

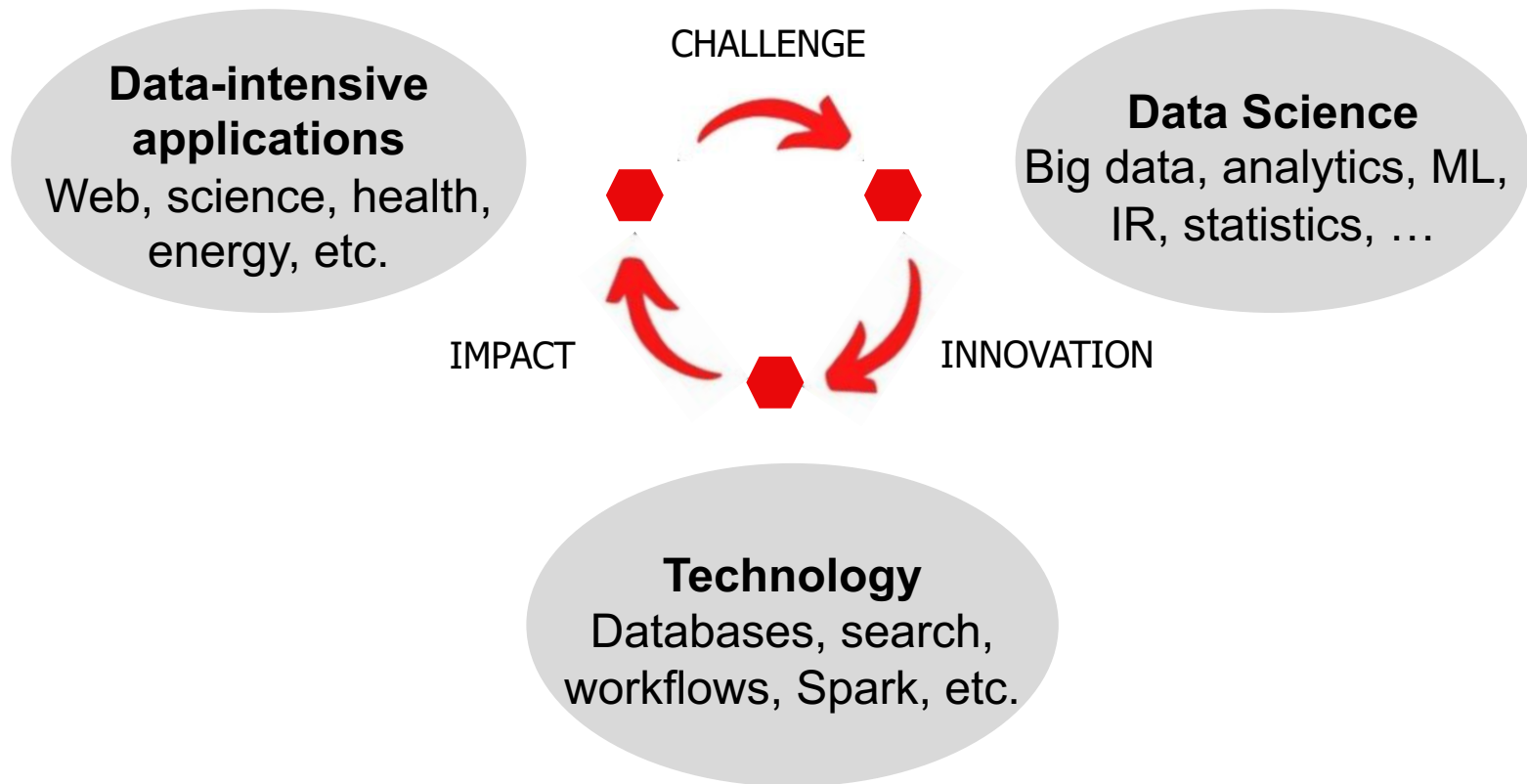
Data Science and Innovation

Patrick Valduriez

Inria

Data Science and Innovation

virtuous circle



Outline

- Technological innovation
- Some success stories in data science
- Innovation at Inria-Brasil
- Conclusion

