Moving away from SAS

An opportunity to modernise the practices of statisticians

2024-11-29

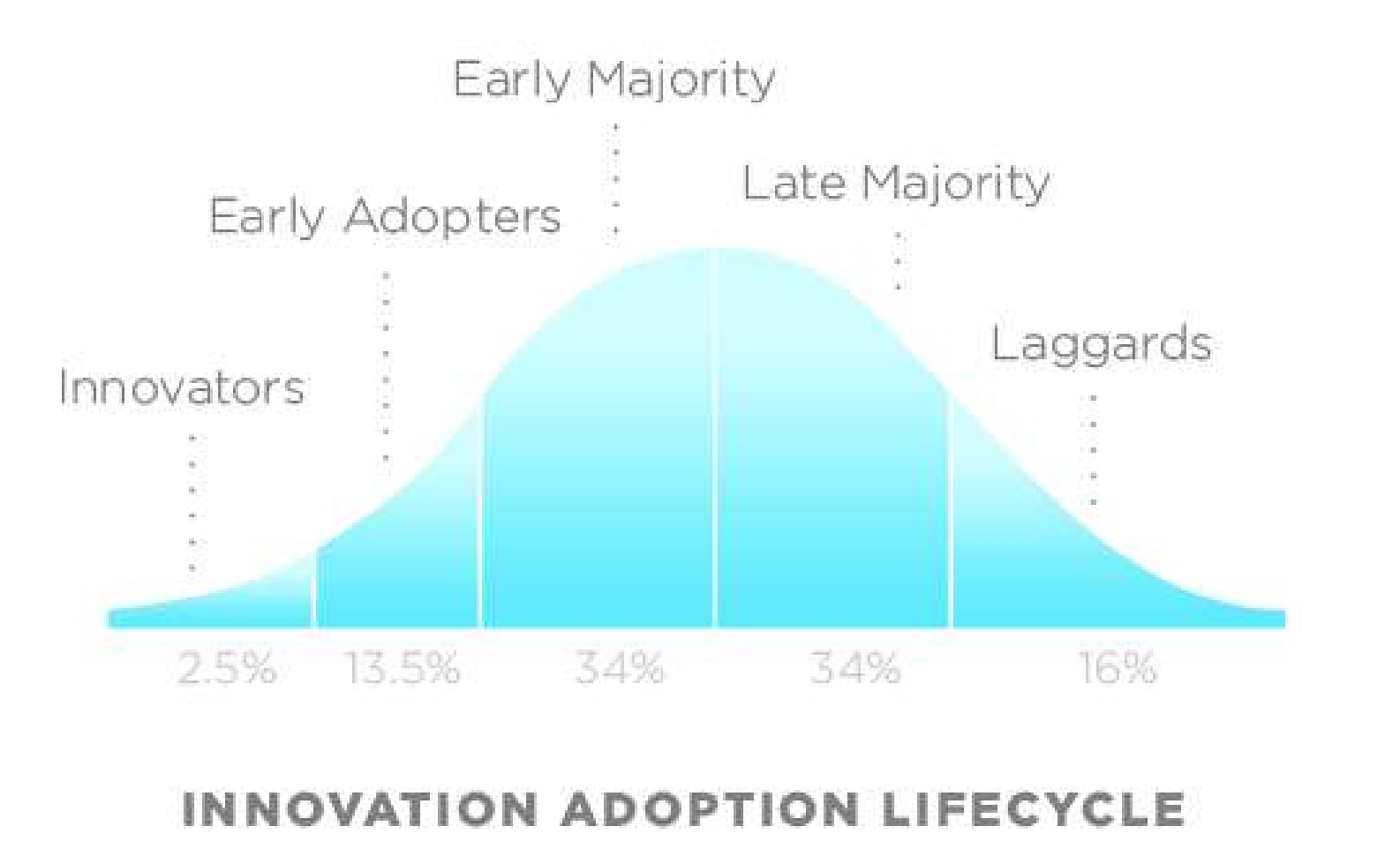
Outline

- Our journey to R adoption
- R adoption motivators
- The main challenge we faced
- The challenges we're still facing

