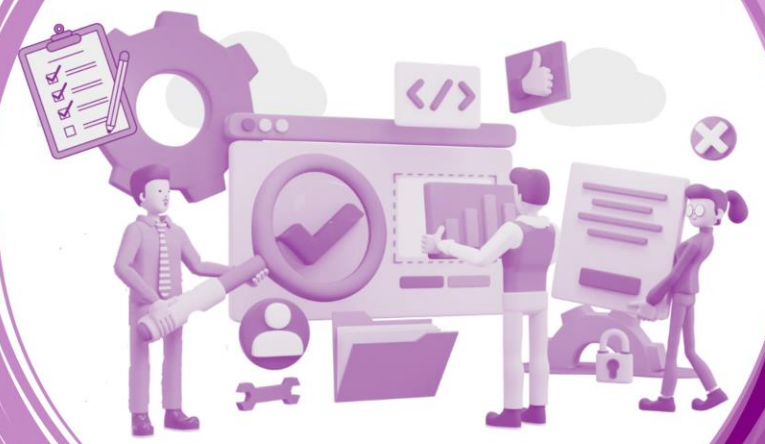
THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

WORLD BANK GROUP
INSTITUTE FOR ECONOMIC DEVELOPMENT

Reproducible Research Repository

Government Analytics and R Training:

Strengthening Public Sector Reporting and Data Analysis



January 13 – January 17, 2024

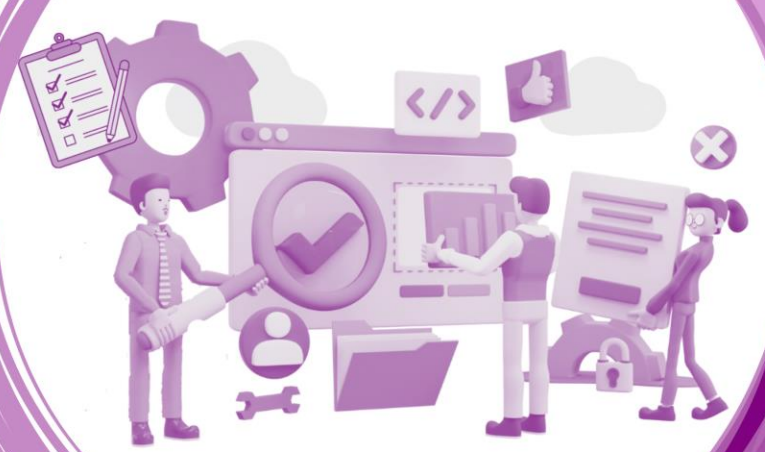
THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

WORLD BANK GROUP
INSTITUTE FOR ECONOMIC DEVELOPMENT

Reproducible Research Repository

Practical tools for Effective Measurement and Analytics

María Reyes Retana

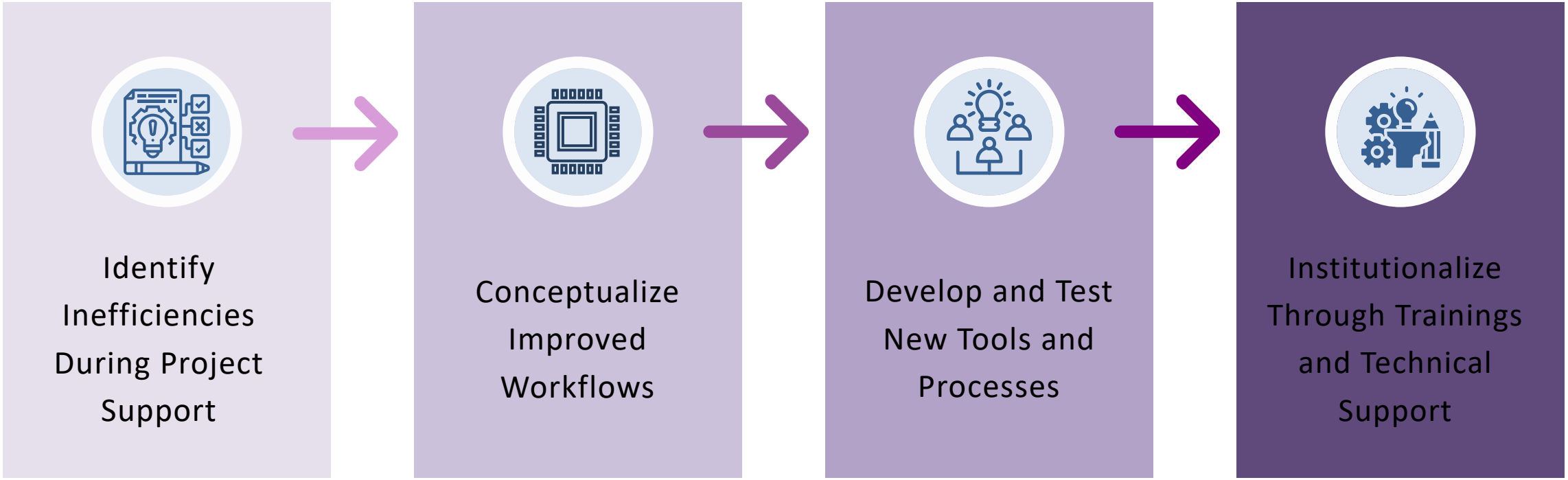


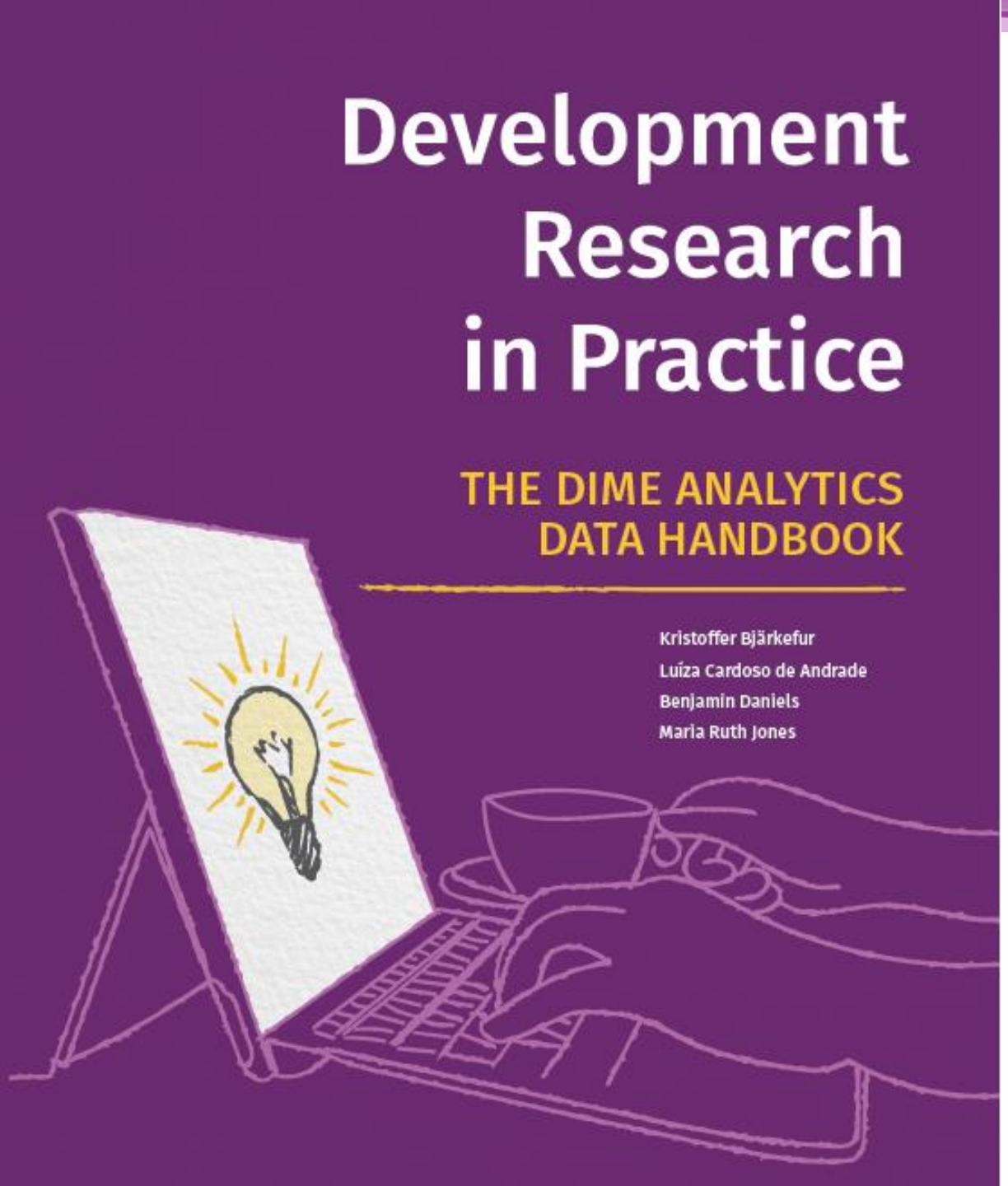
Objectives

- Understand lifecycle of an analytical project
- Learn about publicly-available tools that will make your analytical work more efficient and more credible
- Appreciate why standards for reproducible research are important (and beneficial) for all analytical work

