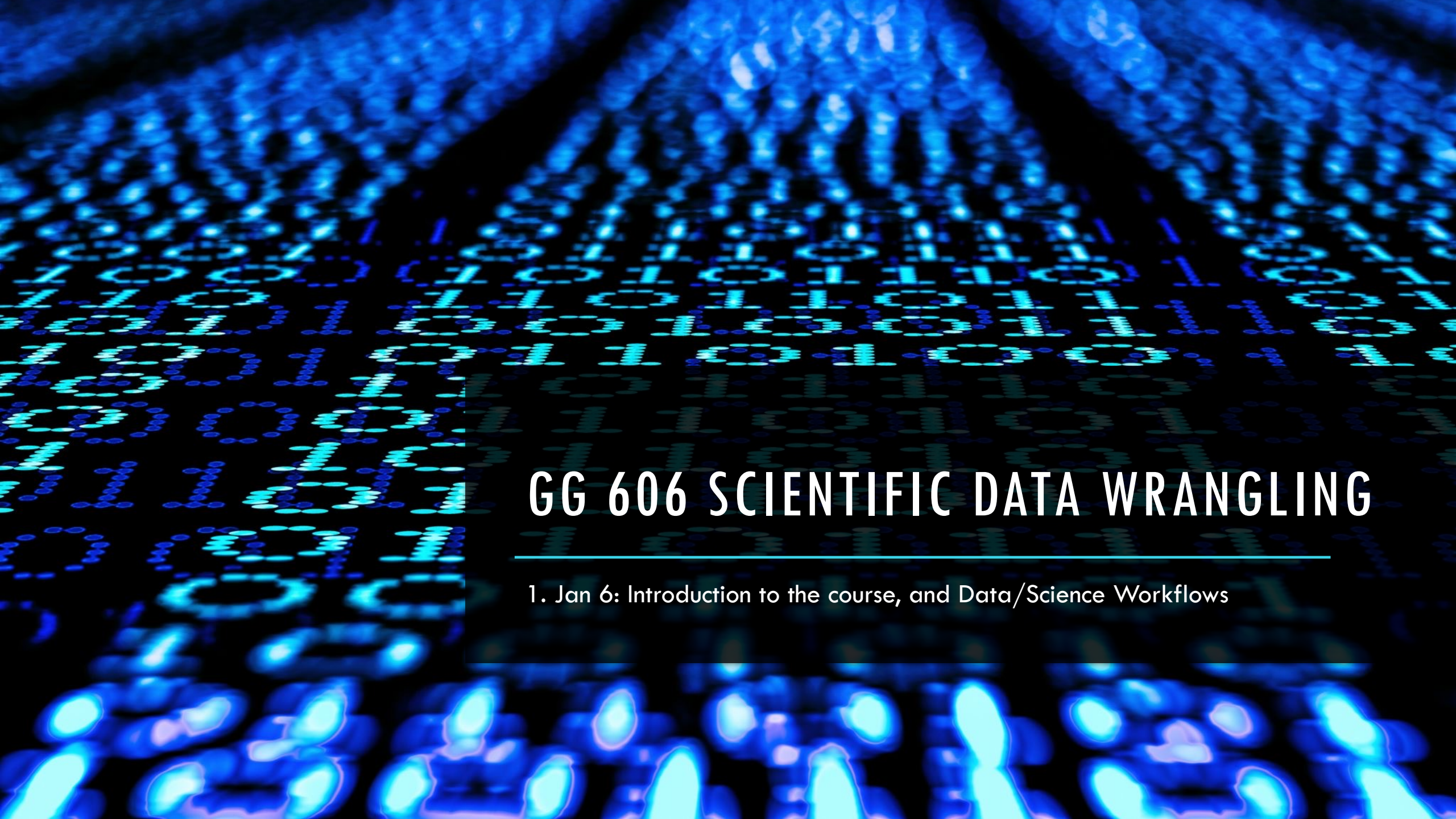


GG 606 SCIENTIFIC DATA WRANGLING

1. Jan 6: Introduction to the course, and Data/Science Workflows

INTRODUCTION TO THE COURSE

- Science → Graduate research, environmental/geographic research
- Data → Data/observations describing environmental features or processes
- Wrangling → making raw data useful for analysis, modelling, and reporting

wrangling

APPROACH TO LEARNING IN THIS CLASS

- learning by doing – code reviews, interactive data analysis, working through examples step-by-step in class

