BANCO DE PORTUGAL
EUROSISTEMA

SPEEDING UP EMPIRICAL RESEARCH: TOOLS AND TECHNIQUES FOR FAST COMPUTING

15-16 DEC 2025

PROGRAMME

15 DEC 2025

14h00 –14h15 **Welcome**
Paulo Guimarães, Banco de Portugal – BPLIM

14h15 – 15h00 **Harnessing the Power of LLMs to Construct Structured Datasets from Unstructured Sources**
David Zarruk, Amazon

One of the major challenges in economic research is efficient data access and processing, particularly when dealing with unstructured data from sources such as press releases, news articles, and public statements. Converting these rich information sources into structured datasets suitable for analysis has traditionally been time-consuming, costly, and often practically infeasible. Recent advances in Large Language Models (LLMs), combined with web-scraping techniques, offer a promising solution to this challenge. This talk will demonstrate how researchers can leverage these technologies to construct structured datasets from unstructured sources, significantly reducing manual effort while expanding research possibilities. Through a practical use case, the session will illustrate how these tools can transform the way economists collect and process data, potentially accelerating the pace of empirical research in economics.

15h00 – 15h45 **Why Your Data Analysis Code Runs So Slowly and What To Do About It**
Jannic Cutura, European Central Bank

This talk — directed at econ grad/PhD students and junior researchers working in statistical departments of public institutions such as central banks — provides a deep dive into what makes data handling and analysis slow, and what can be done about it. The session begins with a beginner-friendly overview of how a computer works, followed by several tricks and strategies that can make data processing faster. After understanding how to optimize code running on a local machine (and the limits of that), the session then turns to cloud-based solutions. By the end of the session, participants will be able to diagnose bottlenecks in their data-handling pipelines and prescribe solutions.

15h45 – 16h15 ***Coffee Break***

16h15 – 17h00 ***collapse* and *fastverse*: Advanced and Fast Statistics and Data Transformation in R**
Sebastian Krantz, CPCS, World Bank, and Kiel Institute

collapse is a large C/C++-based infrastructure package that facilitates complex statistical computing, data transformation, and exploration tasks in R, delivering outstanding performance and memory efficiency. It also implements a class-agnostic approach to R programming, seamlessly supporting vector, matrix, and data-frame-like objects. The *fastverse* extends *collapse* with additional high-performance, low-dependency packages, offering a lightweight and powerful *tidyverse* alternative. This presentation illustrates these capabilities using examples from international trade and spatial analysis.

17h00 – 17h45 **(Pretty) Big Data Wrangling with DuckDB and Polars**
Grant McDermott, Amazon

A new generation of high-performance data engines is transforming how we process large datasets. This session focuses on *DuckDB* and *Polars*, two cutting-edge libraries that deliver exceptional performance while integrating seamlessly into traditional workflows across multiple programming languages. The session will cover real-life examples in R, Python, and Julia, demonstrating how to wrangle hundreds of millions of observations in seconds using only a laptop. Participants will also learn how to scale to contexts where bigger-than-RAM computation is required and discover complementary tools for efficient end-to-end analysis.

BANCO DE PORTUGAL
EUROSISTEMA

16 DEC 2025

10h00 – 10h45 **Parallel and Cross-Language Computing: A Hands-On Workshop for Empirical Researchers**
Miguel Portela/Nelson Areal, Universidade do Minho

This hands-on session demonstrates practical strategies for accelerating computational workflows in empirical research. Participants will learn a systematic approach to performance optimization, beginning with a naive R implementation and progressively applying multiple acceleration techniques. The session covers: (1) native R optimization strategies, (2) integrating C++ via Rcpp, (3) leveraging Julia for high-performance computing, (4) implementing parallelization within R, (5) using GPU acceleration with torch, and (6) scaling to cluster computing for horizontal parallelism. Each approach includes implementation, testing, error handling, and benchmarking to quantify performance gains. The session will also show how AI-assisted coding tools can facilitate cross-language development and code optimization. Attendees will leave with a practical framework for identifying performance bottlenecks and selecting appropriate optimization strategies for their computational research needs.

10h45 – 11h15 ***Coffee Break***

11h15 – 12h00 **Estimating LLRMs (Large Linear Regression Models)**
Alexander Fischer, Trivago

Multiple algorithmic strategies are presented for fitting Large Linear Regression Models (LLRMs) with high-dimensional fixed effects on datasets of multiple tens of millions of observations. The session also introduces open-source software in the R, Python, and Julia ecosystems that implements these strategies. In more detail, the session discusses the following approaches and associated libraries: sparse solvers (*scikit-learn*, *fastreg*), the alternating projections algorithm via the Frisch-Waugh-Lovell theorem (*fixest*, *pyfixest*, *FixedEffectsModels.jl*), data-compression strategies (*duckreg*, *dbreg*), distributed computing (*dask*, *sparkML*), and, finally, moving from the CPU to the GPU (*pyfixest*, *FixedEffectsModels.jl*).

12h00 – 12h45 **Speeding Up Computation Using Julia: An Illustration Using Discrete Choice Demand Models**
Joris Pinkse, Penn State University

Using empirical examples, the session illustrates how Julia can be used to compute large-scale, relatively complex discrete choice demand models with a reasonable amount of resources. Both advantages and challenges will be discussed, and a comparison with alternative approaches will be provided.