Technological Innovation



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

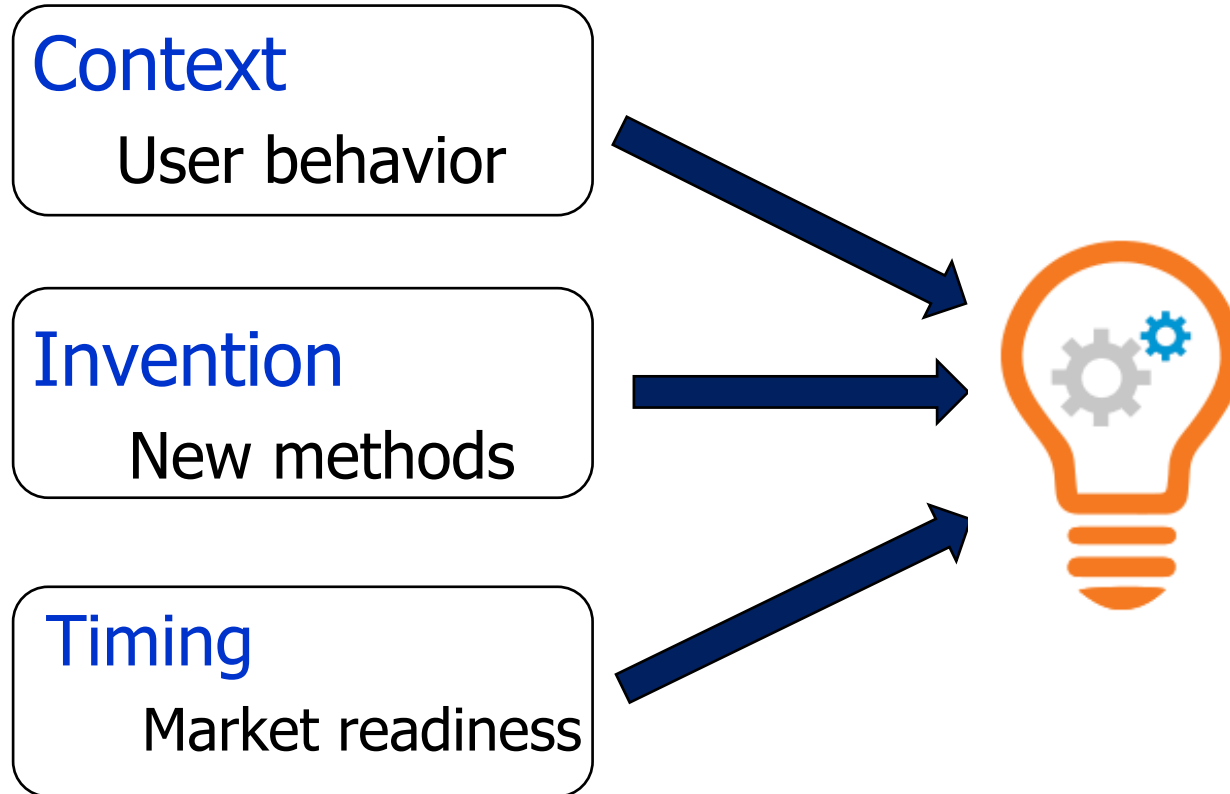
- New technology (as a result of research)
 - E.g. a new code library (implementing a new algorithm)
- Strategies to foster tech innovation
 - Within an organization, the market and customer base are well-known, hence, one can have a formal process, driven by *managers*
 - Within a startup, the context 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, Naughton
Moriarty & McNett

[57] **CLAIM**

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

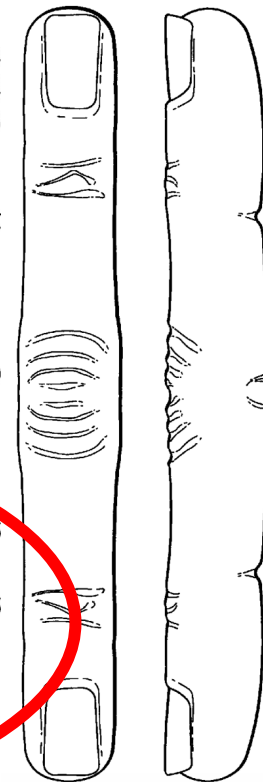
DESCRIPTION

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 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