- case studies – review case studies demoing the good, the bad, and the ugly of data wrangling

- discuss! – bring your ideas, articles, blog posts, code snippets, tweets, youtube links, etc.. there is a rich discourse around current tools and workflows for data processing and analysis, this class needs to be collaborative and interactive.. participate fully!

INTRODUCTION TO THE COURSE

Science → Graduate research, environmental/geographic research

Data Wrangling → domain-independent data science workflows but with special attention to above

COURSE COMPONENTS

1. Review readings w/ very brief lecture/discussions
2. Exercises and code-reviews in class
3. Case studies - code and output reviews in class
4. Assignments
5. Term Projects

ABOUT THE INSTRUCTOR

Dr. Colin Robertson

- Bachelor's, Advanced Diploma, Master's, PhD Degrees in Geography/GIS/Spatial Analysis
- Technical research domains:
  - Spatial-temporal analysis
  - Disease forecasting surveillance methods
  - VGI / citizen science / community-based monitoring
- Applied research domains:
  - Forest ecology
  - Public health
  - Environemntal change / monitoring

I have been using R for a long time and modern ways of working in R, as we will do here, is generally much much better..

[R-sig-Geo] 3D KDE

Colin Robertson colinr23 at gmail.com
Thu May 18 20:04:06 CEST 2006

- Previous message: [R-sig-Geo] GSTAT fit.variogram method
- Next message: [R-sig-Geo] GSTAT
- **Messages sorted by:** [date] [thread] [subject] [author]

```
An embedded and charset-unspecified text was scrubbed...
Name: not available
URL: <https://stat.ethz.ch/pipermail/r-sig-geo/attachment
```

- Previous message: [R-sig-Geo] GSTAT fit.variogram method
- Next message: [R-sig-Geo] GSTAT
- **Messages sorted by:** [date] [thread] [subject] [author]

COURSE RESOURCES

1. Course website is central resource for materials – http://colinr23.github.io/gg606
2. Zoom for remote delivery of classes
3. Readings – free online textbooks (links on course website)
4. Your own computer with R/R-Studio installed

