Introduction to Research
Data Management
Data Processing
http://hdl.handle.net/1969.1/164594
Introduction

Common data manipulations and tools for processing data.

Process your data in a manner that allows you to roll back changes if you make a mistake.
Preparing data for analysis

Processing data by:

• Subsetting
• Merging
• Manipulation
Data transformation

• Normalizing data collected by multiple people and/or instruments.
• Converting data to different units.
• Converting raw data into meaningful values.
De-identification

Removing or obscuring any personally identifiable information from individual records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them.

- Anonymization
- Aggregation
- Masking
- Shuffling
- Perturbation
De-identification tips

● Remove direct identifiers.

● Use pseudonyms or replacements.

● Reduce the precision and detail through aggregation.

● Generalize meaning of detailed text variables.

● Restrict upper or lower ranges to hide outliers.

● Use digital manipulation of audio and image files to remove personal identifiers.

● Avoid over-anonymization and exercise additional care when working as a team.

● Keep master log of all replacements, aggregations, and removals.
Statistics for analysis

- Descriptive statistics are traditionally applied to observational data.
- Conventional statistics are often used to understand experimental data.
Software for data manipulation and analysis

**Microsoft Excel and Google Sheets**: Data entry, manipulation, and graphing.

**OpenRefine**: Working with and cleaning messy data.

**NVIVO**: Powerful qualitative data analysis (QDA).

**SAS**: Advanced analytics, multivariate analyses, and predictive analytics.

**SPSS**: Logical batched and non-batched statistical analysis, data mining and text analytics.

**STATA**: General-purpose statistical analysis, graphics, simulations, regression, and custom programming.

**Matlab**: Numerical computing, matrix manipulations, plotting, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages.

**R**: Statistical computing and graphics, and popular programming language for doing stats.

**Python (NumPy, SciPy, Pandas)**: Object oriented programming language with several data analysis libraries.
Documentation reminder

Workflows allow you to give a precise and reproducible description of your procedure.

Show and describe:

• Inputs
• Outputs
• Transformations
Informal workflows

● Well-described version history.
● Commented scripts.
● Flow charts.
## Version history

- Showing changes you’ve committed over time.
Commented scripts

- Explaining the input, output and function of code.

```python
def trim(docstring):
    if not docstring:
        return ''

    # Convert tabs to spaces (following the normal Python rules)
    # and split into a list of lines:
    lines = docstring.expandtabs().splitlines()

    # Determine minimum indentation (first line doesn't count):
    indent = sys.maxint
    for line in lines[1:]:
        stripped = line.lstrip()
        if stripped:
            indent = min(indent, len(line) - len(stripped))

    # Remove indentation (first line is special):
    trimmed = [lines[0].strip()]
    if indent < sys.maxint:
        trimmed += [line[indent:].strip() for line in lines[1:]]

    return '
'.join(trimmed)
```
Flow charts

- **Inputs or outputs**
  Include data, metadata, or visualizations.

- **Analytical processes**
  Include operations that change or manipulate data in some way.

- **Subroutines**
  Predefined processes that specify a fixed multi-step process.

- **Decisions**
  Specify conditions that determine the next step in the process.
Data import into R

QC and data cleaning

Clean combined data

Analysis: mean and SD

Summary Stats

Graph production

Graph

Raw dataset 1

Raw dataset 2

Data in R format

Clean combined data

Summary Stats

Graph production

Analysis: mean and SD

Clean combined data

Analysis: mean and SD

Summary Stats

Graph production

Graph
Formal workflows

- **Kepler**: Designed to help scientists, analysts, and computer programmers create, execute, and share models and analyses.

- **Taverna**: A suite of tools used to design and execute scientific workflows and aid in silico experimentation.

- **VisTrails**: Scientific workflow and provenance management system that provides support for simulations, data exploration and visualization.
Conclusion

• Discussed preparing data for analysis and documenting data processing using workflows.
References and resources

● DataOne. "Lesson 09: Analysis and workflows" [module](https://www.dataone.org/education-modules)

● Kepler [website](https://kepler-project.org)

● Taverna [website](http://www.taverna.org.uk)

● VisTrails [website](https://www.vistrails.org/index.php/Main_Page)