Introduction: DIME Analytics

DIME Analytics takes advantage of the concentration and scale of research at DIME to identify constraints and inefficiencies in the research production cycle and develop and iteratively test solutions





- Trains users of development data how to handle data effectively, efficiently, and ethically
- Compiles DIME best practices into a single narrative covering the full research production cycle
- Case study of a real DIME impact evaluation provides concrete examples throughout
- Available for free on the World Bank Open Knowledge Repository or purchase on Amazon

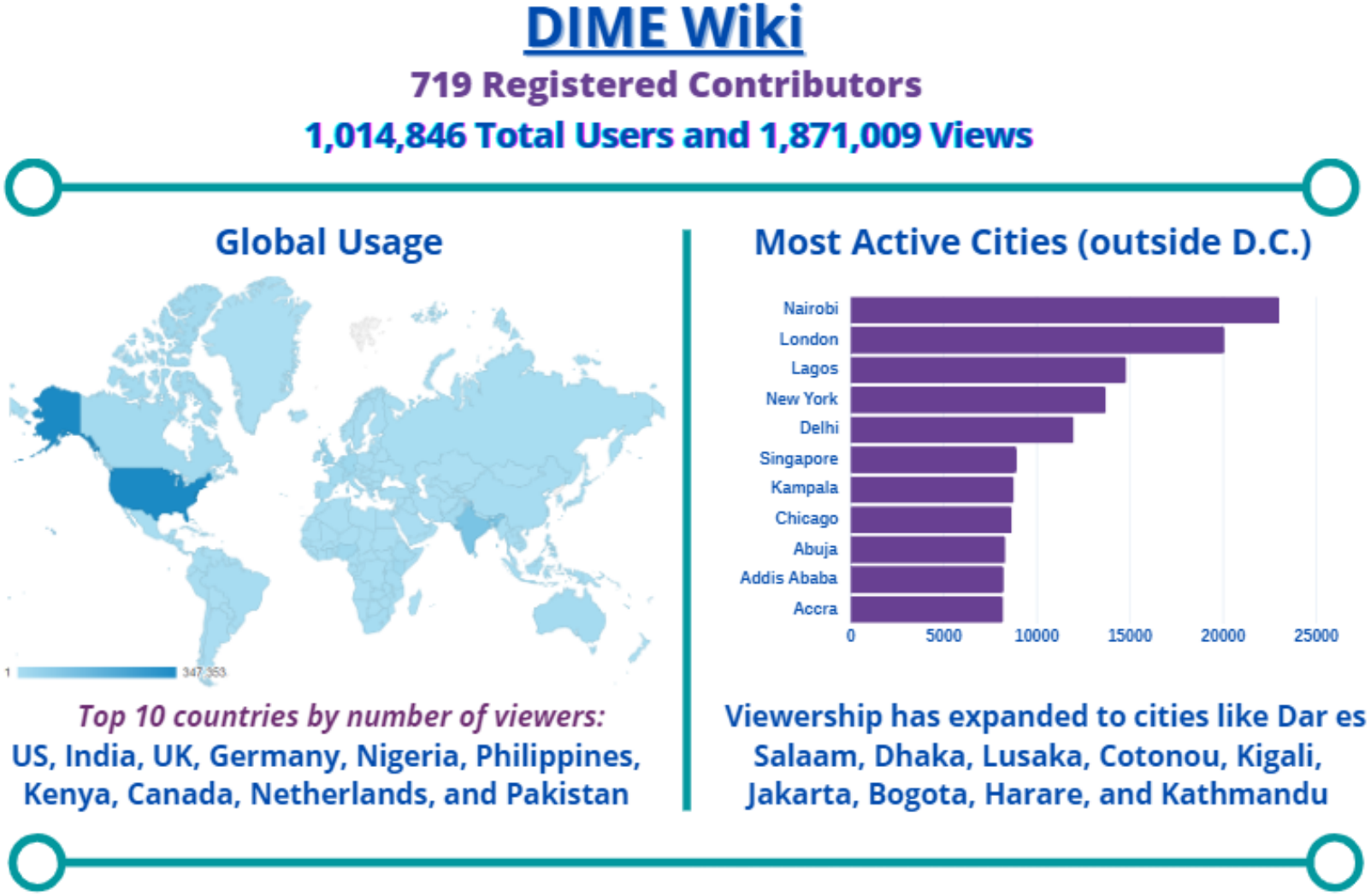
DIME Wiki
(https://dimewiki.worldbank.org/)

One-stop shop for impact evaluation research solutions

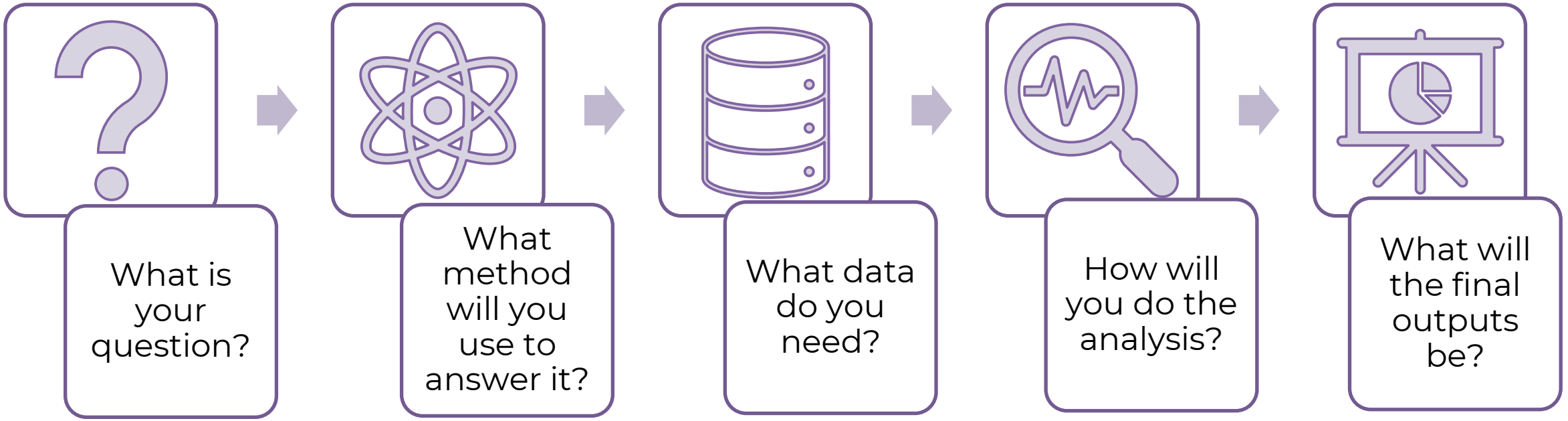
Practical focus, includes implementation details, recommended resources

