Soil Infrared Spectroscopy

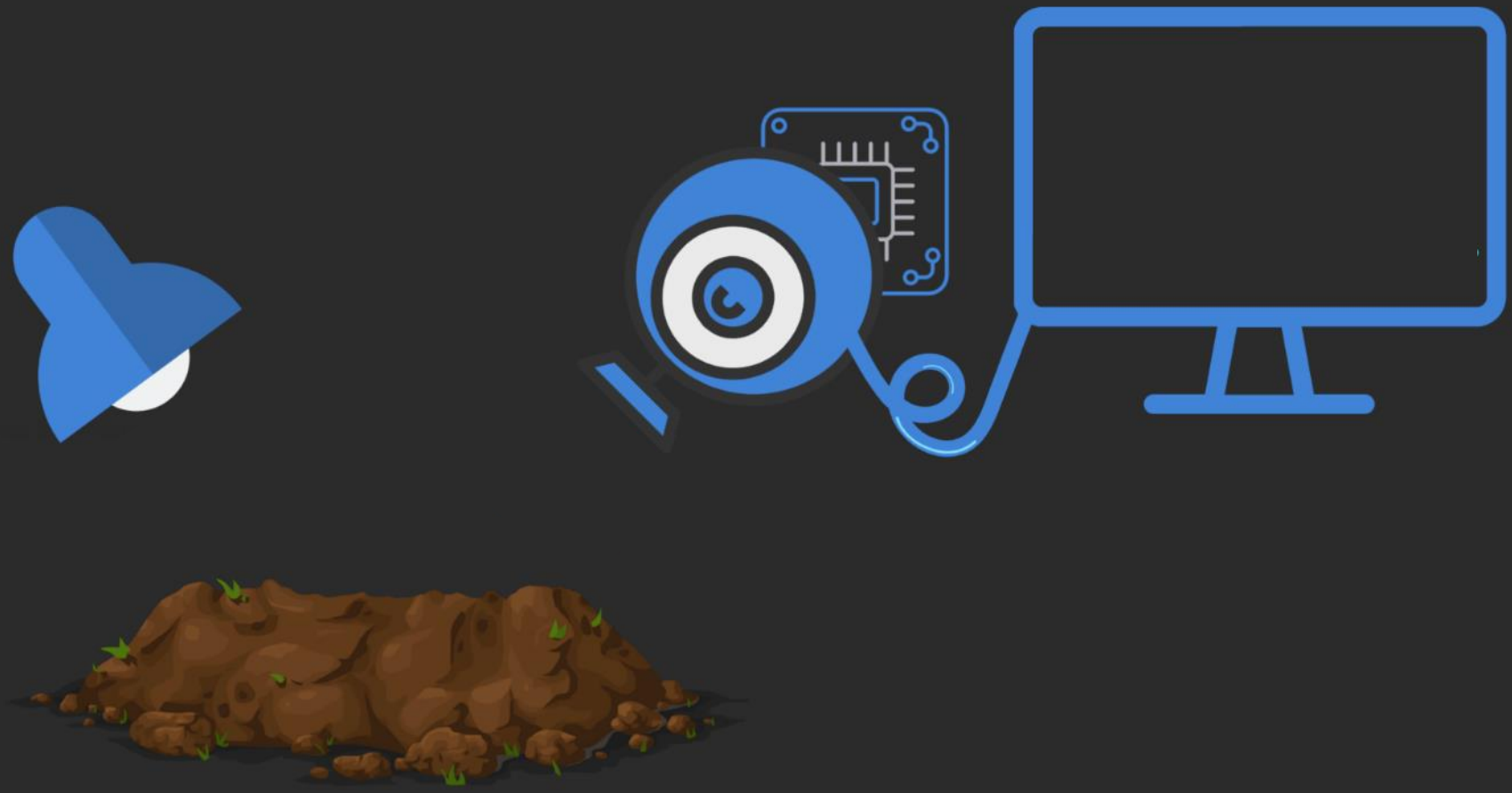
TropiRes – Summer School – 2024

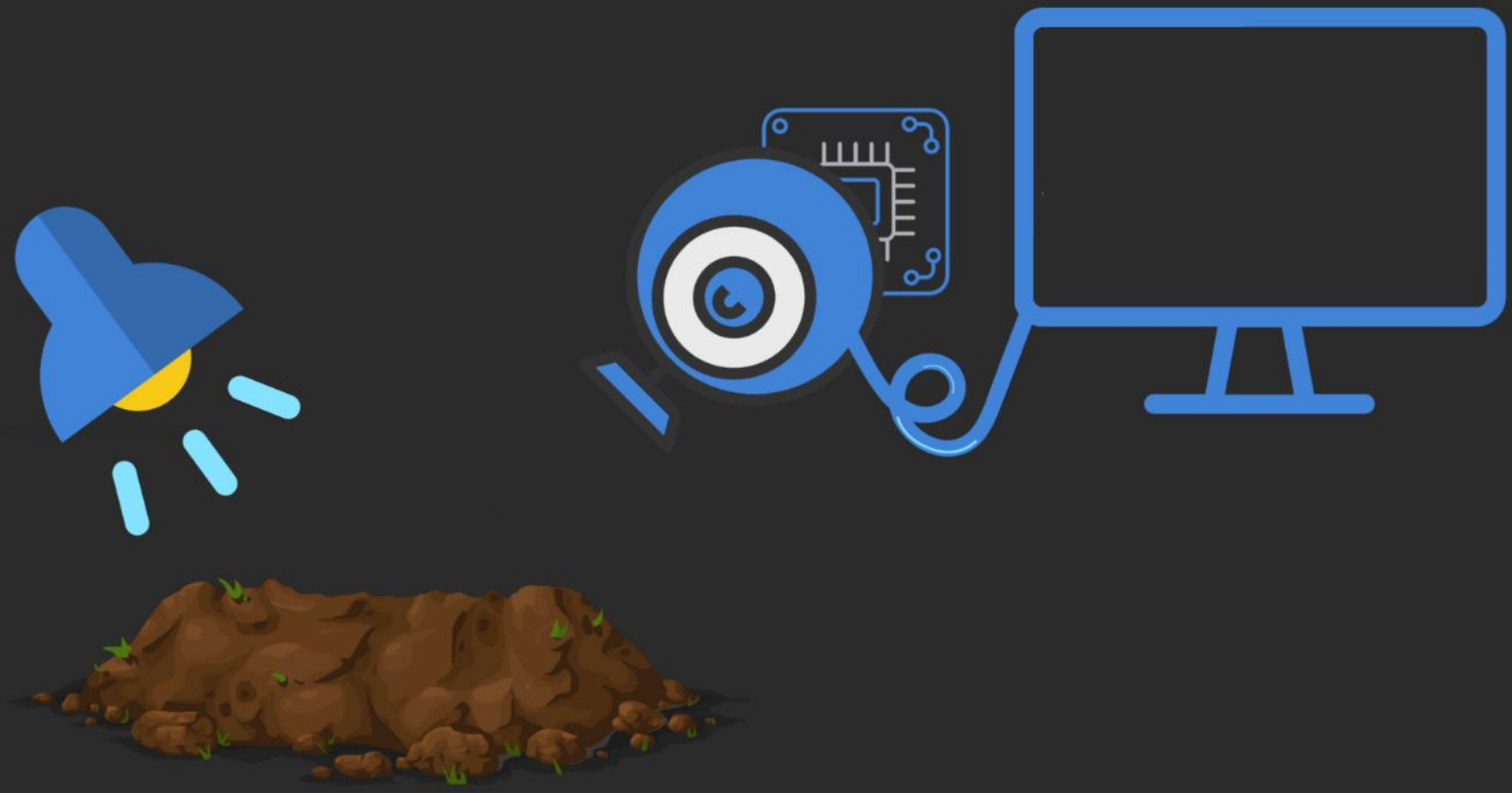
Module 2

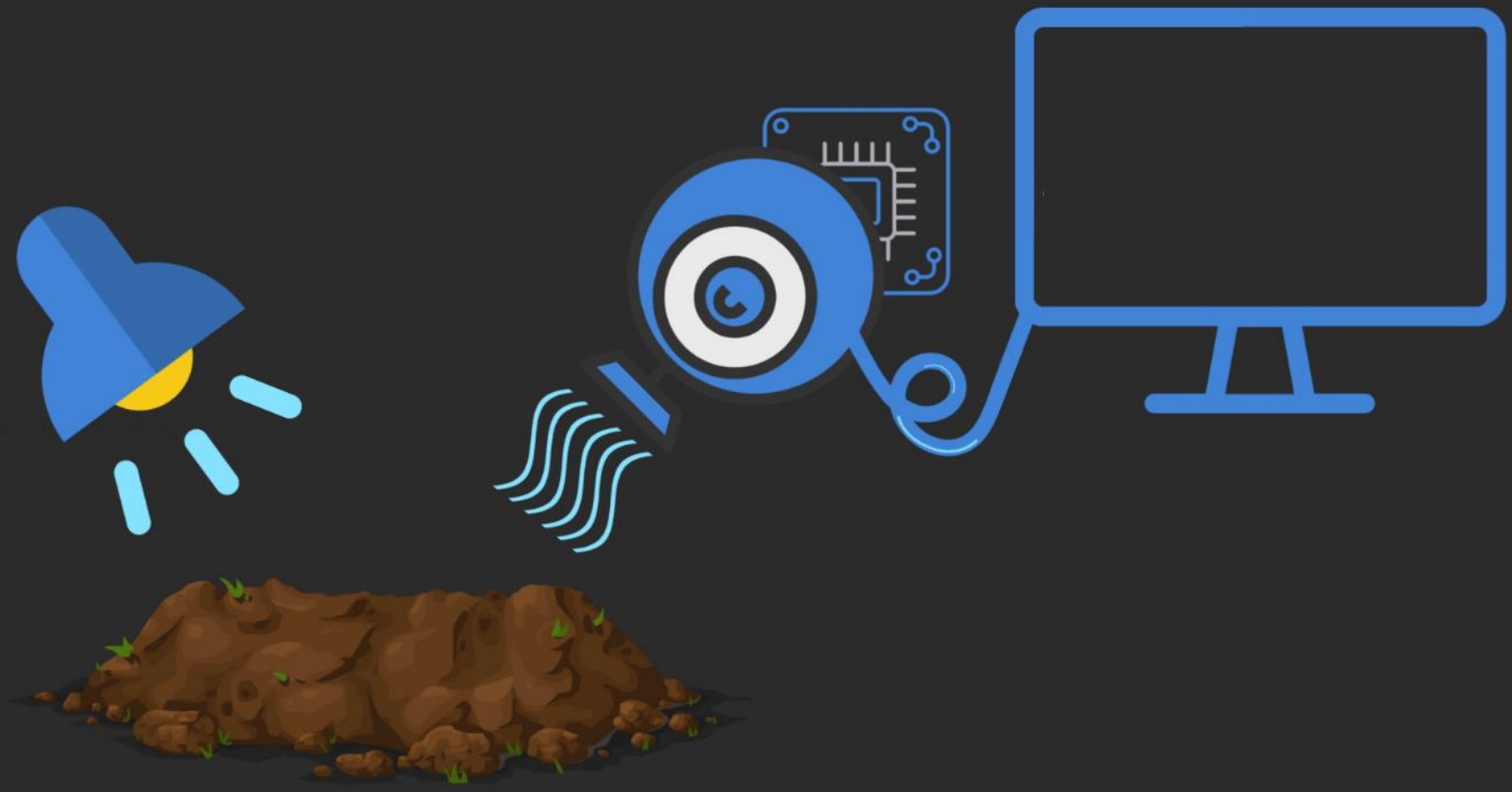
Leo Ramirez-López, Laura Summerauer, Moritz Mainka

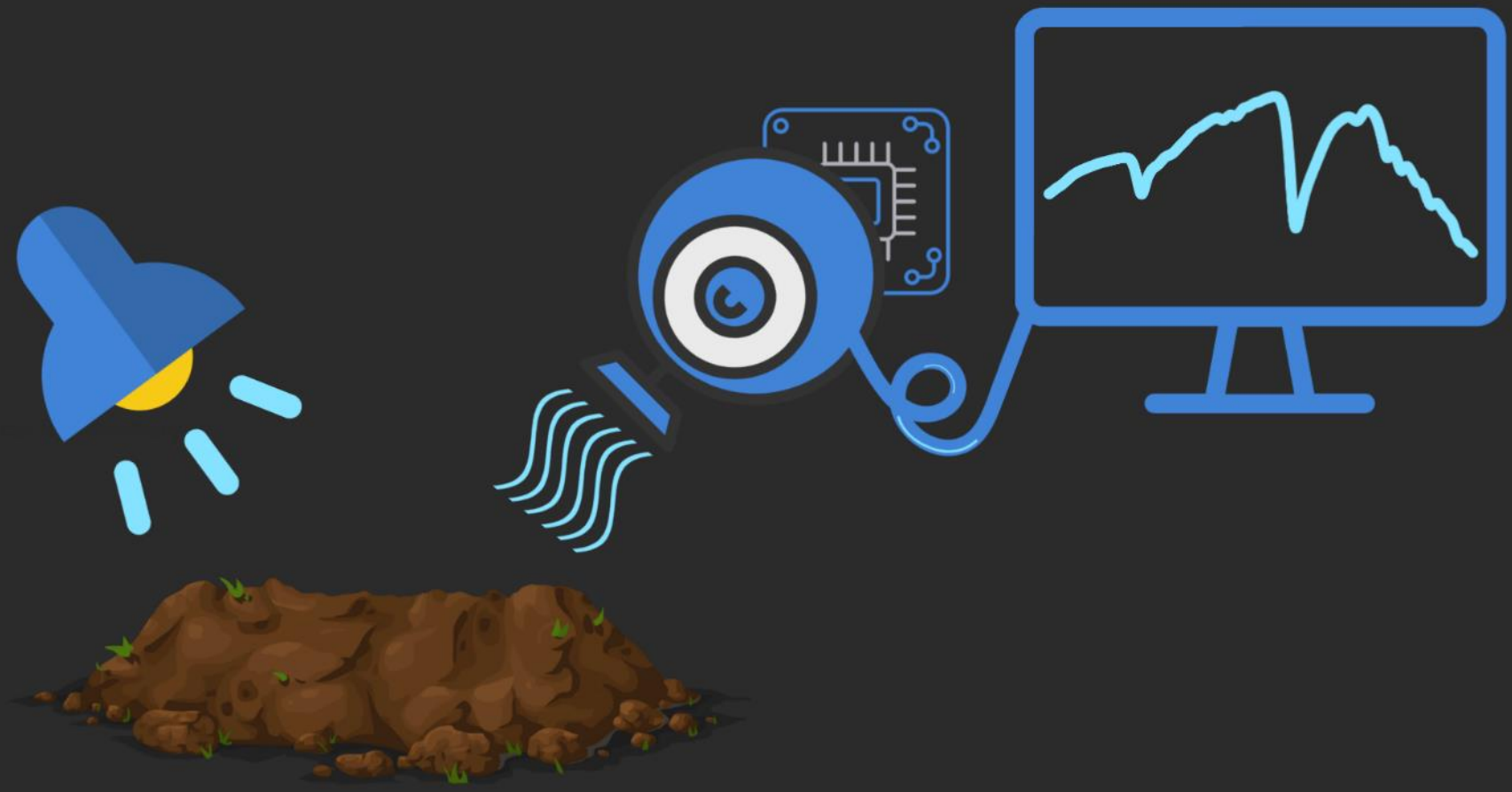
NIR/IR Spectroscopy – An equation for success

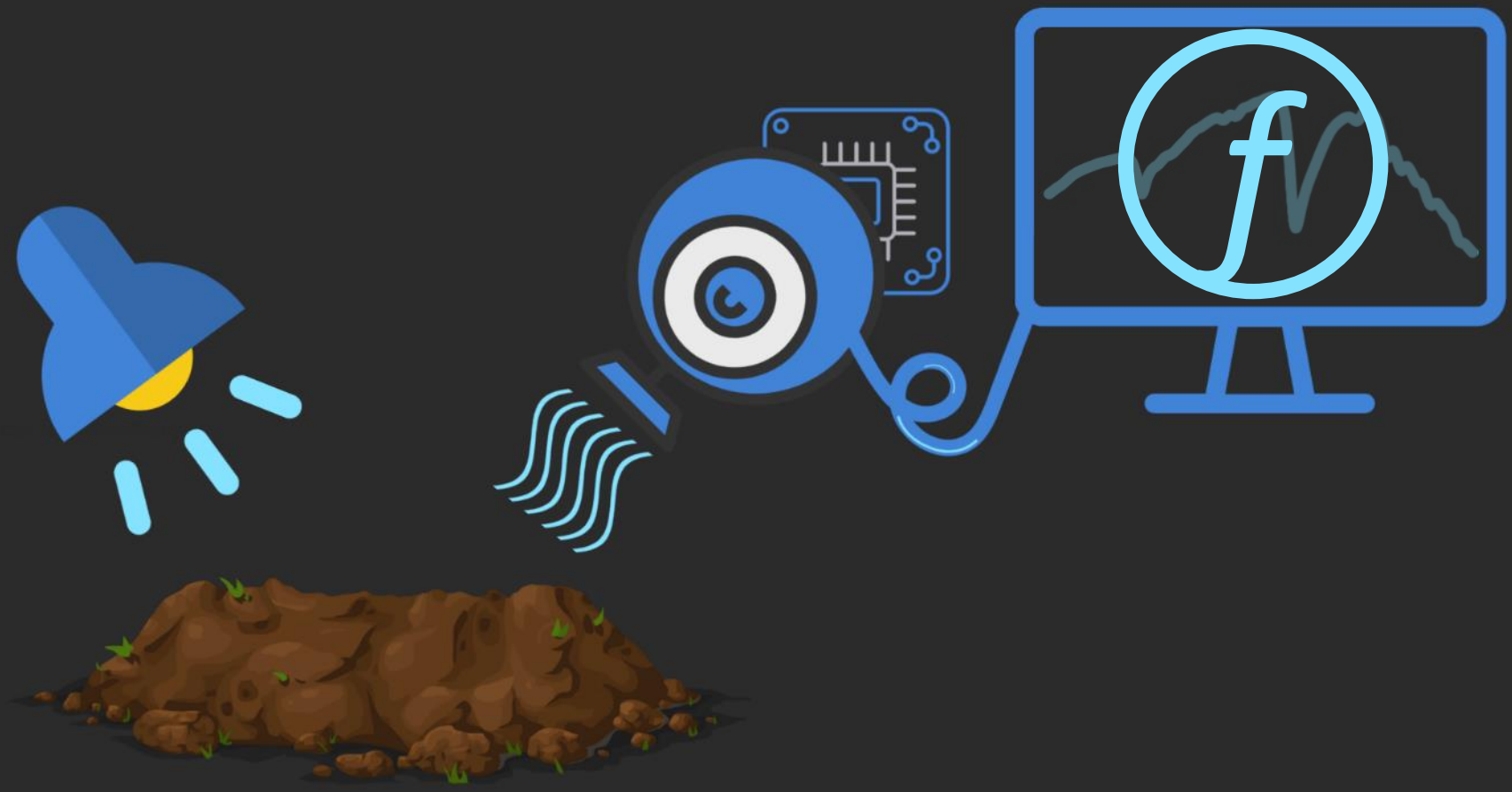
$$\text{Success} = f(\text{hardware, software, data, models, people, automation, ...})$$

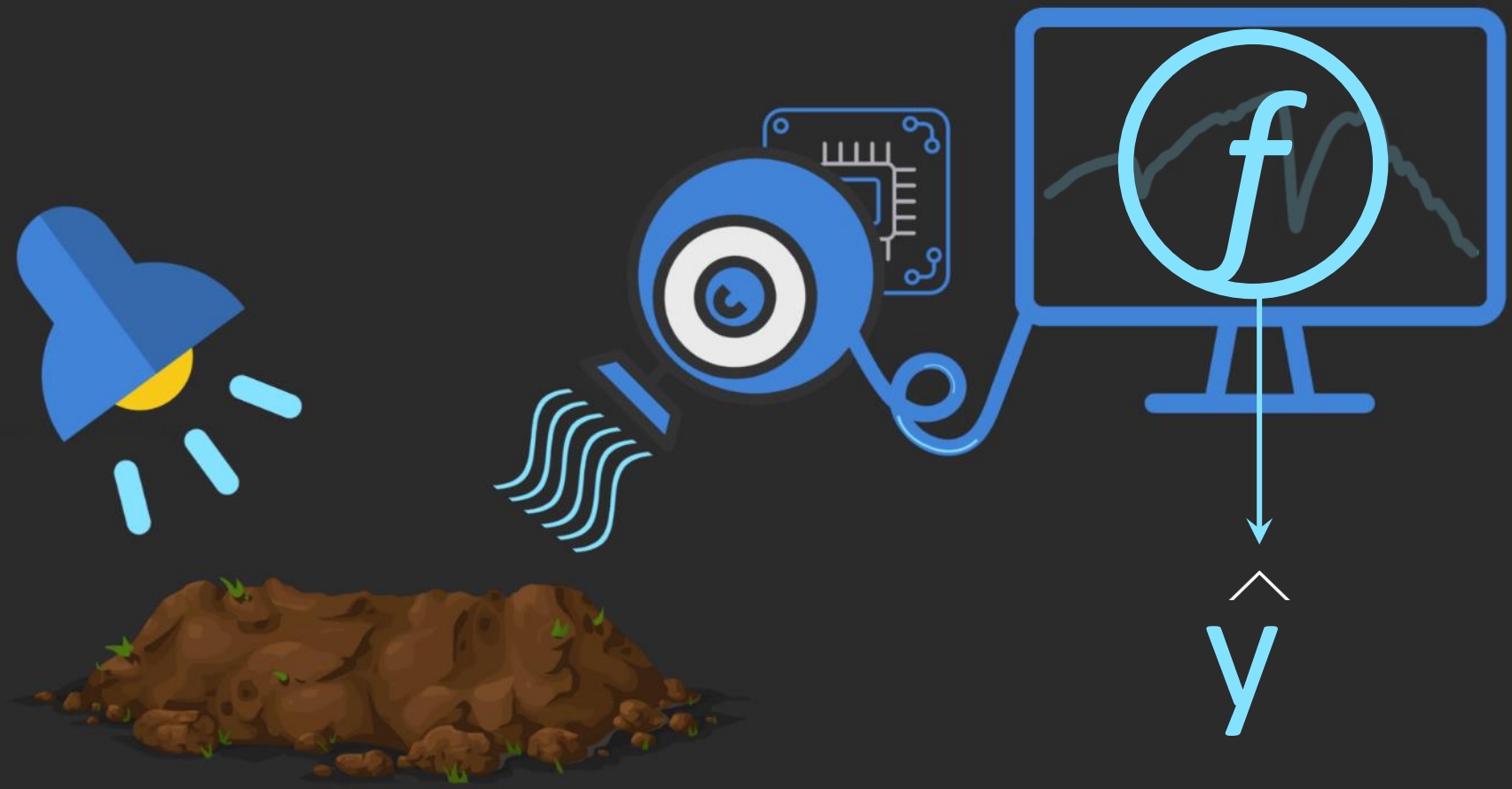




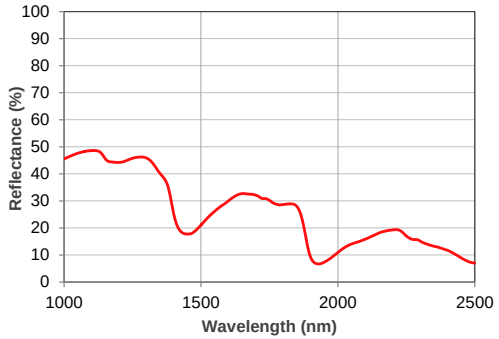




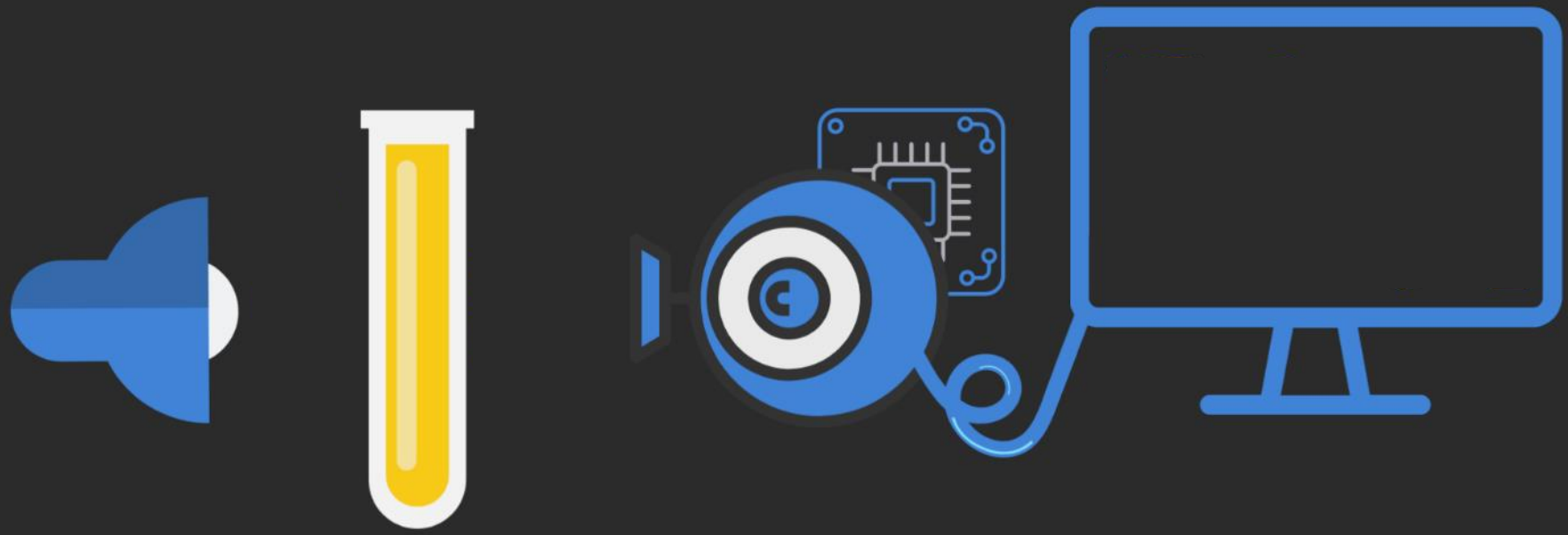


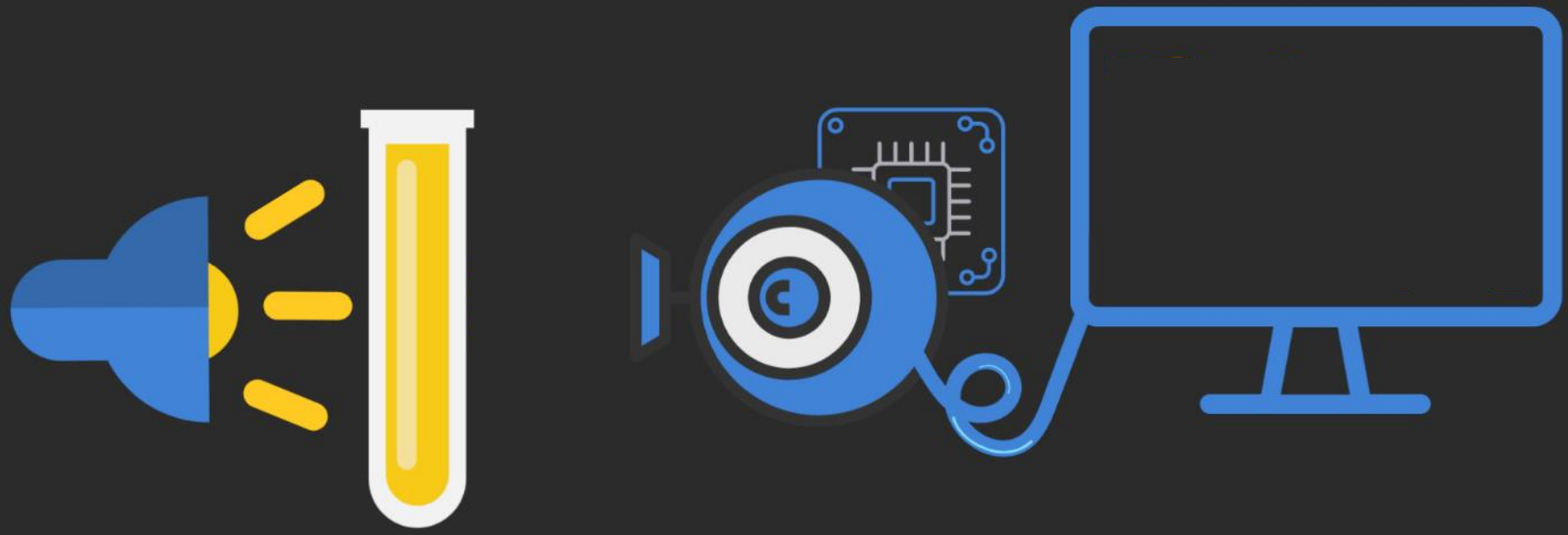

f
ŷ

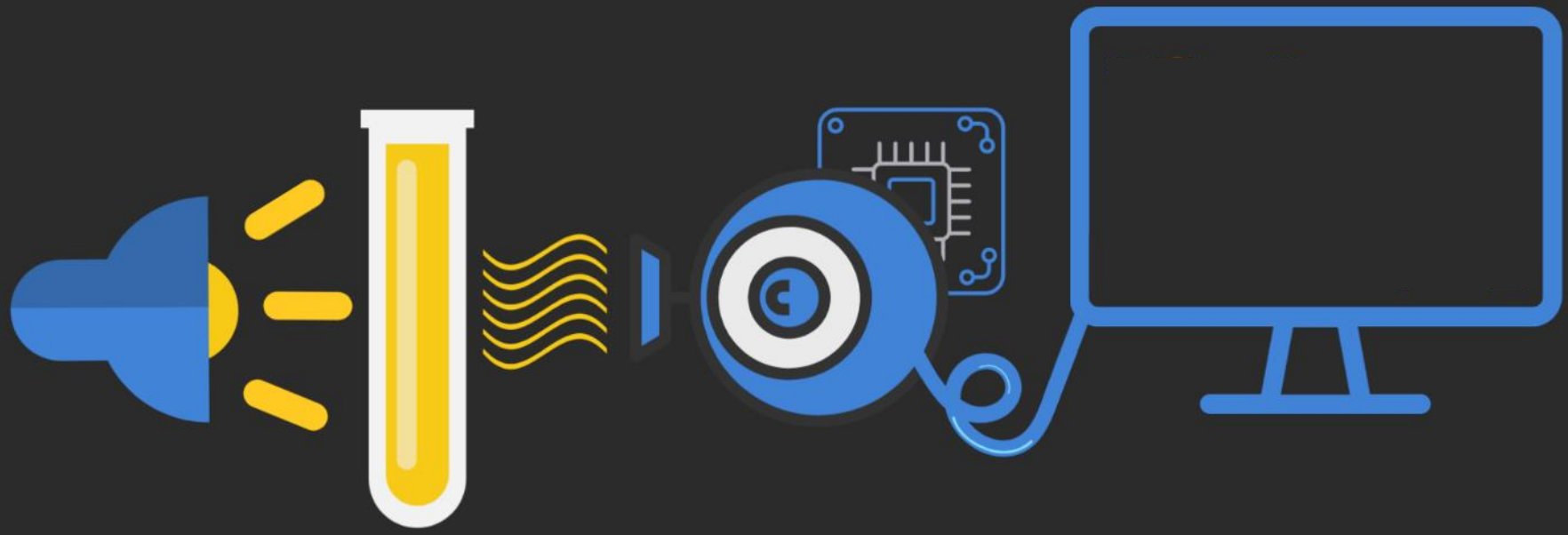
translating spectral information that we have into chemical information that users need!

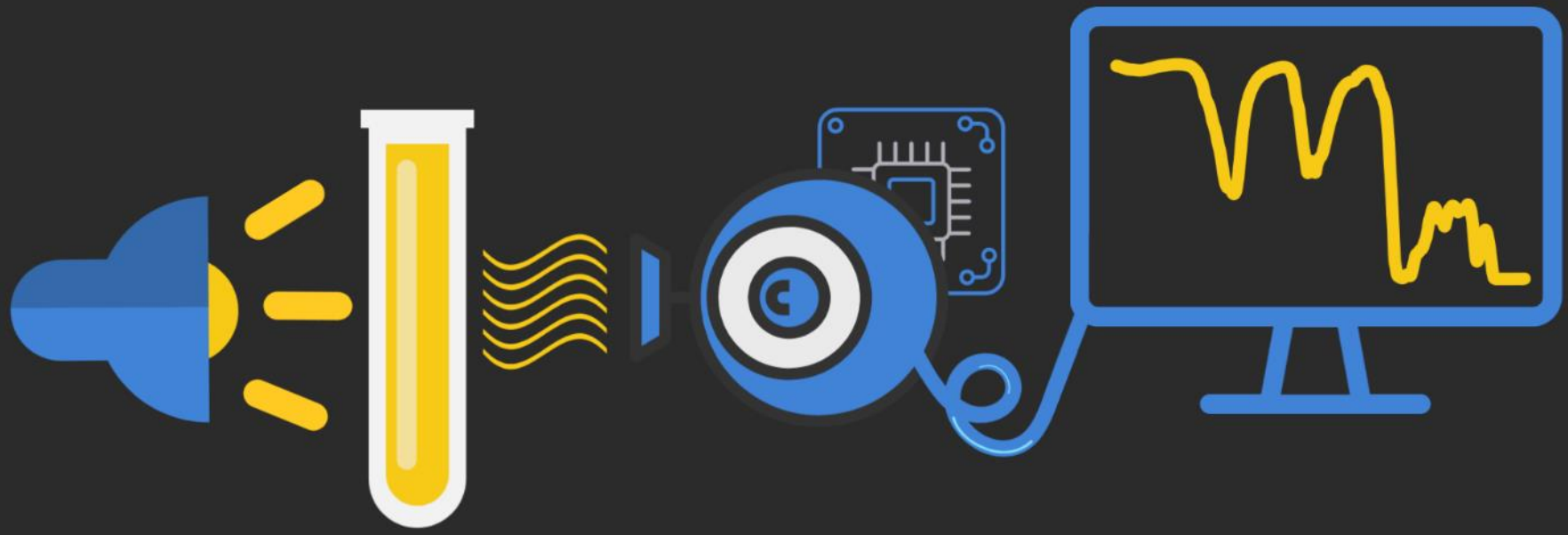
$$\hat{y} = f(\text{  }, \ldots) + \varepsilon$$

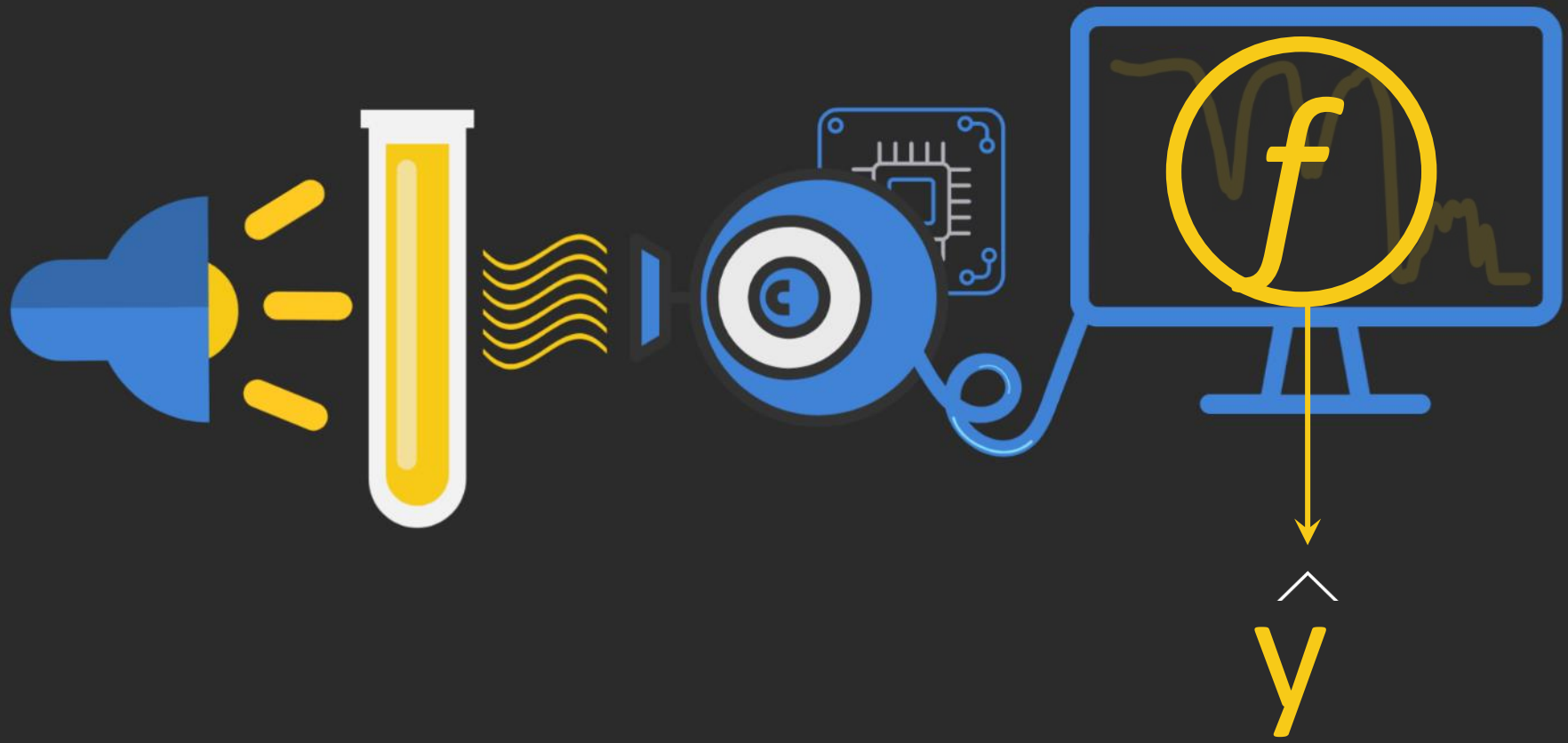
Function that can be approximated by using chemometrics/machine learning



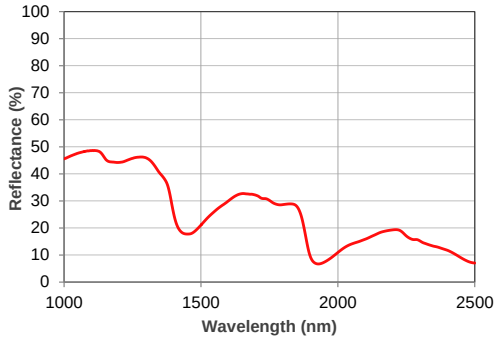







f
ŷ

translating spectral information that we have into chemical information that users need!

$$\hat{y} = f(\text{  }, \ldots) + \varepsilon$$

Function that can be approximated by using chemometrics/machine learning

transmission

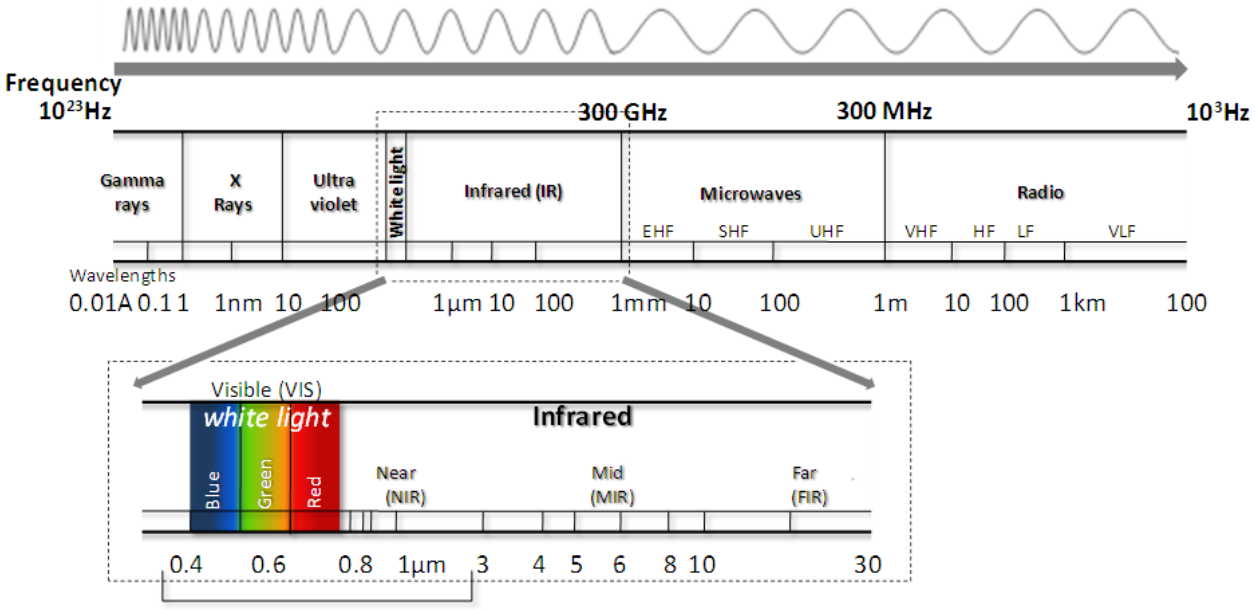
reflection

absorbance

transmission

reflection

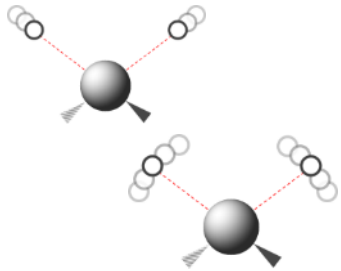
absorbance $[\log_{10}(1/\text{reflection})]$



NIR/IR spectroscopy is based on the measurements of the spectral response of objects in the infrared region of the electromagnetic spectrum.

