FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

NIRS Workshop

USAID
FROM THE AMERICAN PEOPLE

Cornell University

Agenda

1. What is NIRS, applications, spectrometers
2. Analysis and workflow
3. Example code activity

Objectives

- Understand how NIRS can be implemented into your program
- Clean and prepare raw data for analysis
- Train a model to predict phenotypes from spectral data

Introductions

Who we are:

- Marjorie Hanneman, Erin Farmer, Sam Herr
- Plant Breeding and Genetics PhD students at Cornell University, Mike Gore's lab
- Training in phenomics, bioinformatics, and genomics
- Within ILCI, Marjorie works on using NIRS for sorghum grain quality predictions with CACCIA

What is Near Infrared Spectroscopy (NIRS)?

Current use of NIRS in Plant Breeding

Low-cost, handheld near-infrared spectroscopy for root dry matter content prediction in cassava

Jenna Hershberger, Edwige Gaby Nkouaya Mbanjo, Prasad Peteti, Andrew Ikpan, Kayode Ogunpaimo, Kehinde Nafiu, Ismail Y. Rabbi, Michael A. Gore

First published: 31 March 2022 | https://doi.org/10.1002/ppj2.20040 | Citations: 5

Determination of protein, total carbohydrates and crude fat contents of foxtail millet using effective wavelengths in NIR spectroscopy

Jing Chen [a b], Xin Ren [a b], Qing Zhang [a b], Xianmin Diao [c], Qun Shen [a b]

Phenomic selection and prediction of maize grain yield from near-infrared reflectance spectroscopy of kernels

Holly M. Lane, Seth C. Murray, Osval A. Montesinos-López, Abelardo Montesinos-López, José Crossa, David K. Rooney, Ivan D. Barrero-Farfan, Gerald N. De La Fuente, Cristine L. S. Morgan

First published: 06 April 2020 | https://doi.org/10.1002/ppj2.20002 | Citations: 17

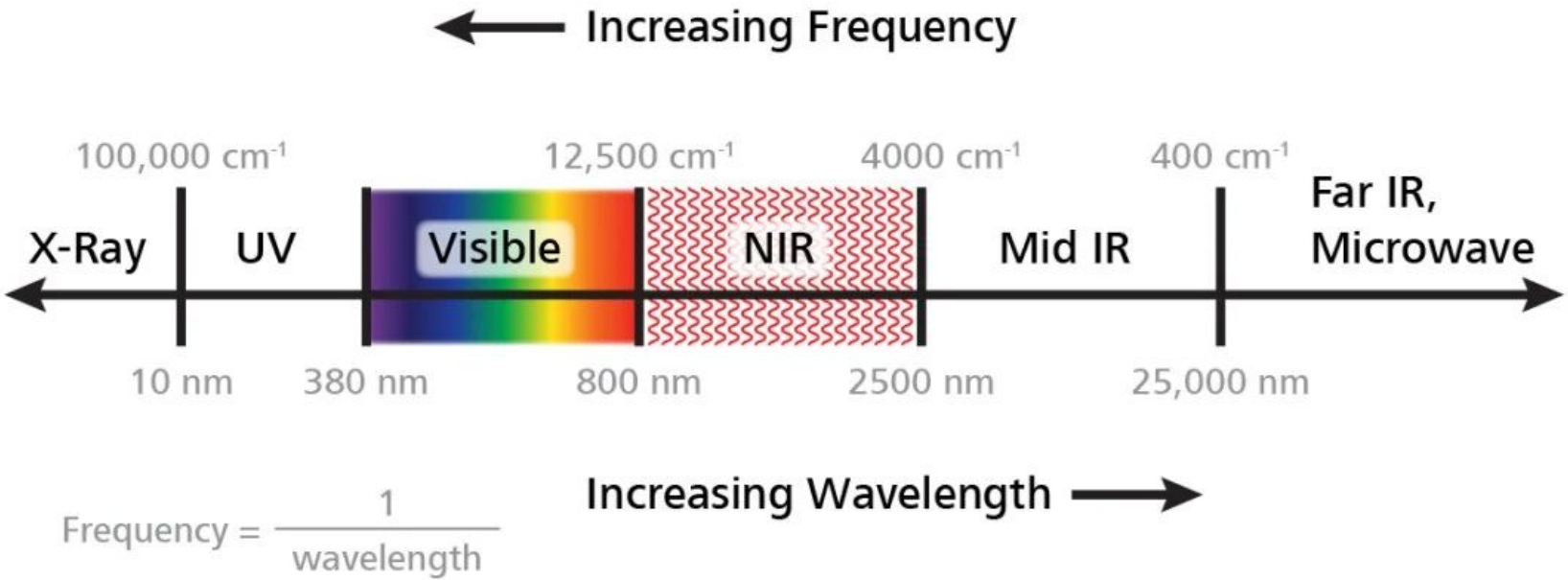
Fourier transform near-infrared spectroscopy (FT-NIRS) application to estimate Brazilian soybean [*Glycine max* (L.) Merril] composition