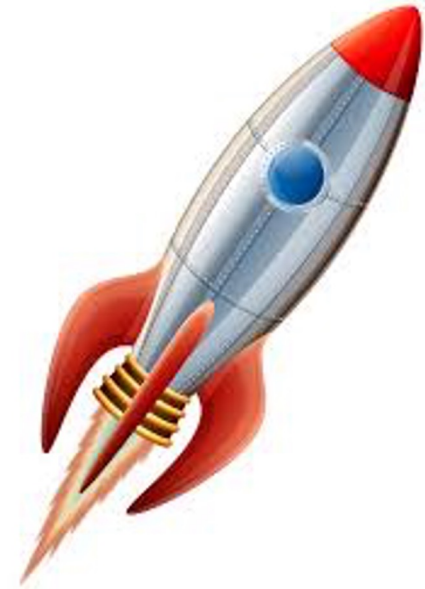
Critique of Patents

- 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

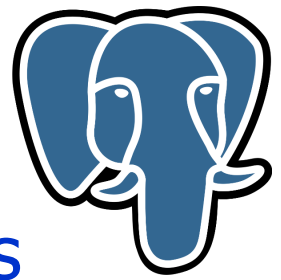
- **Protect and leverage artifacts**
 - Artefacts: open source software, open source hardware, open data, ...
 - The invention is described in research papers, white papers, ...
 - The license specifies how the artifact can be used
 - Many different licenses with different constraints for the users
 - 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



- 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

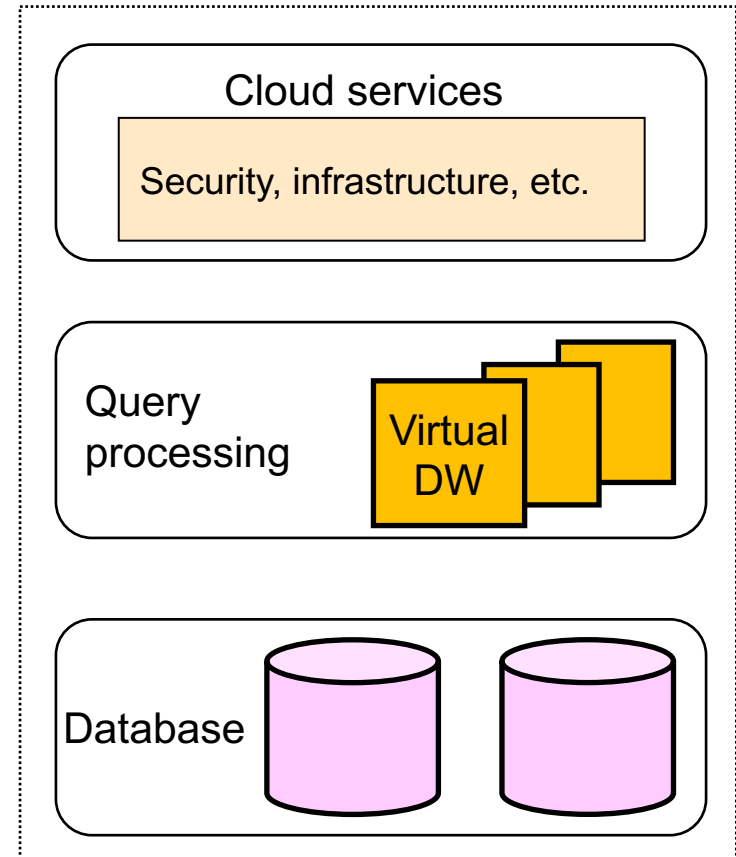


- The next generation of relational databases
 - Postgres (Post-Ingres) project at UCB (1985-1995)
 - The Postgres Next Generation Database Management System. Michael Stonebraker and Greg Kemnitz, Commun. ACM, 1991
- The first open source database
 - Abstract data types
 - Makes the DBMS extensible with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