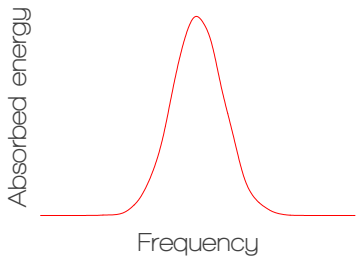
Near-infrared spectroscopy (NIR)...

How does it work?



A given molecule absorbs energy (light) if the frequency of the light energy matches the vibrational frequency of the functional group of the molecule.

The amount of energy absorbed can be different for each frequency (wavenumber). This usually results in energy absorption peaks.




ProxiMate


NIR-Online


TerraSpec

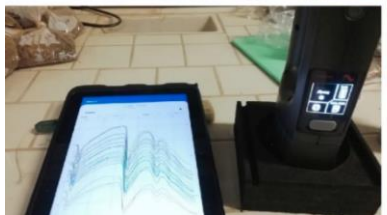

TANGO


MPA II
BRUKER


FOSS

Device name	Manufacturer	Spectral range (nm)	Price (USD)
		Bechtop	
ASD FieldSpec*	Malvern Panalytical	350 – 2500	~68100
NIRS DS2500 L	FOSS Analytics	400 – 2500	~75000
		Handheld	
SR-3500	Spectral Evolution	350 - 2500	~49700
MicroNIR 1700	VIAVI Solutions	950 - 1650	~20000
trinamiX	trinamiX/BASF	1450 - 2450	~8733
NeoSpectra	Si-Ware Systems	1300 - 2600	~3550
Agrocares	Agrocares	1300 - 2600	~6800
STS-VIS	Ocean Insight	350 - 830	~3408
NIRONE Sensor S	Spectral Engines	1750 - 2150	~3635
microPHAZIR	Thermo Fisher	1596 - 2396	?
SCiO	Consumer Physics	740 - 1070	~450
NIRscan	Texas Instruments	900 - 1700	~3000
Tellspec	Tellspec	900 - 1700	~320 - 1500

*a portable option is also available




NeoSpectra

trinamix

Spectral Evolution


ASD FieldSpec


FOSS


Agrocares

Device name	Manufacturer	Spectral range (nm)	Price (USD)
		Bechtop	
ASD FieldSpec*	Malvern Panalytical	350 – 2500	~68100
NIRS DS2500 L	FOSS Analytics	400 – 2500	~75000
		Handheld	
SR-3500	Spectral Evolution	350 - 2500	~49700
MicroNIR 1700	VIAVI Solutions	950 - 1650	~20000
trinamiX	trinamiX/BASF	1450 - 2450	~8733
NeoSpectra	Si-Ware Systems	1300 - 2600	~3550
Agrocares	Agrocares	1300 - 2600	~6800
STS-VIS	Ocean Insight	350 - 830	~3408
NIRONE Sensor S	Spectral Engines	1750 - 2150	~3635
microPHAZIR	Thermo Fisher	1596 - 2396	?
SCiO	Consumer Physics	740 - 1070	~450
NIRscan	Texas Instruments	900 - 1700	~3000
Tellspec	Tellspec	900 - 1700	~320 - 1500

*a portable option is also available


NeoSpectra

trinamix

Spectral Evolution


ASD FieldSpec


FOSS


Agrocares

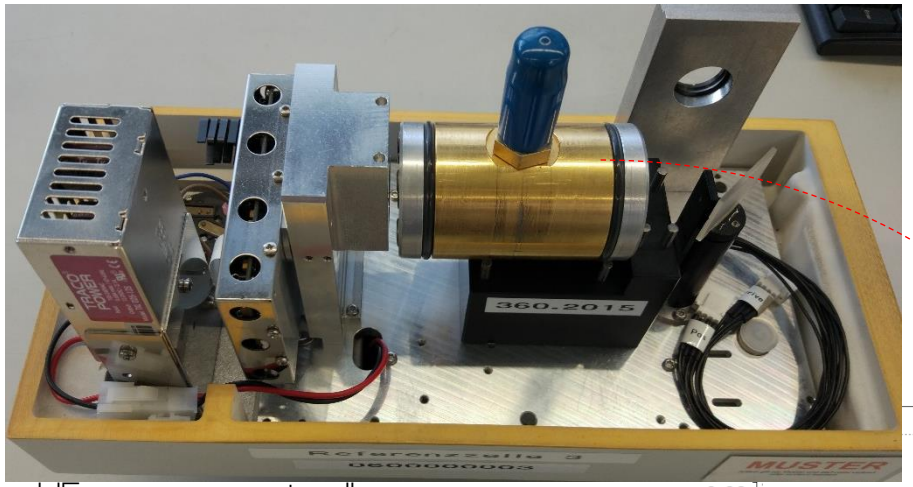
Photometric alignment/calibration

Photometric alignment/calibration

Wavelength alignment

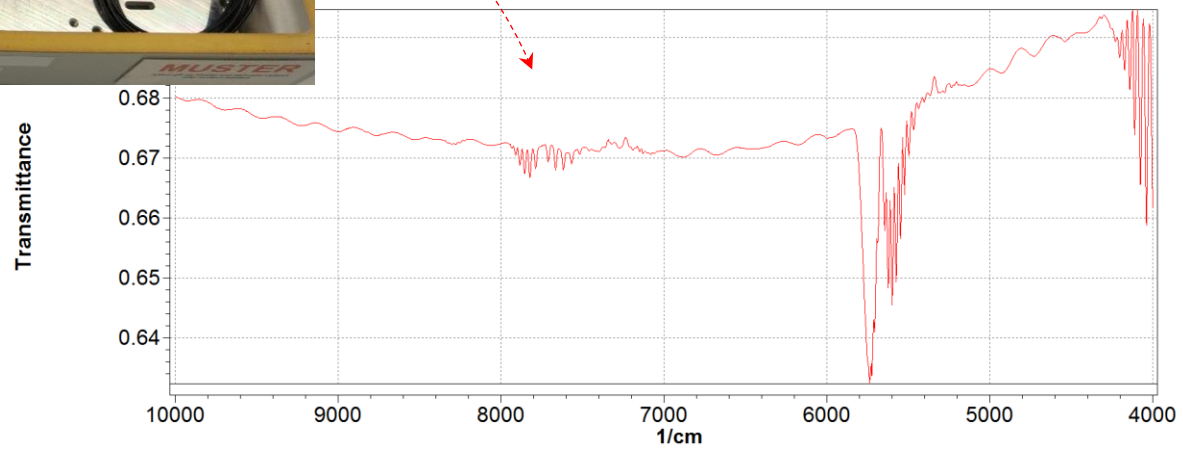
1. Hydrogen fluoride alignment

Some instruments are factory standardized using a wavelength standardization cell containing Hydrogen Fluoride (HF) gas. The spectral signature of this gas consist in very well defined peaks which can be measured with high precision and accuracy.



HF measurement cell

HF spectrum for wavelength alignment

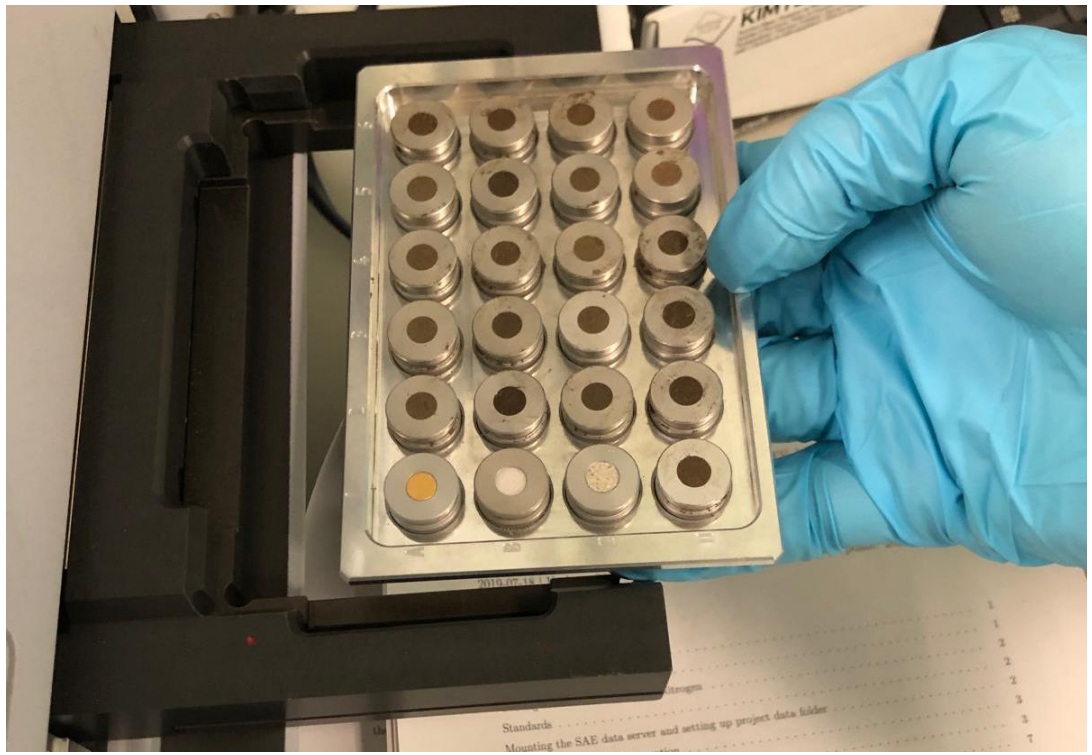


Let’s scan our first sample

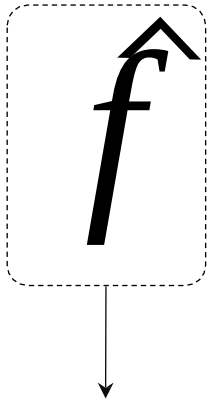
Download Neo-Spectra COLLECT

Download Link GooglePlay

Scanning in the IR spectrometer…



NIR Applications

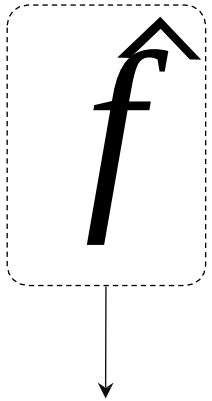


Function that can be approximated by using chemometrics/machine learning

We currently use our own BUCHI software for that!

These functions are device-, product- and property-specific.

NIR Applications



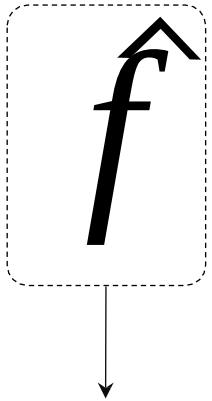
Function that can be approximated by using chemometrics/machine learning

We currently use our own BUCHI software for that!

These **functions are device-, product- and property-specific.**

For each product we build and group a set of functions that quantify relevant product properties (e.g. moisture, protein, fat). **This group of functions is called application.**

NIR Applications

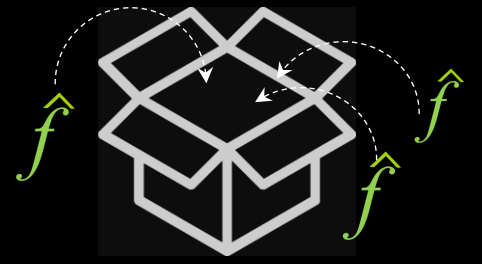


Function that can be approximated by using chemometrics/machine learning

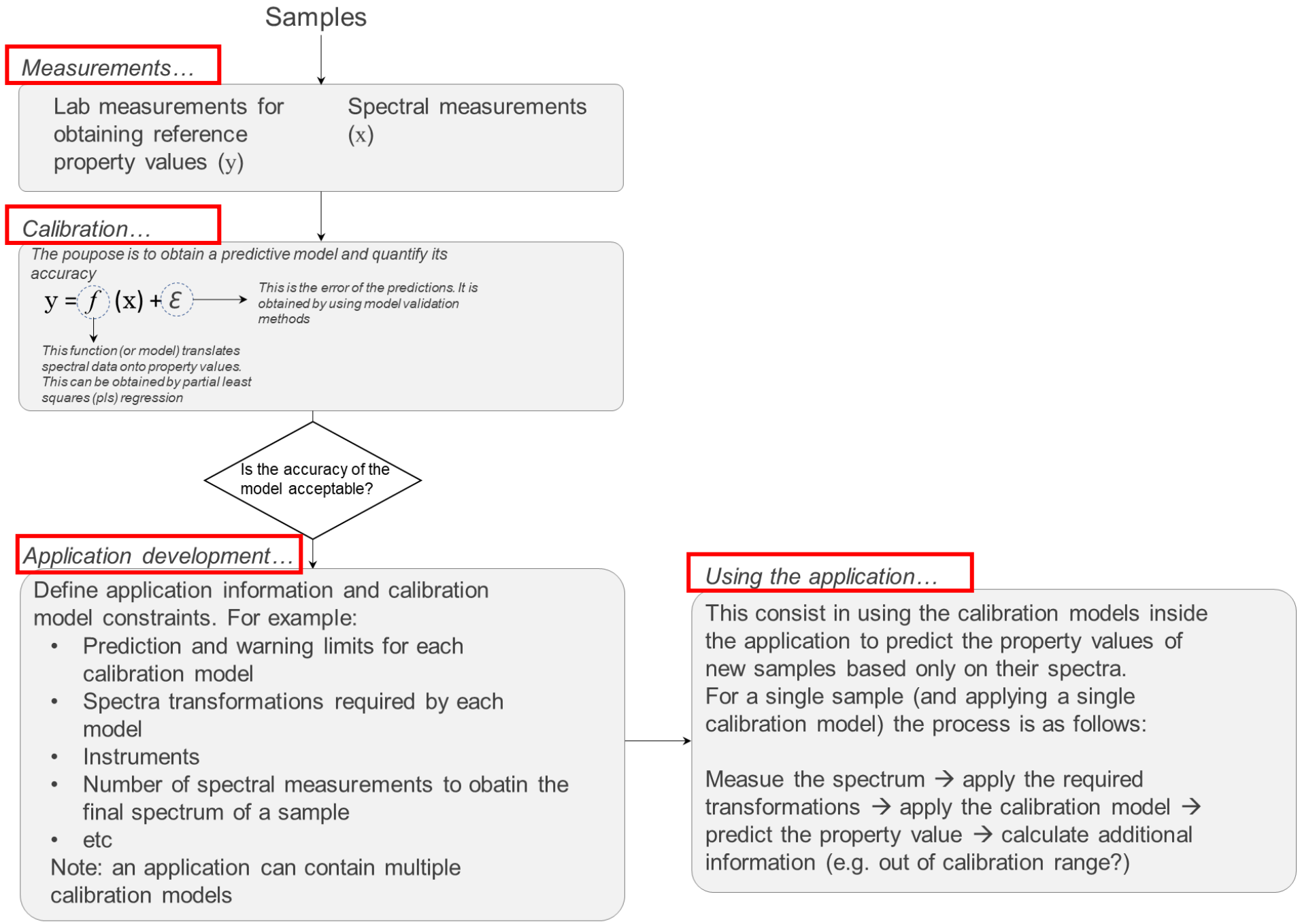
We currently use our own BUCHI software for that!

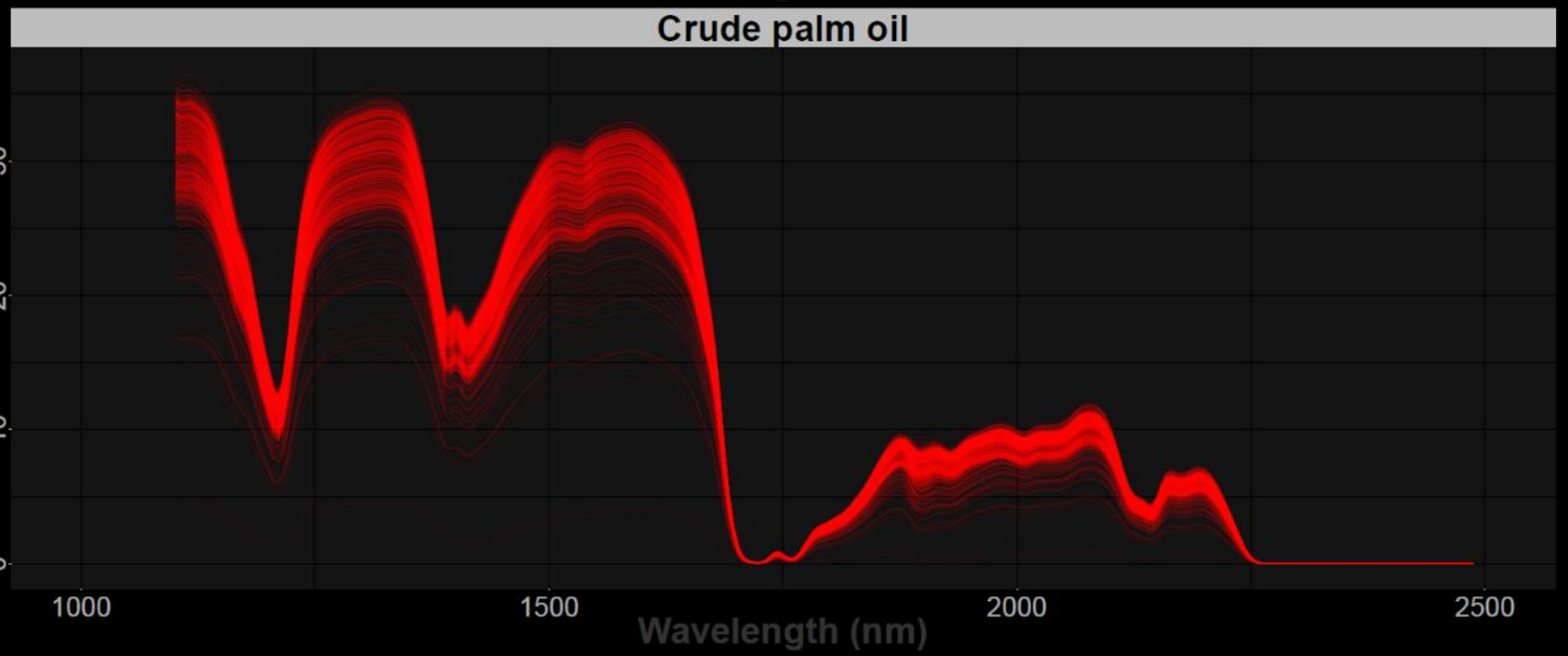
These **functions are device-, product- and property-specific.**

For each product we build and group a set of functions that quantify relevant product properties (e.g. moisture, protein, fat). **This group of functions is called application.**

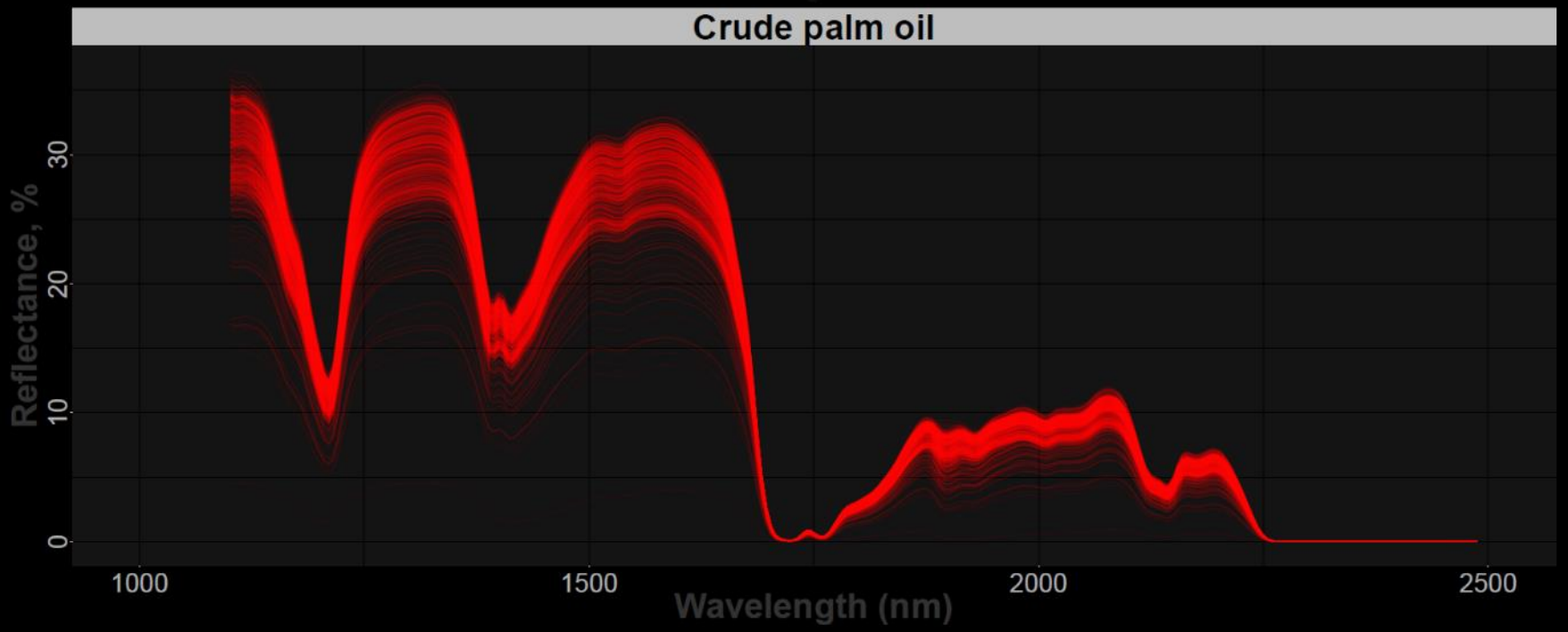


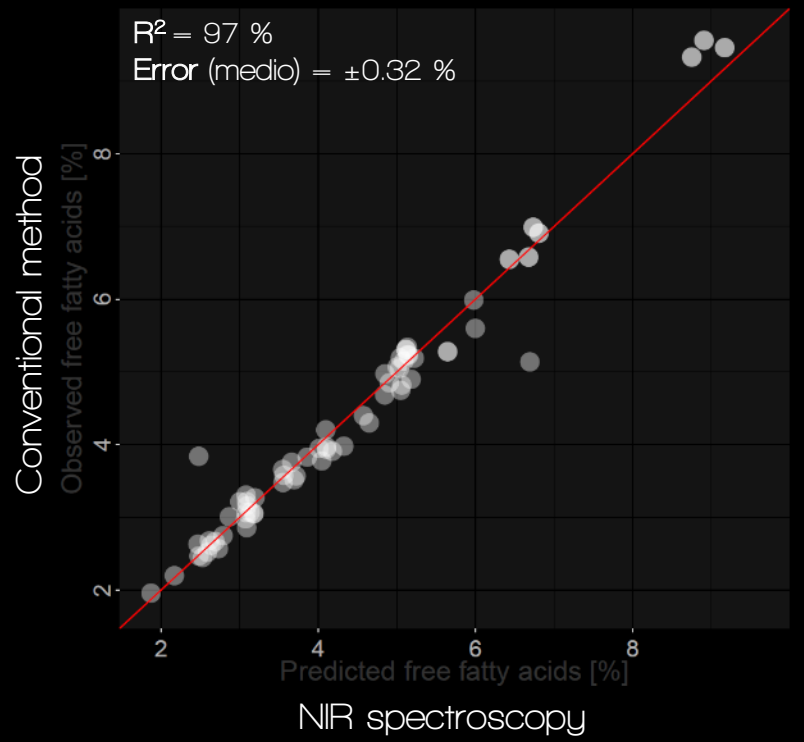
From average user's perspective, **a NIR device without an application can be perceived as useless.**

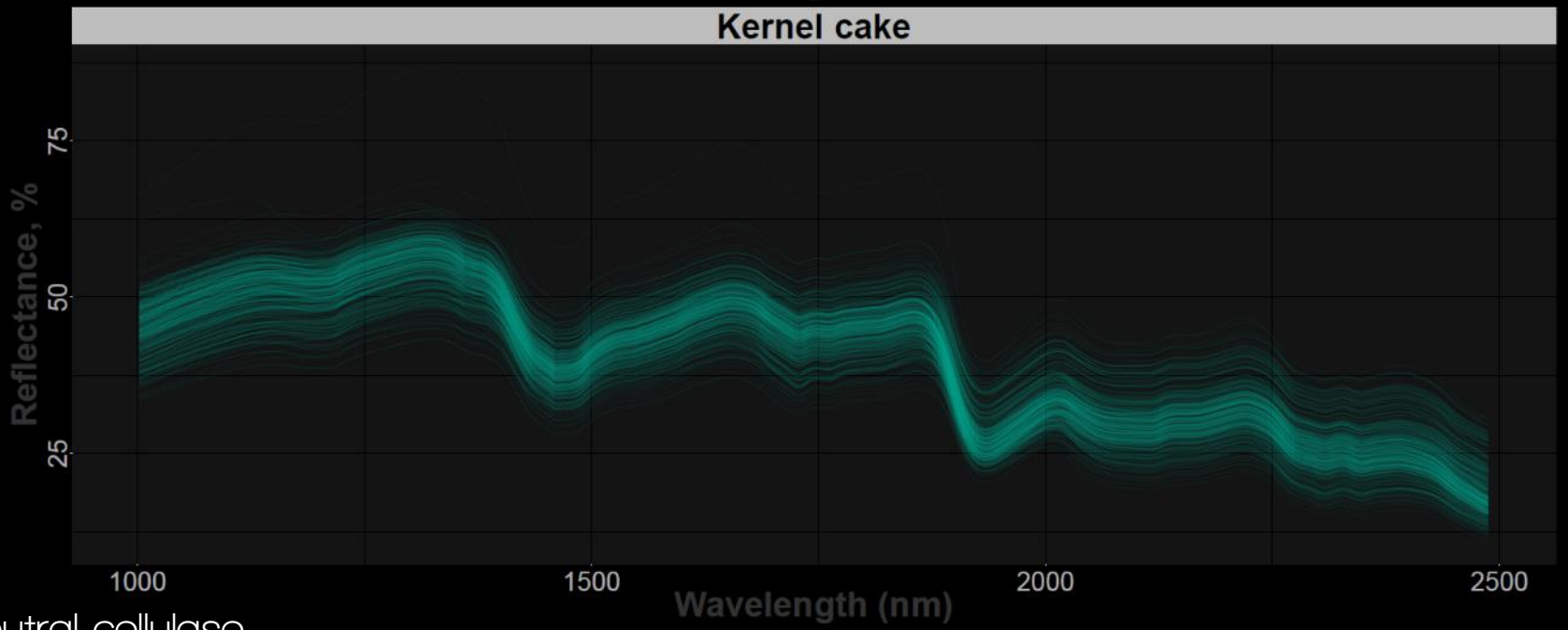

Samples
Measurements…
Lab measurements for obtaining reference property values (y)
Spectral measurements (x)
Calibration…
The poupose is to obtain a predictive model and quantify its accuracy
y = f (x) + ε
This is the error of the predictions. It is obtained by using model validation methods
This function (or model) translates spectral data onto property values. This can be obtained by partial least squares (pls) regression
Is the accuracy of the model acceptable?
Application development…
Define application information and calibration model constraints. For example:
• Prediction and warning limits for each calibration model
• Spectra transformations required by each model
• Instruments
• Number of spectral measurements to obatin the final spectrum of a sample
• etc
Note: an application can contain multiple calibration models
Using the application…
This consist in using the calibration models inside the application to predict the property values of new samples based only on their spectra.
For a single sample (and applying a single calibration model) the process is as follows:
Measue the spectrum → apply the required transformations → apply the calibration model → predict the property value → calculate additional information (e.g. out of calibration range?)