Daniela Souza Ferreira [a], Juliana Azevedo Lima Pallone [a], Ronei Jesus Poppi [b]

Agricultural product	Spectral range
Apple and apple purees	800–2,500 nm
Apple	350–2,500 nm
Tangerine	700–1,100 nm
Elderberry	800–2,500 nm
Apple	800–2,500 nm
Hami melons	550–950 nm
Grapes	800–1,100 nm
Pineapple	740–1,070 nm
Olives	2,307–2,348 nm
Pomegranate	400–1,100 nm
Persimmon	1,000–2,500 nm
Apple	550–950 nm
Grape	450–2,500 nm
Tomato	400–1,100 nm 900–1,700 nm

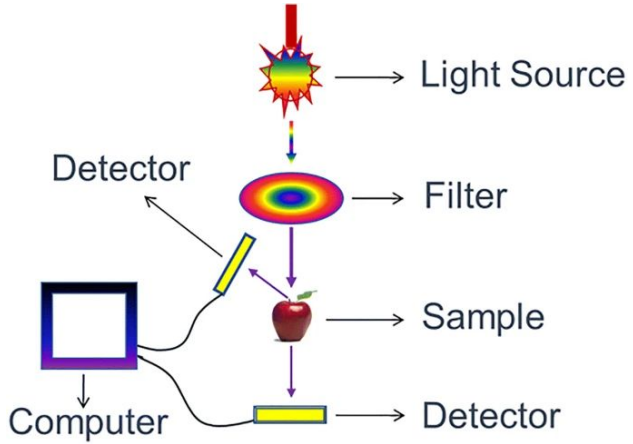
Pandesilvam et al 2020

Physics Review

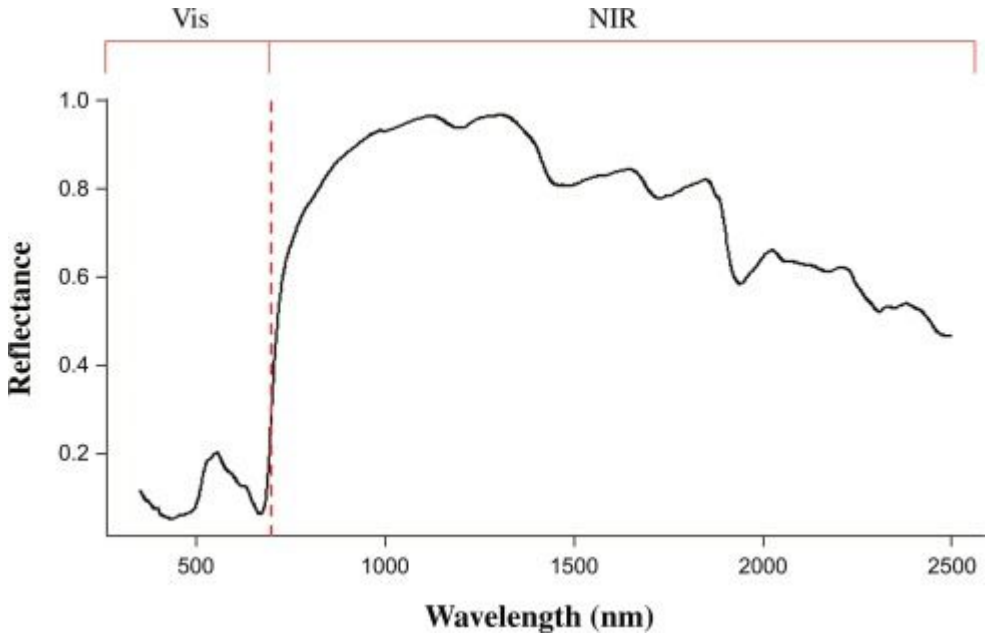


Metrohm

How does NIRS work?