Live resource, continually updated

Global usership

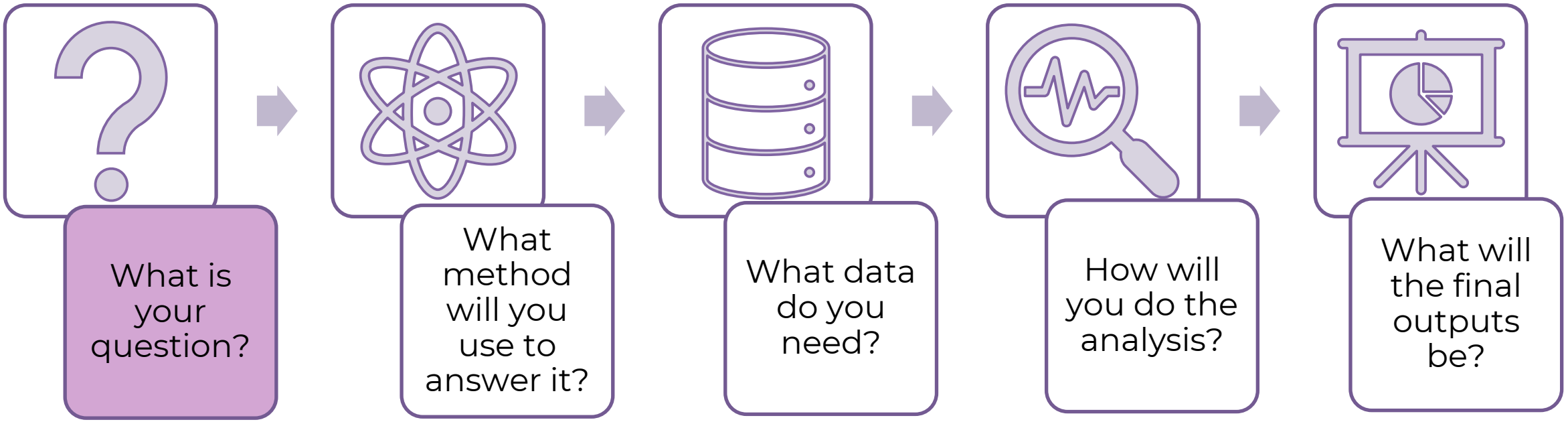


Lifecycle approach to credible research



- **Following best practices for transparent and reproducible research at all stages increases efficiency and credibility**
- **Today we will explain the whole lifecycle, but the practical part will focus on What data do you need? and What method will you use?**

Lifecycle approach to credible research



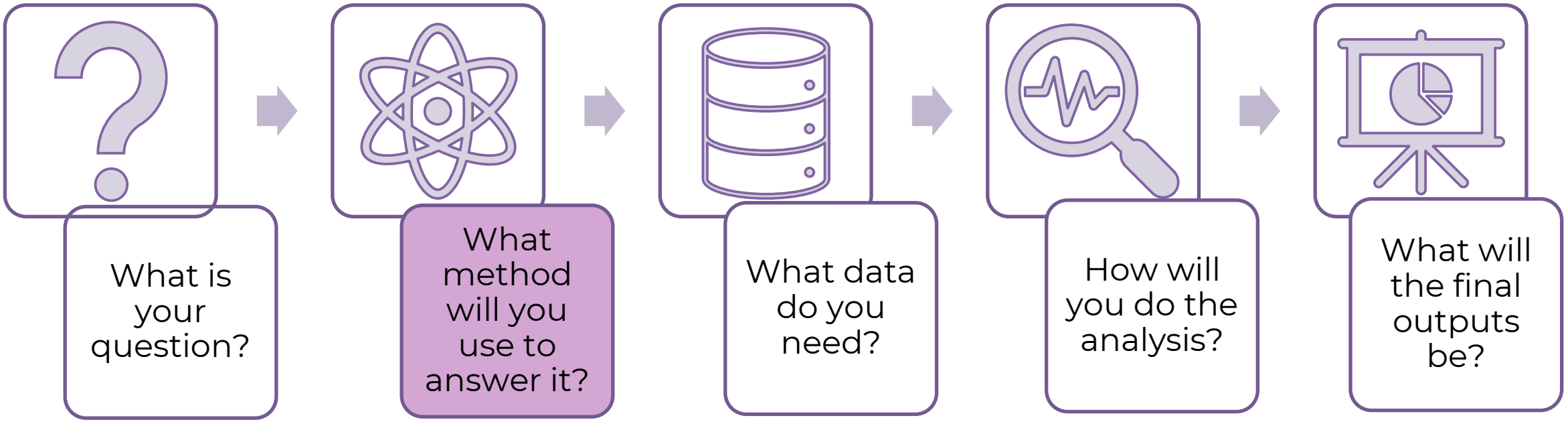
Following best practices for transparent and reproducible research at all stages increases efficiency and credibility

Question

We already started discussing this yesterday

- Define the problem
  - What needs to be solved and why is it important?
- Formulate your question
  - Is the question specific/ focused enough to be answerable?
- Justify your contribution
  - What's already known, and what does your question add?
- Ensure it's actionable
  - What do you want to change based on your findings?

Lifecycle approach to credible research



Following best practices for transparent and reproducible research at all stages increases efficiency and credibility

Method

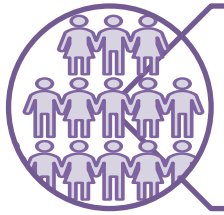
Is your question descriptive, or causal?

- What is the current evidence on best practices for teacher training programs?
- What is the impact of the new teacher training program on literacy rates for primary students?

If causal inference …

- Impact Evaluation in Practice (https://openknowledge.worldbank.org/handle/10986/25030)
- Development Research in Practice, Ch3: Establishing a Measurement Framework & Appendix C: Research Design for Impact Evaluation
- Experimental methods
  - https://dimewiki.worldbank.org/Experimental_Methods
  - https://dimewiki.worldbank.org/Randomized_Evaluations:_Principles_of_Study_Design
- Quasi-experimental methods
  - https://dimewiki.worldbank.org/Quasi-Experimental_Methods
  - https://dimewiki.worldbank.org/Difference-in-Differences
  - https://dimewiki.worldbank.org/Regression_Discontinuity

Method



What is the population of interest for your question? Will you need to construct a sample?

- https://dimewiki.worldbank.org/Sampling

If experimental, what elements will be randomized?

- Development Research in Practice Ch3, "Creating Research Design Variables by Randomization"

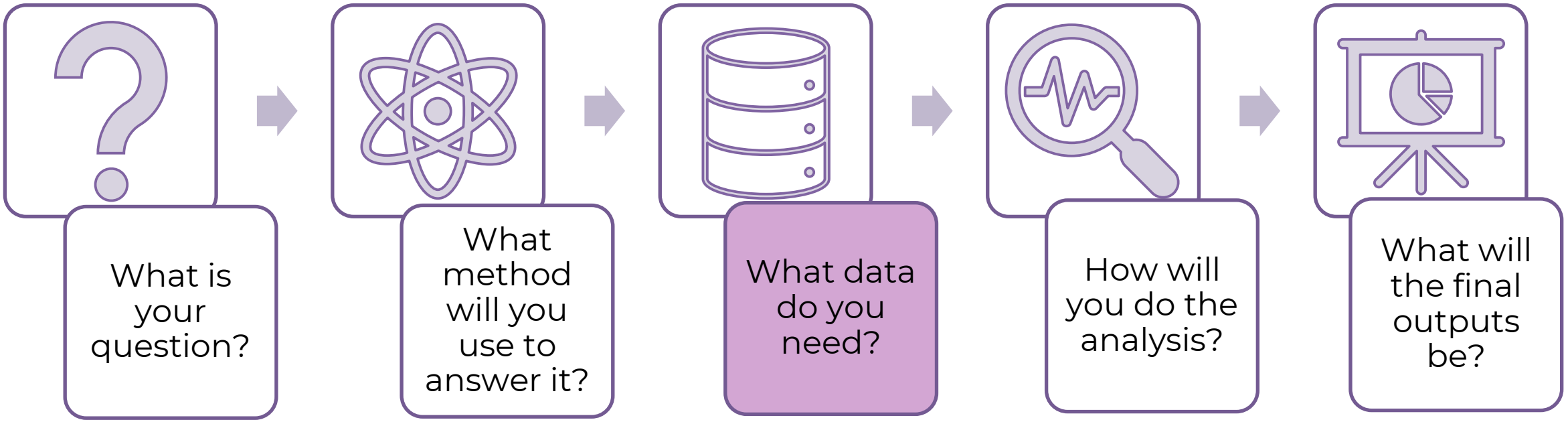
Are there potential ethical concerns?

- https://dimewiki.worldbank.org/Research_Ethics

What steps will you take to protect rights to privacy?

- https://dimewiki.worldbank.org/Protecting_Human_Research_Subjects

Lifecycle approach to credible research



Following best practices for transparent and reproducible research at all stages increases efficiency and credibility

What data would help answer this question?

Map out desired indicators and potential sources

- What data already exists?
- What data would need to be generated?
- Are there non-traditional data sources that might be useful?