Crude palm oil
Reflectance, %
0
10
20
30
1000
1500
2000
2500
Wavelength (nm)

Free fatty acids

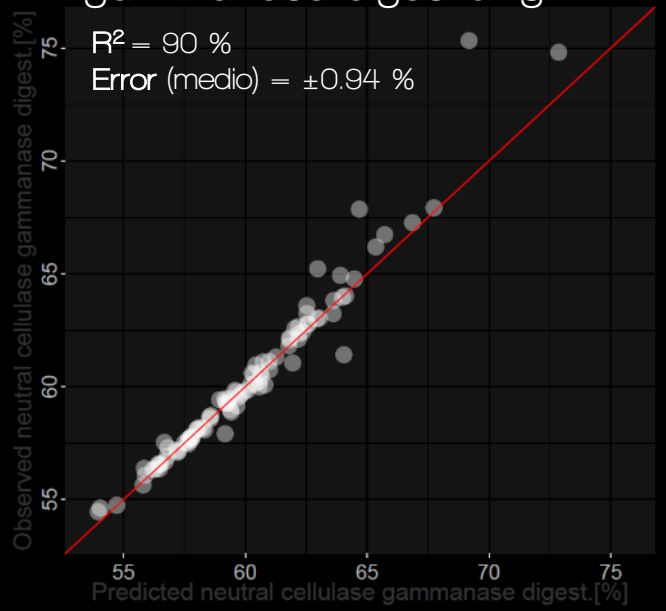




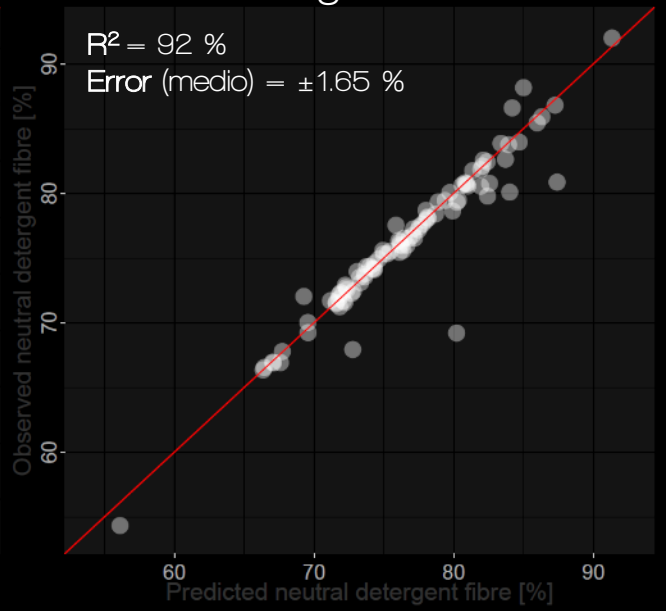


Conventional method

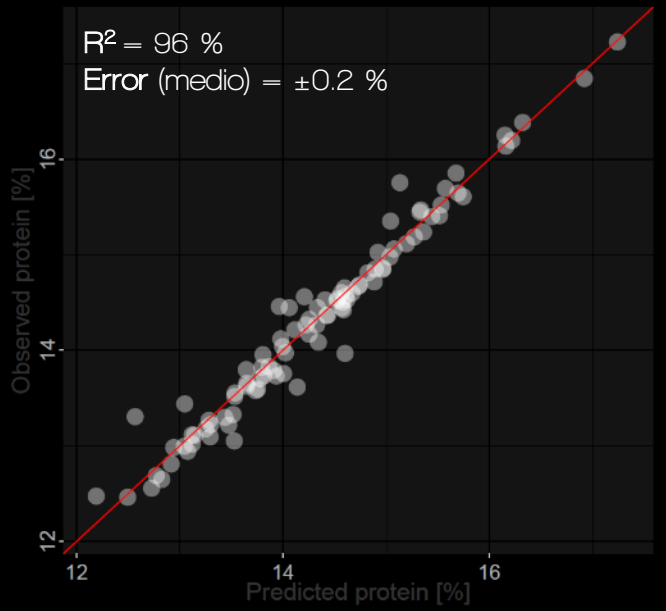
Neutral cellulase gammanase digestibility



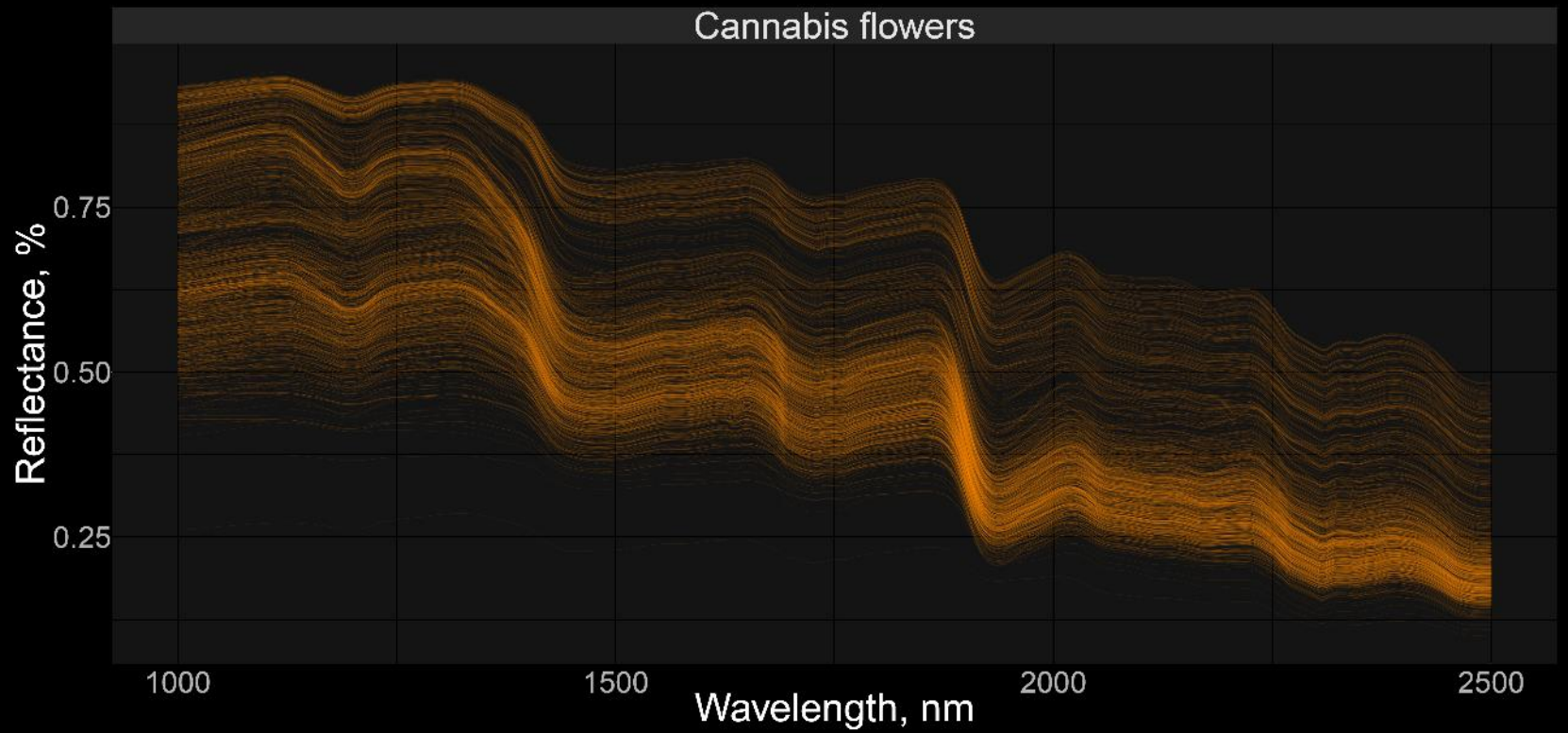
Neutral detergent fibre

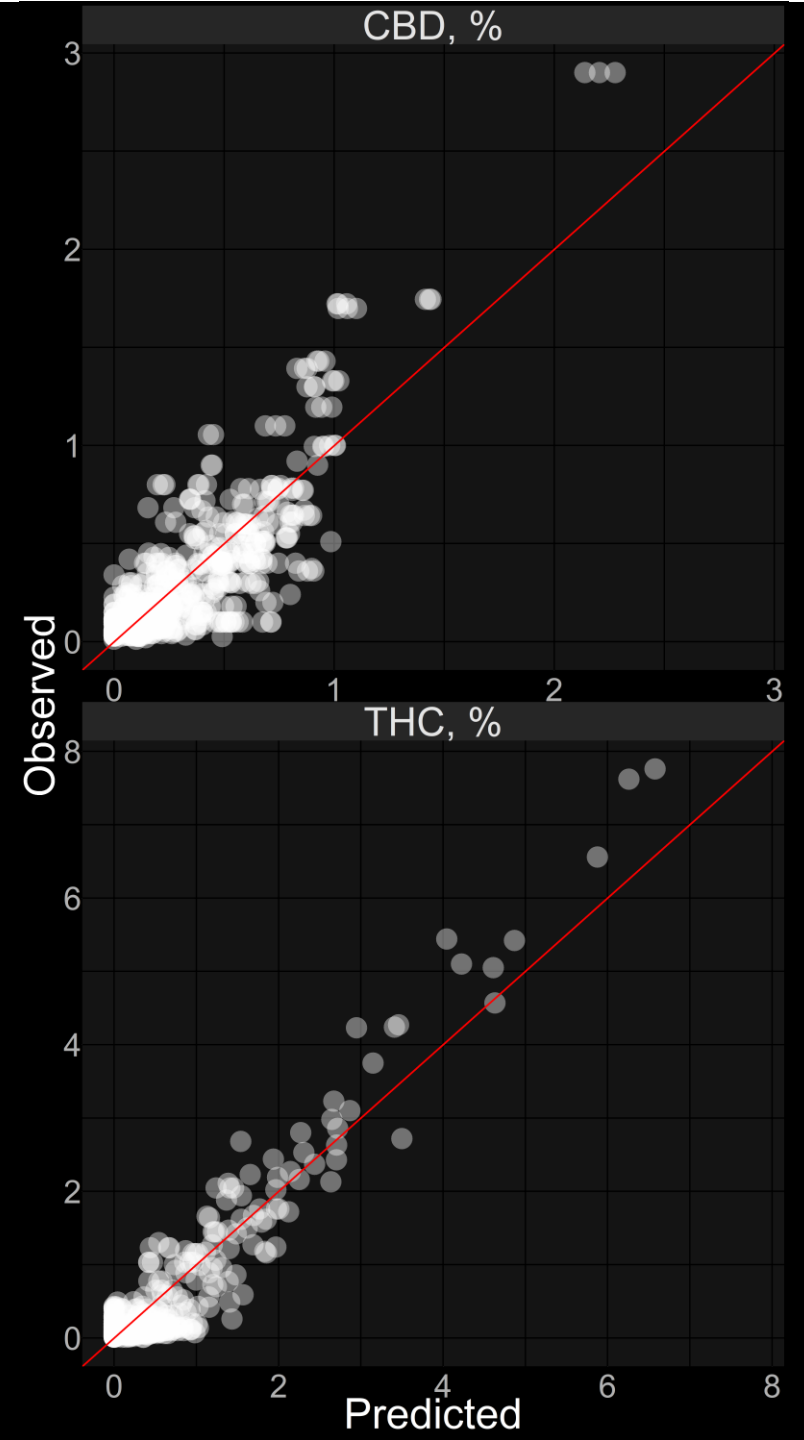


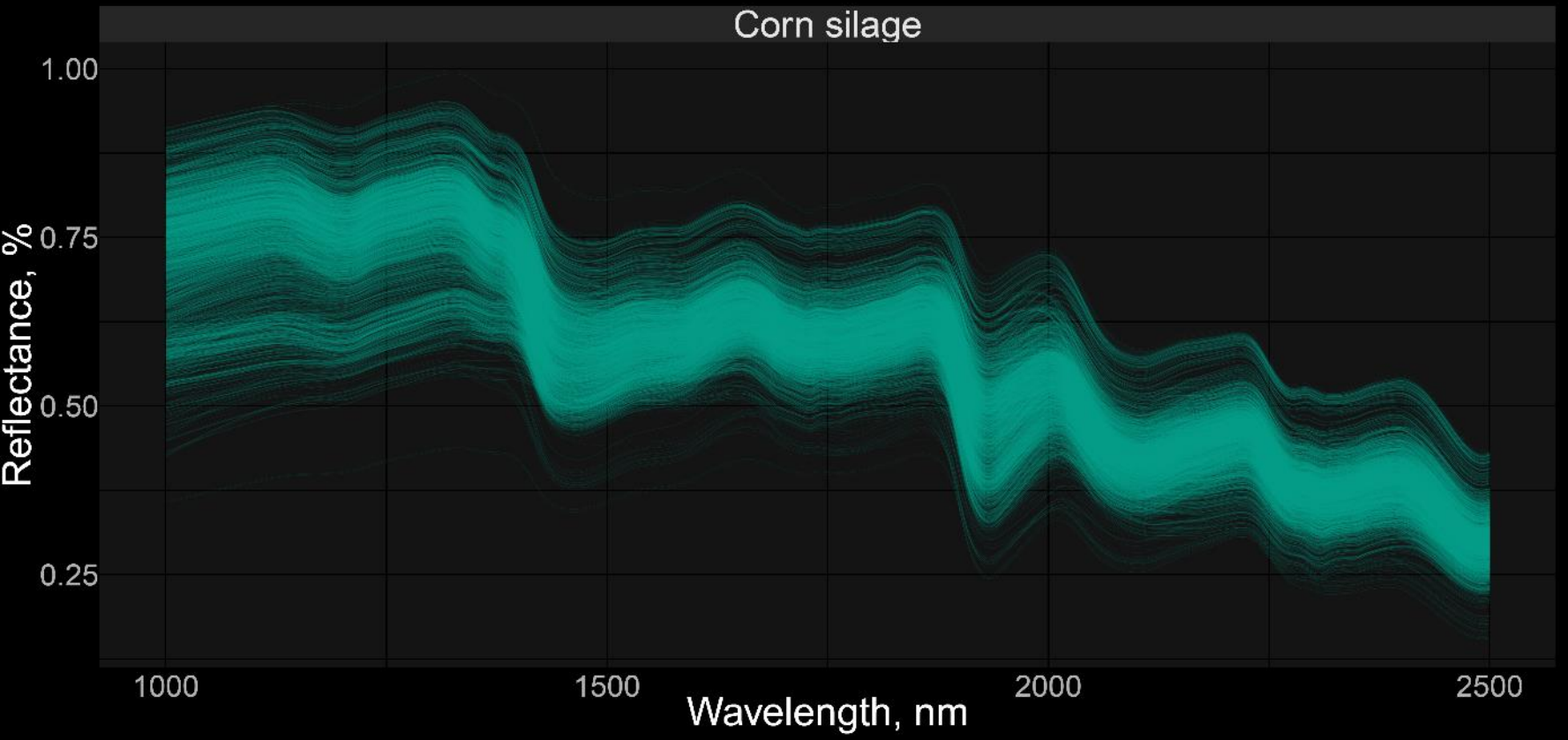
Protein

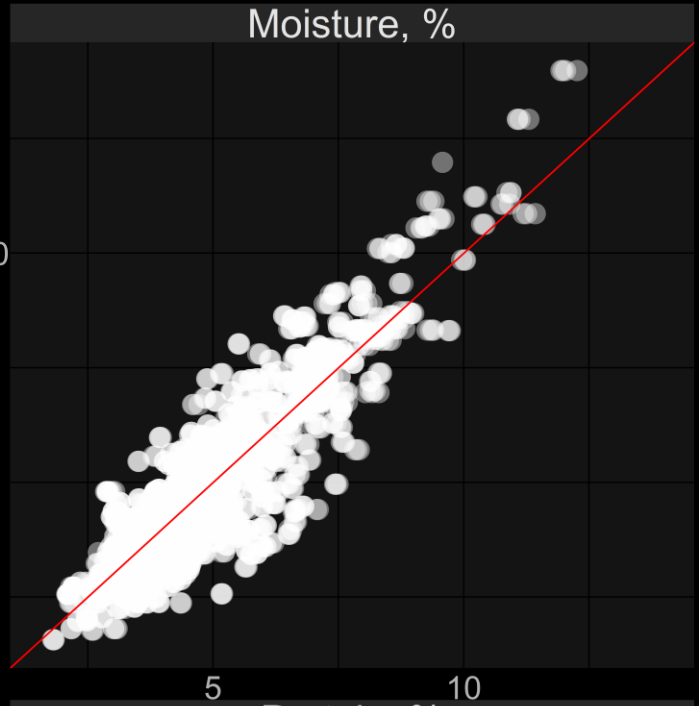


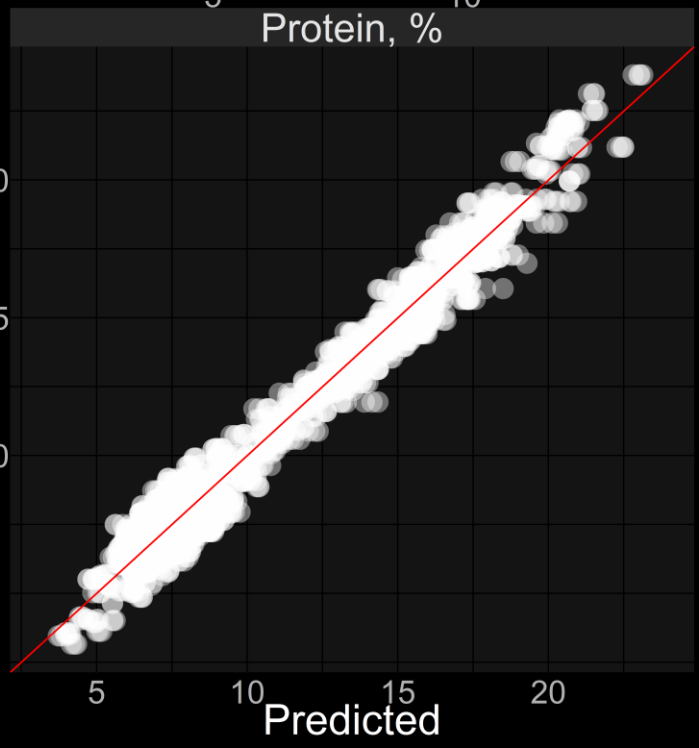
NIR spectroscopy


Cannabis flowers
Reflectance, %
0.75
0.50
0.25
1000
1500
2000
2500
Wavelength, nm

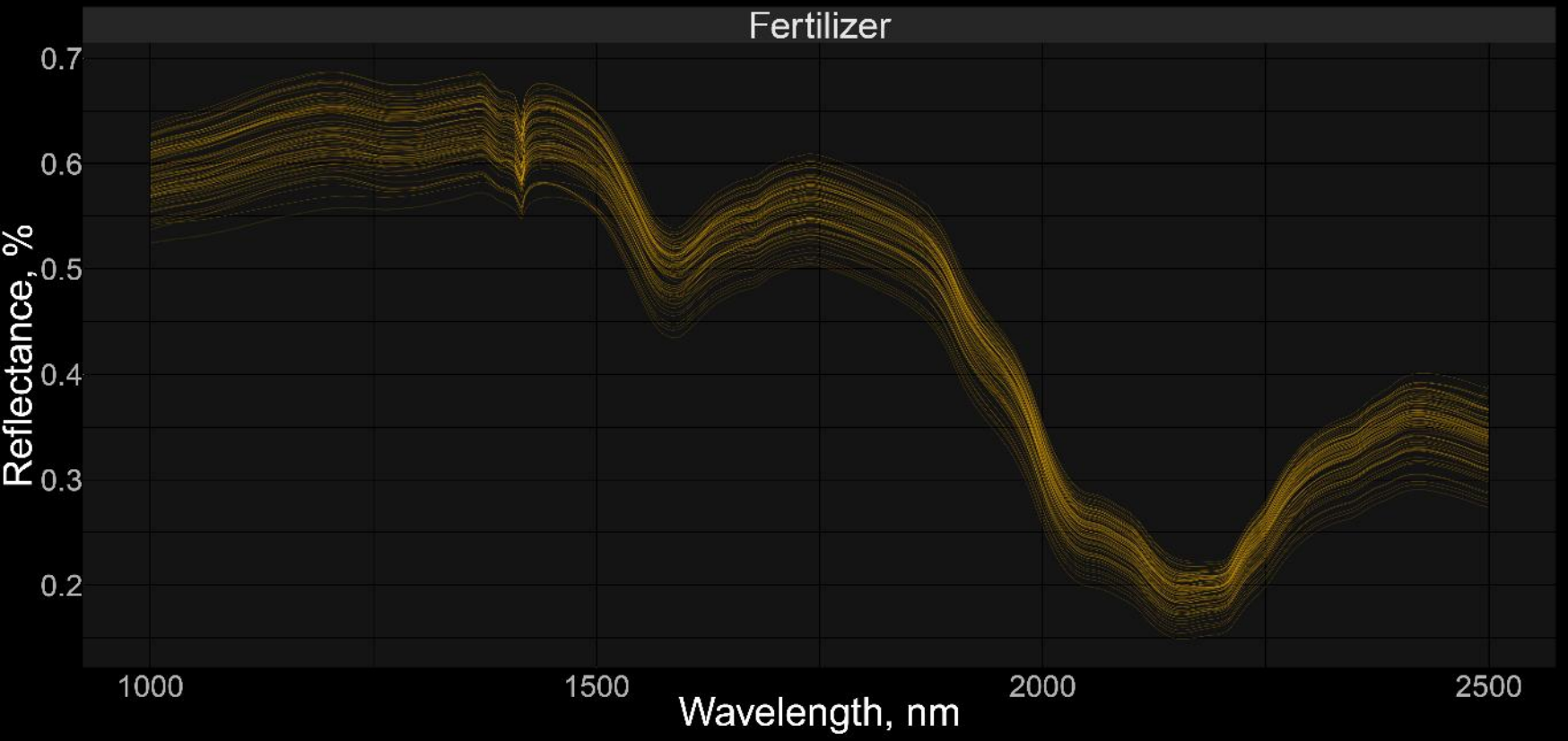

CBD, %
THC, %
Observed
Predicted
0
1
2
3
0
2
4
6
8

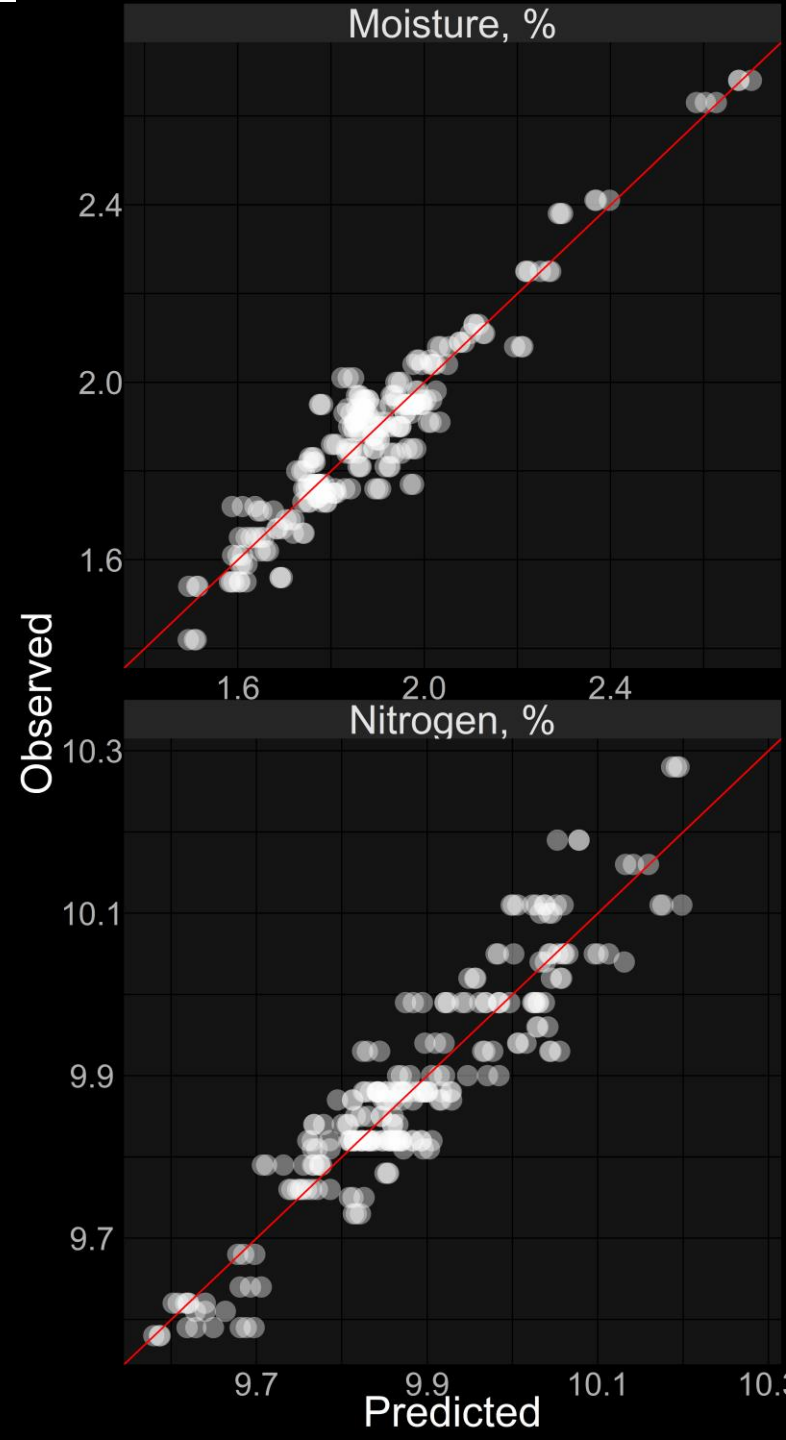

Corn silage
1.00
0.75
0.50
0.25
Reflectance, %
1000
1500
2000
2500
Wavelength, nm

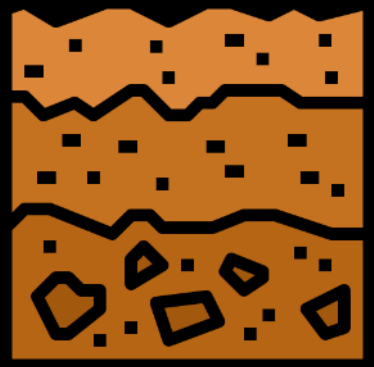

Moisture, %
10
5
5
10

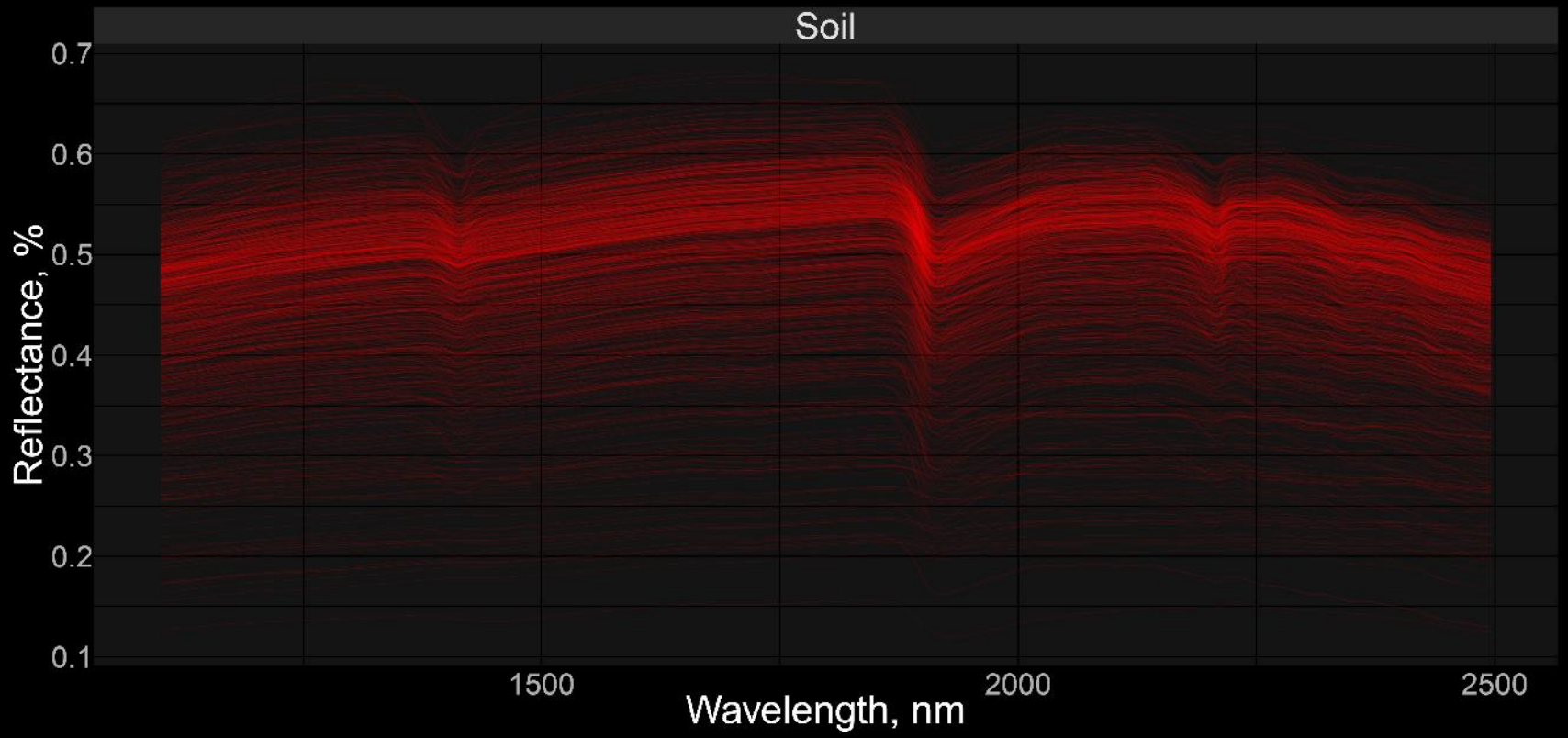

Protein, %
20
15
10
5
5
10
15
20
Predicted

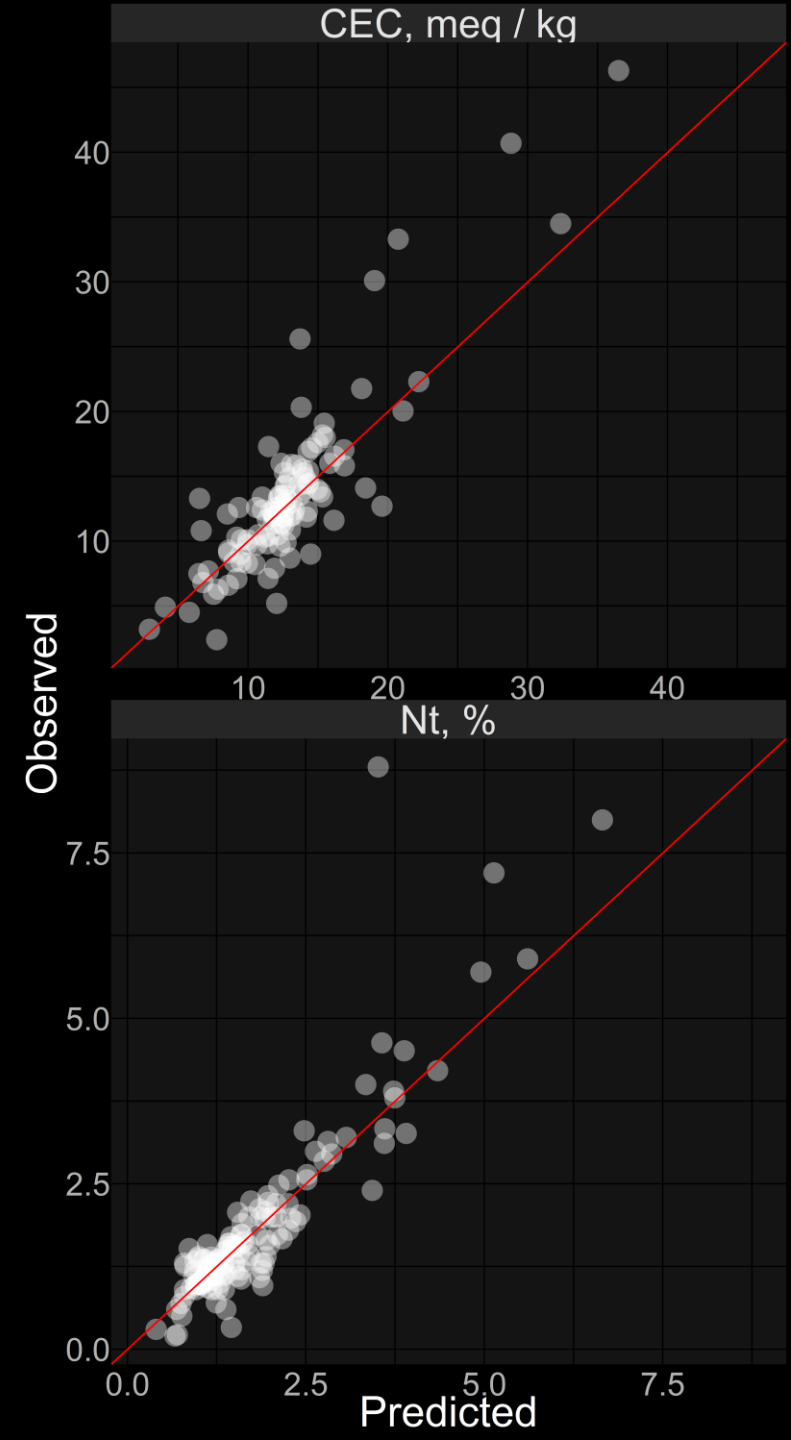
Observed


Fertilizer
Reflectance, %
0.7
0.6
0.5
0.4
0.3
0.2
1000
1500
2000
2500
Wavelength, nm


Moisture, %
2.4
2.0
1.6
1.6
2.0
2.4
Nitrogen, %
10.3
10.1
9.9
9.7
Observed
9.7
9.9
10.1
Predicted




Soil
Reflectance, %
0.7
0.6
0.5
0.4
0.3
0.2
0.1
1500
2000
2500
Wavelength, nm


CEC, meq / kg
40
30
20
10
10
20
30
40
Nt, %
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
Observed
Predicted

NIR/IR Spectroscopy – An equation for success

$$\text{Success} = f(\text{hardware, software, data, models, people, automation, ...})$$

Application Development Roadmap

Collect calibration data

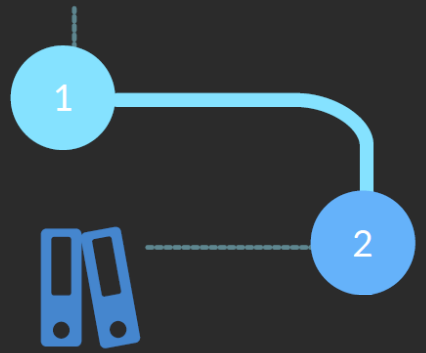
Legacy and/or new data to develop calibrations is collected.

1

Application Development Roadmap

Collect calibration data

Legacy and/or new data to develop calibrations is collected.



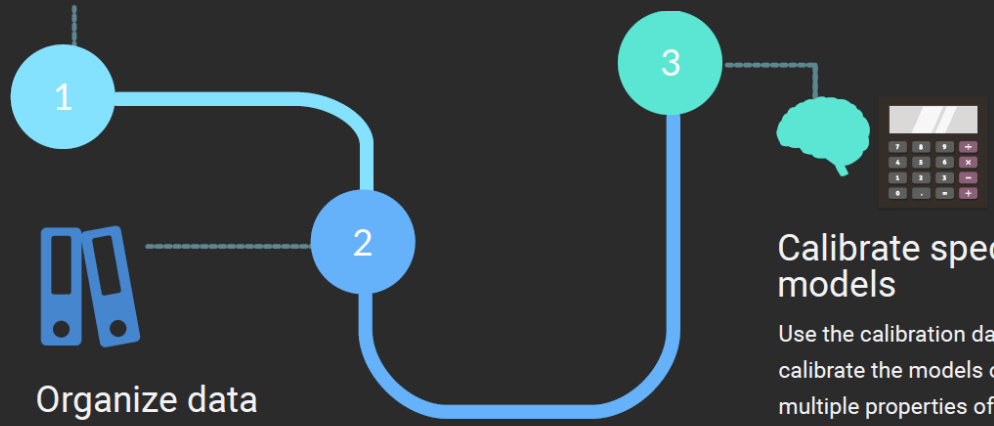
Organize data

cleaning/filtering, sorting, harmonizing, pooling, storing and document all the calibration data gathered.

Application Development Roadmap

Collect calibration data

Legacy and/or new data to develop calibrations is collected.



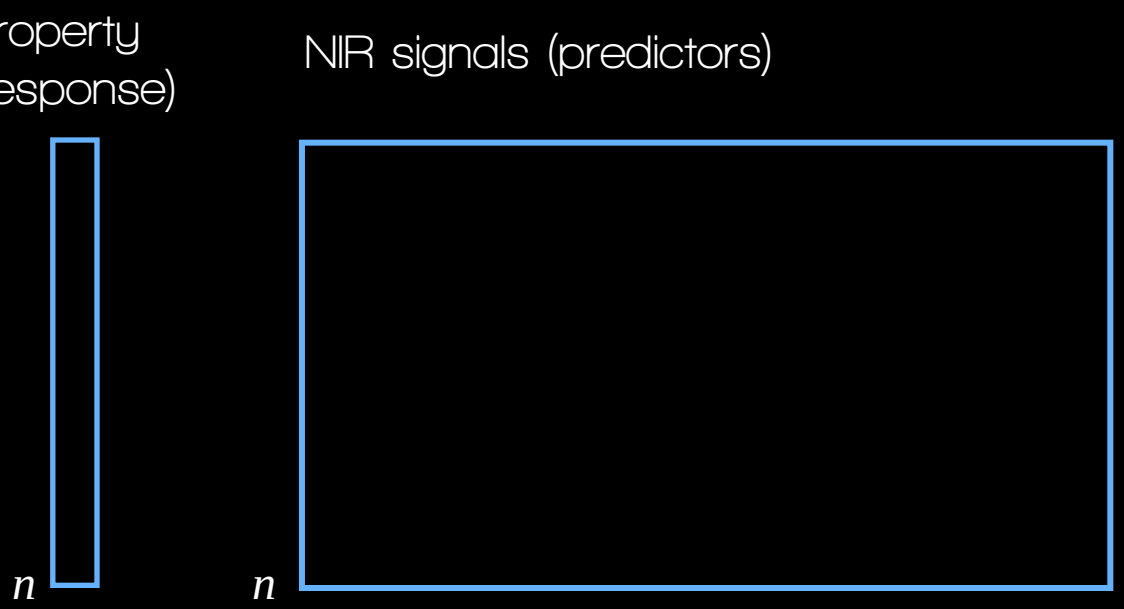
Organize data

cleaning/filtering, sorting, harmonizing, pooling, storing and document all the calibration data gathered.

Calibrate spectral models

Use the calibration data to calibrate the models of the multiple properties of the target product. For each model, several data processing methods are empirically tested to optimize the accuracy and precision of the models.

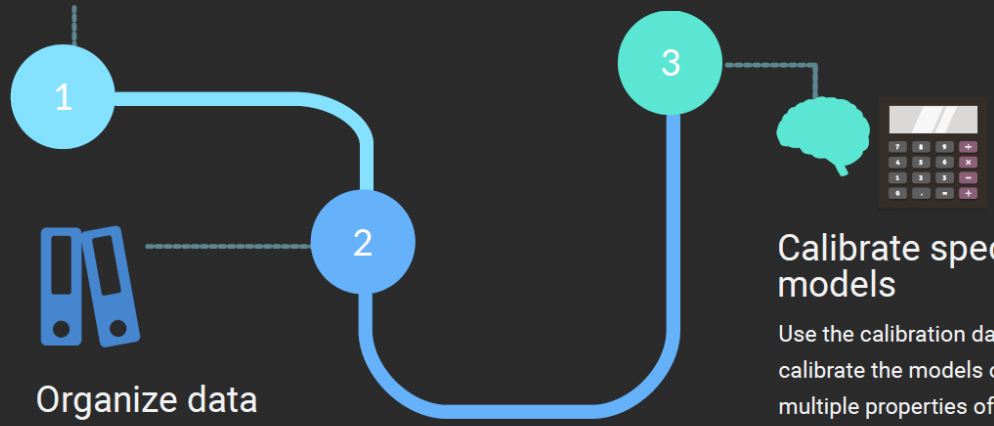
The training/calibration set



Application Development Roadmap

Collect calibration data

Legacy and/or new data to develop calibrations is collected.



Organize data

cleaning/filtering, sorting, harmonizing, pooling, storing and document all the calibration data gathered.

Calibrate spectral models

Use the calibration data to calibrate the models of the multiple properties of the target product. For each model, several data processing methods are empirically tested to optimize the accuracy and precision of the models.