WHY USE R FOR SCIENTIFIC DATA WRANGLING

https://data-flair.training/

EXAMPLES

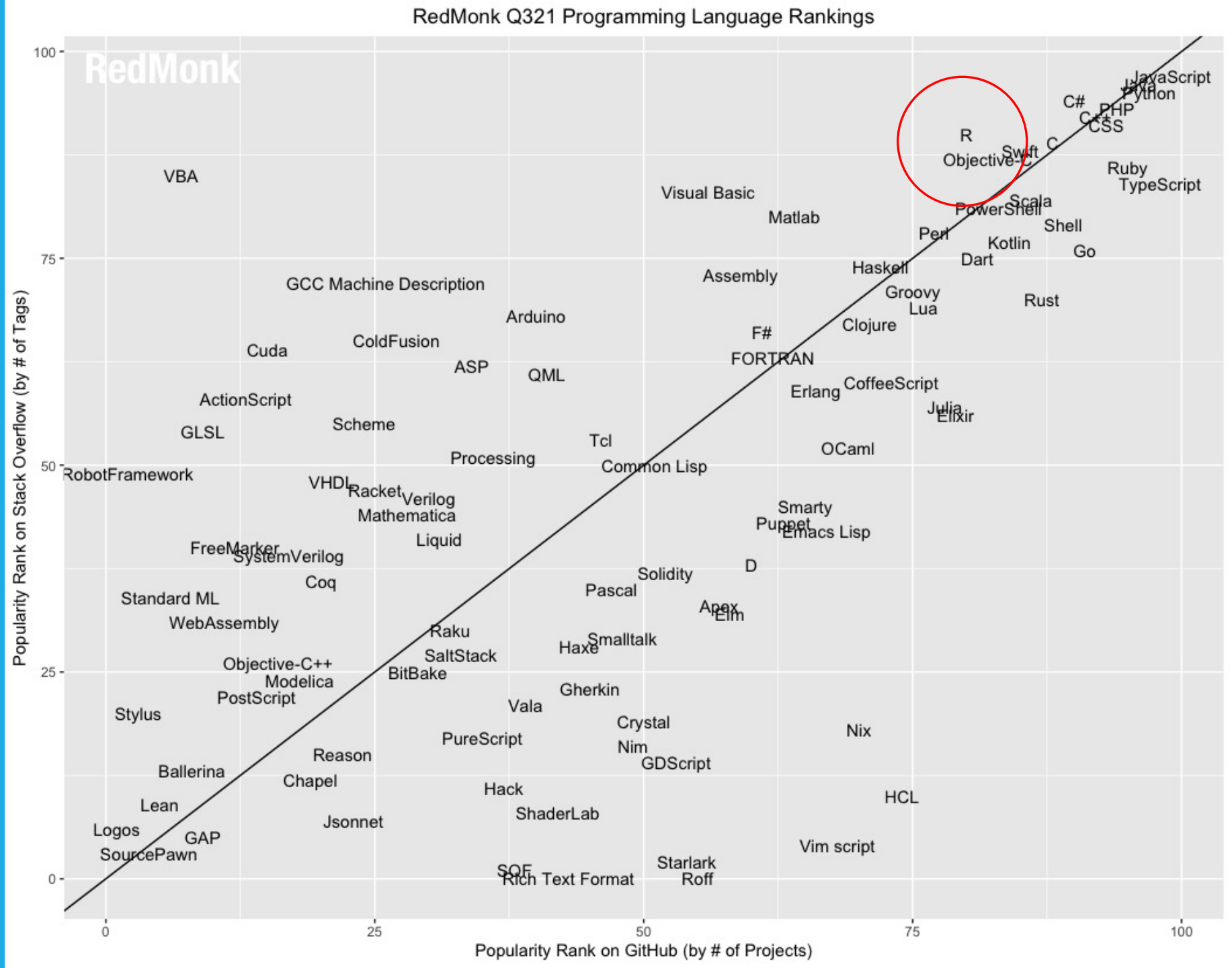
Methods in Ecology and Evolution

- https://besjournals.onlinelibrary.wiley.com/action/doSearch?AllField=r+package&SeriesKey=2041210x

