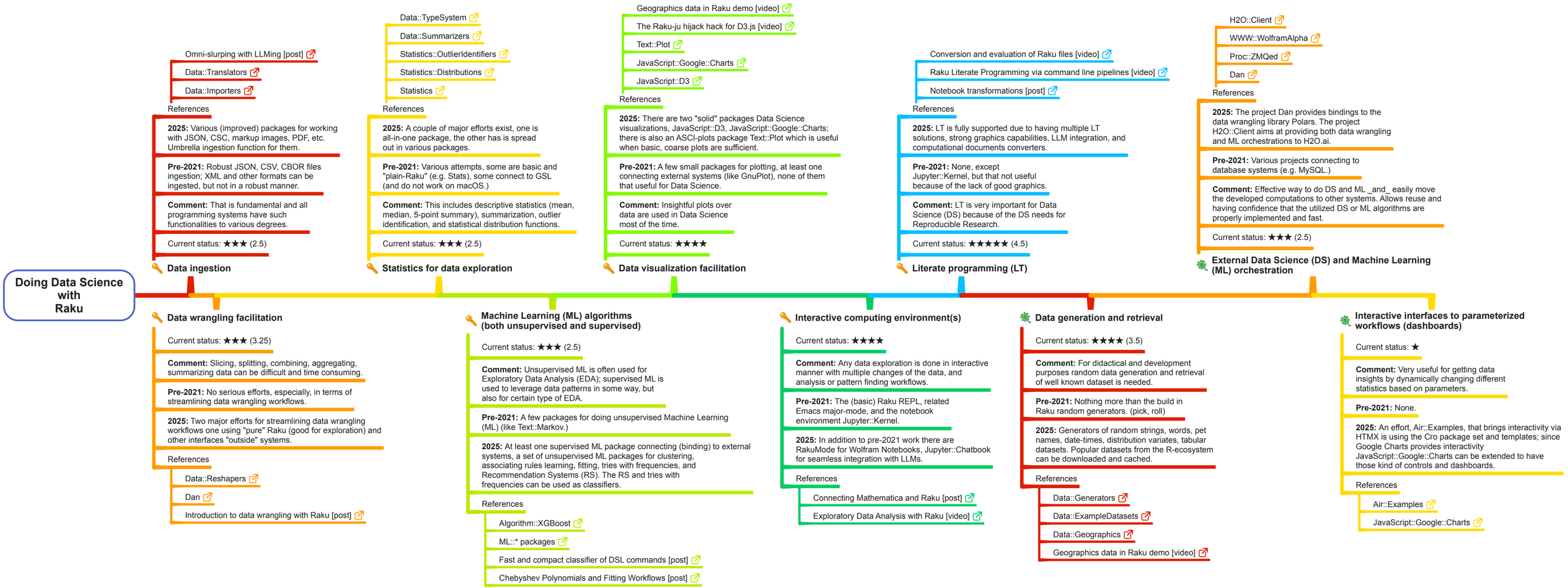
Doing Data Science with Raku
Data ingestion
Omni-slurping with LLMing [post]
Data::Translators
Data::Importers
References
2025: Various (improved) packages for working with JSON, CSC, markup images, PDF, etc. Umbrella ingestion function for them.
Pre-2021: Robust JSON, CSV, CBOR files ingestion; XML and other formats can be ingested, but not in a robust manner.
Comment: That is fundamental and all programming systems have such functionalities to various degrees.
Current status: ★★★ (2.5)
Statistics for data exploration
Data::TypeSystem
Data::Summarizers
Statistics::OutlierIdentifiers
Statistics::Distributions
Statistics
References
2025: A couple of major efforts exist, one is all-in-one package, the other has is spread out in various packages.
Pre-2021: Various attempts, some are basic and "plain-Raku" (e.g. Stats), some connect to GSL (and do not work on macOS.)
Comment: This includes descriptive statistics (mean, median, 5-point summary), summarization, outlier identification, and statistical distribution functions.
Current status: ★★★ (2.5)
Data visualization facilitation
Geographics data in Raku demo [video]
The Raku-ju hijack hack for D3.js [video]
Text::Plot
JavaScript::Google::Charts
JavaScript::D3
References
2025: There are two "solid" packages Data Science visualizations, JavaScript::D3, JavaScript::Google::Charts; there is also an ASCI-plots package Text::Plot which is useful when basic, coarse plots are sufficient.