Find these slides: https://rlesur.github.io/uros2024

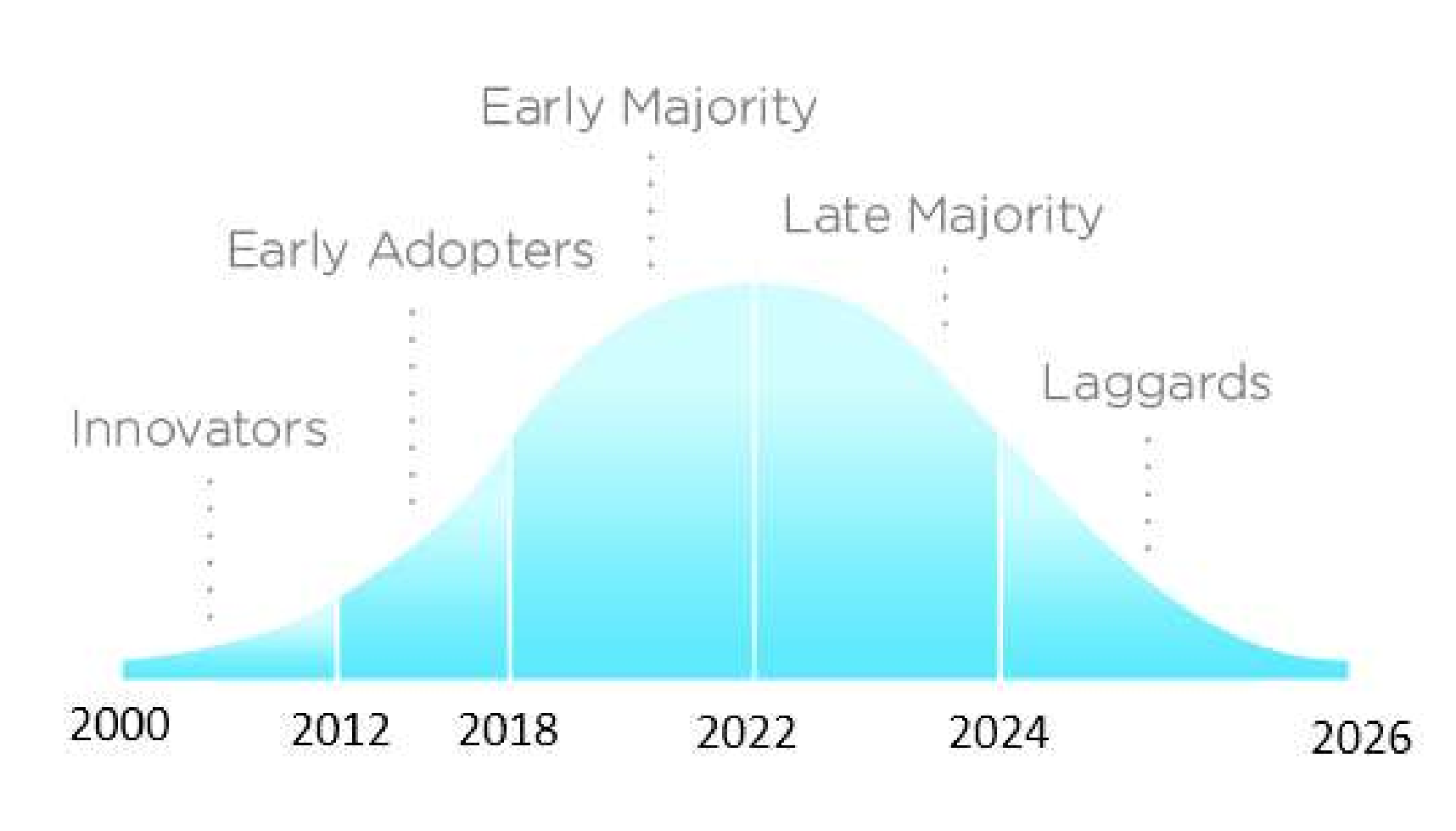
A journey to R adoption in an NSI

Diffusion of Innovations (Rogers, 1962)



A journey to R adoption in an NSI

R adoption at INSEE (France): 1500 statisticians



Innovators' motivators

- **~2000-2012** Innovators' motivators:
  - graphical capabilities (`lattice`)
  - literate programming (`Sweave()`)
  - spatial data analysis (`sp`)

From Innovators to Early Adopters

2012:

- first R user group in Paris (FL\tauR)
- the use of R officially recognized internally
- first RStudio server