Environmental Modelling and Software

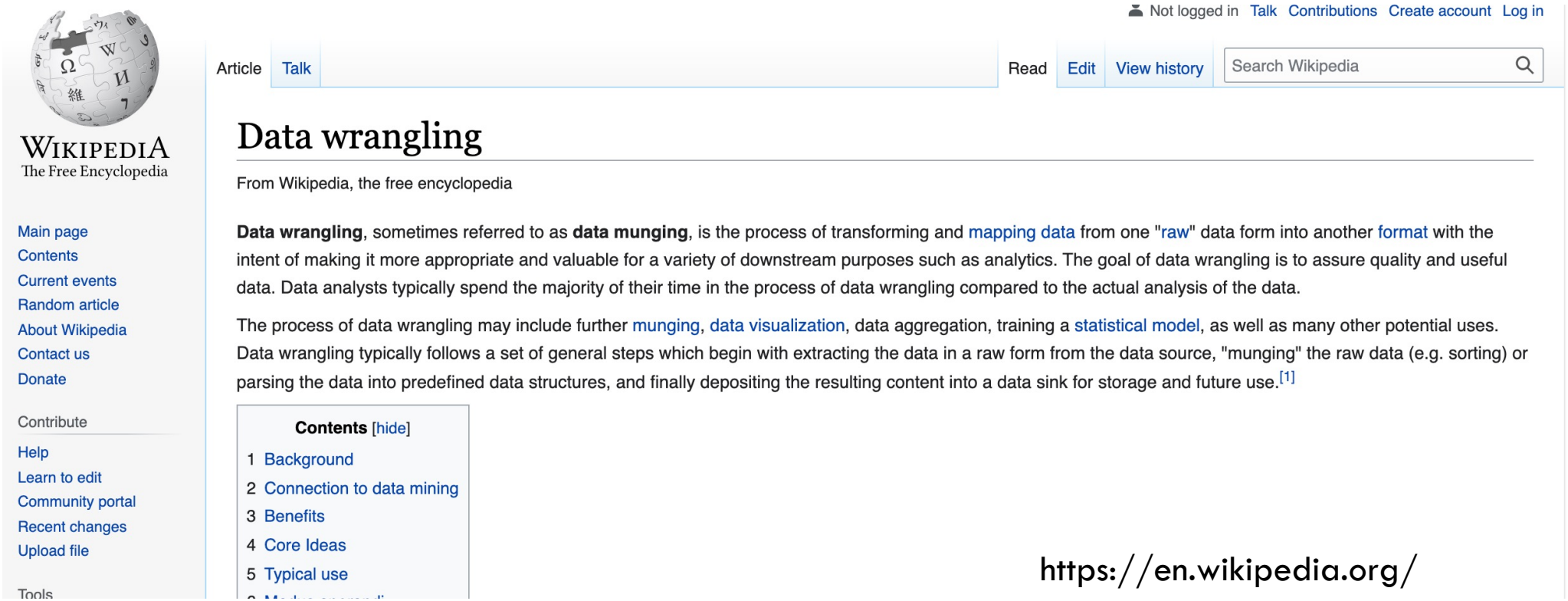
- https://www.sciencedirect.com/search?qs=R%20package&pub=Environmental%20Modelling%20%26%20Software&cid=271872

POPULAR AND INCREASING



https://redmonk.com/

DATA WRANGLING



WIKIPEDIA
The Free Encyclopedia

Not logged in Talk Contributions Create account Log in

Article Talk | Read Edit View history | Search Wikipedia

Data wrangling

From Wikipedia, the free encyclopedia

Data wrangling, sometimes referred to as **data munging**, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. The goal of data wrangling is to assure quality and useful data. Data analysts typically spend the majority of their time in the process of data wrangling compared to the actual analysis of the data.

The process of data wrangling may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses. Data wrangling typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.[1]

Contents [hide]