Resources

- Development Research in Practice, Ch3 “Documenting Data Needs”
- Development Research in Practice, Ch3 “Translating Research Design to Data Needs”
- Development Research in Practice, Ch4 “Acquiring Development Data”
- Manage Successful Field Research lecture, “Establishing a Measurement Framework”

Understand the existing data ecosystem

Project monitoring data

National statistics office

Sector-specific surveys

Other large-scale surveys

Administrative data

... and more!

Resources:

- Manage Successful Field Research lecture, "Acquiring secondary data"
- https://dimewiki.worldbank.org/Secondary_Data_Sources
- Reproducible Research Fundamentals lecture, "Mastering administrative data"
- https://dimewiki.worldbank.org/Administrative_Data

Tools for high-quality surveys

Determine survey mode	• https://dimewiki.worldbank.org/Remote_Surveys • https://dimewiki.worldbank.org/Field_Surveys • https://dimewiki.worldbank.org/Phone_Surveys_(CATI)
Draft + pilot survey instrument	• Design & Pilot a Survey: https://osf.io/bpdgr • https://dimewiki.worldbank.org/Questionnaire_Design • https://dimewiki.worldbank.org/Survey_Pilot
Program + field test survey	• https://dimewiki.worldbank.org/Questionnaire_Programming • https://dimewiki.worldbank.org/Checklist:_Data-focused_Pilot
Train enumerators	• Engaging with data collectors: https://osf.io/bhsp9 • https://dimewiki.worldbank.org/Enumerator_Training
Monitor field work	• Engage with data providers: https://osf.io/y6pjn • https://dimewiki.worldbank.org/Field_Management • https://dimewiki.worldbank.org/Back_Checks
Monitor data quality	• https://dimewiki.worldbank.org/Monitoring_Data_Quality • https://dimewiki.worldbank.org/Data_Quality_Assurance_Plan • https://dimewiki.worldbank.org/High_Frequency_Checks

Development Research in Practice, Ch4 "Collect High-Quality Data Using Electronic Surveys"

Consider Innovative data sources

Satellite data

- https://dimewiki.worldbank.org/Remote_Sensing
- https://dimewiki.worldbank.org/Geo_Spatial_Data

Social media data

- https://dimewiki.worldbank.org/Innovative_Data_Sources

Mobile phone data

- https://dimewiki.worldbank.org/Telecom_Data

Crowdsourcing

- https://dimewiki.worldbank.org/Crowd-sourced_Data

Lecture: "Integrated Data Systems for Monitoring & Evaluation"

Repurposing Administrative Data

Why use administrative data to measure components of the production function?

- **Much government activity is recorded in some way as a standard part of public process.**
- **These records are valuable in themselves and can be used for government analytics when digitized.**
- **Yet government is already digitizing its records at unprecedented levels.**
- **And what these records show is that the public administration is fundamentally diverse.**

Question: Can you think of data that you are collecting as part of the process, but perhaps is not being utilized?

Evidence in Latin America and the Caribbean
A survey of the application of analytics to government administrative data in LAC

Targeted public officials who were domain experts in specific MIS or digital systems within government

Digital government agency coordinated data collection within a specific country

One country-level questionnaire focused on initiatives related to analytical skills. Six questionnaires per country explored specific MIS

All online enumeration

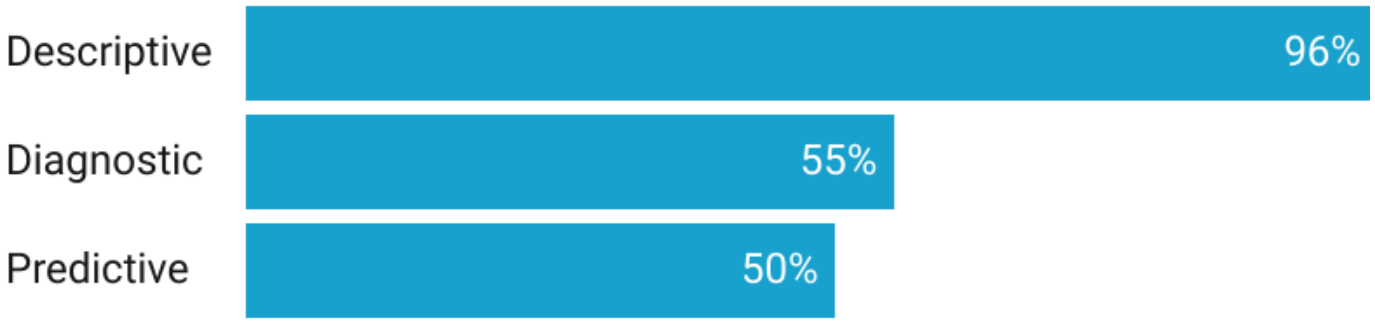
Data collection took place from November 2023 to March 2024

Data

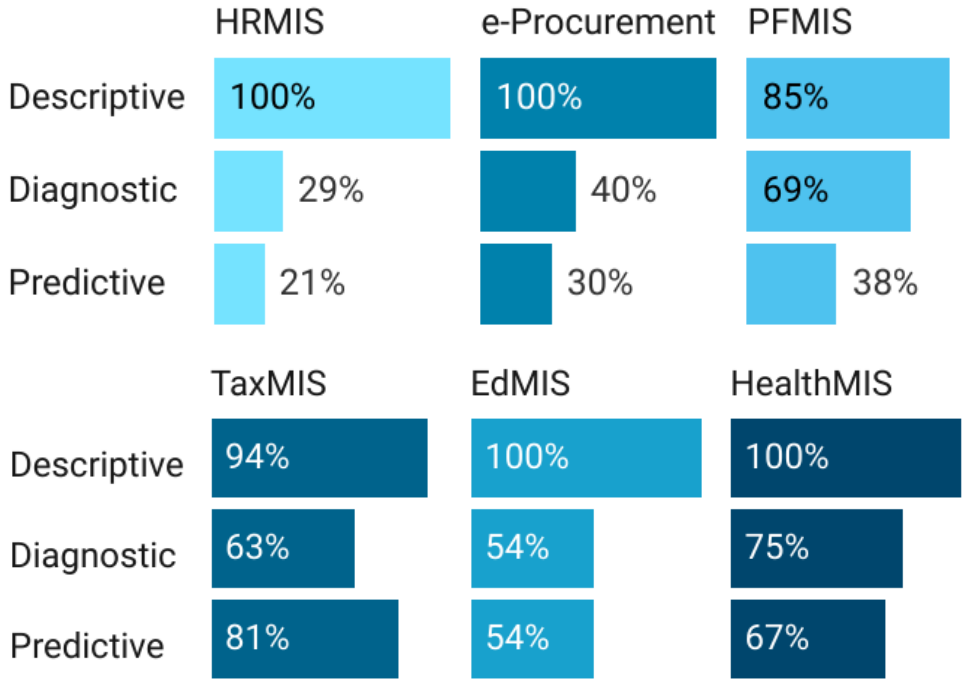
- 84 MIS-questionnaires from 19 different countries
  - EdMIS (ministry of education): 13 questionnaires
  - HealthMIS (ministry of health): 14 questionnaires
  - HRMIS (ministry of civil service): 14 questionnaires
  - PFMIS (ministry of finance/econ): 15 questionnaires
  - TaxMIS (tax authority): 16 questionnaires
  - ProcurementMIS (procurement agency): 12 questionnaires
- 16 Capabilities-questionnaires

Descriptive analytics is the prevailing type of analytics conducted with MIS data

Type of analytics performed (%)