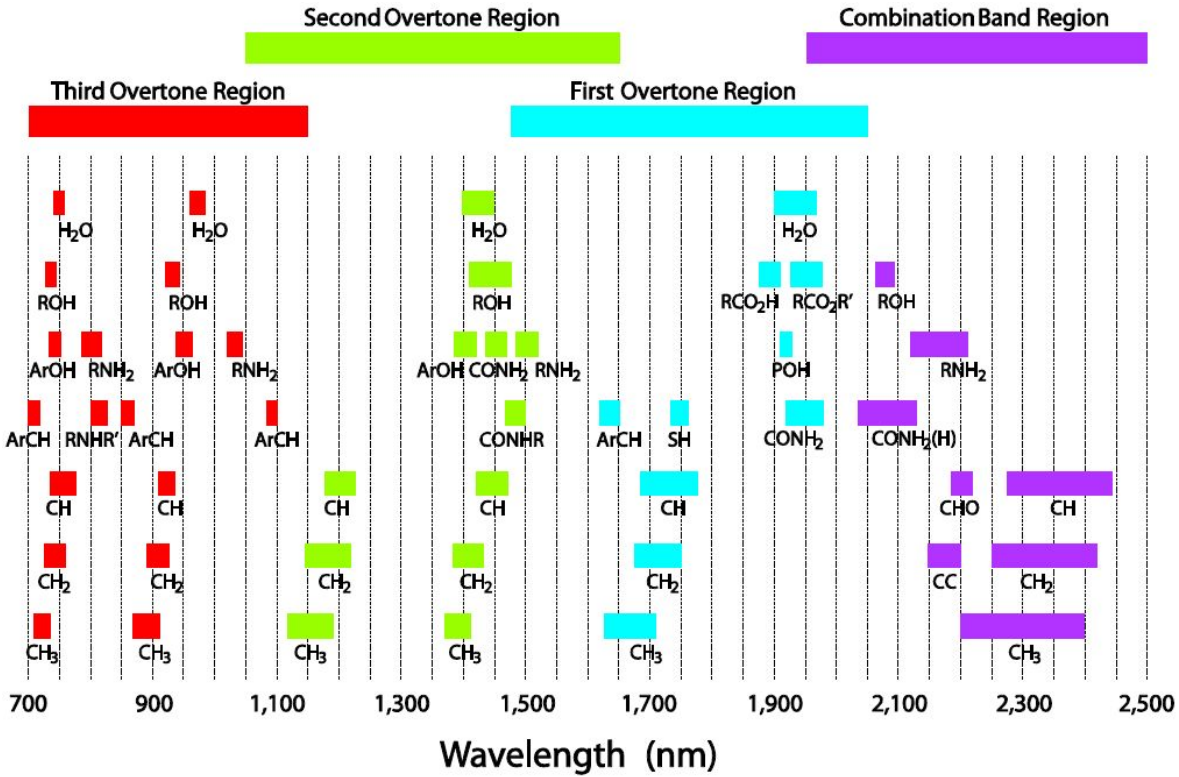


Chandrasekaran et al. 2019



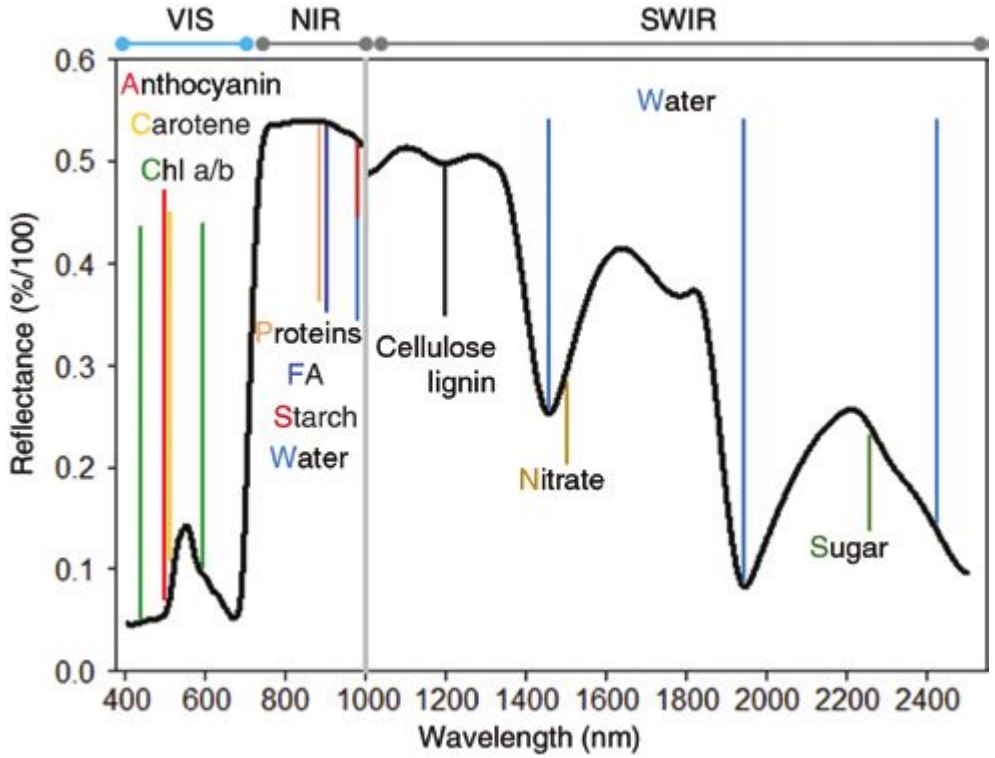
Prananto et al.2020

How does NIRS work?



https://www.nirlab.com/blog/nir-spectroscopy-nutshell/

What traits can we capture with NIRS?



Kuska et al. 2018

Applications of NIRS

Predicting phenotypes that are difficult/expensive to measure

Potential to phenotype in the field

Reduce need for sample transportation

Fast and non-destructive

Secondary phenotypes

Can measure different states (solid vs liquid)

Laboratory methods can be influenced by several variables

Building a training data set

- 3-5 scans/sample
- Population coverage of phenotypic variation and sample size
- Consistent conditions for sample collection
  - Using the same spectrometer for all samples
  - Sample processing (ground vs. whole, fresh vs. dry, etc.)