The training/calibration set

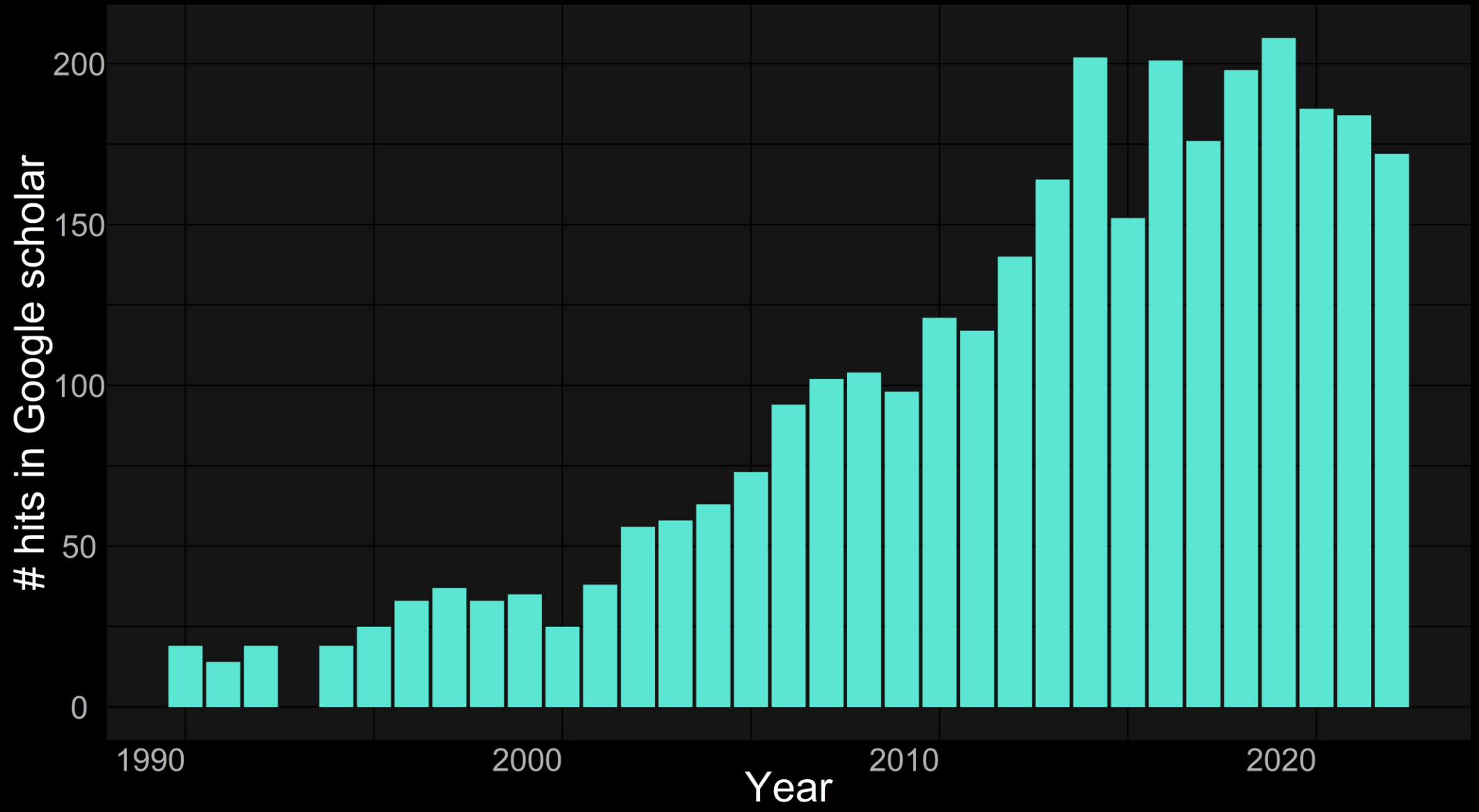
Property (response)

NIR signals (predictors)

The calibration model $\hat{f}$

33 years

~3166 scientific publications



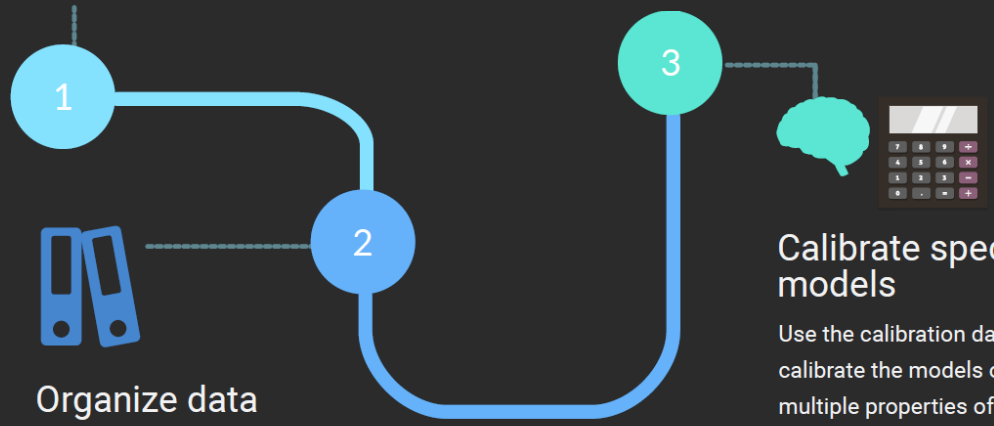
Search terms:

`Soil + spectroscopy` (in the title of the paper)

Application Development Roadmap

Collect calibration data

Legacy and/or new data to develop calibrations is collected.



Organize data

cleaning/filtering, sorting, harmonizing, pooling, storing and document all the calibration data gathered.

Calibrate spectral models

Use the calibration data to calibrate the models of the multiple properties of the target product. For each model, several data processing methods are empirically tested to optimize the accuracy and precision of the models.

and then?

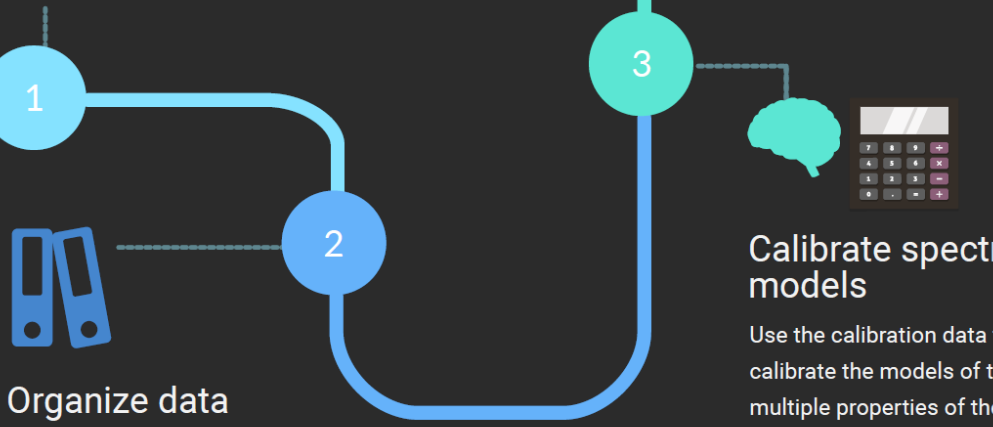
Application Development Roadmap

Pack into a predictive application

All the spectral models are packed into an application. All the technical and user-required documentation is generated along with the necessary licensing information.

Collect calibration data

Legacy and/or new data to develop calibrations is collected.



Organize data

cleaning/filtering, sorting, harmonizing, pooling, storing and document all the calibration data gathered.

Calibrate spectral models

Use the calibration data to calibrate the models of the multiple properties of the target product. For each model, several data processing methods are empirically tested to optimize the accuracy and precision of the models.

Application Development Roadmap

Pack into a predictive application

All the spectral models are packed into an application. All the technical and user-required documentation is generated along with the necessary licensing information.

Collect calibration data

Legacy and/or new data to develop calibrations is collected.

1

2

3

4

5

Organize data

cleaning/filtering, sorting, harmonizing, pooling, storing and document all the calibration data gathered.

Calibrate spectral models

Use the calibration data to calibrate the models of the multiple properties of the target product. For each model, several data processing methods are empirically tested to optimize the accuracy and precision of the models.

Validation test

A specialist (different from the application developer), test that the application can be used. In this step the application is also tested with real samples to verify if reasonable results returned.

Application Development Roadmap

Collect calibration data

Legacy and/or new data to develop calibrations is collected.

1

2

Organize data

cleaning/filtering, sorting, harmonizing, pooling, storing and document all the calibration data gathered.

3

Calibrate spectral models

Use the calibration data to calibrate the models of the multiple properties of the target product. For each model, several data processing methods are empirically tested to optimize the accuracy and precision of the models.

4

Pack into a predictive application

All the spectral models are packed into an application. All the technical and user-required documentation is generated along with the necessary licensing information.

5

Validation test

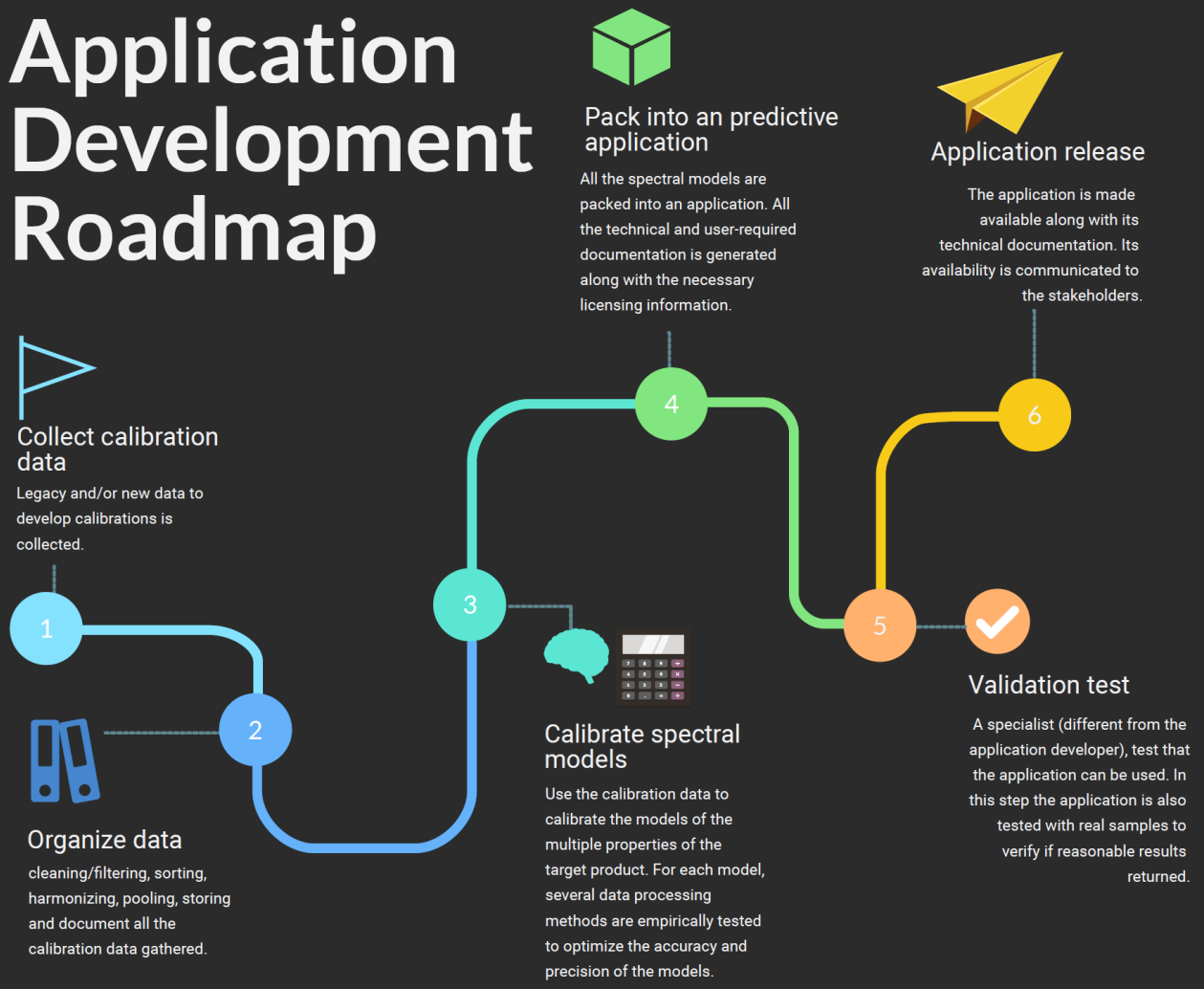
A specialist (different from the application developer), test that the application can be used. In this step the application is also tested with real samples to verify if reasonable results returned.

6

Application release

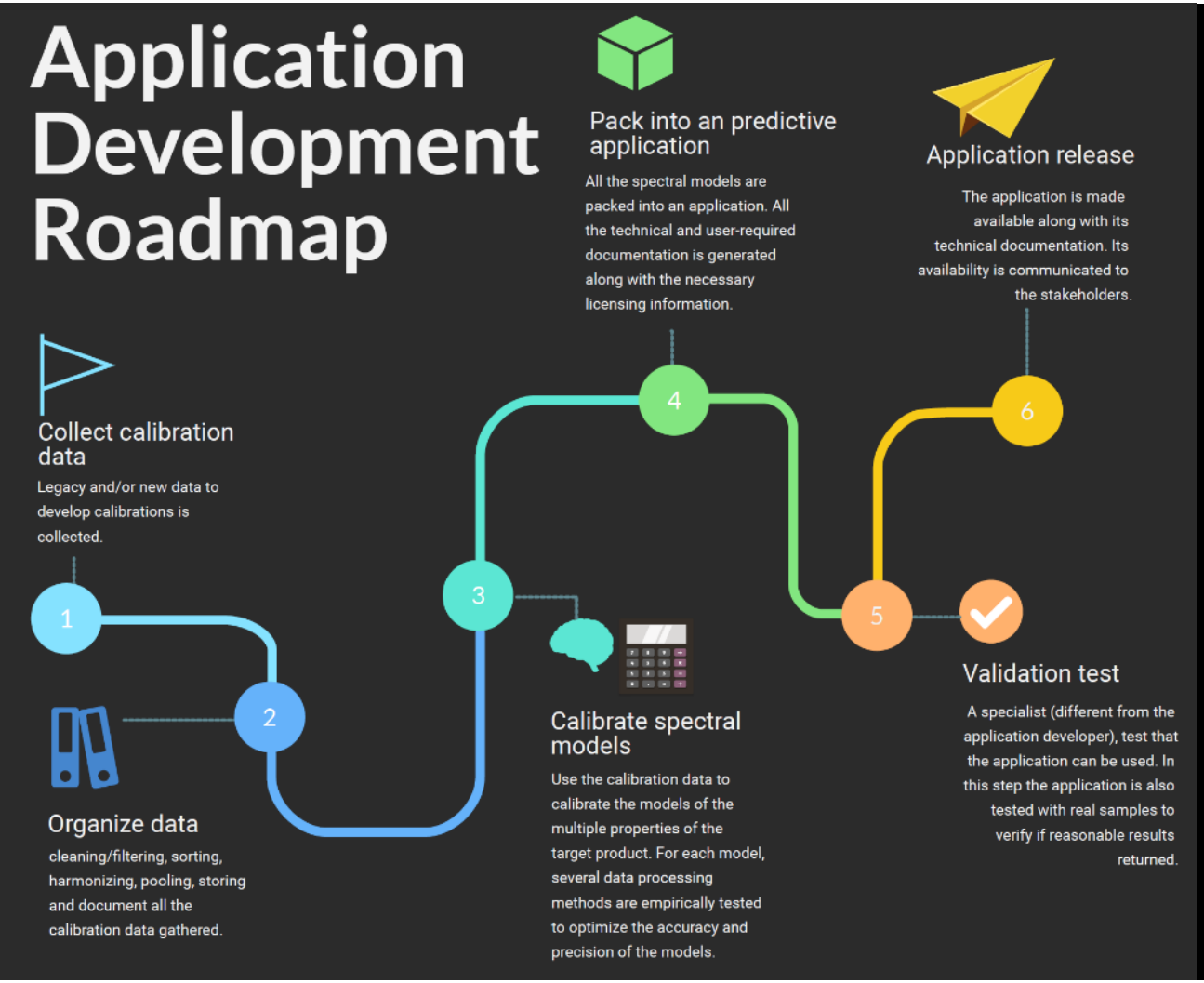
The application is made available along with its technical documentation. Its availability is communicated to the stakeholders.

Application Development Roadmap



Challenges

- It is expensive.
- Labs should focus on routine analyses (not on learning and dealing with chemometrics).
- For end users/labs, many time-consuming iterations to get useful applications creates frustration.
- Chemometrics is a complex subject which requires time and effort to be learned/taught.



Challenges

- It is expensive to calibrate, although it is extremely cost-efficient if it is released for routine analysis.
- Labs should be focused on routine analyses (not on learning and dealing with chemometrics)
- For end users/labs, many time-consuming iterations to get useful applications create frustration
- Chemometrics is a complex subject which requires time and effort to be learned/taught

Exercise

- Introduction to R
- Read in spectral data and visualization (1st script)

How to name your samples?

- Material name:
  - Core_[NAMEOFYOURGROUP]
- Sample names:
  - [REPLICATE NUMBER]_[DEPTH increment]
  - Ex.: 01_0010
  - 01_1020
  - 02_1020

C:/Users/stevens/Dropbox2/Dropbox/SPECcourse - RStudio

File Edit Code View Project Workspace Plots Tools Help

SPECcourse

pres.Rmd | Day1.Rmd

Knit HTML | Run | Chunks

```
67
68  * The expension of __R__ is driven by the increasing number of available functions/packages.
69
70  ![](http://r4stats.files.wordpress.com/2012/04/fig_10_cran.png)
71
72  source: [http://r4stats.com/articles/popularity/](http://r4stats.com/articles/popularity/)
73  ----------------------------------------
74
75  * But, of course, __R__ has also drawbacks:
76   * No GUI (altough solutions exist, such as [Rcmdr](http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/) for data import/export
    and analysis or [DeduceR](http://www.deducer.org/pmwiki/index.php?n=Main.ScreenShots) for graphs)
77   * Steeper learning curve than Excel, SAS or JMP (No GUI...)
78   * Cryptic documentation and error messages
79   * Often, __R__ makes us believe it worked as intended (see Chapter 8 of the [R Inferno](http://www.burns-stat.com/pages
    /Tutor/R_inferno.pdf))
80
81  ```{r}
82  # Exemple:
83  head(iris) # the iris dataset is a data.frame
84  head(iris[,1:2]) # the two first column of the data.frame
85  head(iris[,1]) # subsetting the 1st column returns a vector, not a data.frame!
86  head(iris[,1,drop=F]) # use [,,drop=F] to override the default behavior
87  ```
88
```

111:1 (Top Level) R Markdown

1

Console C:/Users/stevens/Dropbox2/Dropbox/SPECcourse/

```
11 2005  31 263  32    799          342
12 2006  42 281  47    967          397
13 2007  49 321  60   1220          415
14 2008  45 368 106   1300          431
15 2009  54 371 153   1330          447
16 2010  75 375 187   1620          539
17 2011  74 345 199   1710          556
> save(GS_data,file="GS_hits.Rdata")
>
> ggplot(data=melt(data=GS_data,id="year"),aes(x=year,y=value,colour=variable)) + geom_point() + geom_line() + labs(x="Year",y="Num
Google Scholar") + theme_bw() + scale_color_discrete("Software")
Warning messages:
1: Removed 16 rows containing missing values (geom_point).
2: Removed 16 rows containing missing values (geom_path).
> ggsave("figures/GScholar_spectroscopy.png")
Saving 6.61 x 6.17 in image
Warning messages:
1: Removed 16 rows containing missing values (geom_point).
2: Removed 16 rows containing missing values (geom_path).
> rm() # remove objects
> save() # save in Rdata format
Error in save() : 'file' must be specified
>
```

2

Workspace | History

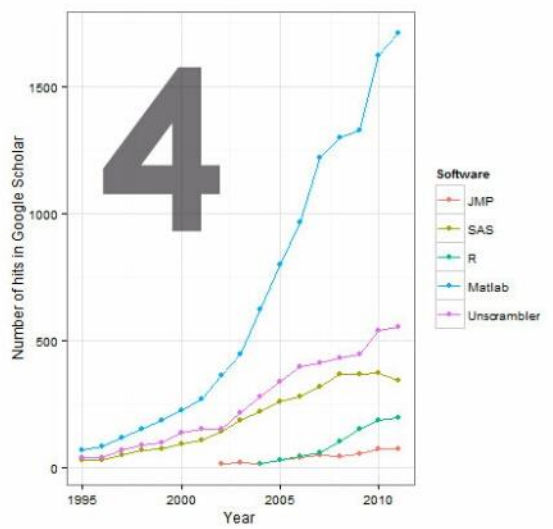
To Console | To Source

```
h1 <- xpathSApply(doc, "//div[@id='gs_ab_md']", xmlValue)
h2 <- as.numeric(sub(",","",strsplit(h1, " ")[[1]][2]))
h2
}
}
colnames(hits) <- labels
data.frame(year=from:(to-1),hits)
}
GS_data <- GScholar_Scraper_hits(inputs,labels)
GS_data
save(GS_data,file="GS_hits.Rdata")
ggplot(data=melt(data=GS_data,id="year"),aes(x=year,y=value,colou
ggsave("figures/GScholar_spectroscopy.png")
rm() # remove objects
save() # save in Rdata format
```

3

Files | Plots | Packages | Help

Zoom | Export | Clear All



Signal processing

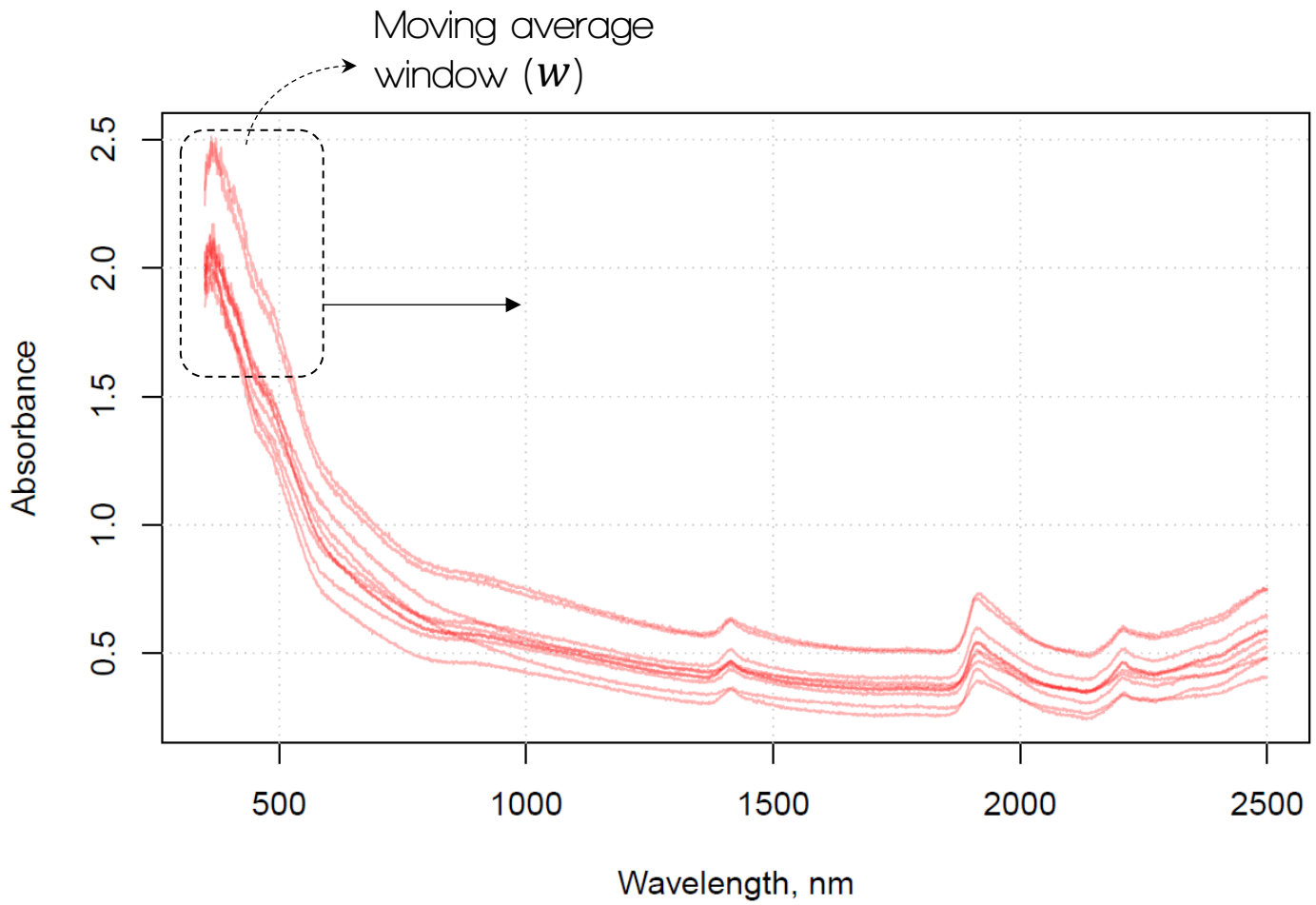
Signal processing aims to improve quality of the spectra before modeling (e.g. by removing undesired spectral variation). In this respect, it improves also the results of quantitative and qualitative spectral models.

The methods must be carefully chosen in order to avoid removing important physical information.

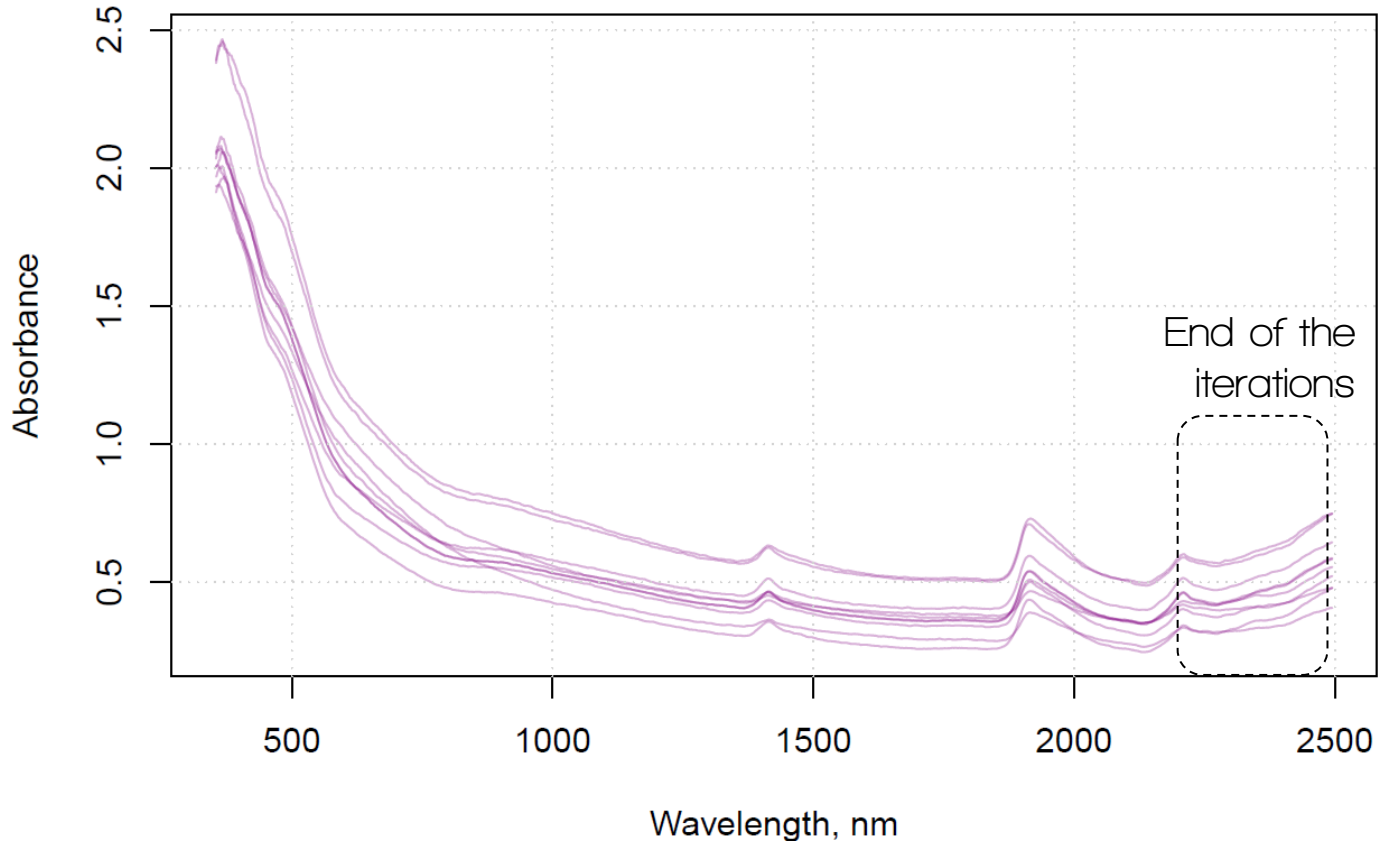
The most common signal processing methods are:

- Splice correction
- Spectra transformation (e.g. reflectance to absorbance)
- Noise removal
- Resampling
- Differentiation
- Scatter correction
- Centering and scaling
- Continuum removal

Moving average



Moving average

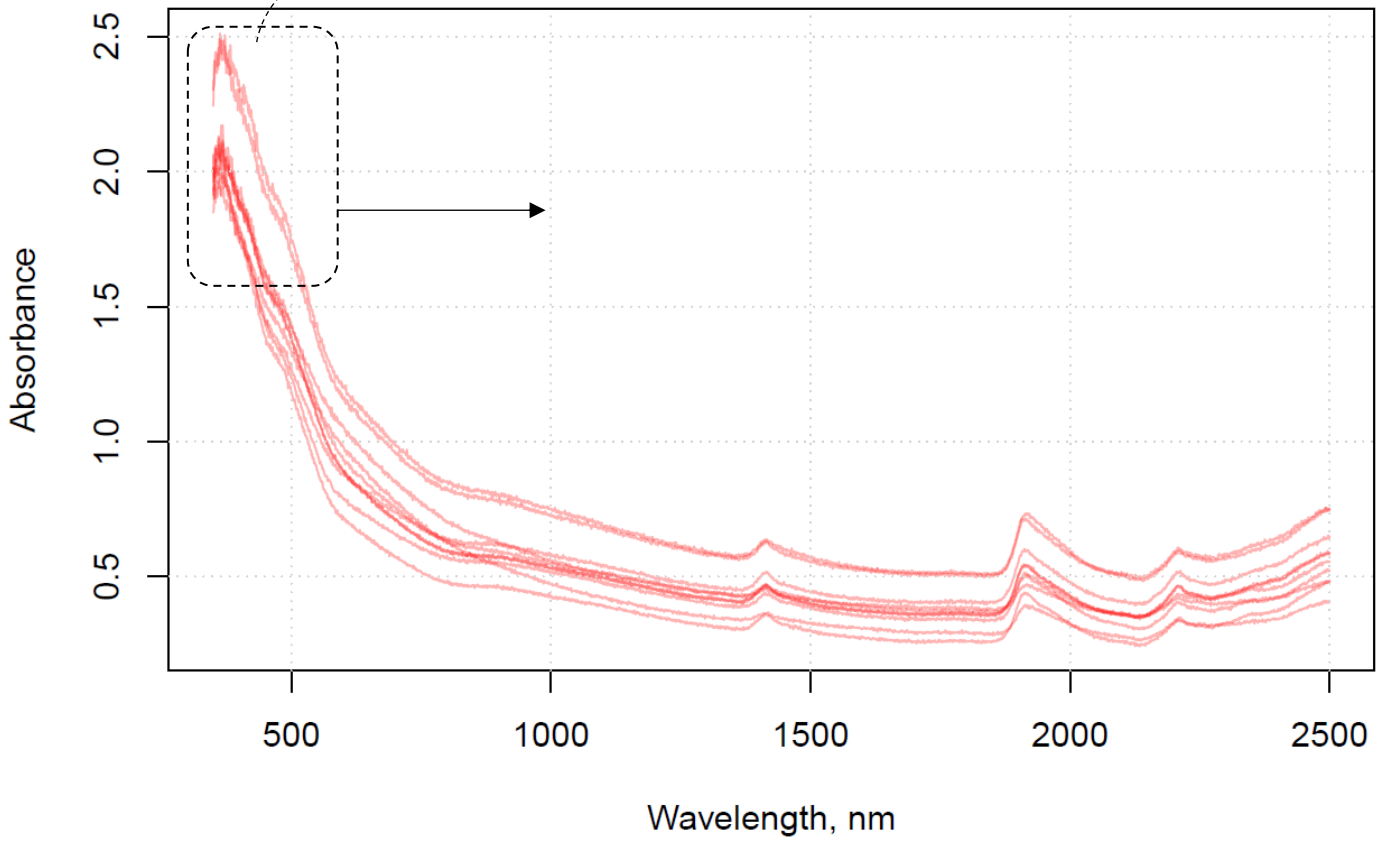


Savitzky–Golay filter

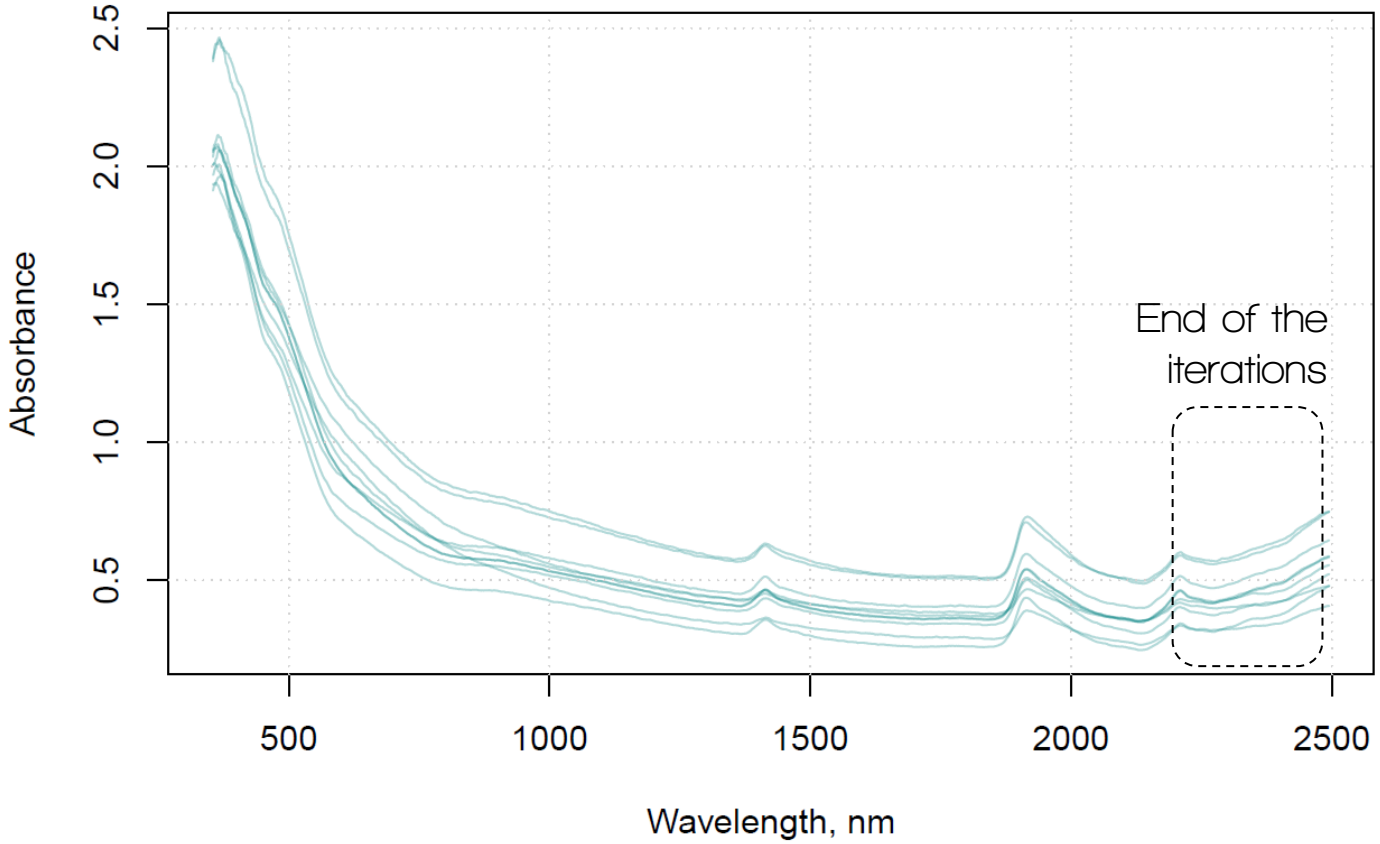
Polynomial order

Moving Polynomial fit window (w)

$$x_{sg} = b_p\, x^{(p)} + \cdots + b_1\, x + b_0$$

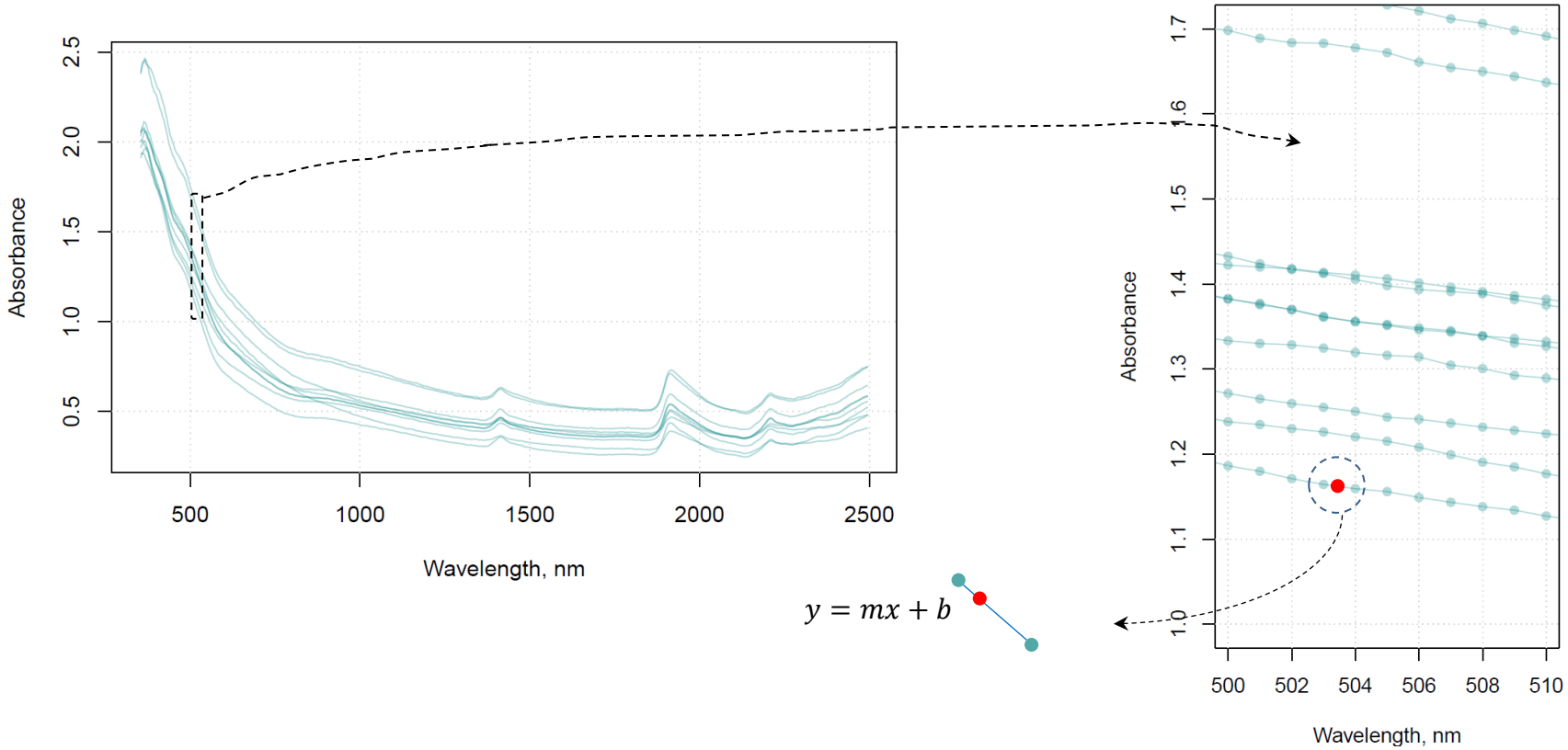


Savitzky–Golay filter



Resampling (also known as interpolation)

