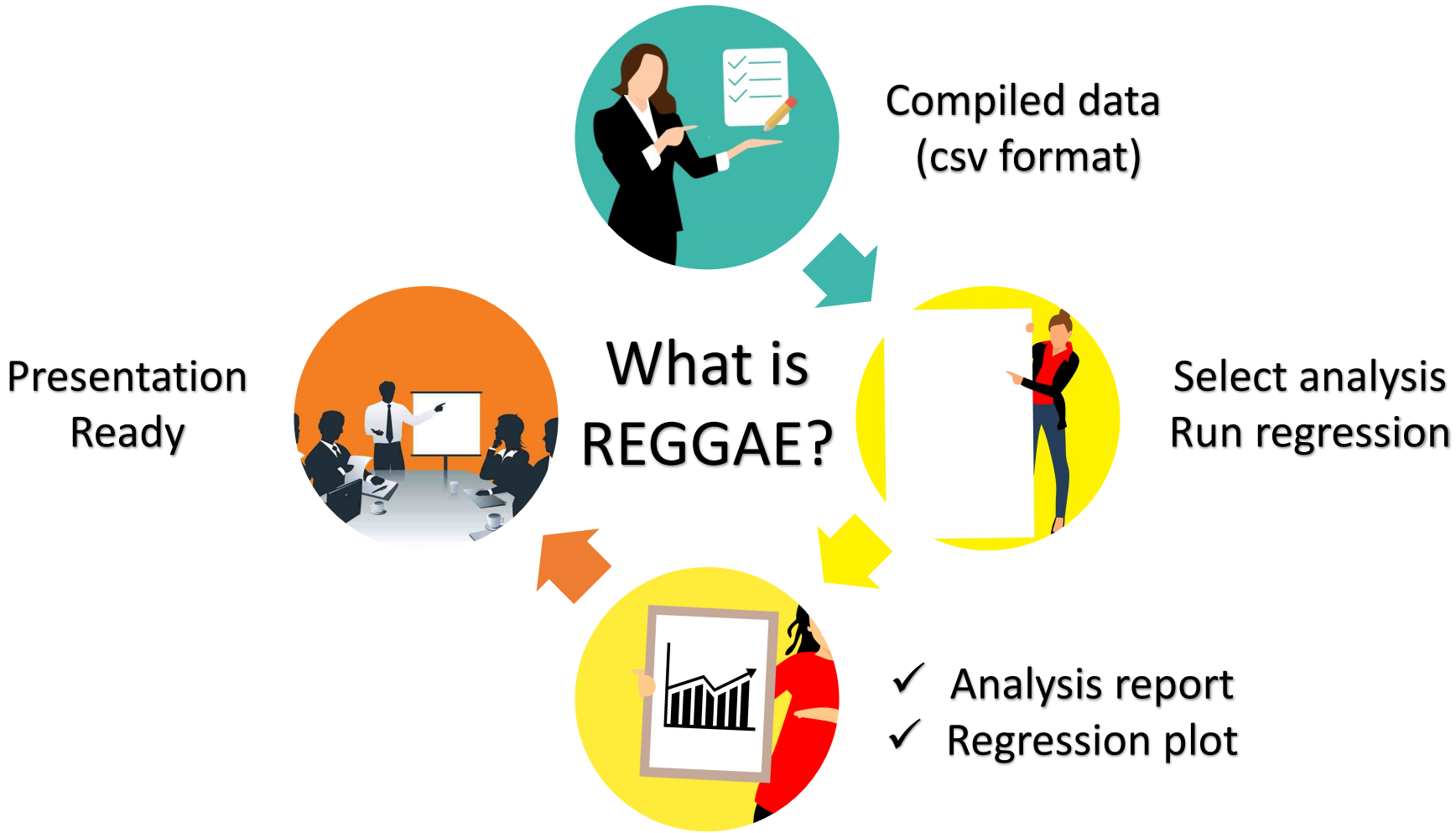

How-To-
Instructions
Regression
Generator &
Analyzer
REGGAE
Coded by Liliana C. Gallegos
Updated


Compiled data
(csv format)
Select analysis
Run regression
✓ Analysis report
✓ Regression plot
Presentation
Ready
What is
REGGAE?