1 Background
2 Connection to data mining
3 Benefits
4 Core Ideas
5 Typical use

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

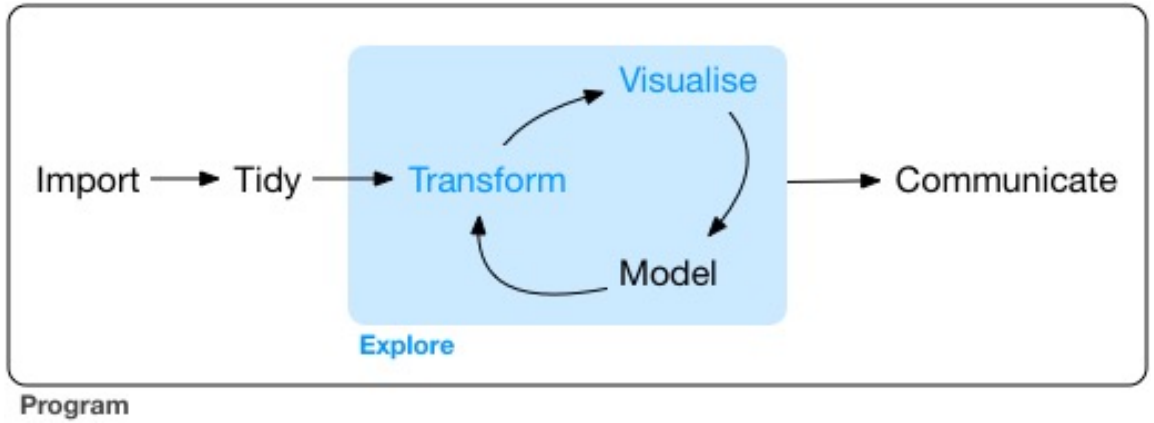
Tools

https://en.wikipedia.org/

APPROACH TO DATA WRANGLING

Data rarely arrives in a useable form, must always be filtered, transformed, examined, etc. before it can generate insights

Data exploration is the art of looking at your data, rapidly generating hypotheses, quickly testing them, then repeating again and again and again



CRAN – R PACKAGE ARCHIVE

- Resource for installing r packages contributed by the community
- There are stringent checks to ensure they are compatible and will install
- no checks on accuracy of actual code/calculations
- There are 1000s of packages

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

Available CRAN Packages By Name

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A3	Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
aaSEA	Amino Acid Substitution Effect Analyser
AATtools	Reliability and Scoring Routines for the Approach-Avoidance Task
aba	Automated Biomarker Analysis
ABACUS	Apps Based Activities for Communicating and Understanding Statistics
abbreviate	Readable String Abbreviation
abbyyR	Access to Abbyy Optical Character Recognition (OCR) API
abc	Tools for Approximate Bayesian Computation (ABC)
abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)
ABC.RAP	Array Based CpG Region Analysis Pipeline
abcADM	Fit Accumulated Damage Models and Estimate Reliability using ABC
ABCanalysis	Computed ABC Analysis
abcdeFBA	ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
ABCoptim	Implementation of Artificial Bee Colony (ABC) Optimization
ABCp2	Approximate Bayesian Computational Model for Estimating P2
abcrf	Approximate Bayesian Computation via Random Forests
abcrlda	Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis
abctools	Tools for ABC Analyses
abd	The Analysis of Biological Data

highly idealized

ENVIRONMENTAL DATA ANALYTICS WORKFLOW

go to the field, collect data

analyze data, gain insights

make the world a better place

what other things can happen or steps are missing from this?

DEMO – LETS ALL ENSURE WE HAVE R/R-STUDIO

- Open up R-Studion or install it if you have never used it
- Work through the examples in the Transform section of the readings, answer the questions in 5.2.4 Exercises

5.2.4 Exercises

1. Find all flights that
    1. Had an arrival delay of two or more hours
    2. Flew to Houston (`IAH` or `HOU`)
    3. Were operated by United, American, or Delta
    4. Departed in summer (July, August, and September)
    5. Arrived more than two hours late, but didn't leave late
    6. Were delayed by at least an hour, but made up over 30 minutes in flight
    7. Departed between midnight and 6am (inclusive)
2. Another useful dplyr filtering helper is `between()`. What does it do? Can you use it to simplify the code needed to answer the previous challenges?
3. How many flights have a missing `dep_time`? What other variables are missing? What might these rows represent?
4. Why is `NA ^ 0` not missing? Why is `NA | TRUE` not missing? Why is `FALSE & NA` not missing? Can you figure out the general rule? (`NA * 0` is a tricky counterexample!)

WORKFLOW: PROJECTS

- Store code not environments
  - much better to replicate your analysis when you go away and come back than saving workspaces with objects
- Use R project files to keep all the files associated with a project together — input data, R scripts, analytical results, figures
- Understand working directories (it is just a folder)

GG 606 SCIENTIFIC DATA WRANGLING

1. Jan 6: Introduction to the course, and Data/Science Workflows