Content from different reference book suggested by Dr. Anita Morey Mam and mentioned in CDA Syllabus file, Presentations, Notes provided in 2023

NOTES ON BASIS OF CDA PPT REPRODUCIBLE RESEARCH 51 SLIDES

Reproducible Research

- Reproducible research involves publishing data analyses and scientific claims along with their data and software code, allowing others to verify findings and build upon them.
- The importance of reproducibility has grown due to complex data analyses with larger datasets and sophisticated computations.
- Reproducibility shifts focus from superficial details to the actual content of a data analysis, enhancing its usefulness.
- It makes analyses more valuable to others by providing access to the data and code used for the analysis.

Computational Reproducibility

- Replicating studies with new independent data is costly and methodologically challenging.
- Computational reproducibility, often called "reproducible research," is suggested to improve the assessment of scientific results' validity and rigor.
- Research is computationally reproducible when others can replicate study results using original data, code, and documentation.

Advantages of Reproducibility

- This approach mirrors the benefits of replicating studies with new data but minimizes the cost of collecting new data.
- While replicating studies remains the gold standard, reproducibility is considered a minimum standard for all scientists.

Principles of Reproducibility

- Researchers can adopt a simple three-part framework to make their current research more reproducible.
- These principles apply to researchers across various sub-disciplines.

Benefits of Reproducible Research

1. Researchers benefit from reproducible research by:
   - Ensuring consistent results upon multiple analyses.
   - Facilitating explanations of work to collaborators, supervisors, and reviewers.
   - Enabling quick and efficient supplementary analyses by collaborators.

2. Reproducible research enables easy modification of analyses and figures:
   - Responding to requests from supervisors, collaborators, and reviewers.

- Saving significant time by updating figures through code changes.

3. Reproducible research simplifies reconfiguration of previous research tasks:
   - Simplifying subsequent projects requiring similar tasks.
   - Enhancing efficiency in iterative research processes.

4. Conducting reproducible research demonstrates rigor, trustworthiness, and transparency:
   - Increases the quality and speed of peer review.
   - Reviewers can directly access analytical processes in manuscripts.
   - Reviewers can cross-check code and methods, catching errors during peer review and reducing post-publication corrections.

Reproducible research benefits researchers, enhances collaboration, and ensures the reliability of scientific findings.

Why Do Reproducible Research?

Protects Against Accusations of Research Misconduct:
- Researchers who openly share code and data are less likely to be accused of research misconduct due to fraudulent practices.
- Fraudulent code and data would be evident to the research community.

Increases Paper Citation Rates:
- Reproducible research leads to higher citation rates for papers.
- Citations extend to code and data in addition to publications.
- Enhances the impact of research by making data and methods accessible.

Benefits the Research Community:
1. Facilitates Learning from Others' Work:
   - Allows researchers to access code and data, aiding in learning complex techniques.
   - Beginners can benefit from experienced researchers' code to perform rigorous analyses.

2. Saves Time and Effort for Experienced Researchers:
   - Experienced researchers can modify existing code more efficiently than writing from scratch.
   - Sharing code accelerates similar analyses for seasoned researchers.

3. Enables Understanding and Reproduction of Work:
   - Others can perform follow-up studies to strengthen evidence.
   - Promotes compatibility and consistency among similar studies.
   - Supports meta-analyses for generalizing and contextualizing findings.

4. Helps Identify and Correct Mistakes:
   - Open access to code and data encourages critical analysis.
   - Co-authors, reviewers, and other scientists can identify and rectify mistakes.

- Prevents mistakes from accumulating over time.

Barriers to Reproducible Research:
- Complexity:
  - Specialized knowledge and tools required for certain analyses.
  - High-performance computing clusters with various programming languages.
  - Proprietary software like SAS or ArcGIS with expensive licenses.

- Technological Change:
  - Rapidly evolving technologies and tools complicate reproducibility.
  - New tools may not be widely available or understood.

- Human Error:
  - Mistakes can occur in scientific research.
  - Open access allows collaborators, reviewers, and others to catch errors early.

- Intellectual Property Concerns:
  - Fear of compromising intellectual property rights may hinder open sharing.
  - Protocols and norms can address these concerns and encourage openness.

Addressing Barriers:
- Complexity:
  - Citations and detailed annotations can reduce knowledge barriers.
  - Thoroughly annotated code and extensive documentation can enhance accessibility.

- Technological Change:
  - Researchers can actively work to bridge the technology gap by providing resources and tutorials.