Setup

- ❑ Install → R and R-studio

```
conda install r-essentials --yes
conda install -c r rstudio --yes
```

- ❑ Run setup.r → installs R packages required

```
Rscript setup.r
```

- ❑ Prepare data.csv & 1st column row names

	A	B	C	D	E	F	G
1	Ligand	dG-dce	BV_lig	Lig_charge	MV_lig	NMR_lig	Pd_charge
2	2-Me-quinoli	6.46	45.99	-0.46284	115.119	-80.1286	0.88276
3	2-pentyl	6.42	45.58	-0.47406	136.551	-88.8872	0.88591

Runs on Dr.Maximus with adapted version:
setup-linux.r
reggae-linux.r

```
~/Desktop/REGGAE » Rscript setup.r        Liliana@x8
 Package already installed:  optparse
 Package already installed:  corrplot
 Package already installed:  bindr
 Package already installed:  MuMIn
 Package already installed:  stats
 Package already installed:  cvq2
 Package already installed:  dplyr
 Package already installed:  car
 Package already installed:  ggplot2
(ML-env)
```

New packages required to run v3!

Statistical analysis options

- ❑ Feature selection and model — –m , --model
  options: full, stepwise, dredge, mincorr
- ❑ Build a model — –b , --buildmodel
  input: y,x1,x2,xn
- ❑ Define y-response — –y , --yresponse
- ❑ Split into Train/ Test datasets — –r , --randsample
  value between 0 and 1; 0 = predefined test/train; 1 = full data
- ❑ Cross Validation — –q , --crossvalidation
  q^2 values: leave-one-out, k-fold, external
- ❑ Diagnostics — –d , --diagnostics
  F-value comparisons, collinearity, outlier testing
- ❑ Pairwise Correlation — –c , --corrplot

mincorr = this modeling option automatically removes the highly correlated values for you to build a reduced model

0 = choosing this split value allows you to split your data based on user labeled 'train' and 'test' samples within a column

Outliers are now defined as samples with Rstudent values > 4. It will only plot the top 5 larges residues considered 'outliers'.

A pearson correlation plot prints and saves in your working folder when you choose this option along with verbose (-v).

NEW options

- ❑ Plant your own seed
  default = 42
- ❑ Adjust K-fold value
  default = 5
- ❑ X-variables are now being scaled for all analysis
- ❑ Plots are now in higher resolution = 300

–s , --seed

Want to plant your own seeds? Now you can!
Make your garden and enjoy the fruits of nature.

–k , --kfoldvalue

You can now adjust your k-value for cross validation.
Note: -q option must also be selected.

Let's REGGAE, mon!

```
(ML-env)
~/Desktop/REGGAE-master » Rscript reggae-v2.r -h                                Liliana@x86_64
Usage: reggae-v2.r [options]


Options:
        -v, --verbose
                Print extra output [defaul = FALSE]

        -i CHARACTER, --inputfile=CHARACTER
                Requires input data file in csv format. NOTE: uses first column as row names.

        -o CHARACTER, --outputfile=CHARACTER
                Optional: output file name [default = REGGAE-analysis-output.txt]

        -m MODEL, --model=MODEL
                Types of linear regression model include: full, stepwise, dredge, and mincorr.

        -b BUILDMODEL, --buildmodel=BUILDMODEL
                Build linear model from input variables.

        -y YRESPONSE, --yresponse=YRESPONSE
                Requires defining the y-response variable for given dataframe. Required to run model

        -r RANDSAMPLE, --randsample=RANDSAMPLE
                Train:Test random split - [default = 1] gives no split; Select 0 for pre-defined Tra

        -s SEED, --seed=SEED
                Optional: specify the seed for random sample split. [Default seed = 42]

        -q, --crossvalidation
                Performs leave-one-out CV and K-fold CV on training (or full) data; external validat

        -k KFOLDVALUE, --kfoldvalue=KFOLDVALUE
                To adjust k-fold value - [default = 5]

        -d, --diagnostics
                Diagnostics include: F-value comparisons, collinearity diagnostics, and outlier test

        -c CORRPLOT, --corrplot=CORRPLOT
```

–v , --verbose

Rscript reggae.r –i data.csv –m stepwise –y dE –r 0.8 –q –d –v

Predictive Model & Report

```
o Linear regression model:

Call:
lm(formula = y ~ E_proton_n + C_shielding + N_shi
    E_deproton_n + fk_n + E_lumo_e + fk_e, data =

Residuals:
     Min       1Q   Median       3Q      Max
-13.1433  -1.2093   0.0322   1.7769   5.4302

Coefficients:
              Estimate Std. Error t value Pr(>|t
(Intercept)  114.107655  14.906858   7.655 2.34e-
E_proton_n     0.228047   0.038627   5.904 6.52e-
C_shielding    0.434323   0.092527   4.694 9.78e-
N_shielding   -0.039849   0.006584  -6.053 3.41e-
Ci_chrg       13.547379   8.114844   1.669 0.0985
E_deproton_n  -0.213743   0.052533  -4.069 0.0001
fk_n          12.950495   5.929968   2.184 0.0316
```