12h45 – 13h00 **Closing Remarks**

BIOGRAPHIES:

Alexander Fischer, Trivago

Alexander Fischer is a Data Science Manager at Trivago. Outside of work, he builds open-source econometric software in R and Python, these days mostly focusing on *PyFixest*, a Python package for efficient estimation of regression models with high-dimensional fixed effects.

David Zarruk, Amazon

David Zarruk earned his PhD in Economics from the University of Pennsylvania in 2018, where he worked in macroeconomics and high-performance computing. Following his PhD, he worked as an Assistant Professor at ITAM in Mexico City before transitioning to industry as a data scientist at Rappi, one of Latin America's largest startups. At Rappi, he developed the company's first recommendation algorithms, established its experimentation platform, and built credit risk and fraud models for its credit card product. In 2022, David joined Amazon, where he currently works on causal machine learning problems within the Supply Chain Optimization Technologies organization.

Grant McDermott, Amazon

Grant McDermott is a Principal Economist at Amazon, where he helps lead data-driven projects across different parts of business. He is an advocate of reproducible and open science and maintains several software packages related to econometrics and data science. Before returning to the private sector, he was a faculty member at the University of Oregon, where he continues his research affiliation as a Courtesy Assistant Professor.

Jannic Cutura, European Central Bank

Jannic Cutura is an Economist turned Data Engineer who works as a Staff Engineer at the European Central Bank's data-lake for banking supervision. Prior to his current position he worked as a Software Engineer in stress testing and as a Research Analyst in the ECB's Financial Stability and Monetary Policy divisions. He holds a master's in data engineering from DSTI School of Engineering and a master's and PhD in quantitative economics from Goethe University Frankfurt, during which he conducted research projects at the BIS, the IMF, and Columbia University.

Joris Pinkse, Penn State University

Joris Pinkse is a Professor of Economics at Penn State University. His research interests include econometrics, industrial organization, and antitrust economics. Prior to coming to Penn State, he was an Associate Professor of Economics at the University of British Columbia. His work has been published in Econometrica, the Review of Economic Studies, the Journal of Econometrics, and other journals. He has served on the editorial boards of five economics journals, including Econometrica and the Journal of Econometrics. He is also the author of the *GruMPS* computing package for demand estimation and the 2014 recipient of the Raymond Lombra Award for Distinction in the Social or Life Sciences. Originally from the Netherlands, he has lived in Belgium, the United Kingdom, Canada, and the United States.

Miguel Portela, Universidade do Minho

Miguel Portela holds a PhD in Economics from the University of Amsterdam. He is Full Professor at the University of Minho, Director of the Centre for Research in Economics and Management (NIPE), and Co-Editor of the Portuguese Economic Journal. He is affiliated with NIPE/UMinho, CIPES, and IZA (Bonn), and collaborates with Banco de Portugal. His research interests include labor economics, the economics of education, and applied econometrics, with a particular focus on the Portuguese economy. He has published articles, books, and book chapters, emphasizing publications in Econometrica, Labour Economics, Scandinavian Journal of Economics, Regional Studies, and Studies in Higher Education. He has research collaborations in different countries and leads and integrates research teams whose work has been funded by private and public entities. One highlights his FCT project *"It's All About Productivity: Contributions to the Understanding of the Sluggish Performance of the Portuguese Economy."* He has also written reports to define public policies on the Minimum Wage, Education and Employment in the Portuguese labor market. In addition, he has experience in consulting, both for private and public institutions.

Nelson Areal, Universidade do Minho

Nelson Areal is an Associate Professor in the Department of Management at the School of Economics and Management, University of Minho, and a member of the Centre for Research in Economics and Management (NIPE). He holds a PhD in Finance from Lancaster University (2006). Throughout his academic career, he has held several leadership roles, including Head of the Management Department, Director of the master's in finance, and Director of the PhD in Management. He has collaborated as a consultant with public and private entities. His research interests include risk measurement and forecasting evaluation of derivative instruments through numerical methods, performance evaluation, text as data, and socially responsible investment. He has experience working with high-dimensional financial data.

Sebastian Krantz, CPCS, World Bank, and Kiel Institute

Sebastian Krantz is an Infrastructure Analytics Consultant at CPCS and the World Bank, focusing on emerging markets. He obtained his PhD in Quantitative Economics from the Kiel Institute for the World Economy in 2024 with a dissertation titled *"Africa's Economic Transformation: A Big Data Perspective."* During his PhD and his previous engagement at the Ministry of Finance, Planning and Economic Development of Uganda, he engaged extensively in low-level software engineering. His R package *collapse* evolved from a collection of functions to support complex economic analysis with statistically advanced, flexible, and fast computations, into one of the largest and best performing R packages. It's statistical power and class-agnostic flexibility are unique in the open-source ecosystem, letting it cater to rigorous scientific and computational demands. In late 2021, he started the *fastverse* project and meta-package to promote high-performance and low-dependency R packages for statistical computing. It offers research and industry professionals a compelling alternative to the popular *tidyverse*.