Pre-2021: A few small packages for plotting, at least one connecting external systems (like GnuPlot), none of them that useful for Data Science.
Comment: Insightful plots over data are used in Data Science most of the time.
Current status: ★★★★
Literate programming (LT)
Conversion and evaluation of Raku files [video]
Raku Literate Programming via command line pipelines [video]
Notebook transformations [post]
References
2025: LT is fully supported due to having multiple LT solutions, strong graphics capabilities, LLM integration, and computational documents converters.
Pre-2021: None, except Jupyter::Kernel, but that not useful because of the lack of good graphics.
Comment: LT is very important for Data Science (DS) because of the DS needs for Reproducible Research.
Current status: ★★★★★ (4.5)
External Data Science (DS) and Machine Learning (ML) orchestration
H2O::Client
WWW::WolframAlpha
Proc::ZMQed
Dan
References
2025: The project Dan provides bindings to the data wrangling library Polars. The project H2O::Client aims at providing both data wrangling and ML orchestrations to H2O.ai.
Pre-2021: Various projects connecting to database systems (e.g. MySQL.)
Comment: Effective way to do DS and ML _and_ easily move the developed computations to other systems. Allows reuse and having confidence that the utilized DS or ML algorithms are properly implemented and fast.
Current status: ★★★ (2.5)
Data wrangling facilitation
Current status: ★★★ (3.25)
Comment: Slicing, splitting, combining, aggregating, summarizing data can be difficult and time consuming.
Pre-2021: No serious efforts, especially, in terms of streamlining data wrangling workflows.
2025: Two major efforts for streamlining data wrangling workflows one using "pure" Raku (good for exploration) and other interfaces "outside" systems.
References
Data::Reshapers
Dan
Introduction to data wrangling with Raku [post]
Machine Learning (ML) algorithms (both unsupervised and supervised)
Current status: ★★★ (2.5)
Comment: Unsupervised ML is often used for Exploratory Data Analysis (EDA); supervised ML is used to leverage data patterns in some way, but also for certain type of EDA.
Pre-2021: A few packages for doing unsupervised Machine Learning (ML) (like Text::Markov.)
2025: At least one supervised ML package connecting (binding) to external systems, a set of unsupervised ML packages for clustering, associating rules learning, fitting, tries with frequencies, and Recommendation Systems (RS). The RS and tries with frequencies can be used as classifiers.
References
Algorithm::XGBoost
ML::* packages
Fast and compact classifier of DSL commands [post]
Chebyshev Polynomials and Fitting Workflows [post]
Interactive computing environment(s)
Current status: ★★★★
Comment: Any data exploration is done in interactive manner with multiple changes of the data, and analysis or pattern finding workflows.
Pre-2021: The (basic) Raku REPL, related Emacs major-mode, and the notebook environment Jupyter::Kernel.
2025: In addition to pre-2021 work there are RakuMode for Wolfram Notebooks, Jupyter::Chatbook for seamless integration with LLMs.
References
Connecting Mathematica and Raku [post]
Exploratory Data Analysis with Raku [video]
Data generation and retrieval
Current status: ★★★★ (3.5)
Comment: For didactical and development purposes random data generation and retrieval of well known dataset is needed.
Pre-2021: Nothing more than the build in Raku random generators. (pick, roll)
2025: Generators of random strings, words, pet names, date-times, distribution variates, tabular datasets. Popular datasets from the R-ecosystem can be downloaded and cached.
References
Data::Generators
Data::ExampleDatasets
Data::Geographics
Geographics data in Raku demo [video]
Interactive interfaces to parameterized workflows (dashboards)
Current status: ★
Comment: Very useful for getting data insights by dynamically changing different statistics based on parameters.
Pre-2021: None.
2025: An effort, Air::Examples, that brings interactivity via HTMX is using the Cro package set and templates; since Google Charts provides interactivity JavaScript::Google::Charts can be extended to have those kind of controls and dashboards.
References
Air::Examples
JavaScript::Google::Charts