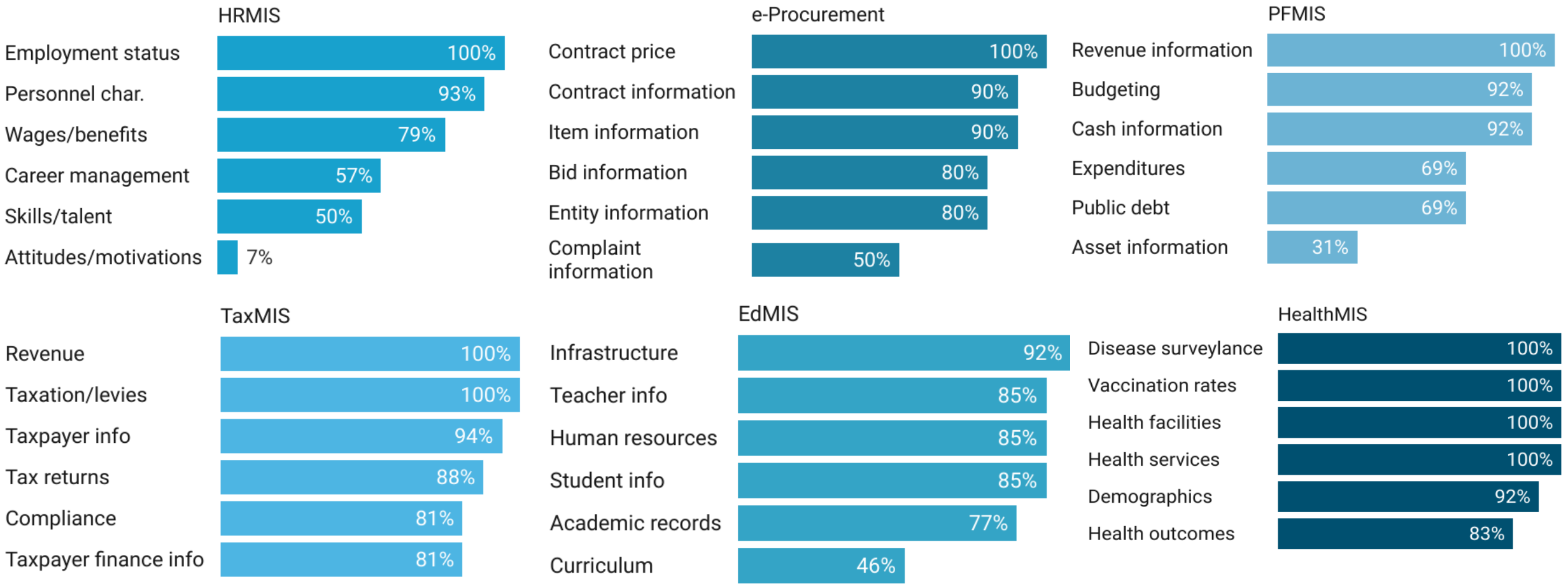
Type of analytics performed by MIS (%)



Source: LAC Gov Analytics Survey – MIS-specific questionnaires
N= 84 respondents from different countries/MIS

Most of the data available in each MIS is used for analytics

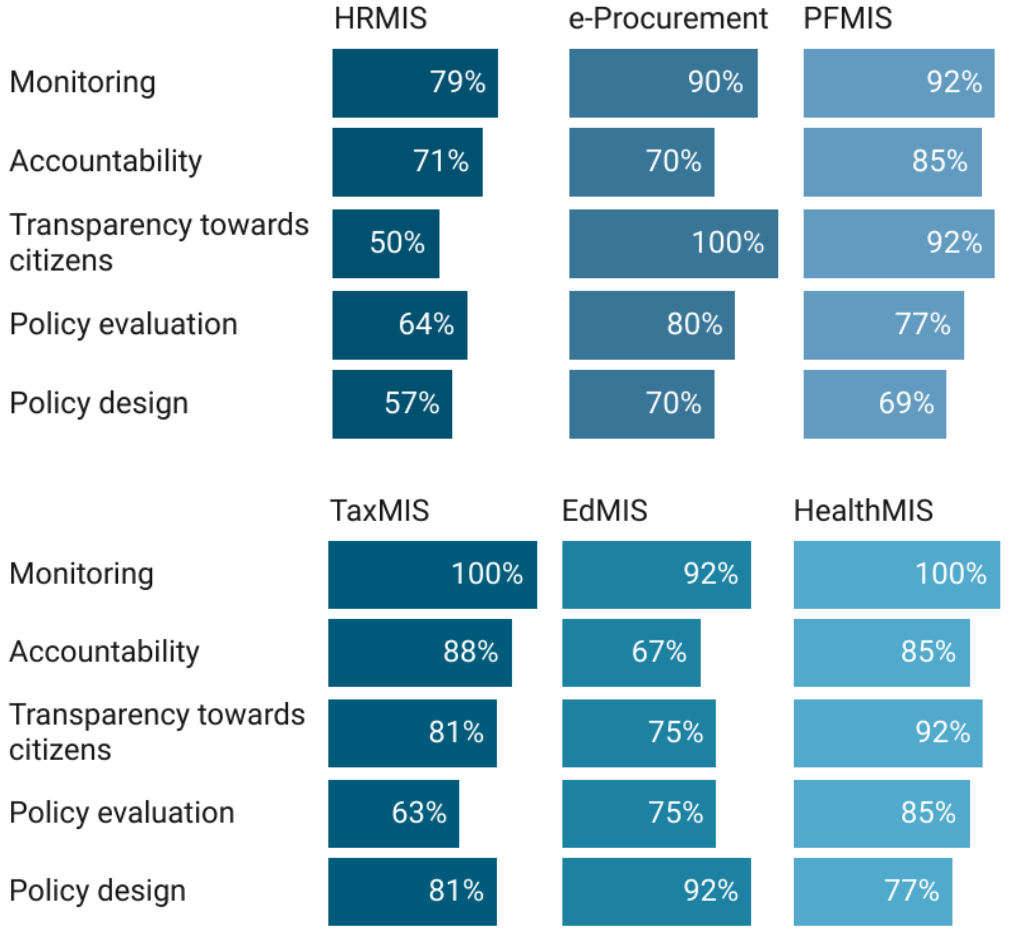
Data elements used for the analysis by MIS (%)



Decision-making facilitated by analytics is predominantly employed for monitoring and operational purposes

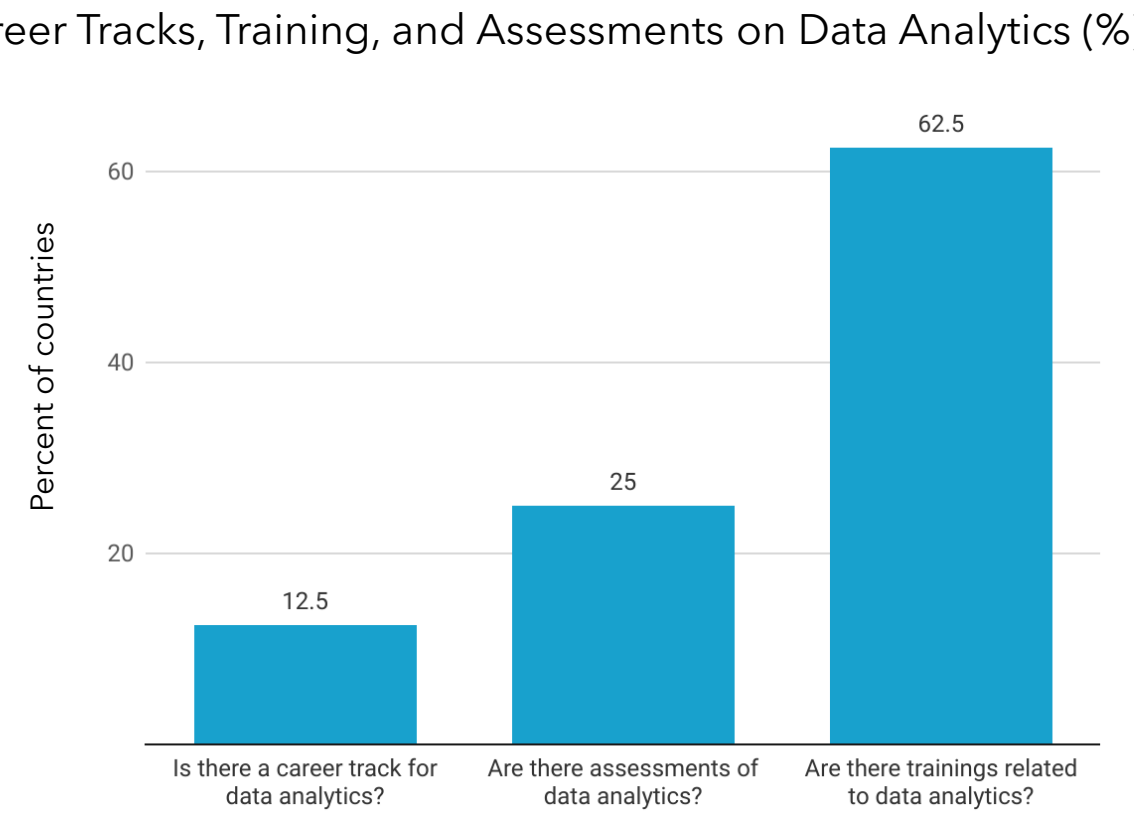
Applications of data analytics (%)