Example: if the first wavelength of the new set of wavelengths (at wich the spectra needs to be resampled) is 504.3 nm, the two wavelengths at the right and left sides of the new point are taken. Then a linear model is fitted between these two points and this models is then used to estimate the intensity at the new wavelength



Differentiation

Advantages

- Reduce of baseline offset
- Helps to resolve absorption overlapping
- Compensates for instrumental drift
- Enhances small spectral absorption features
- Often increases the predictive accuracy for complex datasets

Drawbacks

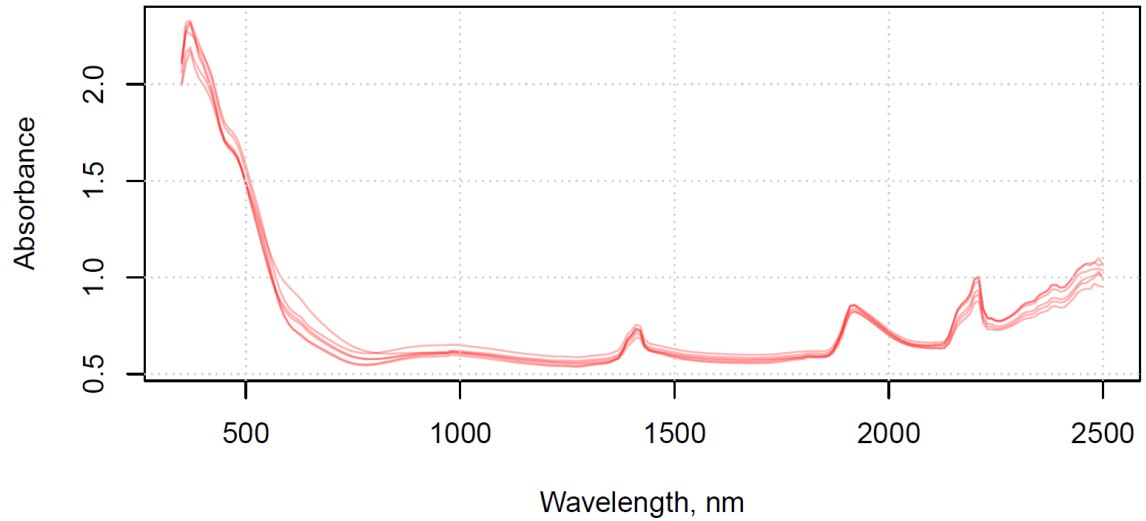
- Risk of overfitting
- Amplifies noise, smoothing required
- Less robust model and increase uncertainty in model coefficients
- Complicate spectral interpretation
- Remove the baseline! (albedo variation)

First and second derivatives of a spectrum can be computed with the finite difference method (difference between to subsequent data points), assuming that the band width is constant:

First derivative $x_i' = x_i - x_{i-1}$

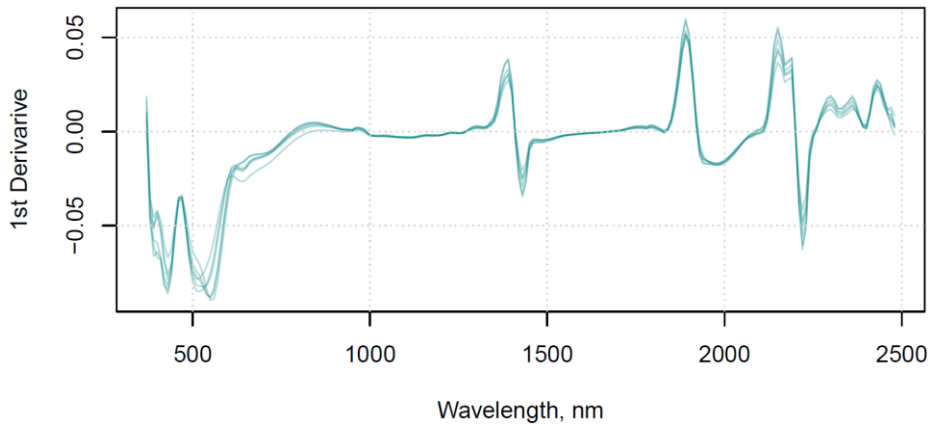
Second derivative $x_i'' = x_{i-1} - 2\,x_i + x_{i+1}$

Original absorbance spectra of a soil profile

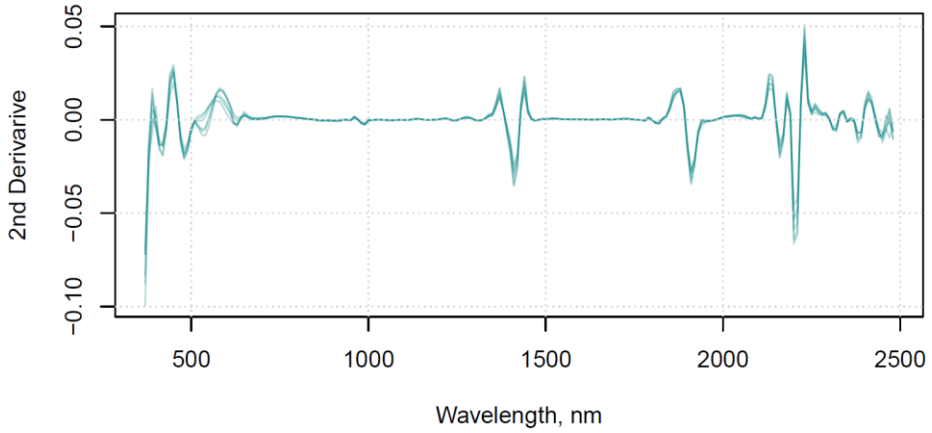


Using the **`savitzkyGolay`** function both the first and second derivative spectra, can be computed

First derivative spectra



Second derivative spectra



Multiplicative Scatter Correction (msc)

- It mitigates the effect of scattered light (which is multiplicative) on the NIR spectra
- msc is a method that assumes for a given set of spectra, the scattering and offset effects can be removed from each spectrum.
- The msc model is determined by regressing each spectrum onto the mean spectrum. The model can be formally written as:

$$x_i = a_i + b_i\bar{x} + \varepsilon_i$$

where $\varepsilon_i$represents the chemical information of the spectrum while a_i *and* b_i represent the offset and the scatter respectively. Both a_i *and* b_i are found by using the Least Squares method. The msc corrected spectrum can be then computed as:

$$\text{msc}(x_i) = \frac{x_i - a_i}{b_i}$$

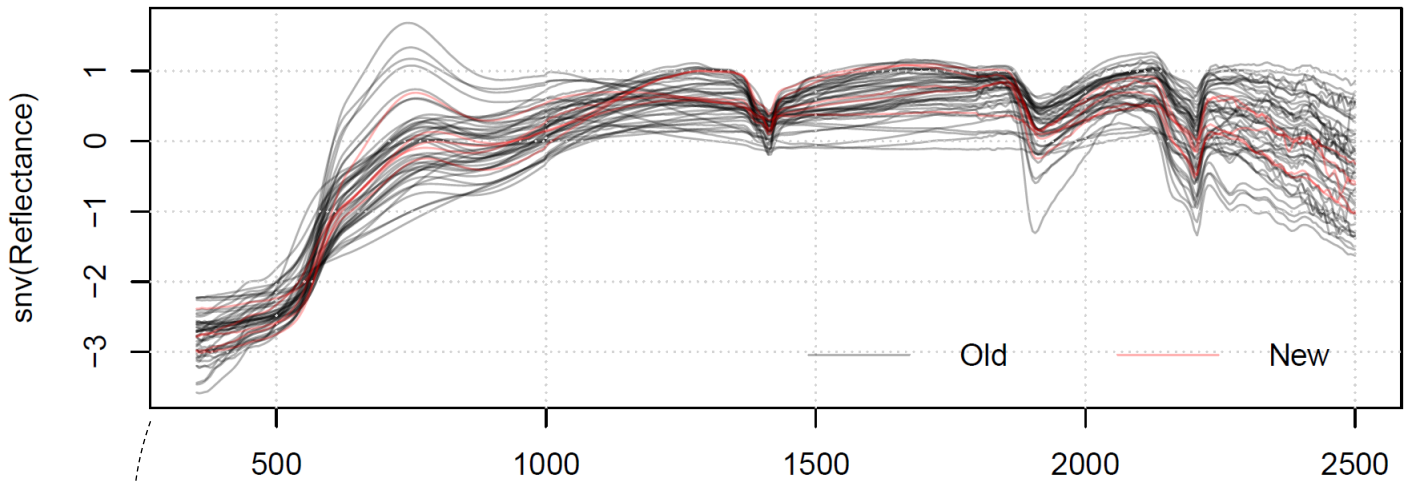
The `msc` function pf the `pls` package can be used to apply the Multiplicative Scatter Correction

Centering and scaling

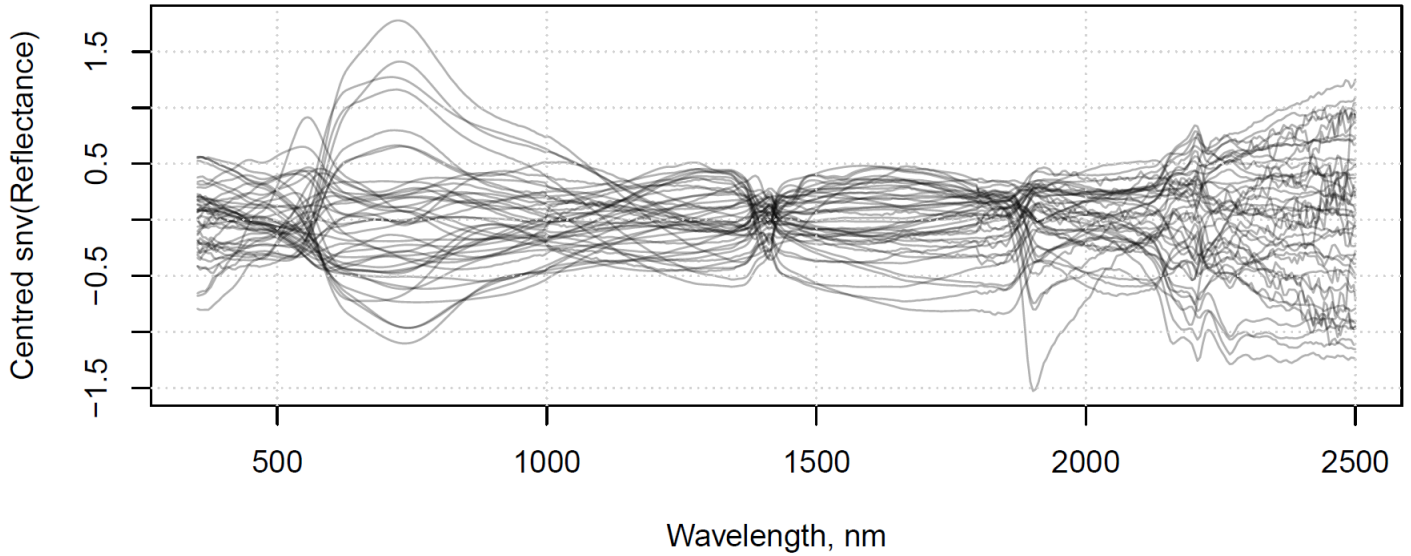
- *Centering:* a given matrix is transform into a matrix with columns that have zero mean. This is achieved by subtracting the mean vector of the matrix columns from each row of the original matrix. From a statistical point of view, *centering* eliminates the need for an intercept in the model and makes the model more simple and interpretable. This operation is **strongly** advised for most regression approaches, although it does not imply that in all cases the model performance will improve
- *Scaling:* a given matrix is transform into a matrix with columns that have the same variance. This is achieved by dividing each row of the original matrix by the vector of standard deviations of the matrix columns. The purpose is to get all the variables . . . on the same scale.
- Both centering and scaling are usually applied after data transformation.

The `scale` function can be used for applying either centering and scaling or both

Centering and scaling

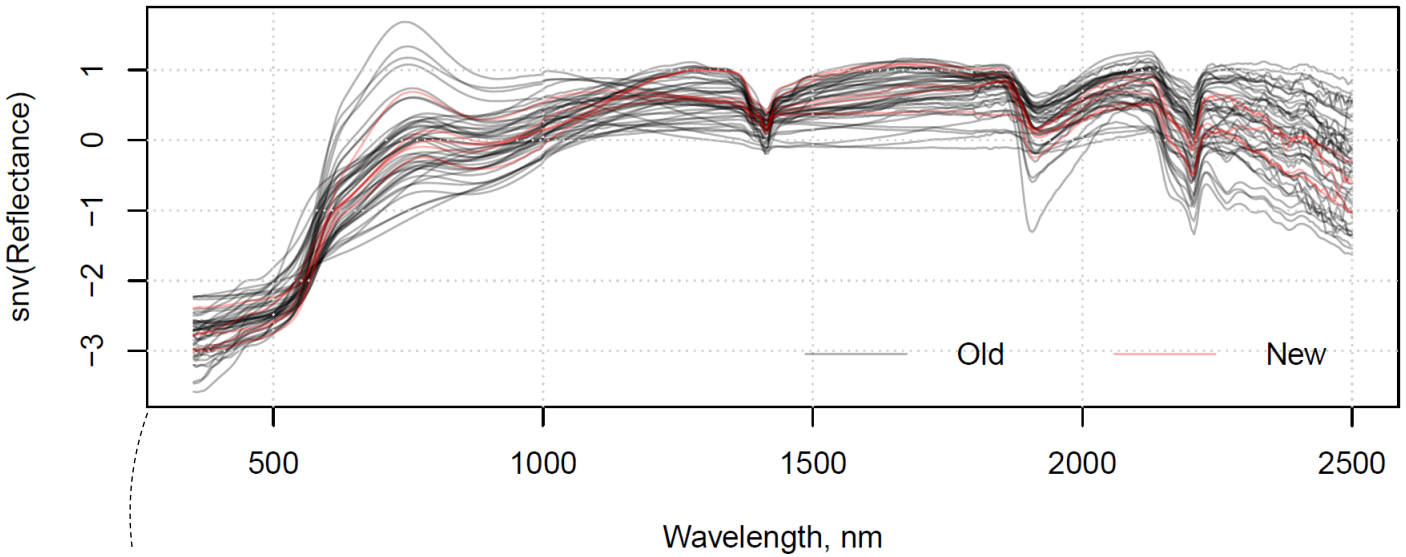


Centering

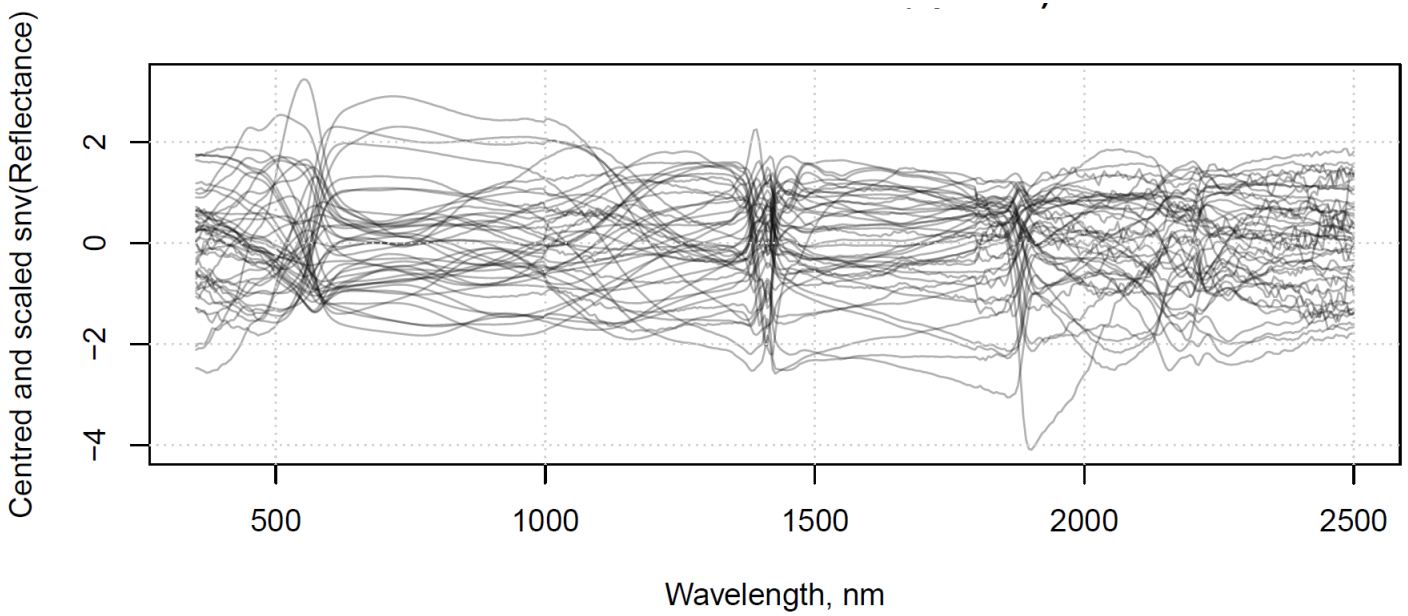


See file: 04_preprocessing_III.R

Centering and scaling



Centering and scaling



See file: 04_preprocessing_III.R

To *scale* or not to *scale*?

- Scaling will down-scale variables with large variation, but can inflate exaggeratedly variables with small variance (e.g. variables with no variation but small noise).
- Decision depends on the data pre-treatment also: is it meaningful?
- For instance, variation in some variables is greater than in other for good reasons (e.g. absorption features in continuum removed spectra, 1st derivative), so that *scaling* is not appropriate and will reduce model interpretability
- *Centering and scaling* is sensitive to outliers. A robust scaling can be computed using the median instead of the mean and the median absolute deviation (MAD) instead of the standard deviation

What and how many samples to choose to build accurate NIR models

How many observations (samples) should be included in a calibration set in order **to efficiently develop the best NIR models**?

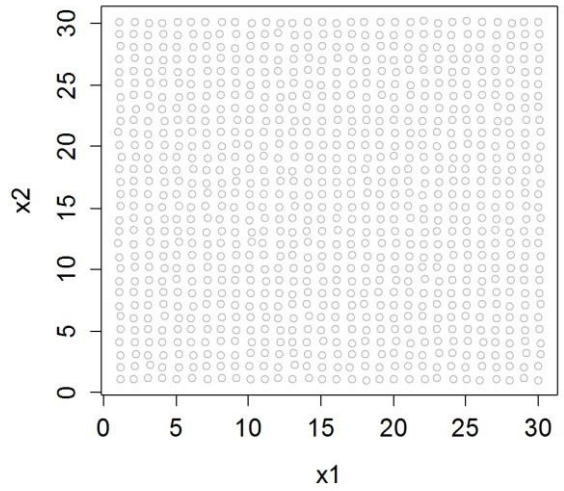
How many samples should be used for calibration?

→ Calibration set size

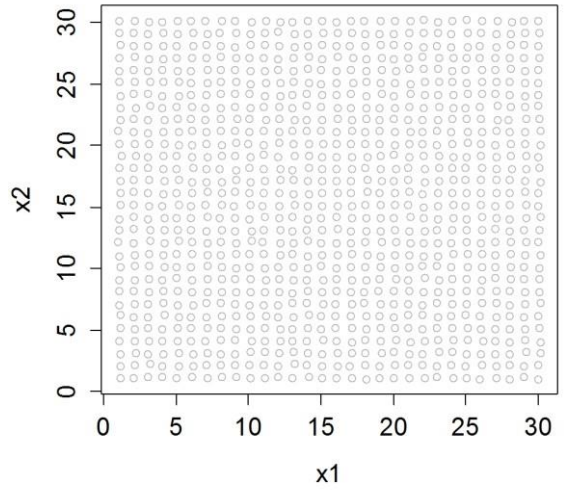
How to select the calibration samples?

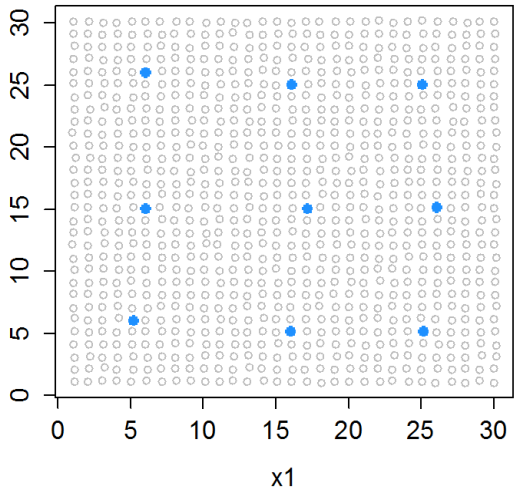
→ Sampling design

Sample representativeness



Sample representativeness





Kennard-Stone sampling (KSS)

The KSS algorithm (Kennard and Stone, 1969) to select a (calibration) subset of m samples (S_{tr}) from a given set of N samples (S) works as follows:

1. Compute the distance matrix between all the samples in X
2. Find in S the two samples s_{tr1} and s_{tr2}, which are farthest apart, allocate them in S_{tr} and remove them from S.
3. Find in S the sample s_{tr3}, which is the most dissimilar to the ones already allocated in S_{tr}. Allocate s_{tr3} in S_{tr} and then remove it from S. Note that the dissimilarity between S_{tr} and each s_i is given by the minimum distance of any sample allocated in S_{tr} to each s_i.
4. Repeat the step $3m - 4$ times in order to select the remaining samples.

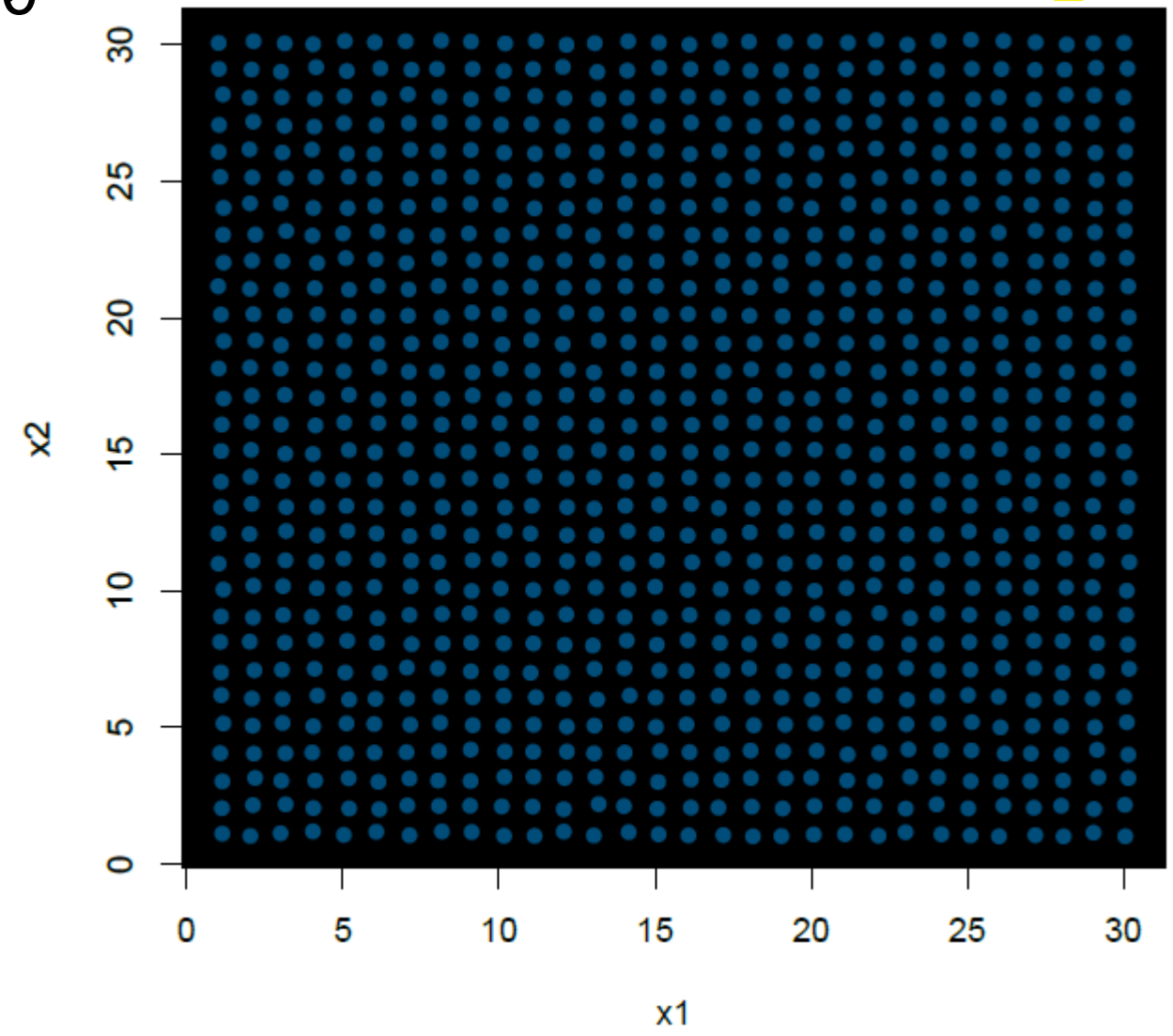
The KSS is usually applied to the PC socres of the spectral data

The **kenStone** function of the **prospectr** package can be used for KSS

Kennard, R.W., Stone, L., 1969. Computer aided design of experiments. Technometrics 11, 137–148.

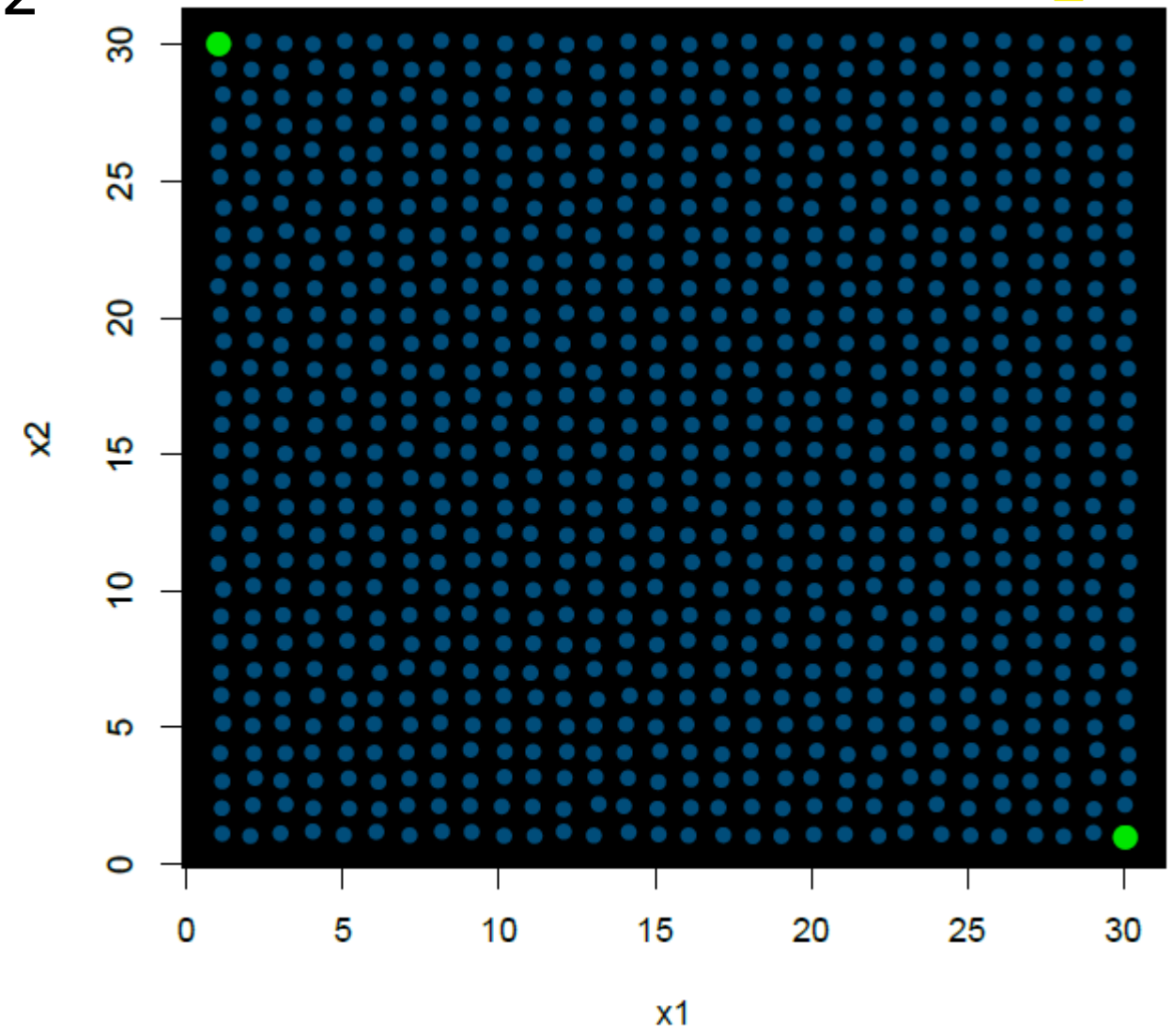
n = 0

Kennard-Stone sampling (KSS)



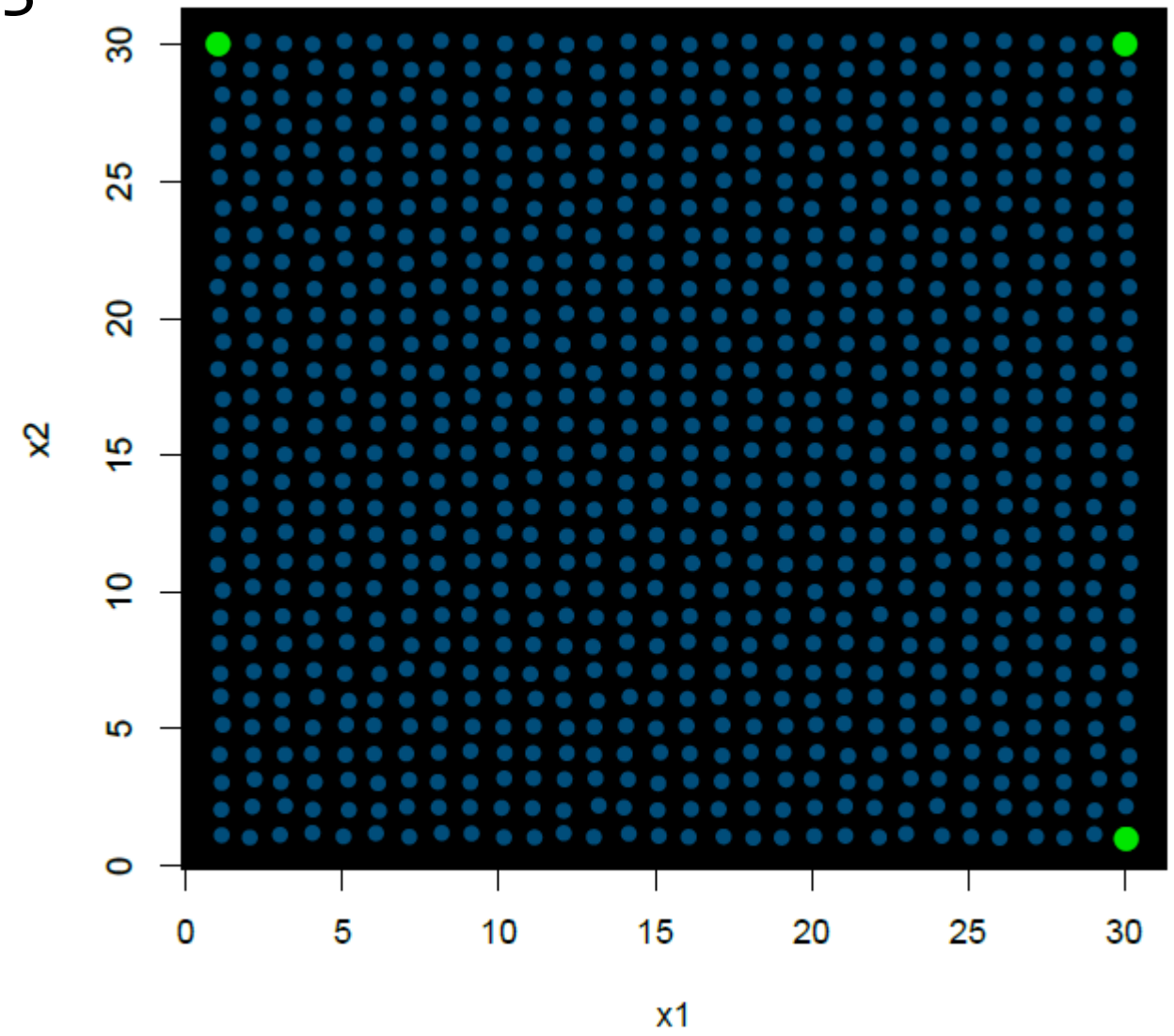
n = 2

Kennard-Stone sampling (KSS)