  - Environmental conditions, such as lighting and humidity
- Need to train a new model for each phenotype, spectrometer, crop combination
- Need measured phenotypes to be accurate for training
- Validation samples in subsequent years

Data Management

- Ensure all scans of the same sample have the same unique identifier
  - C16Mcal_3
  - 23GHTX23
- Include relevant sample information such as date, spectrometer, location, and trial
- Utilize barcodes (QRLabelR) while scanning, can match samples in FieldBook
- Track changes made to spectral data during preprocessing and analysis
- Document data processing steps and analysis procedures to ensure reproducibility and transparency

Spectrometers

Foss Benchtop

- 400 - 2500 nm
- Most accurate but most expensive
- Stays on a lab bench

InnoSpectra

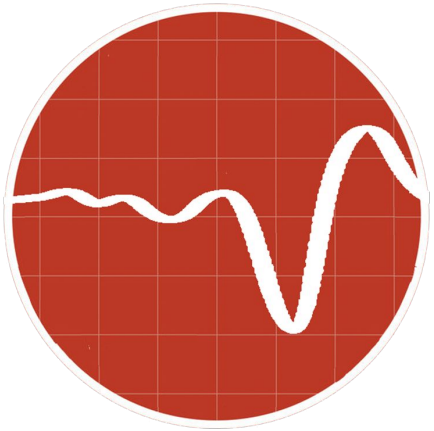
- 900 - 2400 nm
- Handheld, can go in field
- Uses Prospector app