Applications of data analytics by MIS (%)



Source: LAC Gov Analytics Survey – MIS-specific questionnaires
N= 84 respondents from different countries/MIS

Capacity building remains a significant challenge for many countries

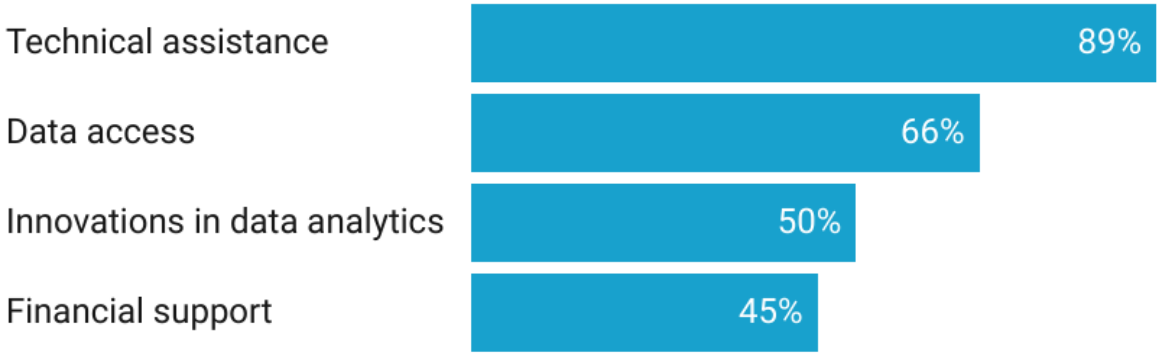


Source: LAC Gov Analytics Survey - Capabilities questionnaire
N= 16 countries

Funding for data analytics is limited and technical assistance is the main driver of collaboration with external analysts

Internal funding opportunities and Collaboration with externals (%)

Main drivers of collaboration with externals (%)



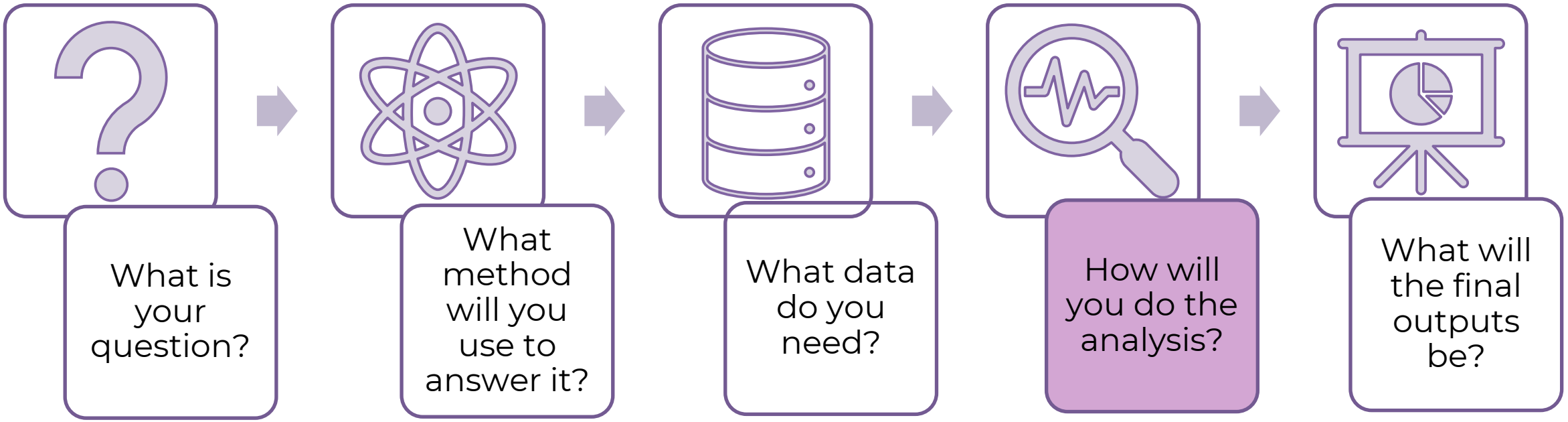
Source: LAC Gov Analytics Survey – MIS-specific questionnaires
N= 84 respondents from different countries/MIS

Conclusions

- **In LAC, governments use of MIS data predominantly focuses on descriptive analytics, with diagnostic and predictive analytics being relatively undeveloped and underutilized.**
- **MIS data are primarily employed for monitoring and accountability, with a secondary focus on policy evaluation and design, primarily leveraging analytics for operational and transactional purposes instead of developing advanced analytics to improve decision-making.**
- **While governments acknowledge the importance of analytical skills, they lack structured career tracks for data analysts and proficiency evaluations for public servants.**
- **While some countries offer analytical training programs, these programs often lack a coherent framework for connecting skills development to practical application.**

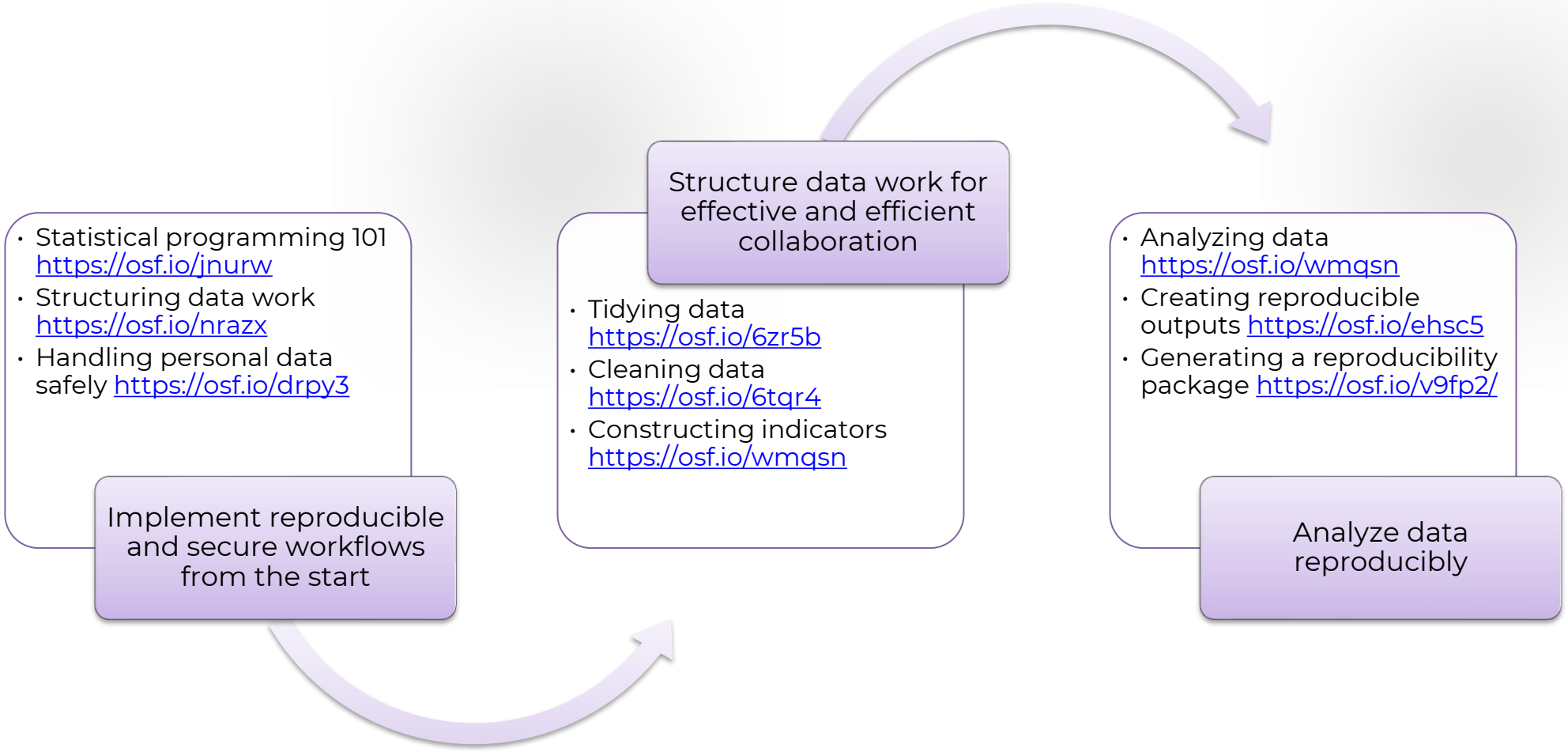
Question: Can you think how this is similar/different in Ghana?