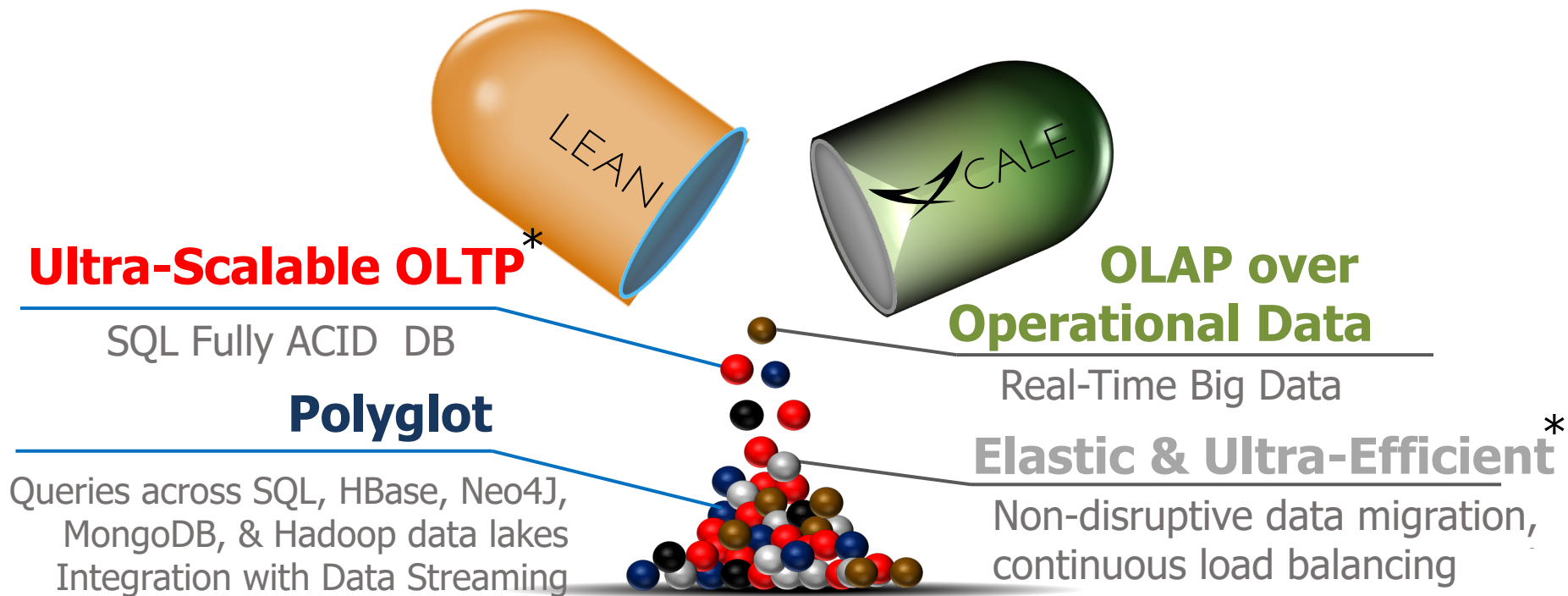
- Created in 2012
 - Founder: Benoit Dageville
 - 2020: largest IPO at Nasdaq ever (\$3.4 billion)
- Cloud agnostic
 - AWS, Azure, Google, ...
- Independent levels
 1. Cloud services
 2. SQL query processing on virtual DWs
 3. Database
- With separate provisioning and invoicing

Cloud data warehouse





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



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

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



- **BR-EU project 2015-2017**

- BR: ITA, LNCC, UFRGS, UFRJ, UFP, Petrobras, IBM Brazil
- EU: BSC, Inria, Lancaster Univ, CIEMAT, Repsol, Total

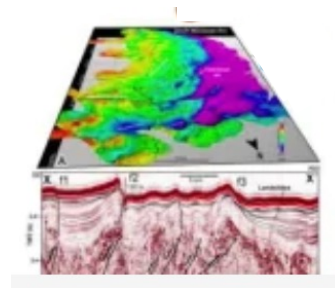
- **Objectives**

- 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

Geophysics for Energy



- **Challenge**

- Improving the efficiency of geophysical exploration codes for different HPC architectures

