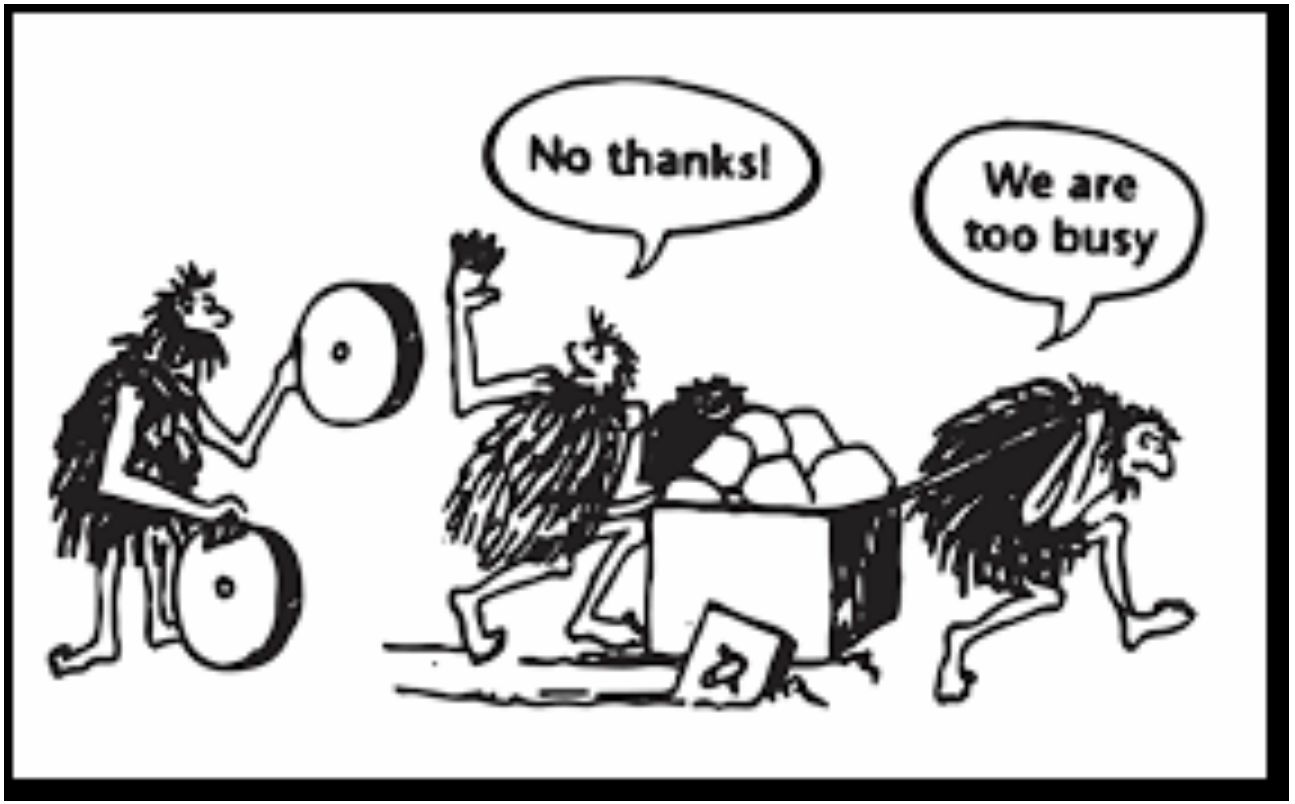
Promoting Innovation



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

Hints to Foster Innovation

- Work with creative, technical 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

Questions?