Lifecycle approach to credible research



Following best practices for transparent and reproducible research at all stages increases efficiency and credibility

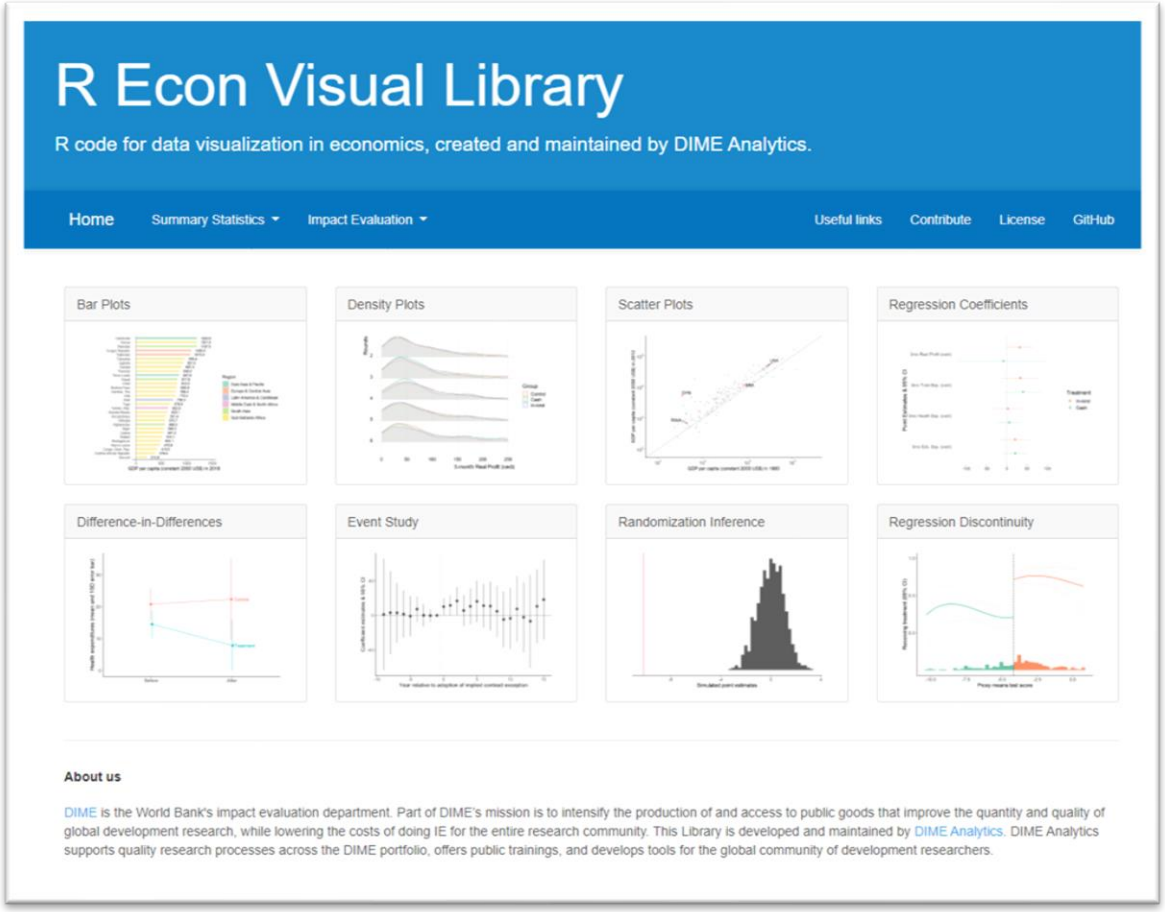
Workflows for efficient, reproducible analytics



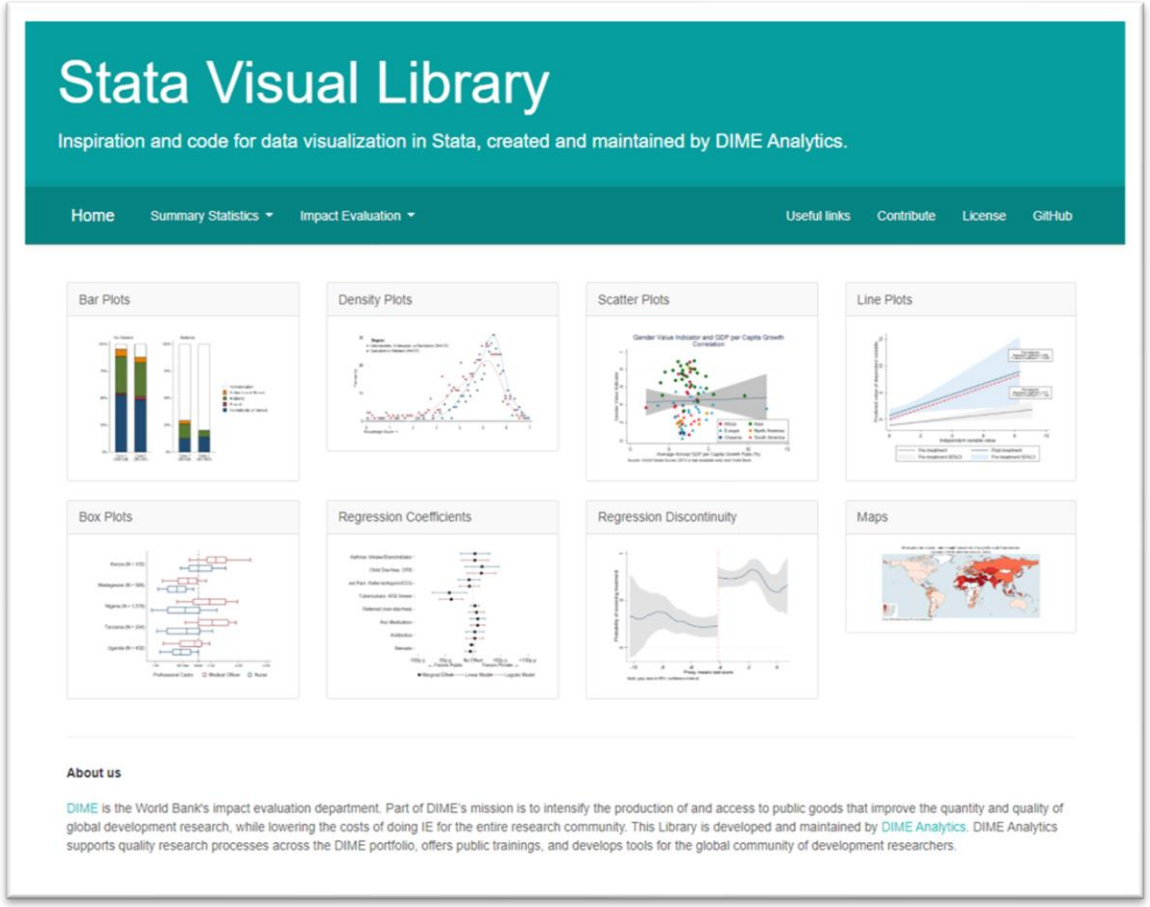
Development Research in Practice

Ch5: Cleaning and Processing Research Data & Ch6: Constructing and Analyzing Research Data