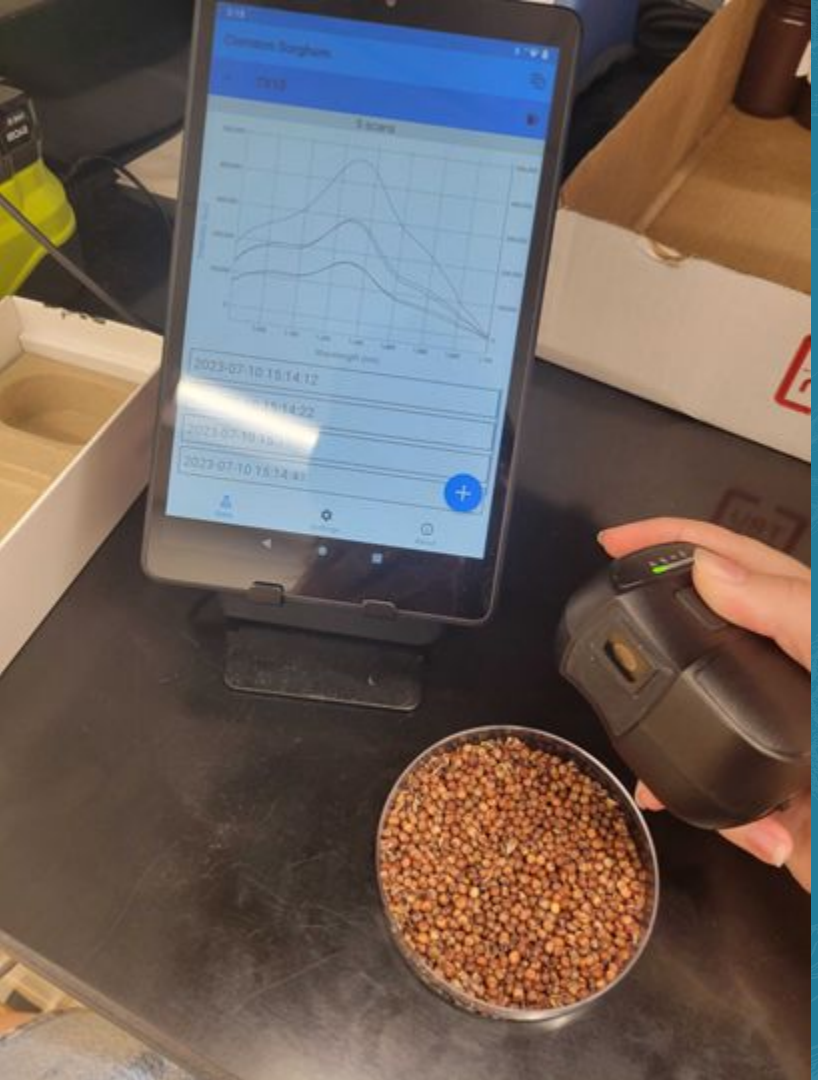
SCiO

- 740 -1070 nm
- Handheld, can go in field
- Proprietary software, expensive
- Smallest range

Spectrometers

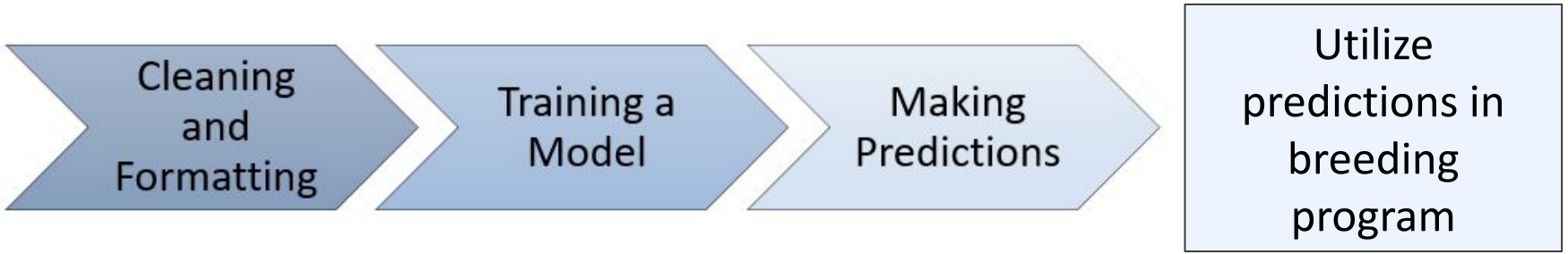
- Prospector App
  - Prospector is an Android app capable of capturing, storing, visualizing, and exporting data from handheld NIR spectrometers. Prospector provides fast, reliable capture of scans for phenotype prediction with a standardized user interface and data export. Prospector allows breeders to rapidly utilize NIRS for phenotyping.