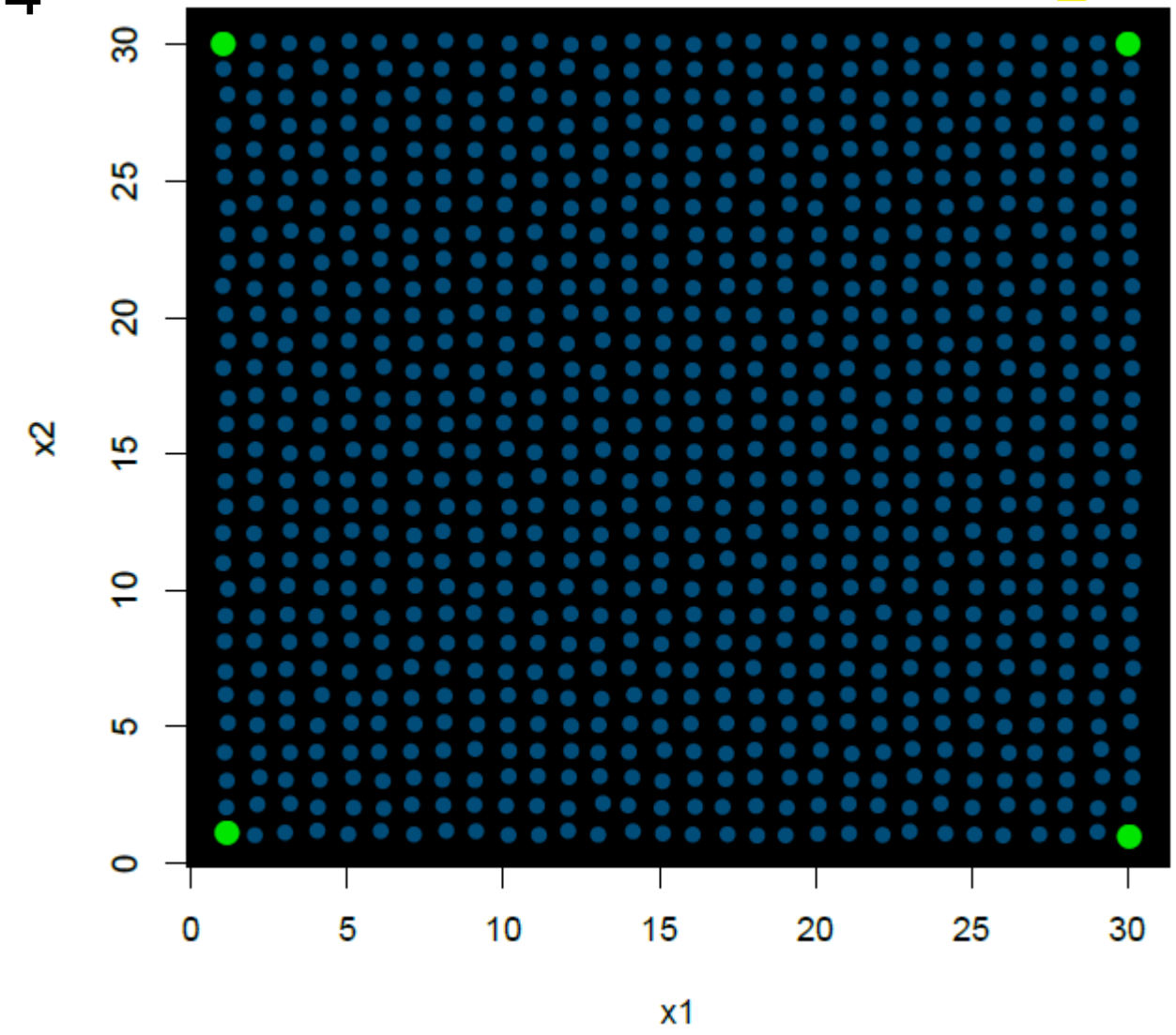
n = 3

Kennard-Stone sampling (KSS)



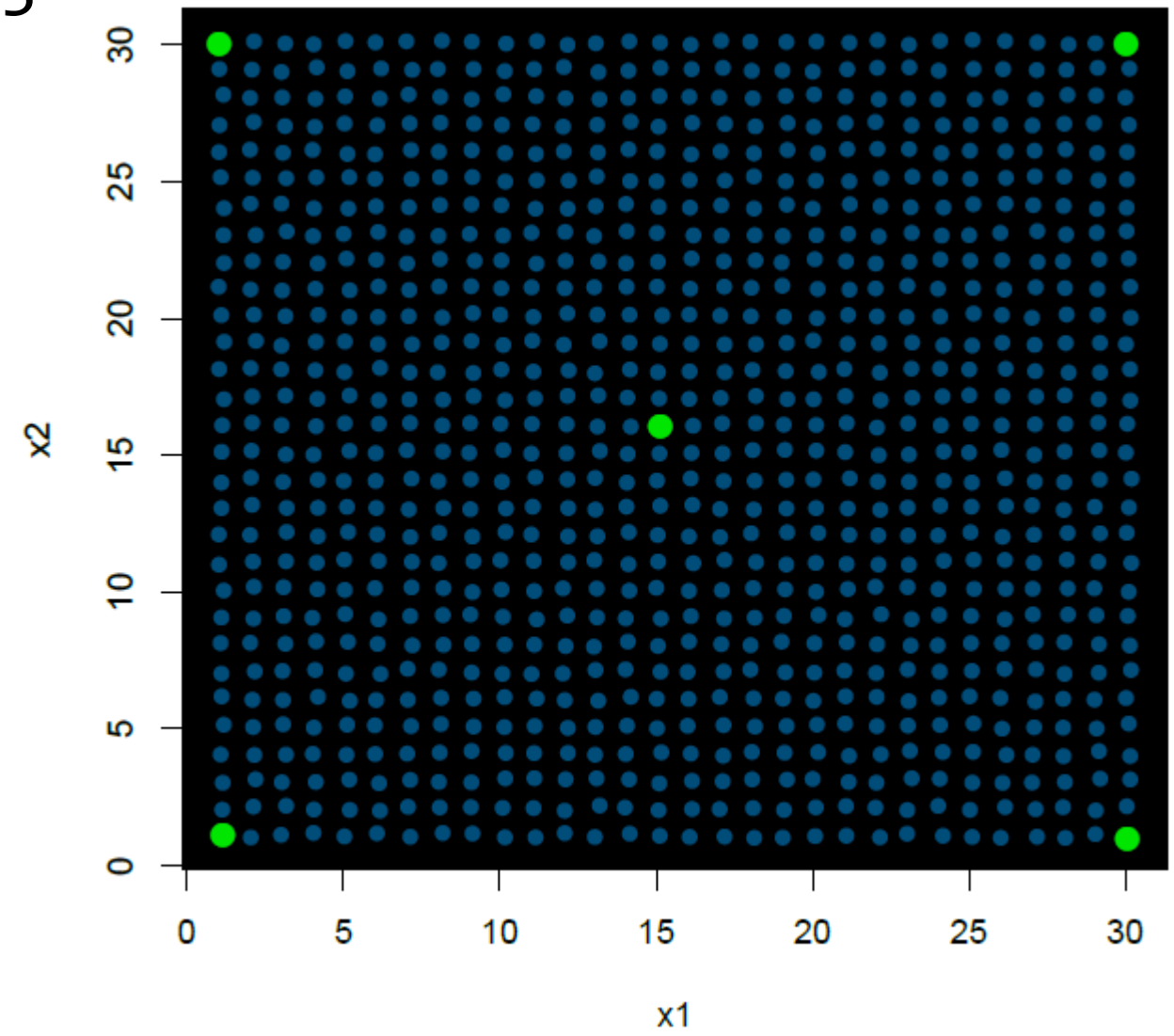
n = 4

Kennard-Stone sampling (KSS)



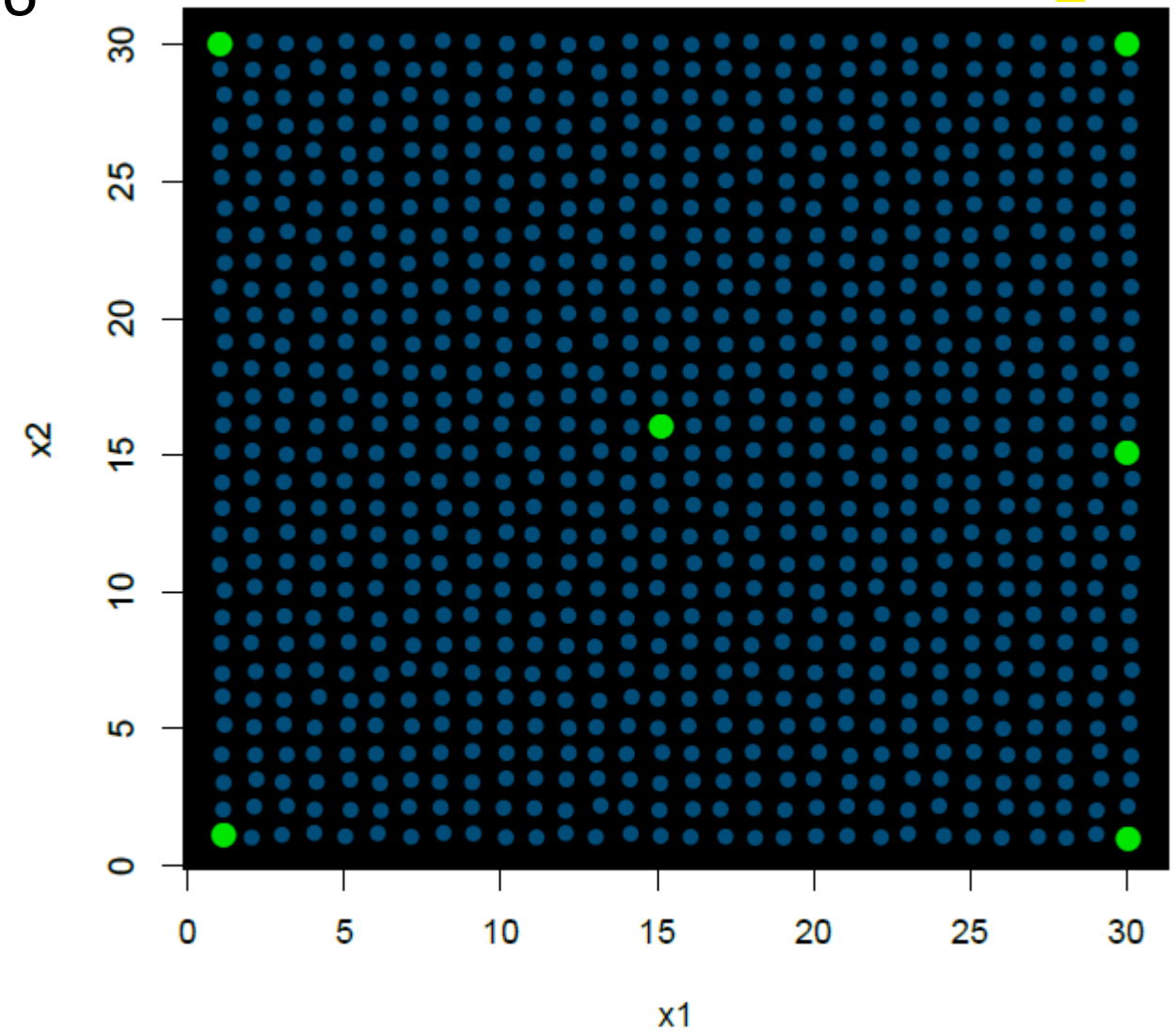
n = 5

Kennard-Stone sampling (KSS)



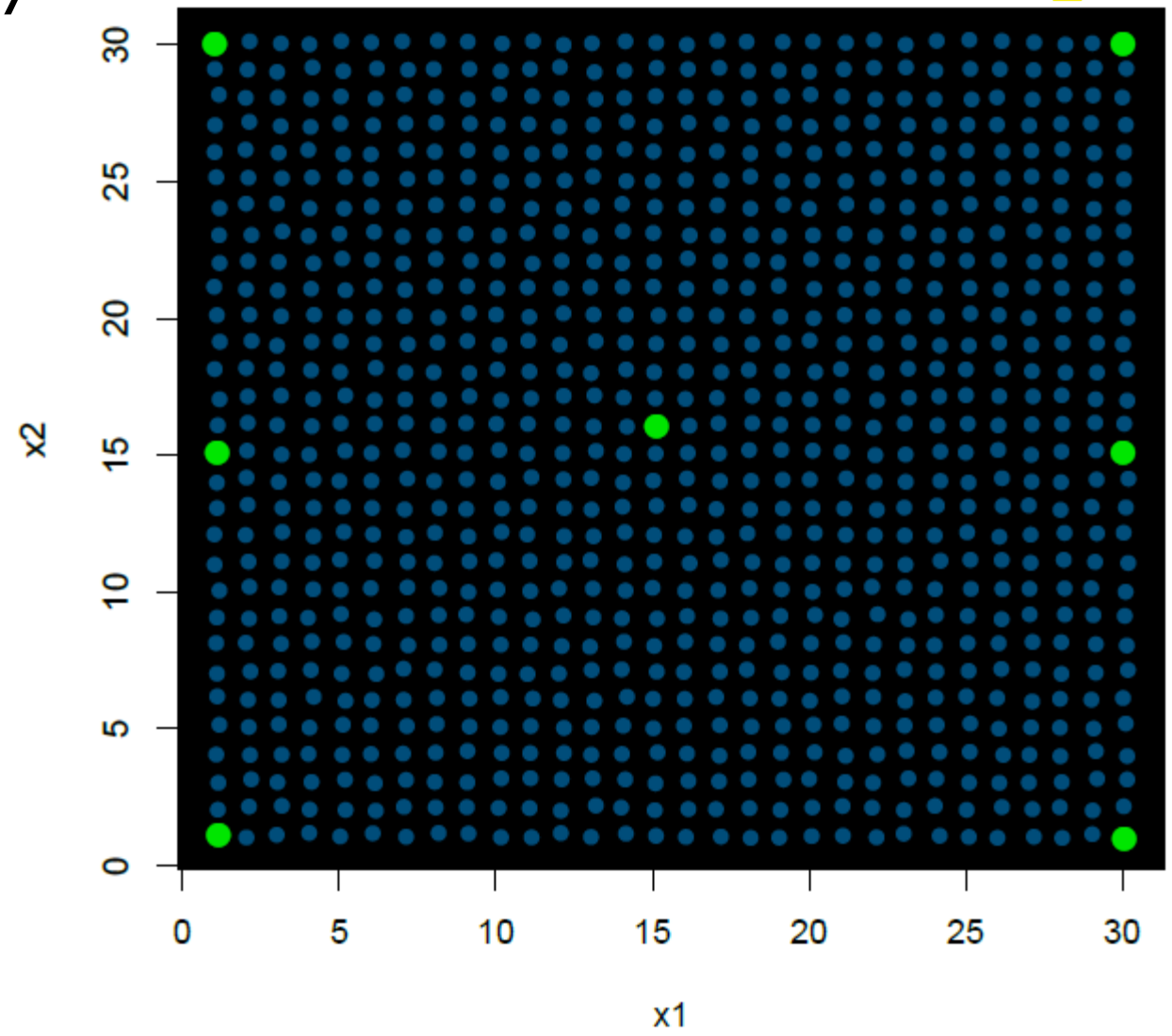
n = 6

Kennard-Stone sampling (KSS)



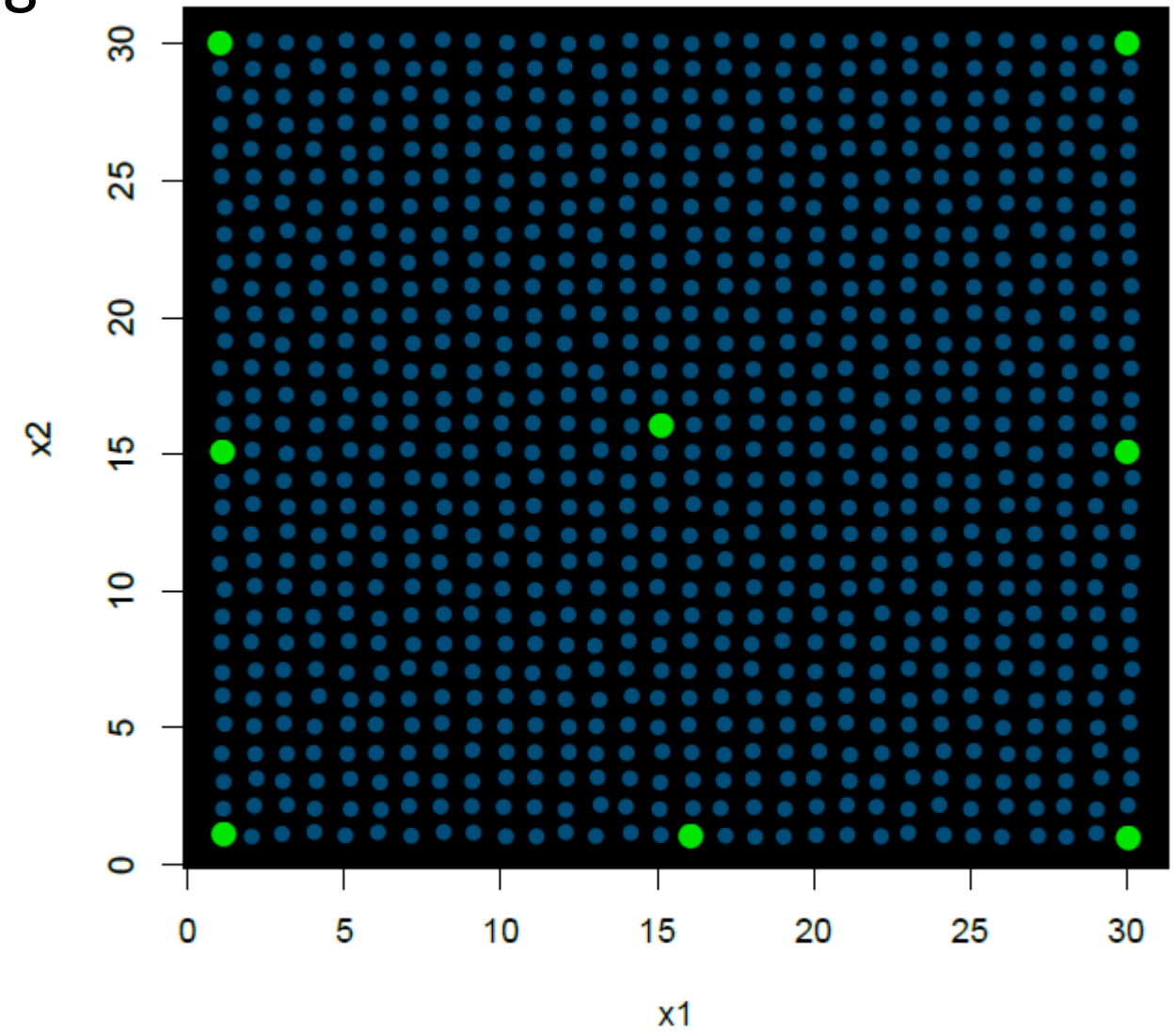
n = 7

Kennard-Stone sampling (KSS)



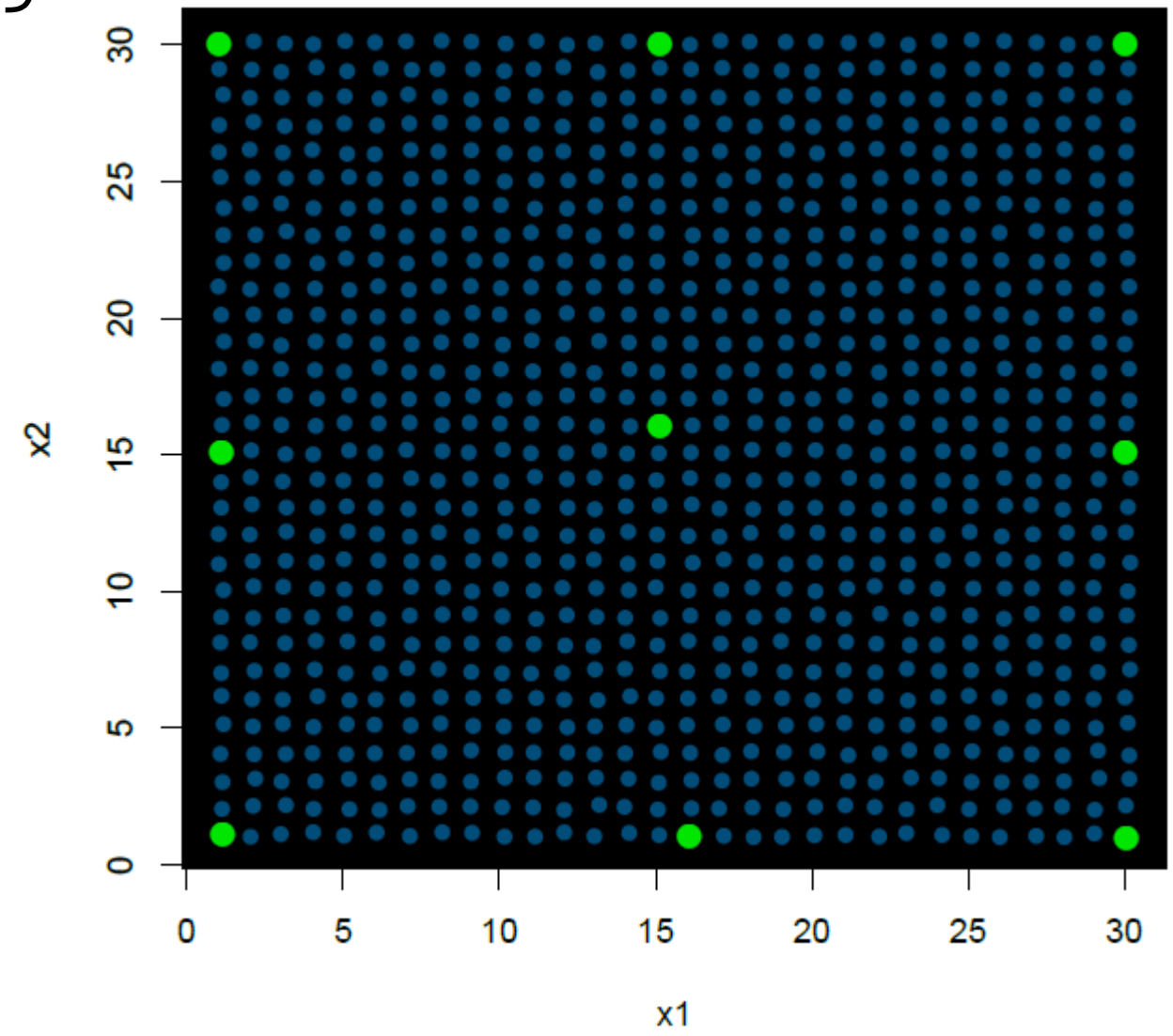
n = 8

Kennard-Stone sampling (KSS)



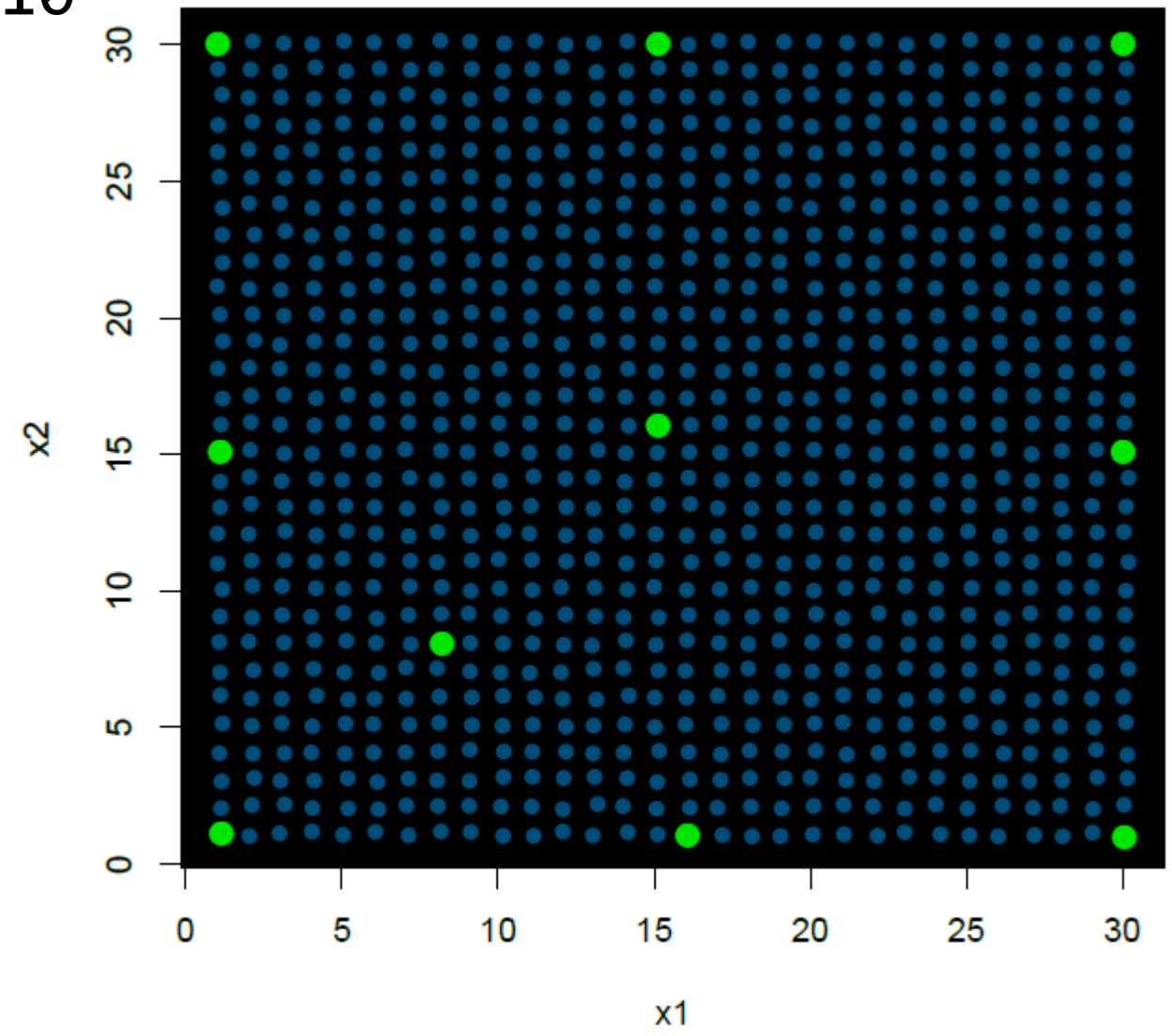
n = 9

Kennard-Stone sampling (KSS)



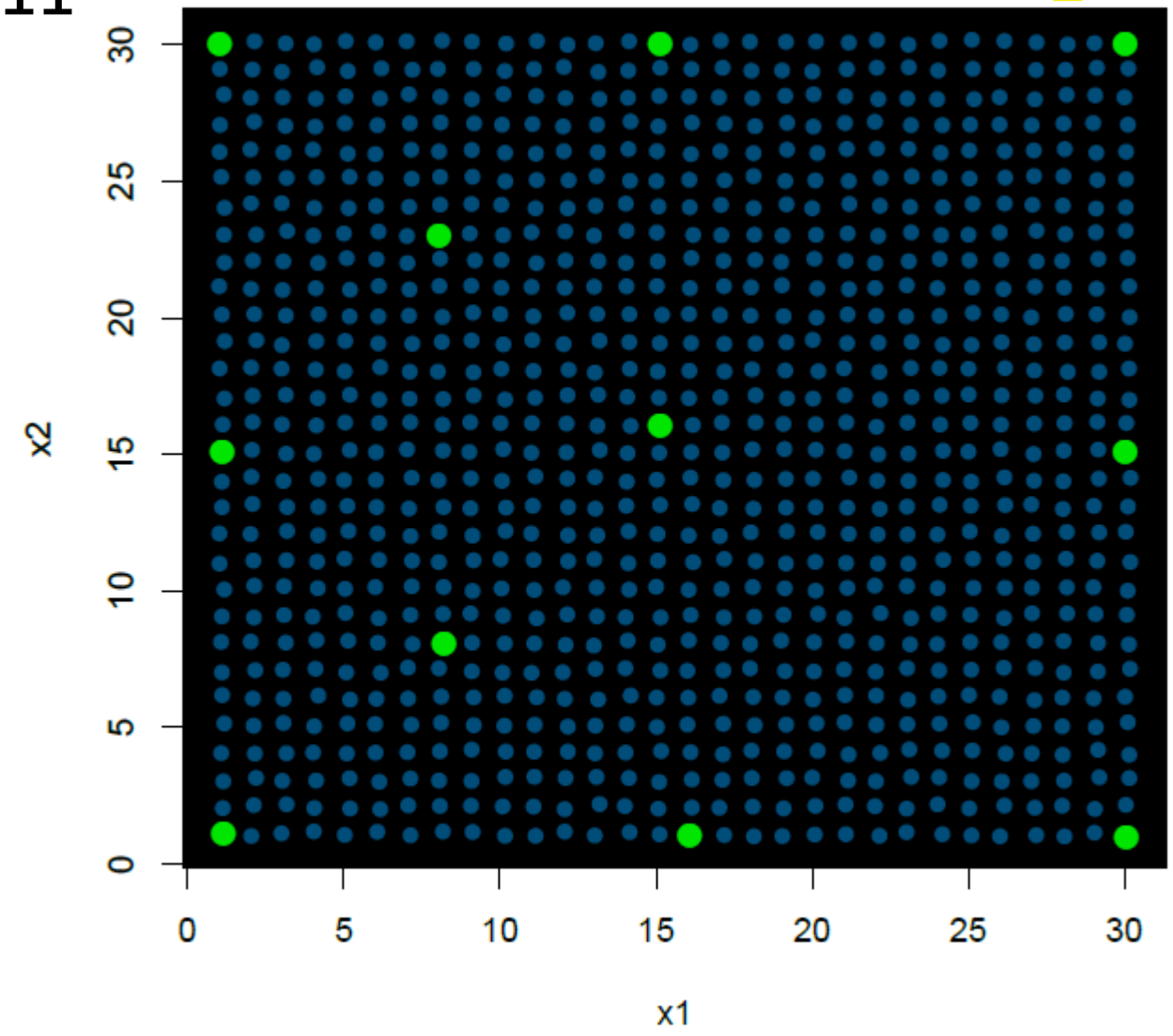
```
n = 10
```

Kennard-Stone sampling (KSS)



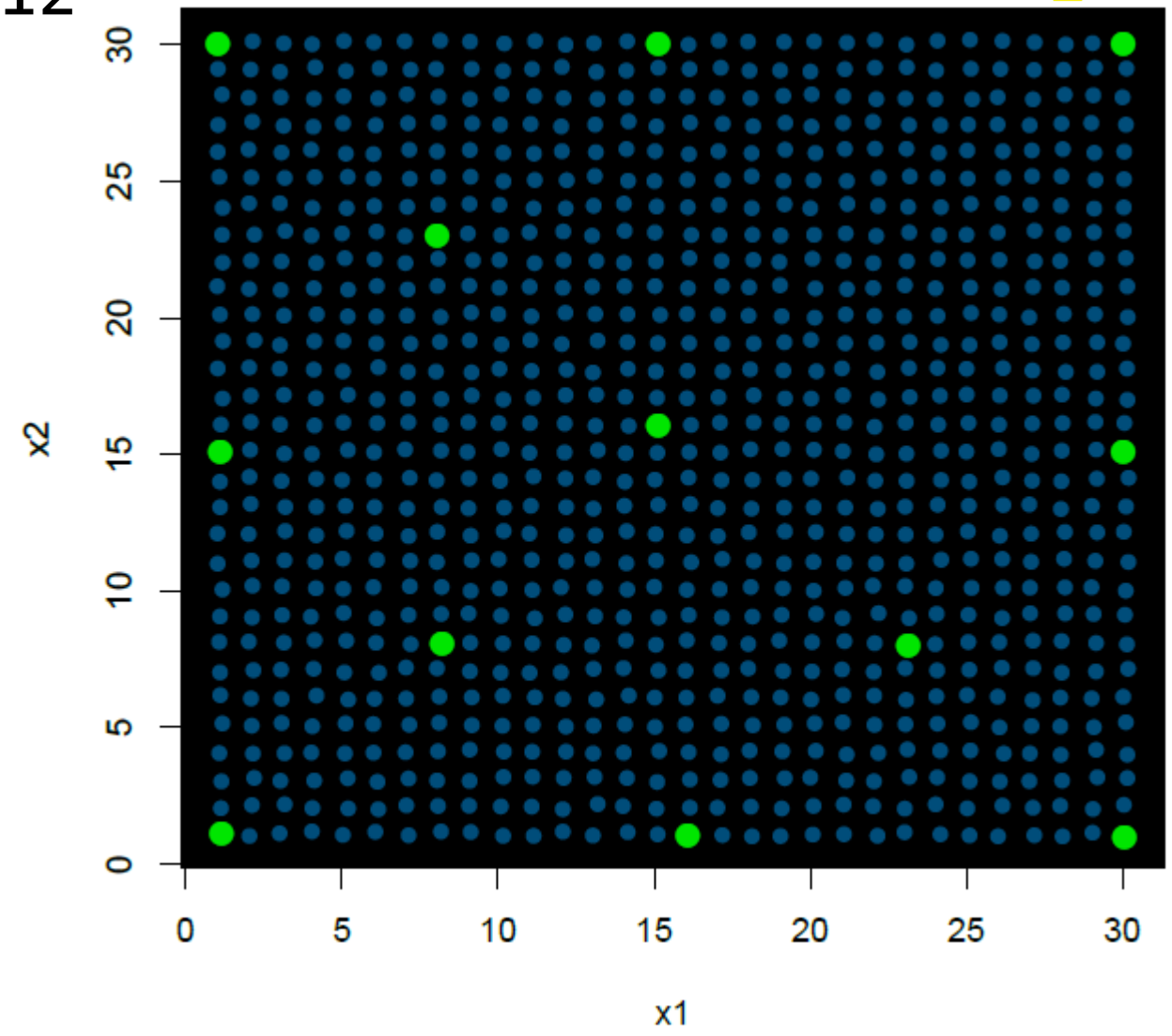
n = 11

Kennard-Stone sampling (KSS)



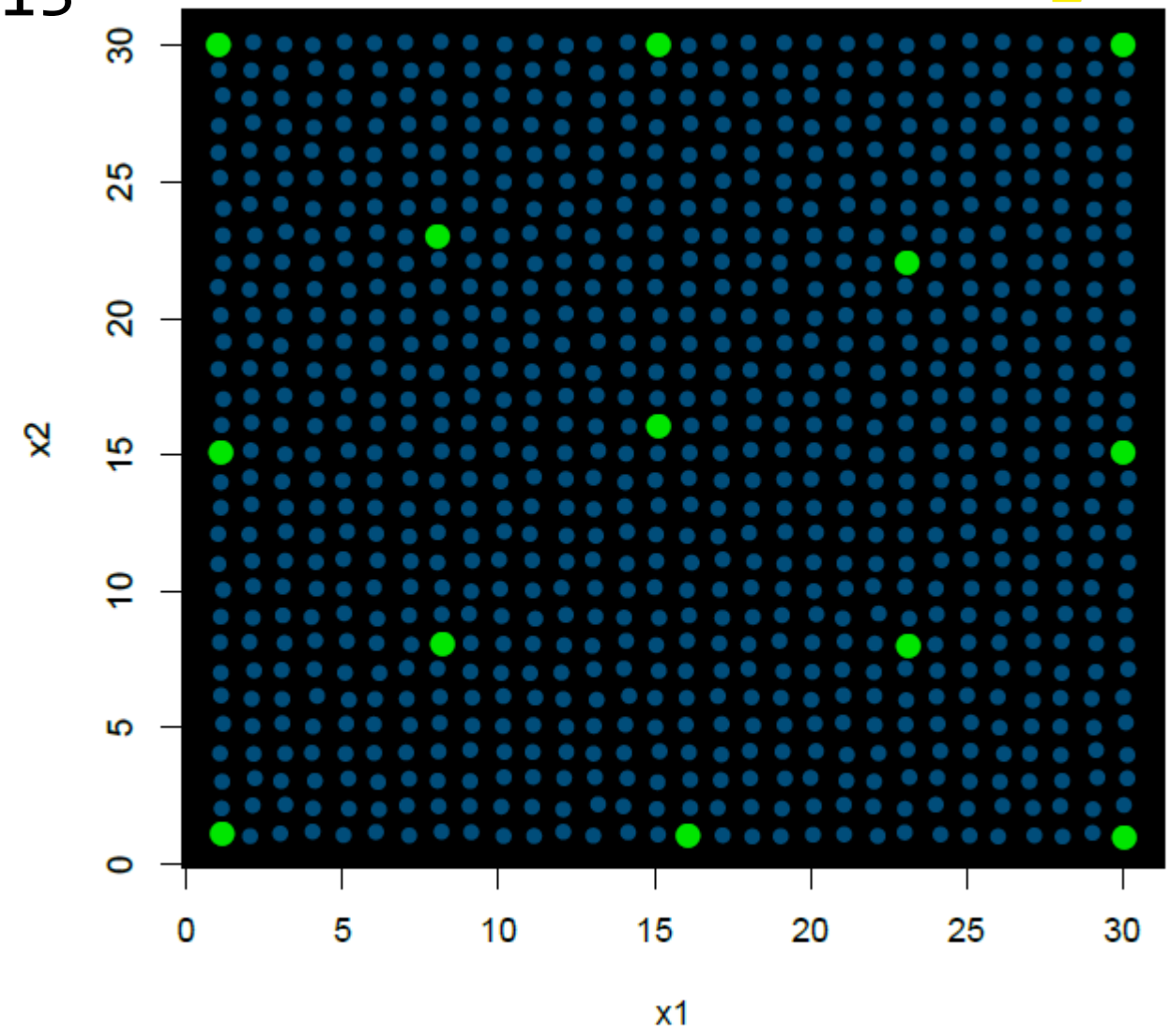
n = 12

Kennard-Stone sampling (KSS)