Early Adopters' Motivators

- **2012-2018** Early adopters' motivators
  - RStudio IDE
  - graphical capabilities (`ggplot2`)
  - literate programming (`knitr`, `rmarkdown`)
  - dataviz web applications (`shiny`)
  - data wrangling (`tidyverse`, `data.table`)
  - first packages released on CRAN (`icarus`, `btb`, `gustave`)

Getting the Early Majority

- **2018**: all open source languages officially recognized
- **2018-2022**: a collective skills upgrade
  - mass trainings in R
  - an open source, community-based documentation on R by and for INSEE statisticians: utilitR...
  - adoption of an internal open source code publication policy
  - prototyping a new working environment for data science: an open source cloud based datalab (Onyxia)

Moving away from proprietary languages

- **2022**:
  - 50% of the statistical scripts already been rewritten in R (voluntary initiatives)
  - decision to use only open source languages from 2026 onwards (motivated by the increase in licensing costs)
  - new internal environment for statisticians and data scientists based on Onyxia (vendor lock-ins have moved from languages to platforms)
  - `parquet` files to store and disseminate data (preference for a cross language file format)

From early majority to full adoption

- **2022-2026**:
  - mass trainings for late majority and laggards
  - new trainings on Good Practices with git and R, Bringing Data Science Projects into Production
  - an active community helping each other
  - rewriting of the remaining codes (thanks to ChatGPT for helping us understand SAS programs)

The main challenge we faced

Lack of alignment between IT and statisticians

- IT teams are risk averse: they are paid to provide **stable** IT environments
- providing a working environment for R and python users is challenging: open source languages are **unstable** by nature
- R/python users need flexibility and IT need stability
  Can we find a way out?

Tips for dealing with IT

Dean Marchiori's post: *"5 tips for dealing with IT"*

- **Have empathy**: understand them and what they're facing
- **Political alignment**: convince your manager to help you with IT
- **Find supporters in IT**
- **Find (safe) workaround**: there is always a grey area between what is allowed and what is not
- **Buy commercially licensed software**

Tips for dealing with IT

Dean Marchiori's post: *"5 tips for dealing with IT"*:

- **Have empathy**: understand them and what they're facing
- **Political alignment**: convince your manager to help you with IT
- **Find supporters in IT**
- **Find (safe) workaround**: there is always a grey area between what is allowed and what is not
- ~~**Buy commercially licensed software**~~ **not always necessary**

How did we align IT and statisticians?

- Political alignment: creation of an innovation team in the IT Department (2018)
- Mutual empathy between IT innovation team and statisticians
- Statisticians' most valuable supporters
- Create workarounds for statisticians: sspcloud.fr

Towards a new deal between IT and statisticians

An official statistical office is a socio-technical system. We need:

- people (business units, methods, IT)
- data
- hardware
- software
- law and regulations

The technology that opened us new horizons

The benefits of containers

- From the outside, they are all identical: IT teams can learn to manage them routinely
- Inside, you can put what you want (any R version, R packages, system dependencies...)
- Containers are now the basis of all data platforms: IT in official statistics offices **must** learn to deal with them
- Containers also used in DevOps platforms

A new deal between IT and statisticians

- IT cares about containers orchestration and possible vulnerabilities inside them
- IT offers a flexible container based platform (many versions of R, python, IDEs...) and offers support
- IT recognizes statisticians as “citizen developers”:
  - statisticians have extensive rights on the infrastructure
  - statisticians are encouraged to apply software development good practices
  - use of git is mandatory
- The DevOps principle applies: **“You build it, you run it”**

The modern statistician: a proposal

The challenges we are still facing

A wall of confusion between statisticians:

- some statisticians (mostly managers) see R a new statistical "tool". *They focus on the product, not the process* (see Sandra's keynote).
- other statisticians (mostly the younger ones) see R as a programming language and recognize themselves as developers. *They care about the process*
- mass trainings on Good Practices with git and R for both R users and managers

The challenges we are still facing

Statistical trainings in university lack solid basis in computer science:

- Academics have become too specialized
- Few academics also have a production experience
- Inspired by The Missing Semester of Your CS Education, we have created Bringing Data Science Projects into Production and proposed it to our academic partners

Conclusion

- Create a community of R users
- Invest in mass trainings
- Invest in your IT department
- IT must consider statisticians as (citizen) developers
- Statisticians must also be trained in computer science
- Don’t forget to train the middle management

Thanks! Questions?