- **Results (Inria, ITA, UFRGS)**

- A set of synthetic models to test exascale solutions for geophysical imaging
- Different high order methods and computational kernels compared to assess the feasibility of executing applications
 - M. Serpa, ... Ph. Navaux. Strategies to improve the performance of a geophysics model for different manycore systems. IEEE Int. Symp. on Computer Architecture and High-Performance Computing Workshops (SBAC-PADW), 2017
 - R. Lorenzoni, M. Serpa, E. Padoin, J. Panetta, Ph. Navaux, J-F. Méhaut. Otimizando o uso do Subsistema de Memória de GPUs para Aplicações Baseadas em Estênceis. Workshop em Desempenho de Sistemas Computacionais e de Comunicação, SBC, 2017

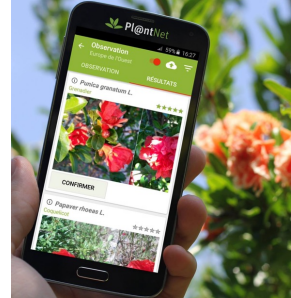
- **Impact**

- Petrobras, Repsol and Total have integrated the results of the project in their in-house codes

Provlake

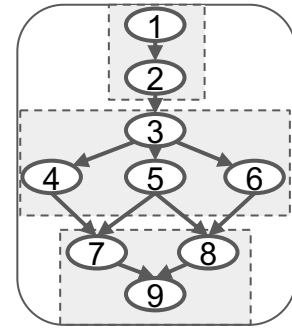


- **Computational Science & Engineering: challenge**
 - Analyzing the data processed by workflows developed by multiple teams globally is critical for reproducibility, understanding and explaining
 - Hard to integrate heterogeneous data and workflows and make runtime data analysis efficient
- **Results (Inria, UFRJ, IBM Brazil)**
 - Provlake: a data management system capable of efficiently querying data across distributed services, databases and workflows by leveraging provenance data
 - Open source: <https://github.com/IBM/multi-data-lineage-capture-py>
 - Renan Souza, SBBD 2021 PhD award
 - R. Souza, P. Valduriez, M. Mattoso, et al. Efficient Runtime Capture of Multiworkflow Data Using Provenance. IEEE Int. Conf. on e-Science, 2019
 - R. Souza, P. Valduriez, Marta Mattoso et al. Workflow Provenance in the Lifecycle of Scientific Machine Learning. Concurrency and Computation: Practice and Experience, 34 (14), 2022
- **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



- Citizen science project - Inria, CIRAD, INRAE, IRD, Telabotanica
 - Goal: help identifying plants and better understanding vegetal biodiversity using AI
 - The Pl@ntNet app (AppStore, Google Play) is used in 200+ countries (30M downloads) with 45K plant species
- Innovations
 - Intensive use of deep learning, content-based IR, parallelism, advanced databases, etc.
 - Innovation Prize 2021, Inria - Académie des Sciences
- Collaboration with LNCC
 - Optimization of Pl@ntNet's performance
 - D. Rosendo, A. Costan, G. Antoniu, M. Simonin, J-C. Lombardo, A. Joly, P. Valduriez. Reproducible Performance Optimization of Complex Applications on the Edge-to-Cloud Continuum, IEEE Cluster 2021
 - ML model selection with the Gypscie framework from LNCC
 - New project with CIRAD and USP to improve our knowledge of the Amazonia flora

Caching of Scientific Workflows



- **Challenge**

- Improving the performance of scientific workflows in a multisite cloud

- **Results (Inria, UFF, CIRAD, INRAE)**

- A distributed solution that leverages the heterogeneous resources available at multiple geo-distributed data centers through adaptive caching
 - DEXA 2020 best paper award by Gaetan Heidsieck, Daniel de Oliveira, Esther Pacitti, et al.
 - G. Heidsieck, D. de Oliveira, E. Pacitti, et al. Cache-aware scheduling of scientific workflows in a multisite cloud. FGCS Journal 2021

- **Impact**

- The solution has been deployed in the OpenAlea workflow system used daily by CIRAD, INRAE, and their partners in the world for high-throughput plant phenotyping

Conclusion

Inria Brasil

A PARTNERSHIP WITH LNCC AND BEYOND

in HPC, AI, Scientific Computing and Data Science



Inria-Brasil and Innovation

- We do have success stories

- But there is room for more!

- Objectives

- Keep doing high quality research in a very competitive environment
- Create an eco-system that implements the innovation virtual circle with industrial partners
 - From Brazil and France
- Get sustained support from Brazil, France and EU to fund projects and keep our top talents