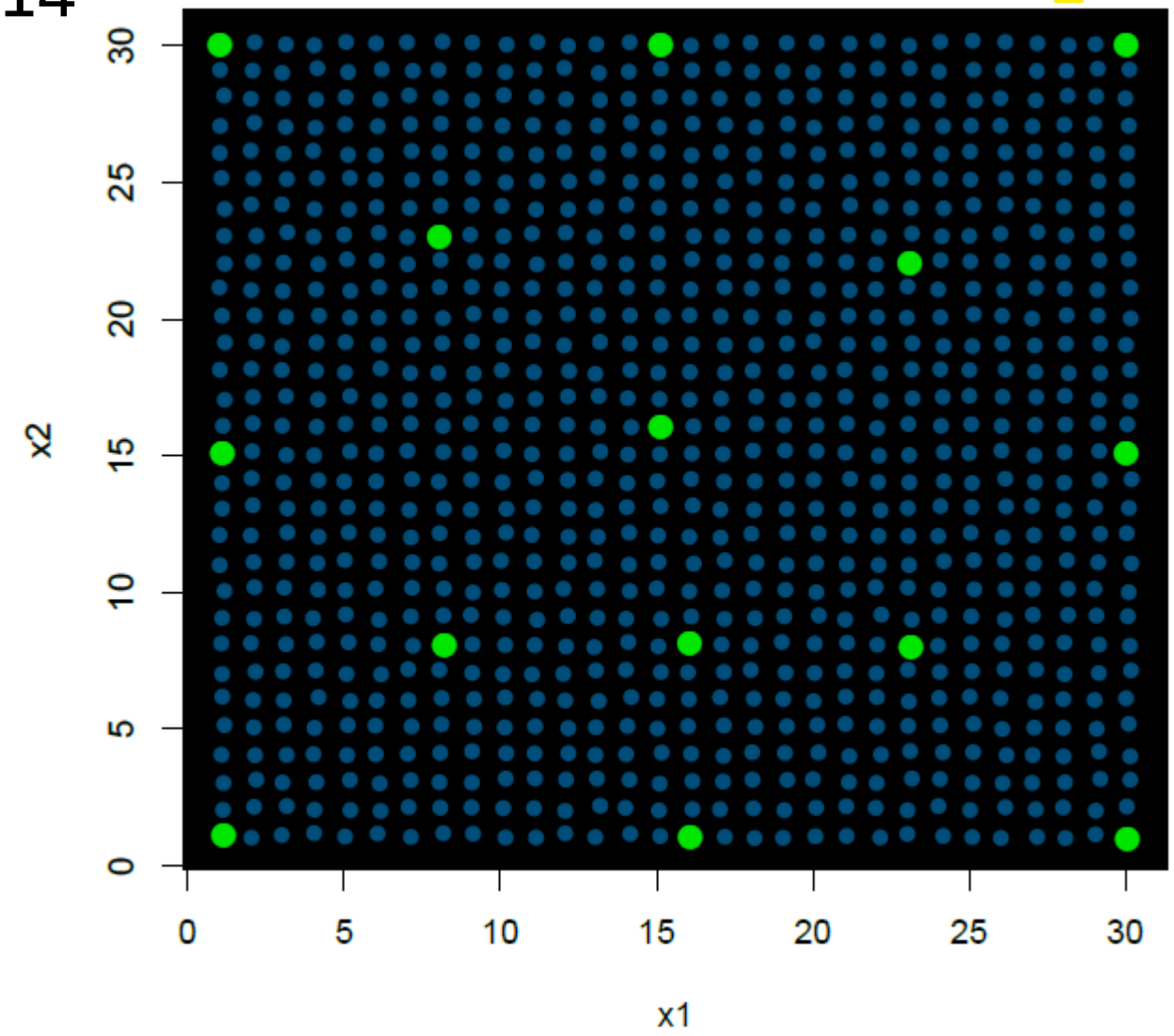
n = 13

Kennard-Stone sampling (KSS)



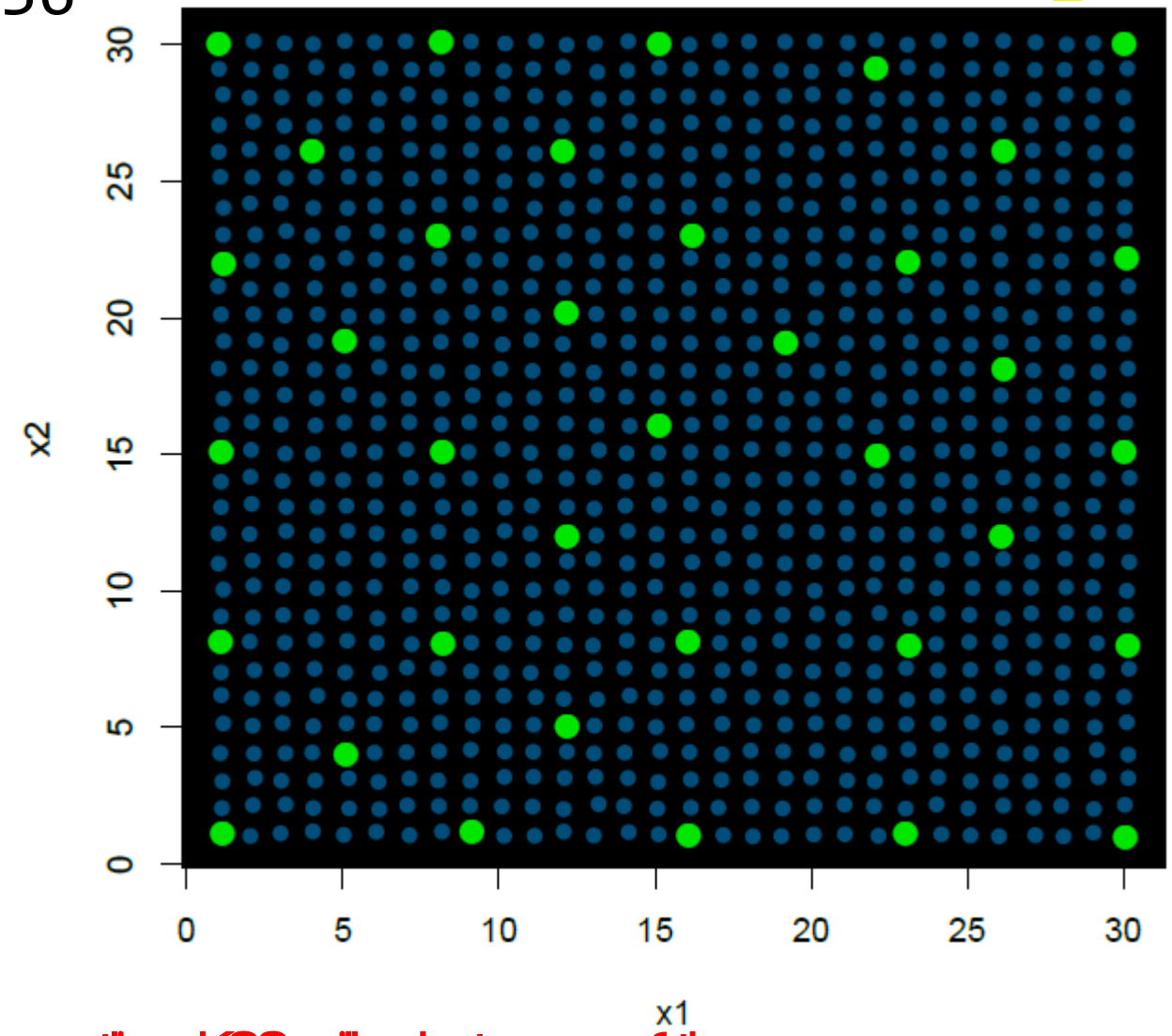
n = 14

Kennard-Stone sampling (KSS)



n = 36

Kennard-Stone sampling (KSS)



If there are outliers KSS will select some of them

K-means sampling (KMS)

The KMS algorithm (Naes, 1987) to select a (calibration) subset of m samples (S_{tr}) from a given set of N samples (S) works as follows:

1. Carry out a k-means clustering on the principal component scores and choose the number of resulting clusters to be equal to the number of desired calibration samples (m),

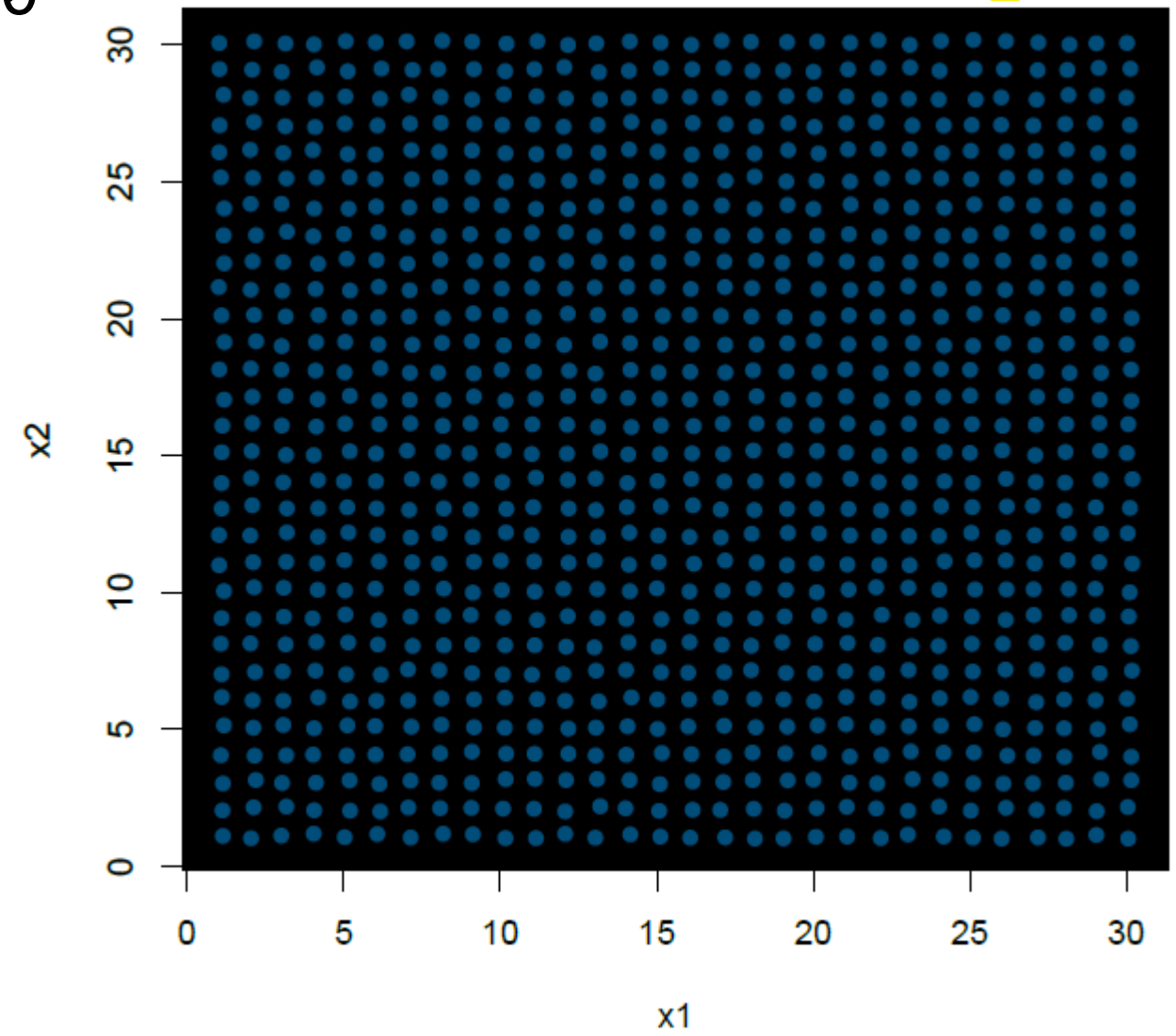
2. Select one sample from each cluster (usually the closest to the center of the cluster).

The **naes** function of the **prospectr** package can be used for KMS

Naes, T., 1987. The design of calibration in near infra-red reflectance analysis by clustering. Journal of Chemometrics 1, 121-134.

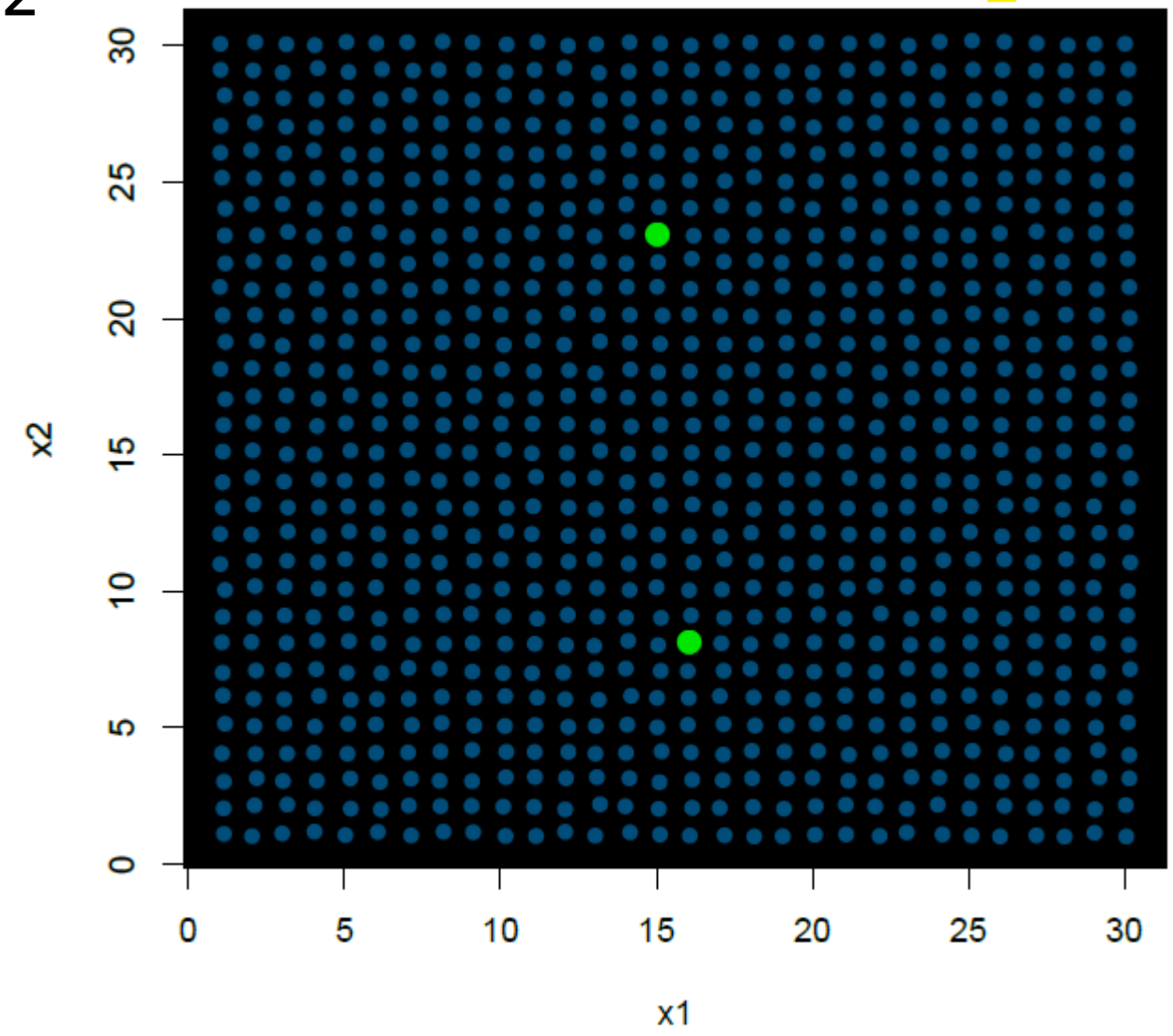
n = 0

K-means sampling (KMS)



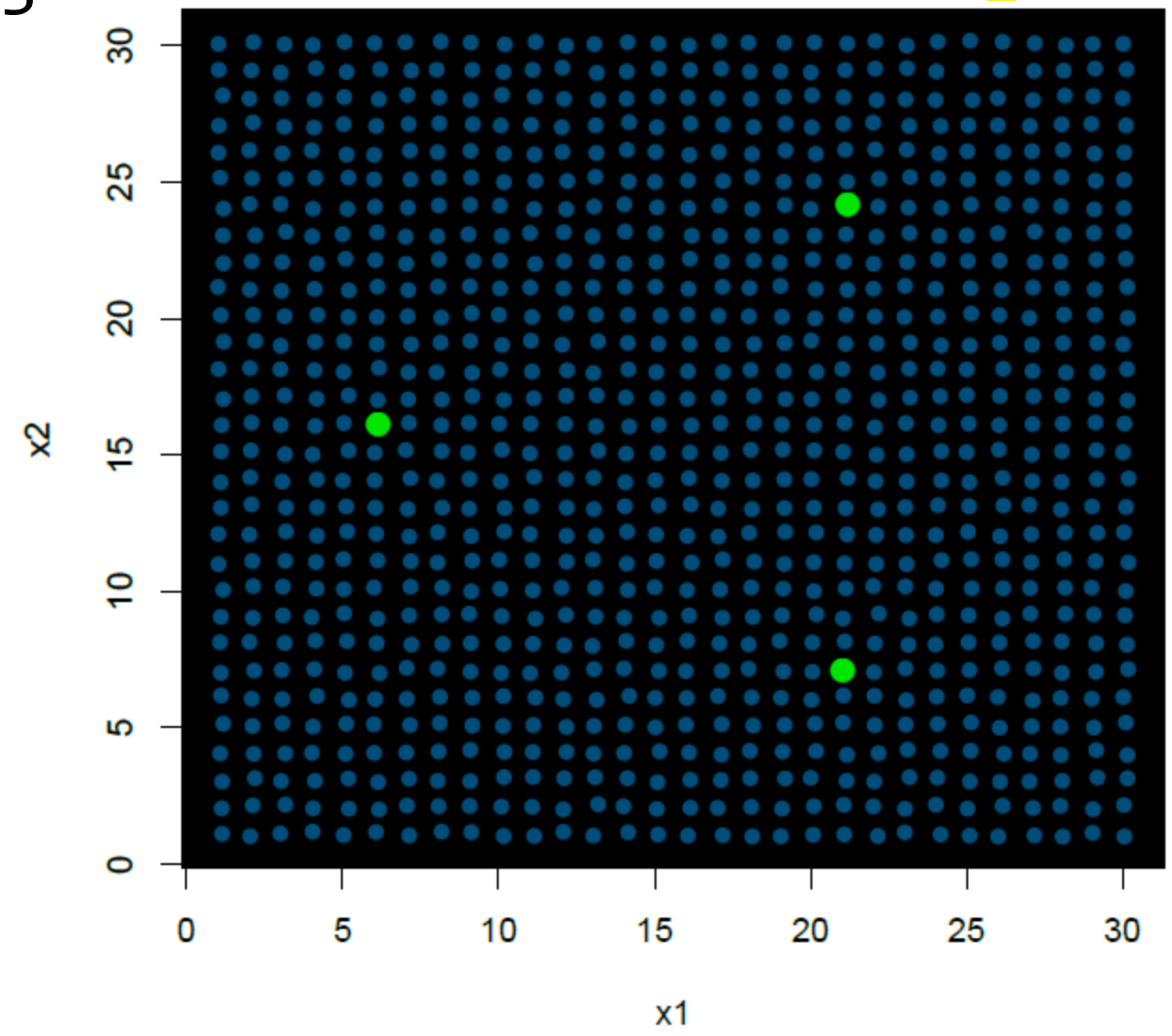
n = 2

K-means sampling (KMS)



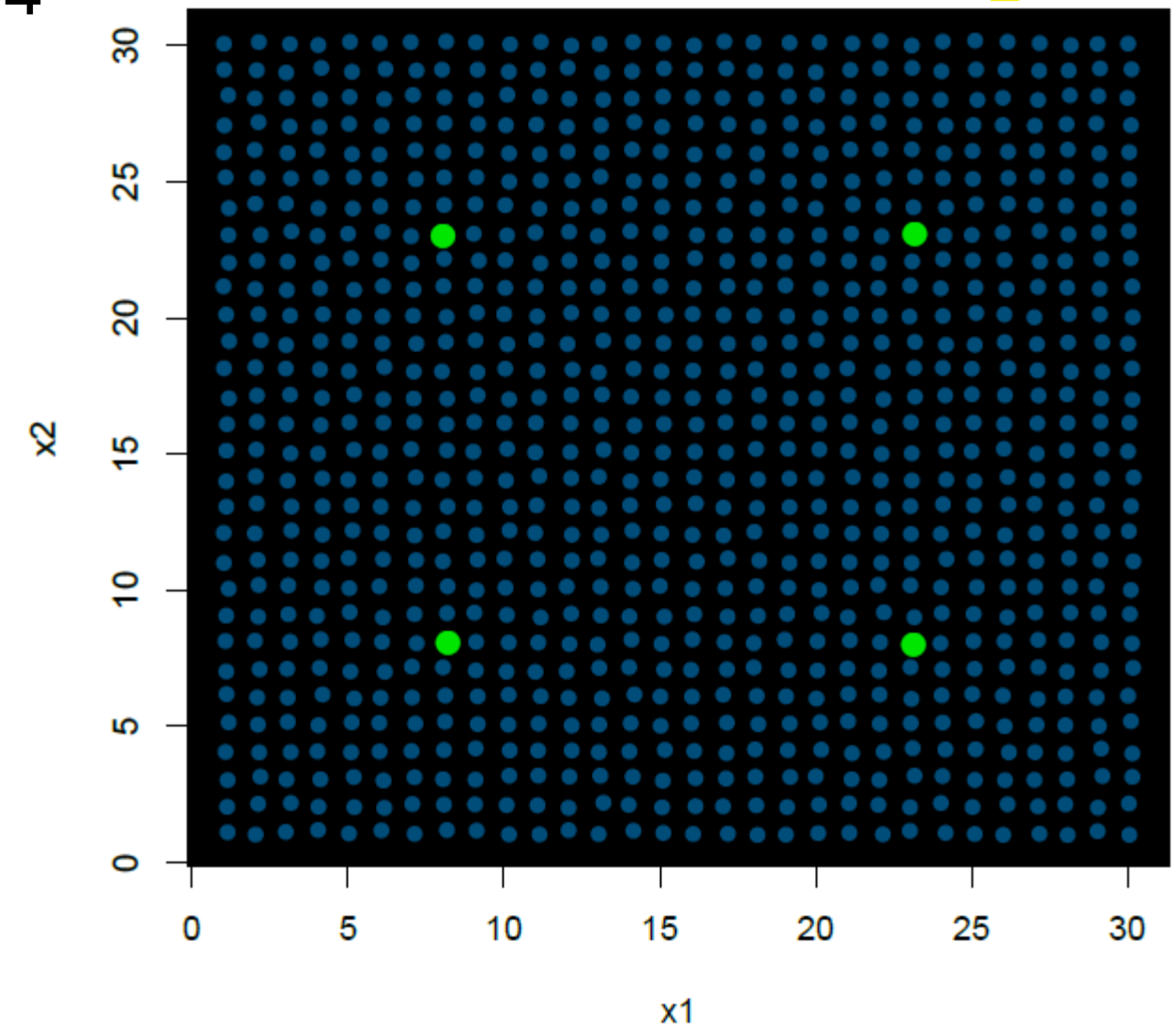
n = 3

K-means sampling (KMS)



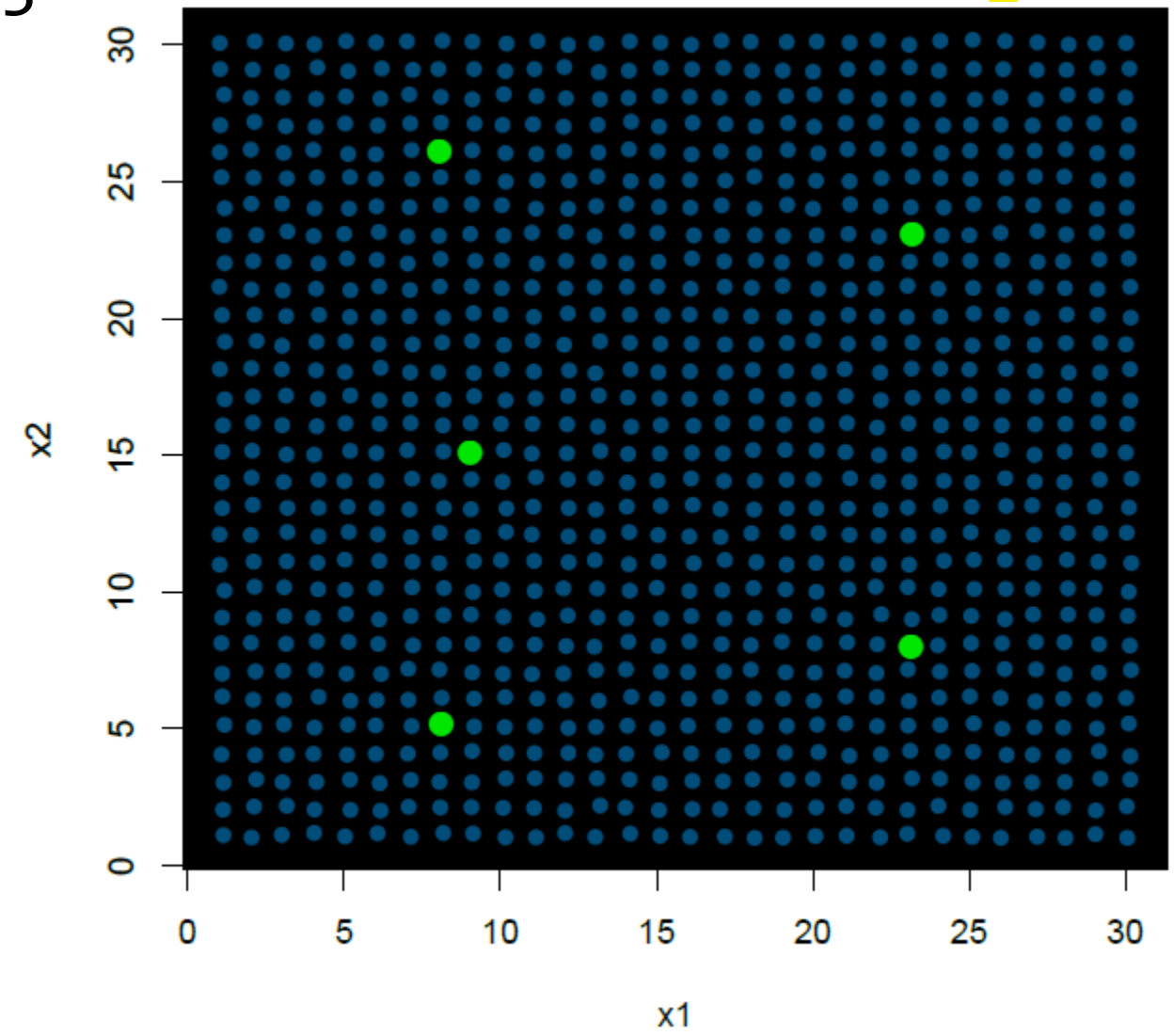
n = 4

K-means sampling (KMS)



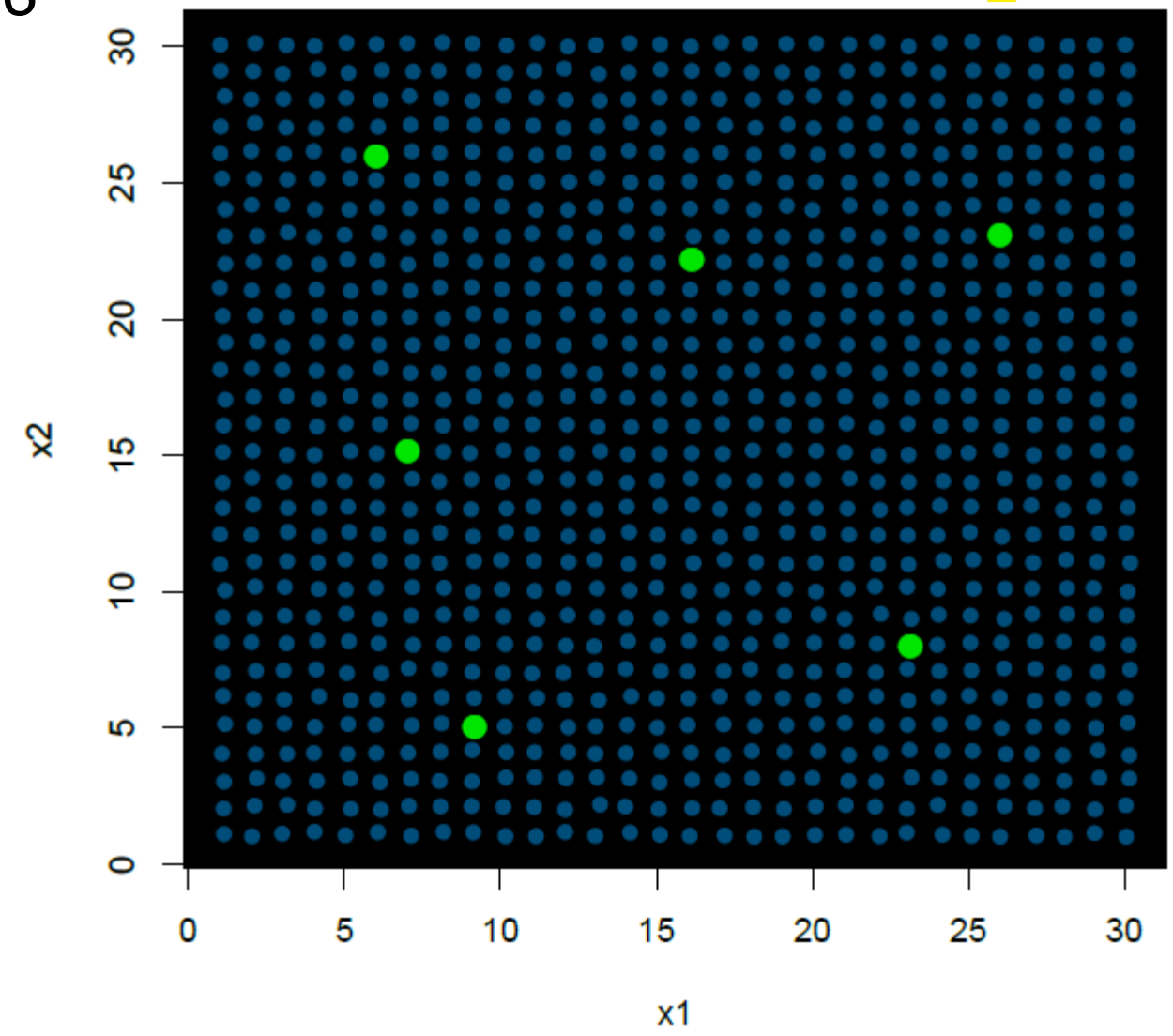
n = 5

K-means sampling (KMS)



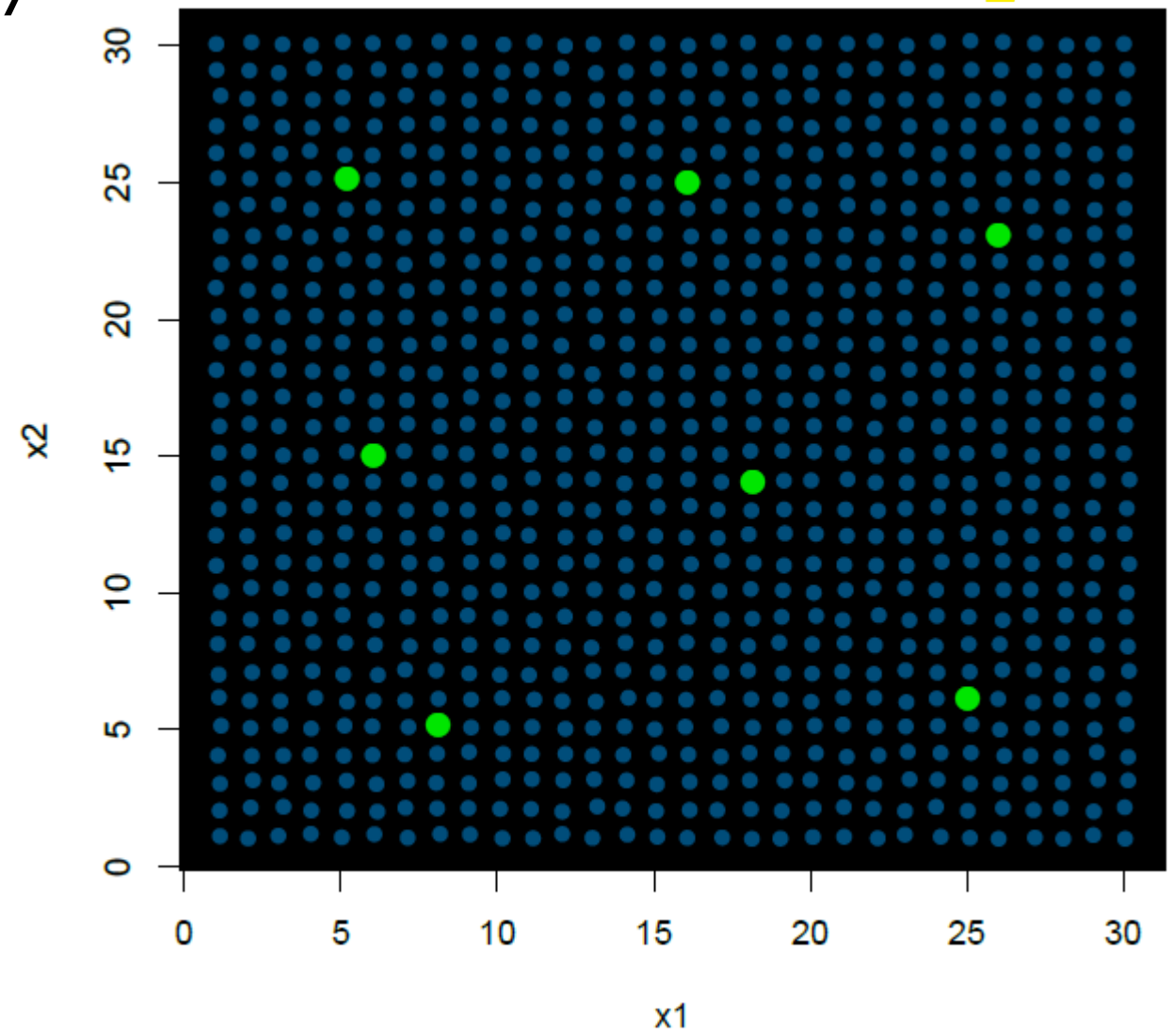
n = 6

K-means sampling (KMS)



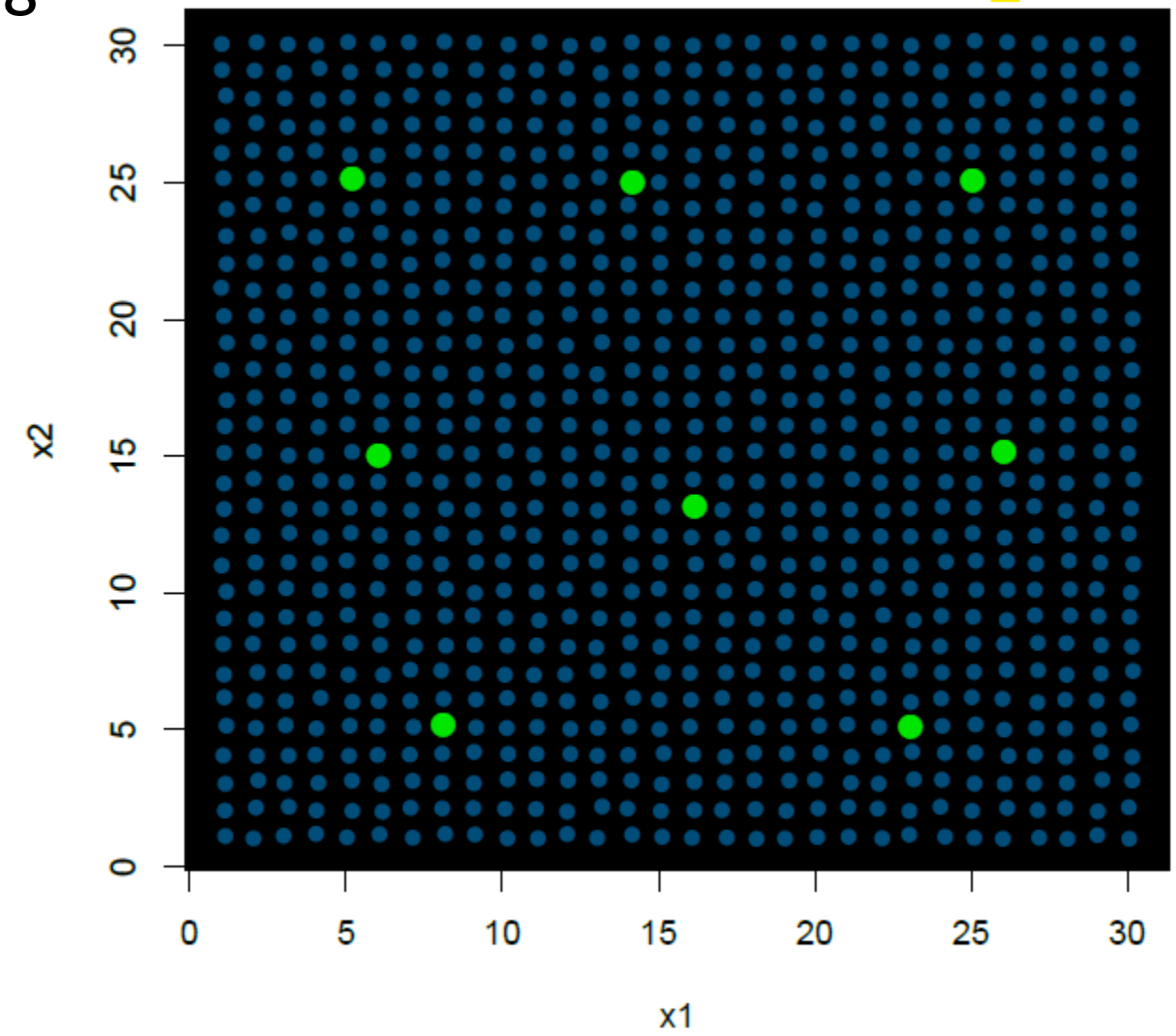
n = 7

K-means sampling (KMS)