Libraries of reproducible visualizations

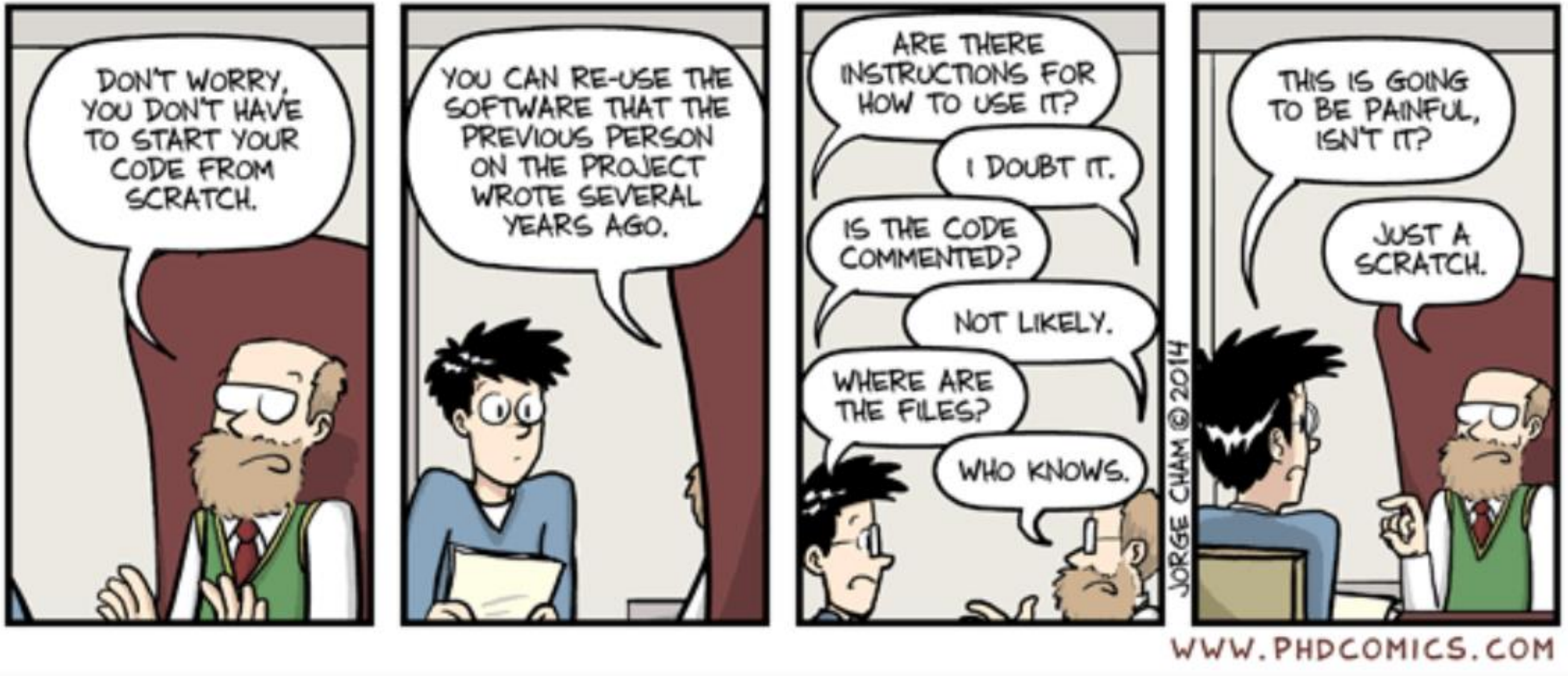


https://worldbank.github.io/r-econ-visual-library/



https://worldbank.github.io/stata-visual-library/

Reproducible practices benefit your colleagues and your future self too



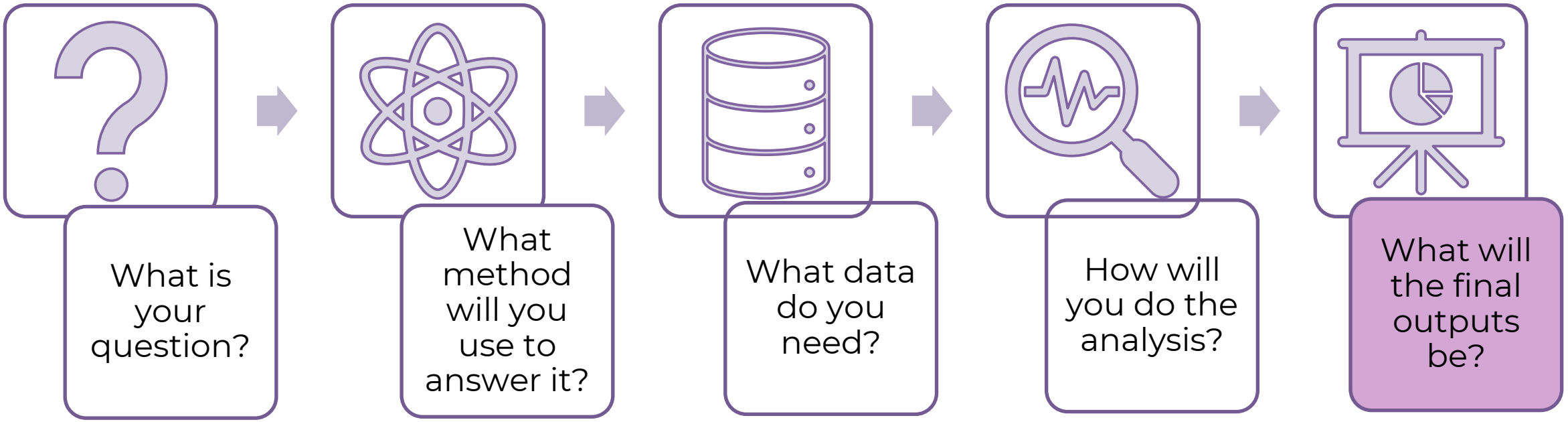
Reproducible analytics useful even for work that will never be published

Tasks can be easily handed over to colleagues

Others can understand and replicate your process or build upon your work

Planning for reproducibility from the outset improves quality and efficiency of your work

Lifecycle approach to credible research



Following best practices for transparent and reproducible research at all stages increases efficiency and credibility

Why should your findings be reproducible?

People who use evidence to make decisions should be able to scrutinize and recreate results easily. This requires:

- Clearly documented data
- Accessible and clearly documented analytical process
- Validation that all results can be reproduced

Gold standard for reproducibility

Computational reproducibility

A third party can reproduce the exact findings using the data, code, and documentation provided by the author

Open data

De-identified data published to a reputable archive with all relevant documentation

Open analytics

All steps to get from original data to final results are clearly documented and easy for others to read and understand

What's not reproducible?

Publishing only the final index, not the analysis that shows how it's created

Data analysis that's done by point and click (rather than code)

Many code files that lack a clear order or organization

Making lots of manual changes to tables or graphs

Tools for Reproducible Research

WB Reproducibility Initiative Resources

- Reproducible Research Repository
- Checklist for reproducibility package content
- How to Create a Reproducibility Package
- Handling User Written Programs Reproducibly
- Creating Reproducible Tables and Graphs
- How Git/GitHub Make Research More Reproducible

Textbooks on reproducible research

- Development Research in Practice: The DIME Analytics Data Handbook
- Transparent and Reproducible Social Science Research: How to Do Open Science

Reproducible Research Standards

- American Economics Association Data and Code Availability Policy
- Guidance and standards from the data editors of top social science journals

WB Blogs on Reproducible Research

- How to Make Sure Your Research Paper is Reproducible?
- What development economists talk about when they talk about reproducibility
- Quality of computer code and reproducibility of economic research
- Save your environment: The (often) overlooked problem of research reproducibility in economics

Exercise

Understand Data Challenges

5 minutes: Work individually to assess and evaluate the key challenges to using the data sets you have identified in the session yesterday.

5 minutes: Small group discussion

- Present your individual challenges to others in your group
- Peer feedback: Are these similar to those you face? What other issues might the analyst face?

10 minutes: Plenary discussion

- Groups nominate one member to present their approach to the plenary session.
- Discuss different approaches and provide feedback

Lifecycle approach to credible research

Day	Key Focus	Lifecycle Question	Session 1 (Theory & Discussion)	Session 2 (Code / Application)
Day 1	Introduction to Government Analytics & Framing	What is your question?	Introduction to Government Analytics	**Introduction to R**
Day 2	Data Wrangling & Repurposing Existing Data	What data do you need? What method will you use?	Identifying and Repurposing Data, Discussion on Data Sources	**Data Wrangling in R**
Day 3	Descriptive Analysis & Tabulation for Reports	How will you do the analysis?	Discussion on Annual Reports	**Summarizing & Tabulating Data in R**
Day 4	**Data Visualization & Final Outputs**	**What will the final outputs be?**	**Finalizing Outputs, Best Practices for Visualization & Reporting**	**Creating Visualizations & Exporting Outputs in R**
Day 5	**Bring Your Own Data**	**Q&A and Problem-Solving**	**Open Q&A, Personalized Feedback on Participant Data**	**Practical Application and Hands-on Guidance**

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

WORLD BANK GROUP
INSTITUTE FOR ECONOMIC DEVELOPMENT

Reproducible Research Repository

Thank you!