Analysis and Workflow

Workflow


Cleaning and Formatting
Training a Model
Making Predictions
Utilize predictions in breeding program

Cleaning and Formatting

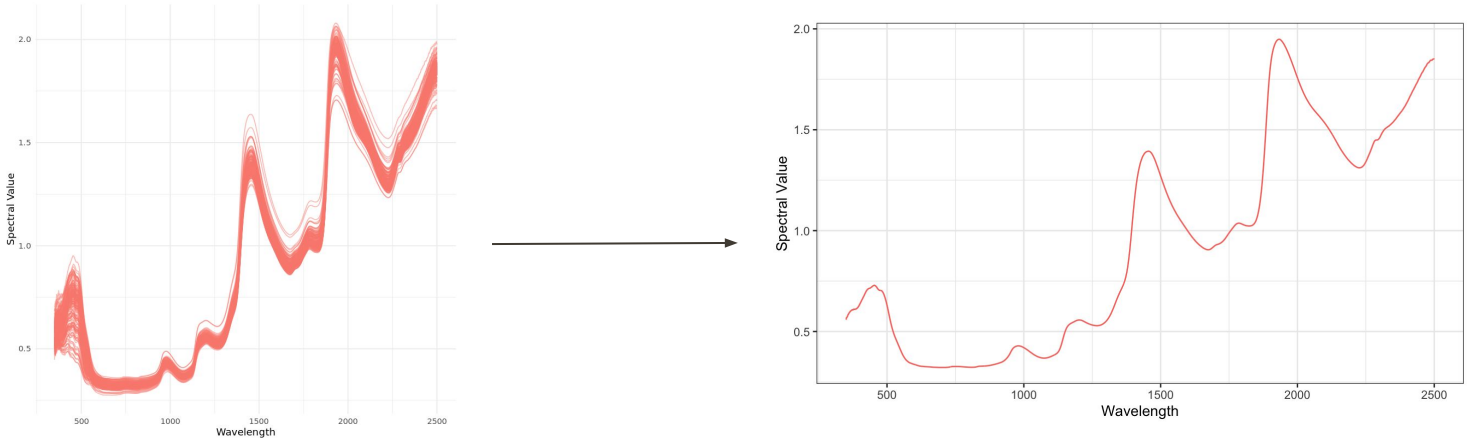
- Raw spectral data from different spectrometers will have different formats
- Need to clean and arrange data for analysis
- Match spectra with reference values so that you have a dataframe with unique identifiers, reference values, and other metadata as columns to the left of spectral values. Spectral column names should start with "X".

```
#>   study.name sample.id DMC.oven   TCC  X350  X351  X352
#>   <chr>      <chr>        <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1 C16Mcal    C16Mcal_1     39.6  1.00 0.488 0.495 0.506
#> 2 C16Mcal    C16Mcal_2     35.5 17.0  0.573 0.568 0.599
#> 3 C16Mcal    C16Mcal_3     42.0 21.6  0.599 0.627 0.624
#> 4 C16Mcal    C16Mcal_4     39.0  2.43 0.517 0.516 0.514
#> 5 C16Mcal    C16Mcal_5     33.4 24.0  0.519 0.548 0.554
#> 6 C16Mcal    C16Mcal_6     32.1 19.0  0.576 0.566 0.589
#> 7 C16Mcal    C16Mcal_7     35.8  6.61 0.530 0.536 0.525
```