- Human Error:
  - Open access and collaborative review help identify and correct mistakes.

- Intellectual Property Concerns:
  - Proper protocols and norms can balance openness with intellectual property rights.

Reproducible research benefits researchers, the scientific community, and the quality and reliability of scientific findings. Overcoming barriers through accessible resources and collaborative efforts is essential for fostering reproducibility.

Barriers to Reproducible Research

Technological Change:
- Hardware and software used for data analysis evolve rapidly.
- Research conducted with outdated tools becomes less reproducible over time.
- For instance, research from previous decades may require entirely new tools for replication today.

- Even minor updates in software can impact the reproducibility of a project.

Mitigation Through Established Tools:
- Use well-documented versions of software tools.
- Careful documentation of software versions is essential.

Human Error:
- Researchers make mistakes in documenting procedures and analyses.
- Incomplete descriptions and documentation can lead to inaccuracies.
- Critical data might be omitted initially but become vital later.

Documentation as a Safeguard:
- Detailed documentation guards against errors and incomplete analyses.
- Record data collection details, decisions, and labeling conventions.
- Data wrangling errors can be mitigated through multiple data backups and thorough documentation.

Intellectual Property Rights:
- Researchers may hesitate to share data and code due to misuse or unethical use.
- Sharing data without proper citation can lead to misinterpretations.
- Researchers might withhold data to protect their future analyses.

Balancing Openness and Protection:
- Emerging tools allow sharing while preserving control and credit.
- Open data sharing is a contentious aspect of reproducible research.

Framework for Reproducible Research

Before Data Analysis: Data Storage and Organization:
- Plan for reproducibility from the start with effective data management.
- Data should be backed up at every stage and stored in multiple locations.
- Backups should include raw and clean analysis-ready data.
- Keep paper copies of data sheets paired with digital datasets.
- Use portable, non-proprietary formats for digital data.

Addressing Technological Change:
- Use well-documented versions of software tools.
- Careful documentation of software versions is essential.

Human Error:
- Thorough documentation of processes guards against errors and incomplete analyses.

Intellectual Property Concerns:
- Emerging tools offer data sharing while safeguarding ownership and credit.

Framework for Conducting Reproducible Research

During Analysis: Best Coding Practices:

- Tidy Data Format: Transform data into a "tidy" format for cleaning and standardization. Tidy data are organized in long format, with consistent structure and informative headers.

- Metadata: Store metadata explaining data cleaning and variable meanings along with the data. Metadata enhances data interpretability and should include data collection details, variable meanings, and coding explanations.

- Organized File Structure: Organize files with informative names and directories. Consistent naming protocols for files and directories enhance searchability and accessibility.

- Version Control: Use version control to document project history and changes. This aids in tracking updates and provides snapshots of data and code.

During Analysis: Coding Practices:

- Use coding scripts for data wrangling and analysis for documentation and repeatability.
- Thoroughly annotate analytical code with comments for clarity and metadata.
- Follow consistent coding styles for readability.
- Automate repetitive tasks using functions and loops.
- Use parameters at the beginning of a script to allow easy adaptation to new data.

Mitigating Technological Change:
- Use established software versions and document dependencies.
- Consider using software containers for reproducibility.

After Analysis: Finalizing and Sharing Results:

- Share input data, scripts, program versions, parameters, and intermediate results publicly.
- Create figures and tables directly from code for dynamic, reproducible documents.
- Use tools like LaTeX for creating dynamic presentations.

Sharing and Archiving Results:

- Automation with Make: Use GNU Make to automate and coordinate command-line processes, making data wrangling, analysis, and document creation a streamlined process.

- Sharing Research: Currently, data and code for replicating research are often found in journal article supplementary materials. Some journals are experimenting with embedding data and code in articles. Authors can also post preprints on preprint servers or postprints on postprint servers to increase access to publications.

- Use of Data Repositories: Data archiving in online repositories is becoming more popular due to technology improvements, large-scale data sets, and encouragement from publishers and funding organizations. Repositories collect and store data for analysis, sharing, and reporting. Researchers can find appropriate repositories through journal recommendations.

- Research Compendia: Archiving data, code, software, and research products together forms a research compendium. These compendia provide a standardized way to organize and share research materials, making it easier for other researchers to reproduce and extend the research.

Three-Step Framework and Check-list Guide for Reproducible Research: This section provides a concise summary of the three-step framework for conducting reproducible research and emphasizes the importance of adopting these practices for improved research transparency and reliability.