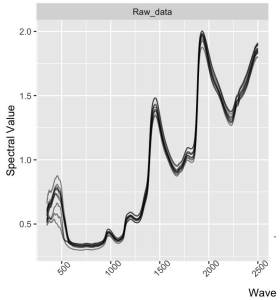
```
#>   unique.id study.name reference  X350  X351  X352  X353
#>   <chr>     <chr>          <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1 C16Mcal_1 C16Mcal         39.6 0.488 0.495 0.506 0.494
#> 2 C16Mcal_2 C16Mcal         35.5 0.573 0.568 0.599 0.593
#> 3 C16Mcal_3 C16Mcal         42.0 0.599 0.627 0.624 0.606
#> 4 C16Mcal_4 C16Mcal         39.0 0.517 0.516 0.514 0.536
#> 5 C16Mcal_5 C16Mcal         33.4 0.519 0.548 0.554 0.549
#> 6 C16Mcal_6 C16Mcal         32.1 0.576 0.566 0.589 0.591
#> 7 C16Mcal_7 C16Mcal         35.8 0.530 0.536 0.525 0.539
```

Cleaning and Formatting

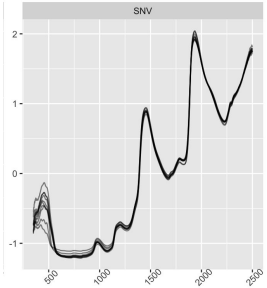
- Visualize and filter spectra
- Remove NAs, outliers, and noise from physical effects
- You will likely have more than one scan per unique identifier. Aggregate the scans by mean or median



Training a Model



1. Pretreatment



2. Model Selection

Models:

- Partial Least Squares regression (PLSR)
- Random Forest
- Support Vector Machine

3. Test models

4. Model Performance

$$\sqrt{\frac{(Predicted - Actual)^2}{Number\ of\ Samples}}$$

Root Mean Square error

Use best fitting model for predictions

Making Predictions

- Input new spectral data from samples with unknown phenotype, and can predict them

Sample ID	1st Wavelength	2nd Wavelength	…	Phenotype
Cassava_1	1.05	8.23		**?**
Cassava_2	5	6.45		**?**
Cassava_n	2	7.6		**?**

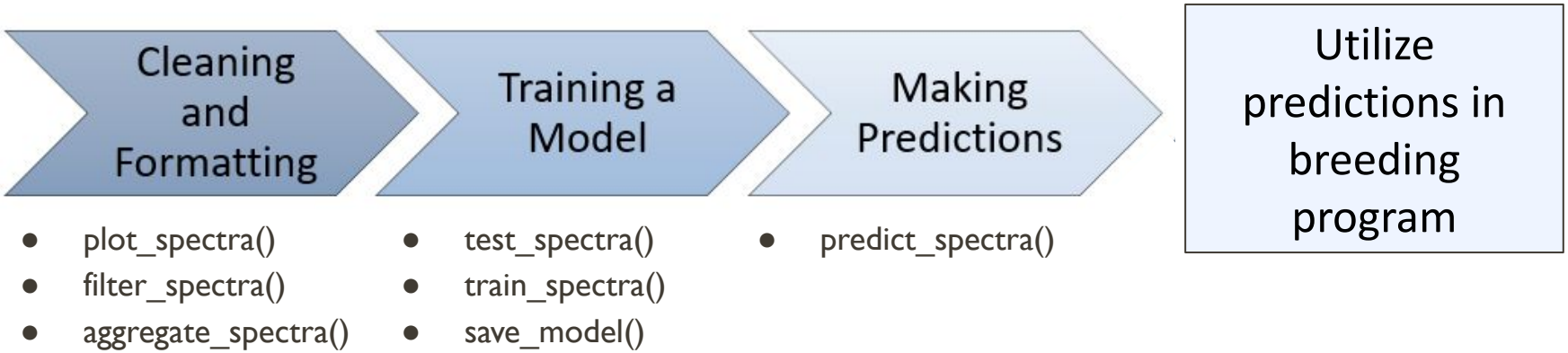
Best fitting model →

Sample ID	1st Wavelength	2nd Wavelength	…	Phenotype
Cassava_1	1.05	8.23		**2**
Cassava_2	5	6.45		**5**
Cassava_n	2	7.6		**3**

Introduction to waves

- A package within R that combines different functions to streamline NIRS analysis
- Analysis of visible and near-infrared reflectance measurements.
- It includes visualization, filtering, aggregation, pretreatment, cross-validation set formation, model training, and prediction functions
- https://gorelab.github.io/waves/index.html

Workflow - waves functions



Example Code

Dry matter example dataset

- Cassava roots were phenotyped for dry matter content (DMC) and total carotenoid content (TCC)
- A portable NIRS device (QualitySpec Trek: S-10016) was used to collect spectral data on both intact and mashed root samples.
- Three spectra per root were taken respectively on the proximal, middle and distal regions of roots
- Raw data file will have the study, sample id, DMC, TCC, and spectra
- Calibration and validation set

Ikeogu UN, Davrieux F, Dufour D, Ceballos H, Egesi CN, et al. (2017) Rapid analyses of dry matter content and carotenoids in fresh cassava roots using a portable visible and near infrared spectrometer (Vis/NIRS). PLOS ONE 12(12): e0188918. https://doi.org/10.1371/journal.pone.0188918

Intro to pipes %>%

input %>% functionName(, ...)

other inputs

functionName(input, ...)

other inputs

Piping in R is like baking

slice <- slice(decorate(bake(mix(ingredients))))

? <- 

%>%

mix( ) %>%

bake( ) %>%

decorate( )%>%

slice( )

By @ArthurWelle / Images from flaticon.com

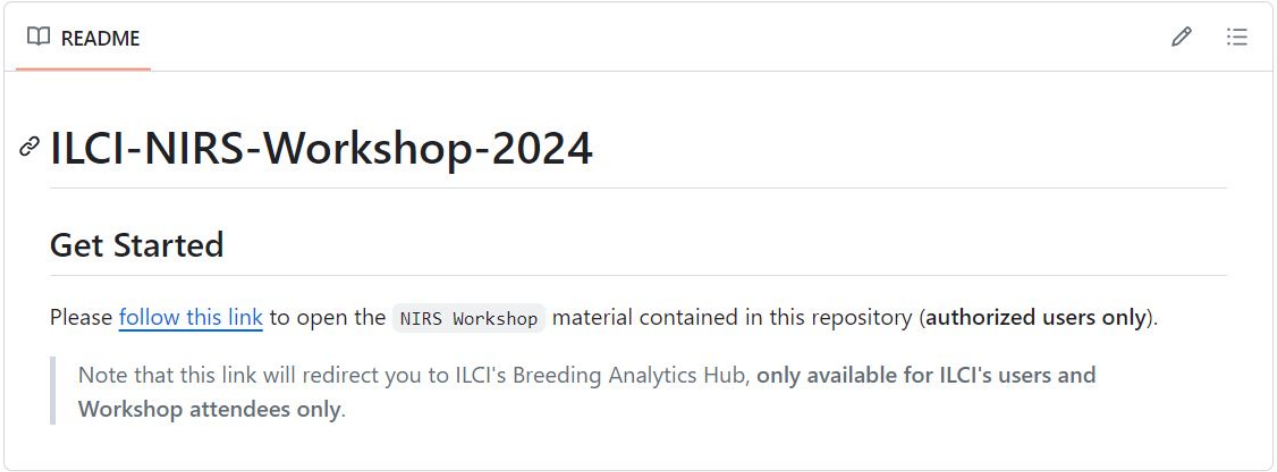
Intro to JupyterHub

https://ilci.cornell.edu/introduction-to-jupyterhub/

Cloud-based interactive coding and documentation environment for data analysis projects

If you pre-registered for this workshop, you will have have an ILCI JupyterHub account

http://tinyurl.com/NIRS2024



The link will take you to the JupyterHub login page. **Use the same email you registered with.**

Input a password of your choosing. Remember this password!

If you have used the Hub before, use your previous password.