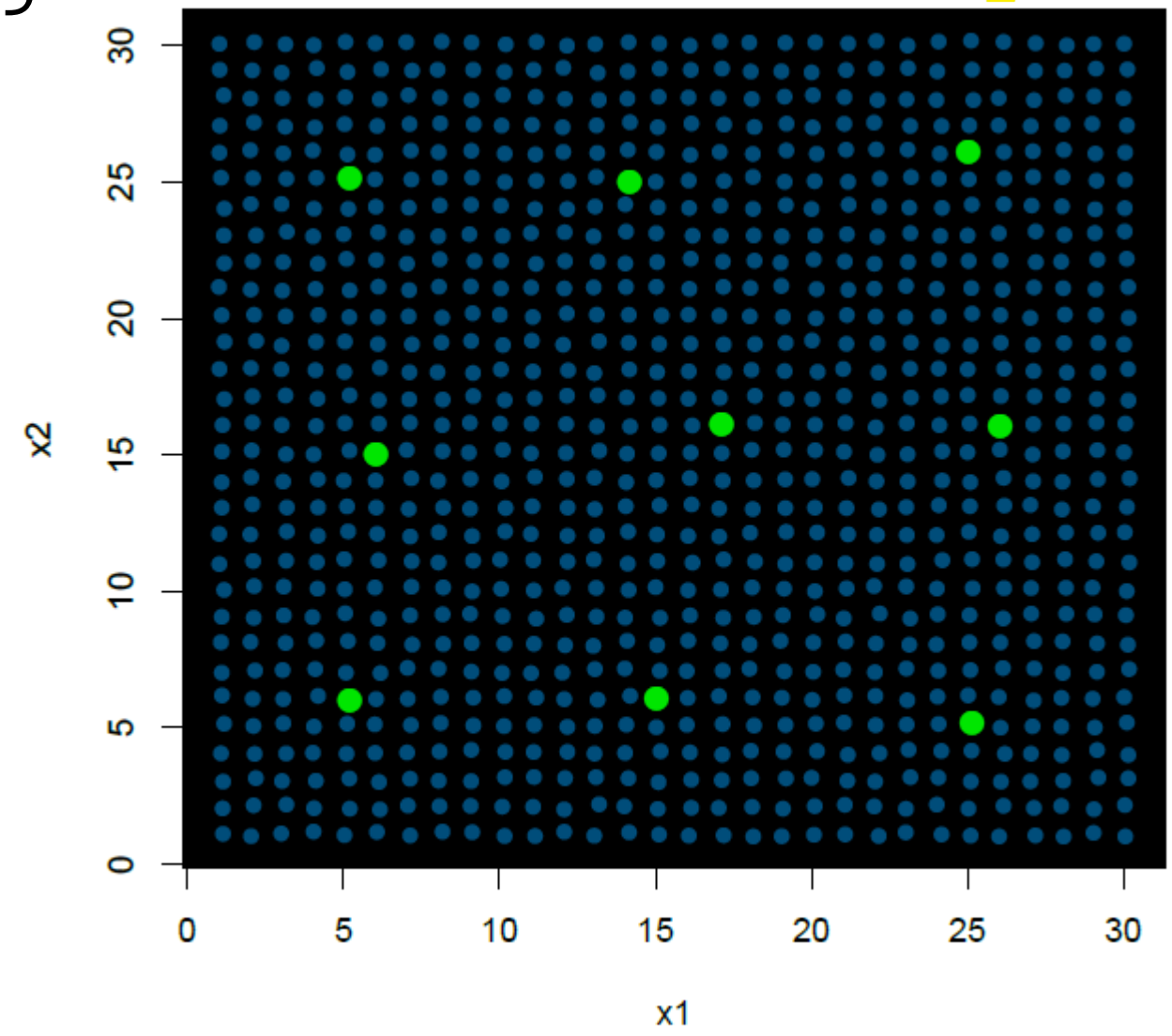
n = 8

K-means sampling (KMS)



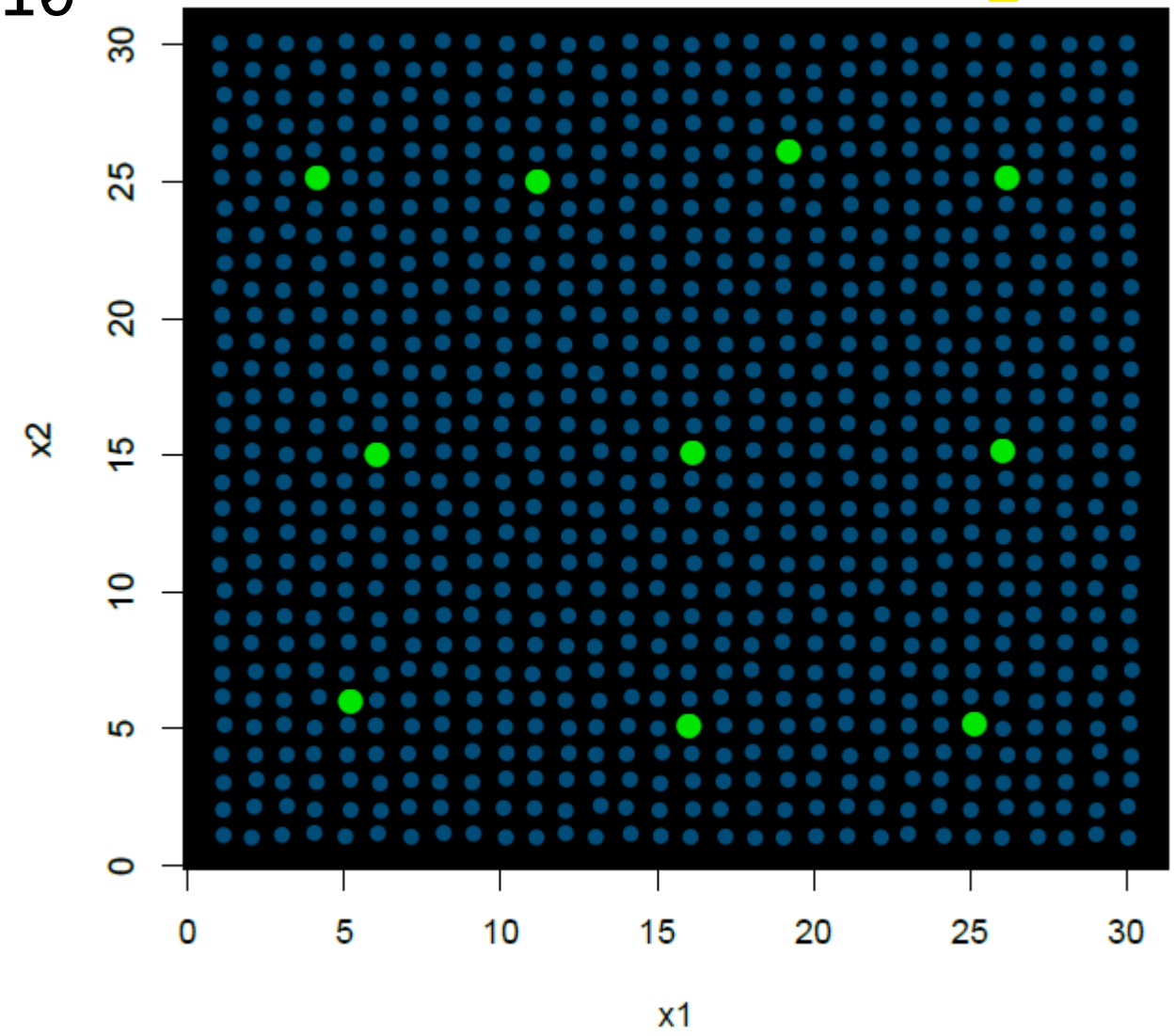
n = 9

K-means sampling (KMS)



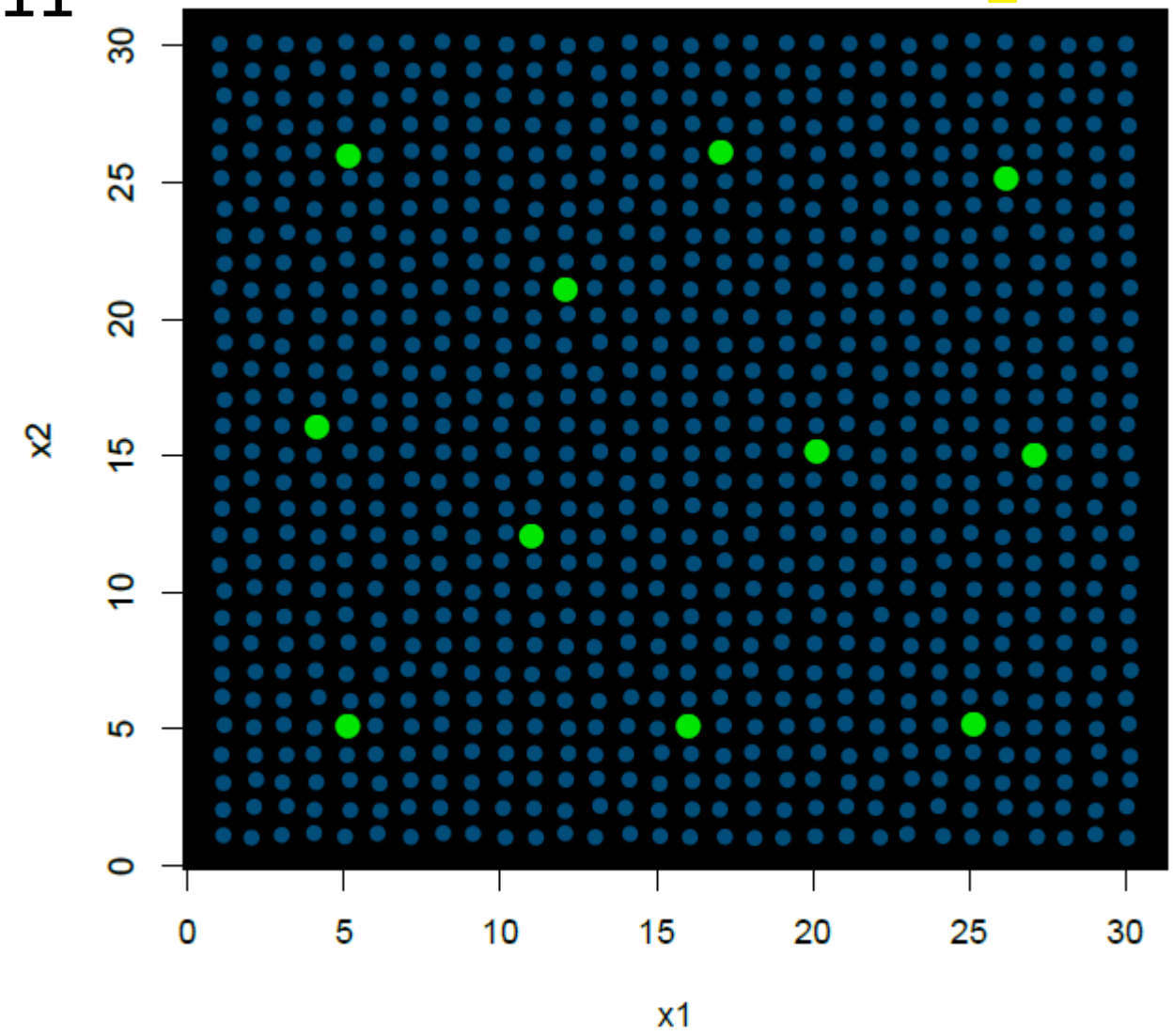
n = 10

K-means sampling (KMS)



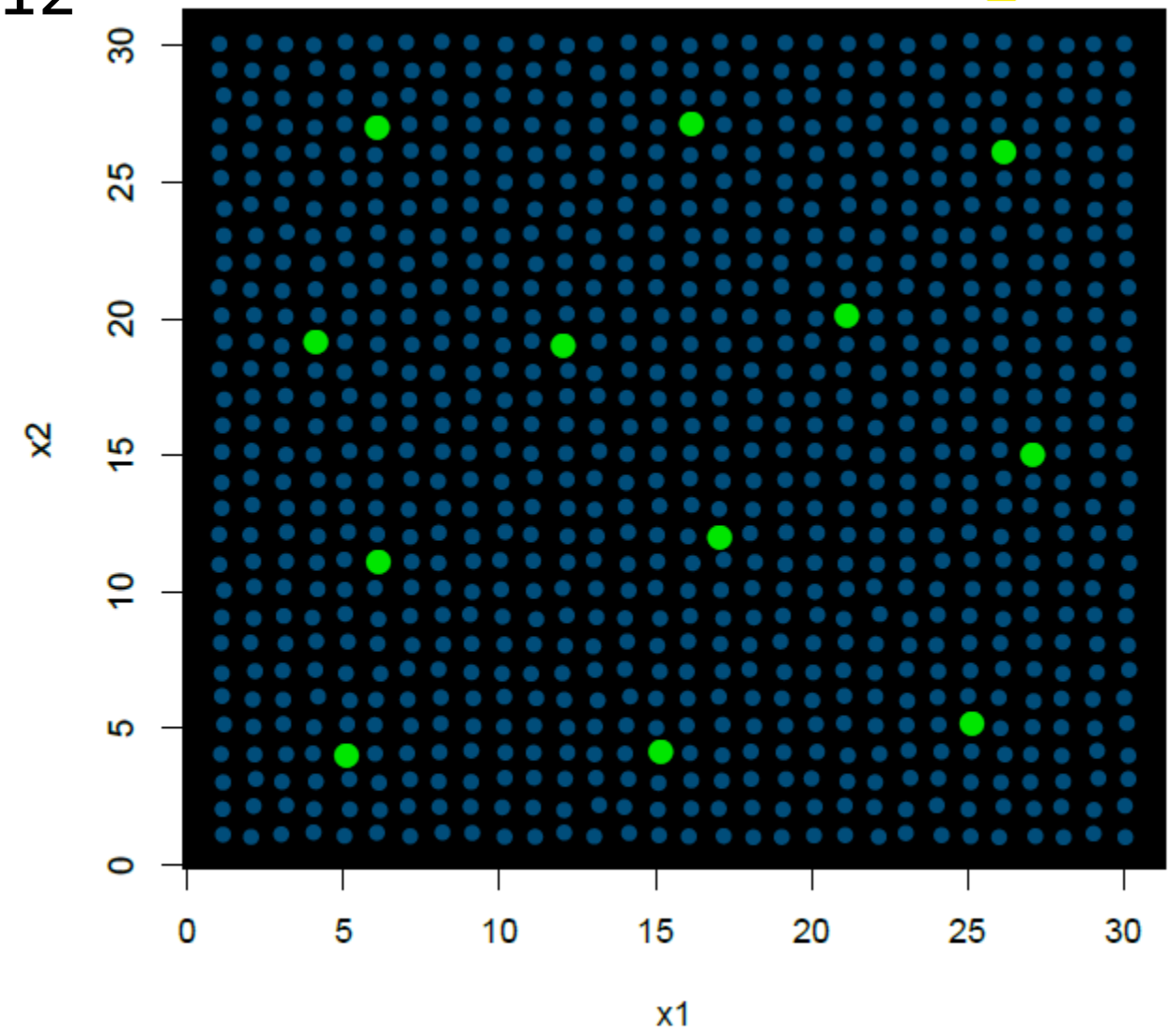
n = 11

K-means sampling (KMS)



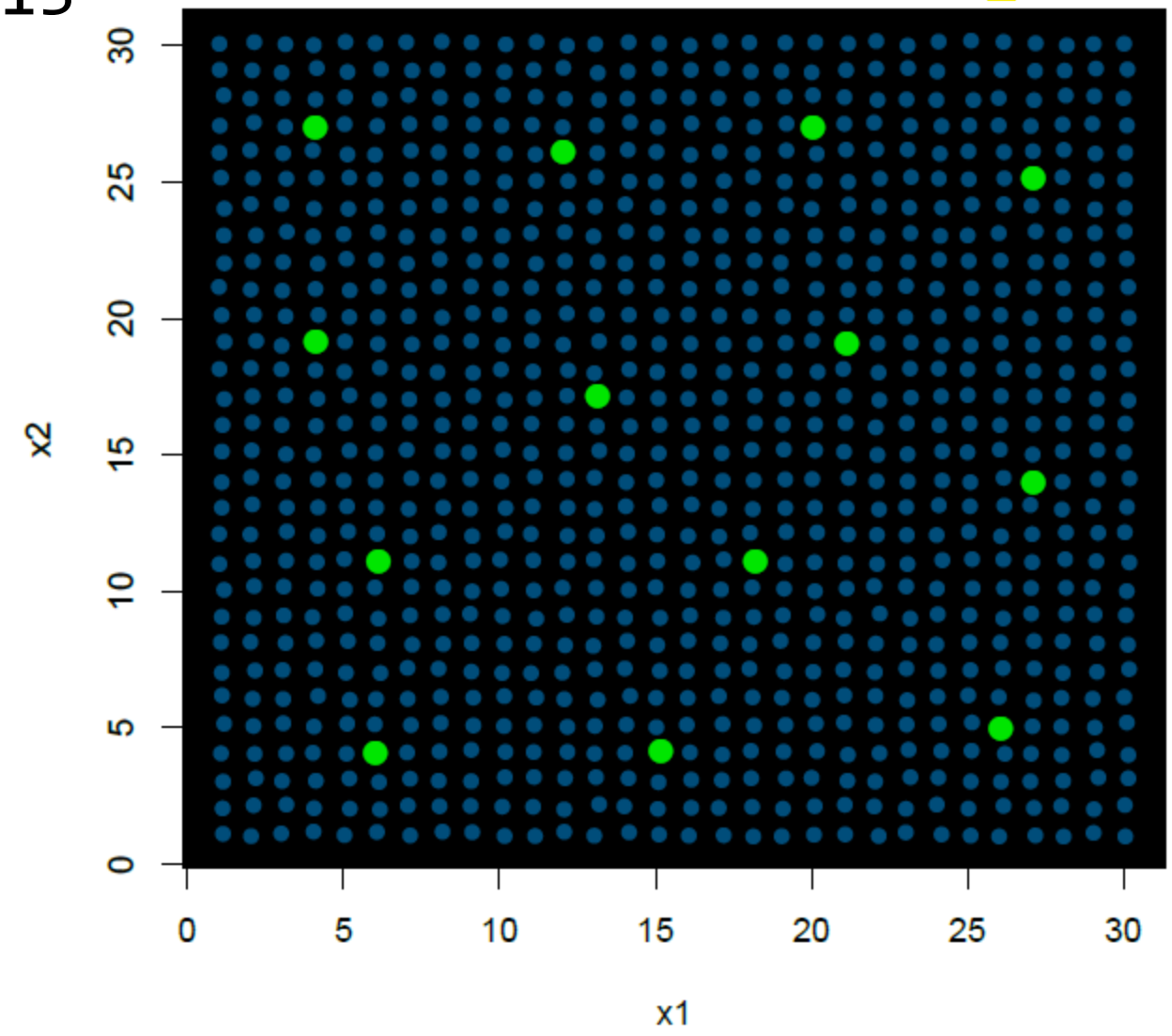
n = 12

K-means sampling (KMS)



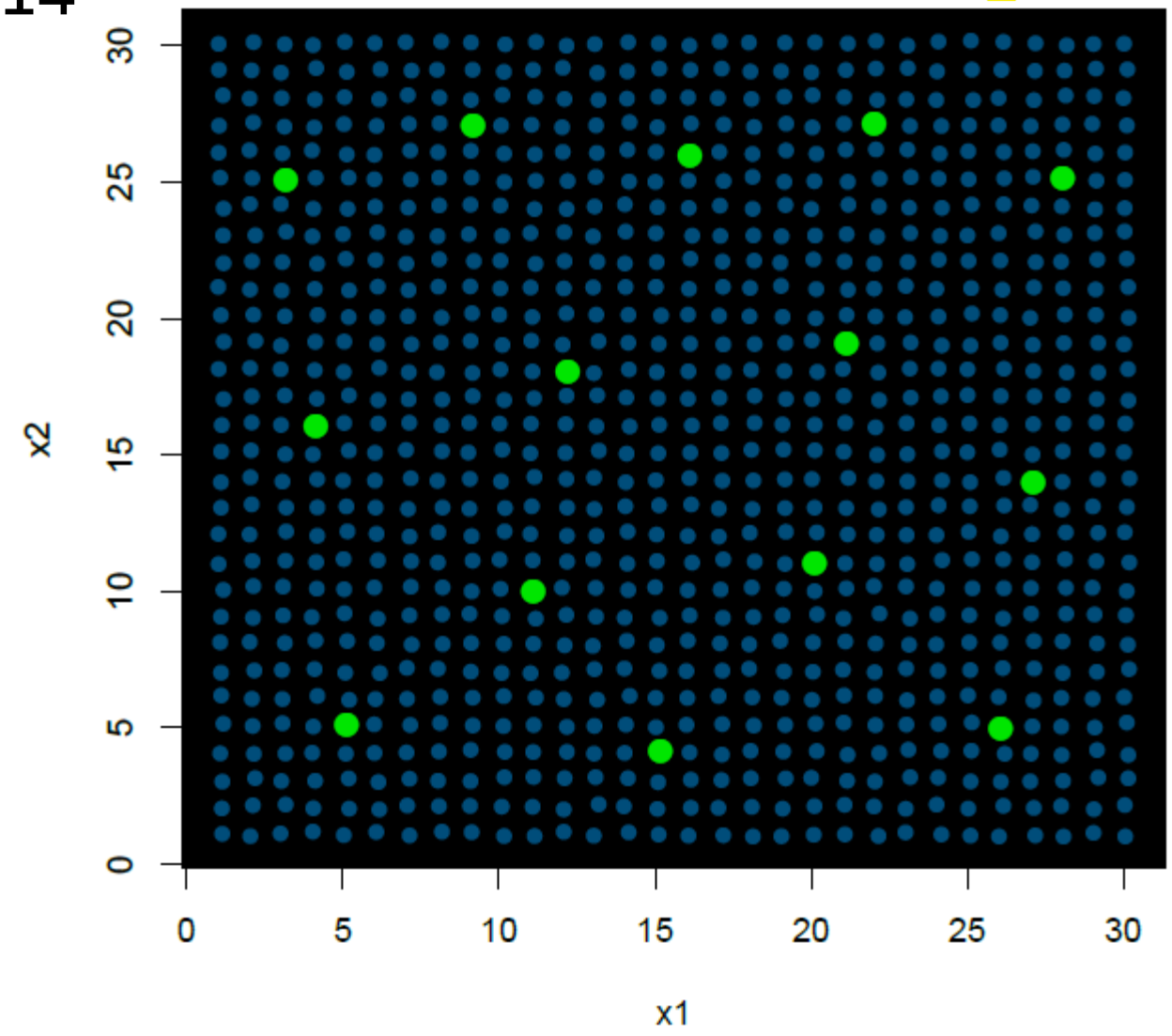
n = 13

K-means sampling (KMS)



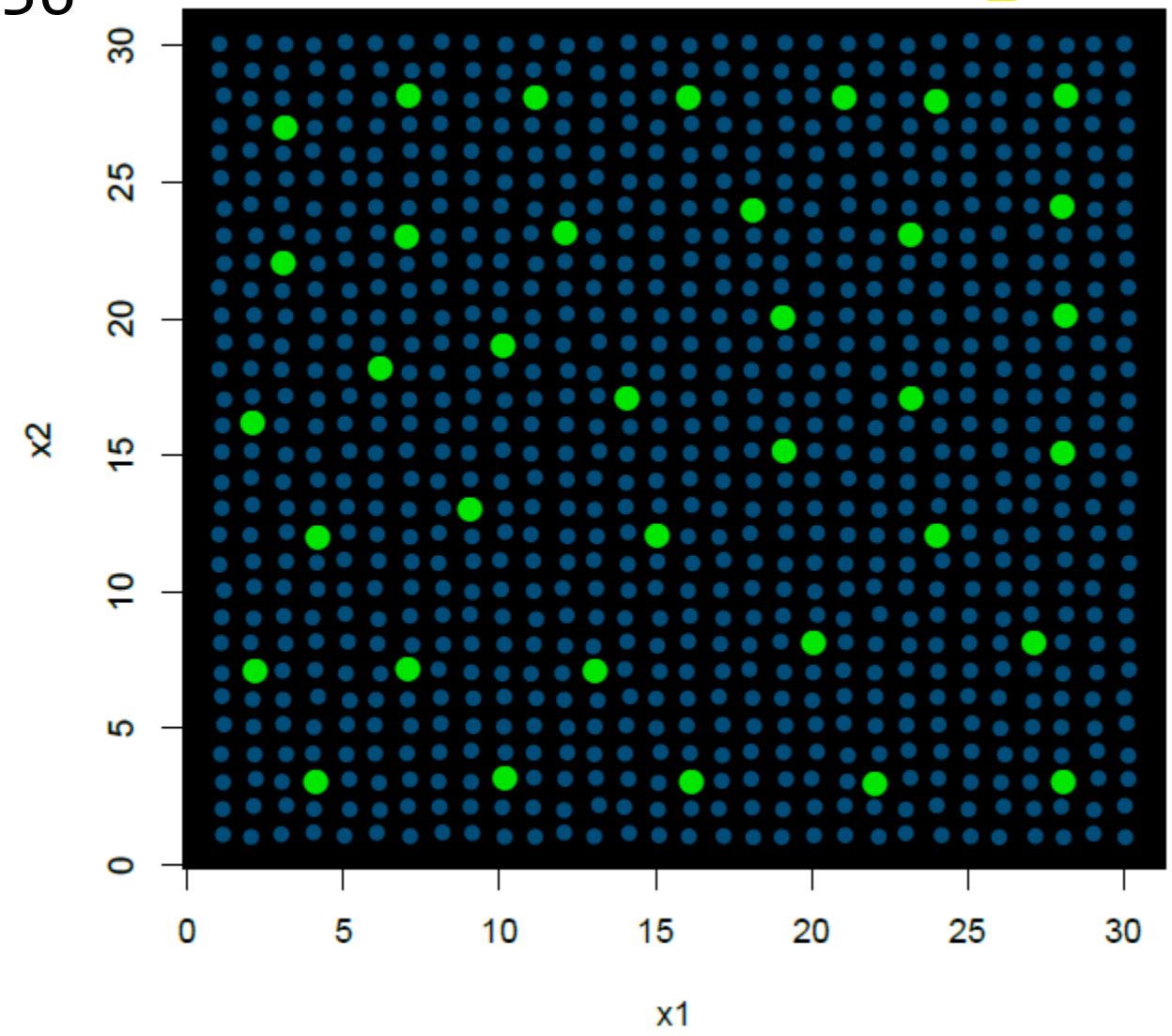
n = 14

K-means sampling (KMS)

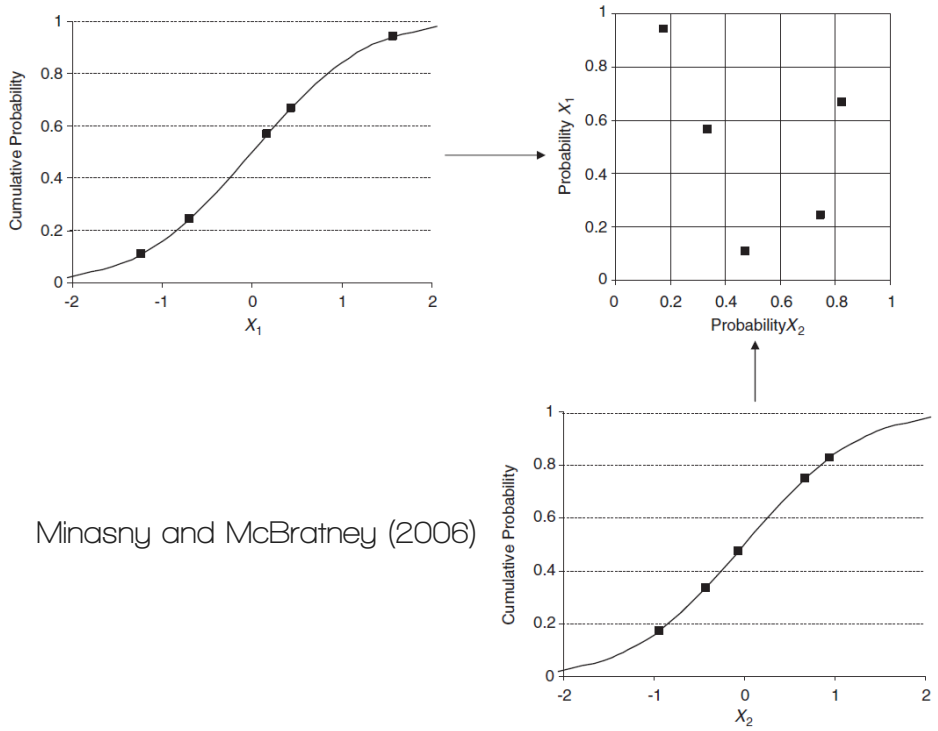


n = 36

K-means sampling (KMS)



Conditioned Latin hypercube sampling (cLHS)



Minasny and McBratney (2006)

The cLHS attempts to cover the multidimensional distribution corresponding to a set of variables by using a stratified random sampling.

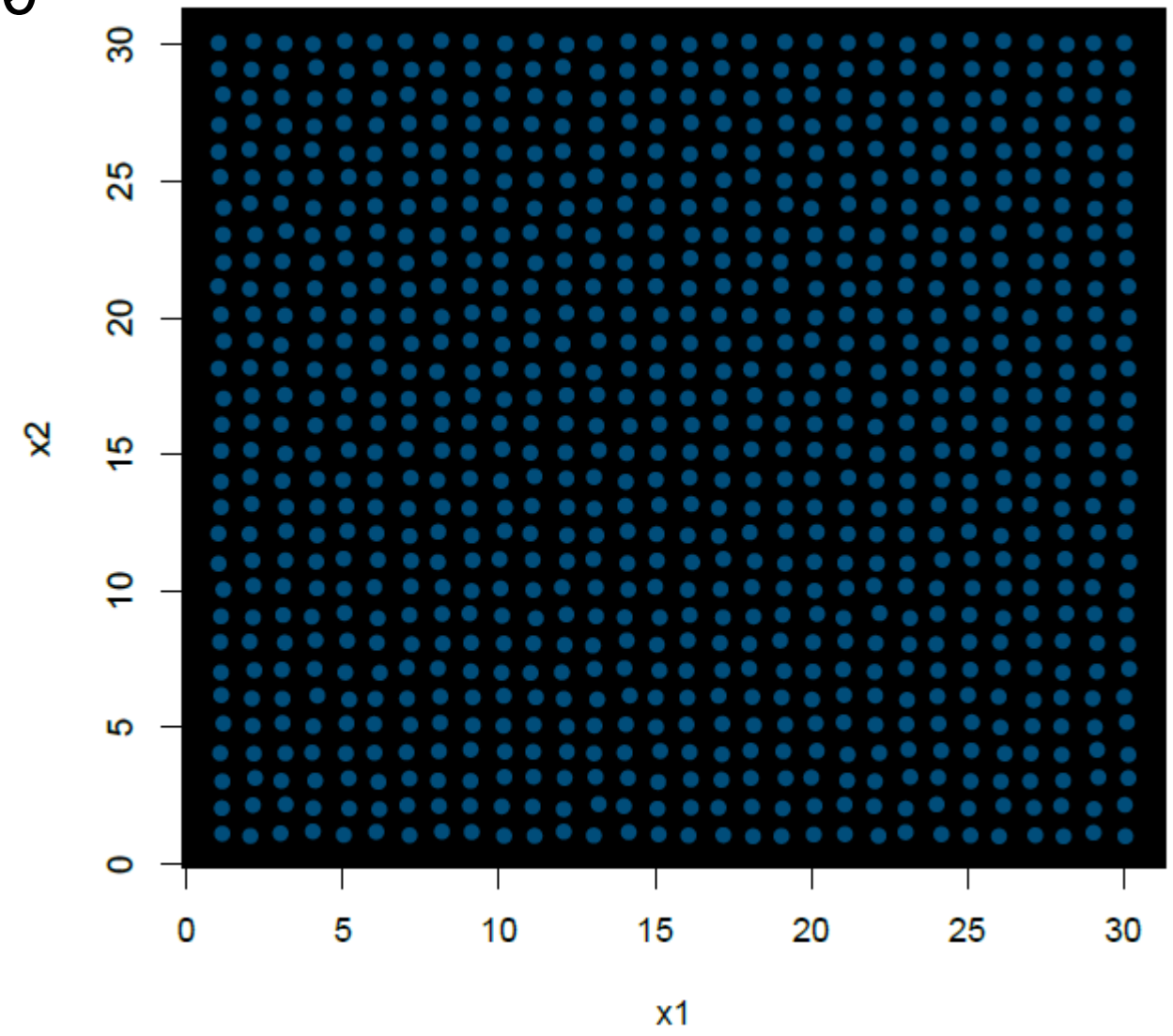
The **clhs** function of the **clhs** package can be used for cLHS

cLHS does not maximize the dissimilarity between samples

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences, 32(9), 1378-1388.

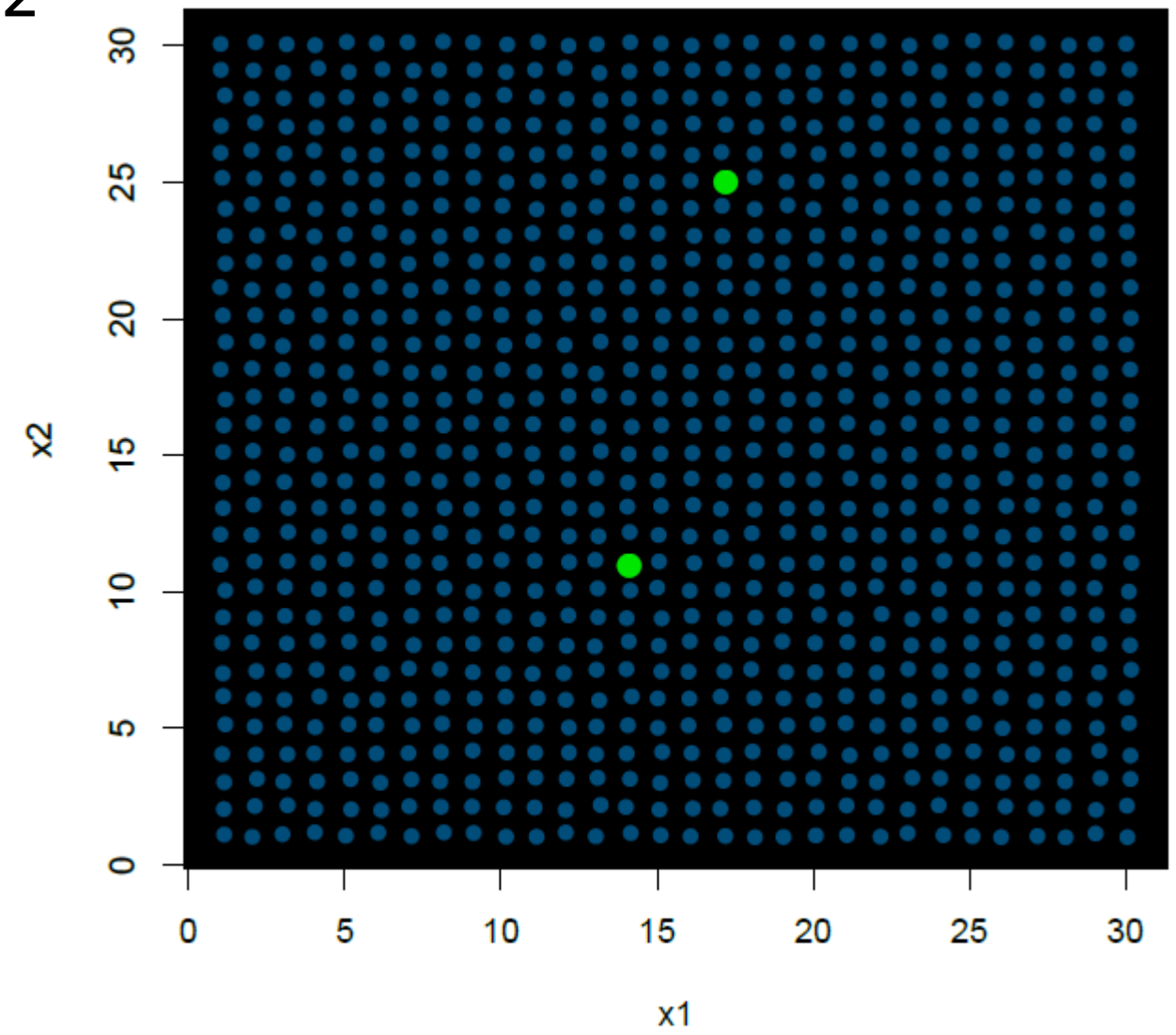
n = 0

contioned Latin Hypercube sampling (cLHS)



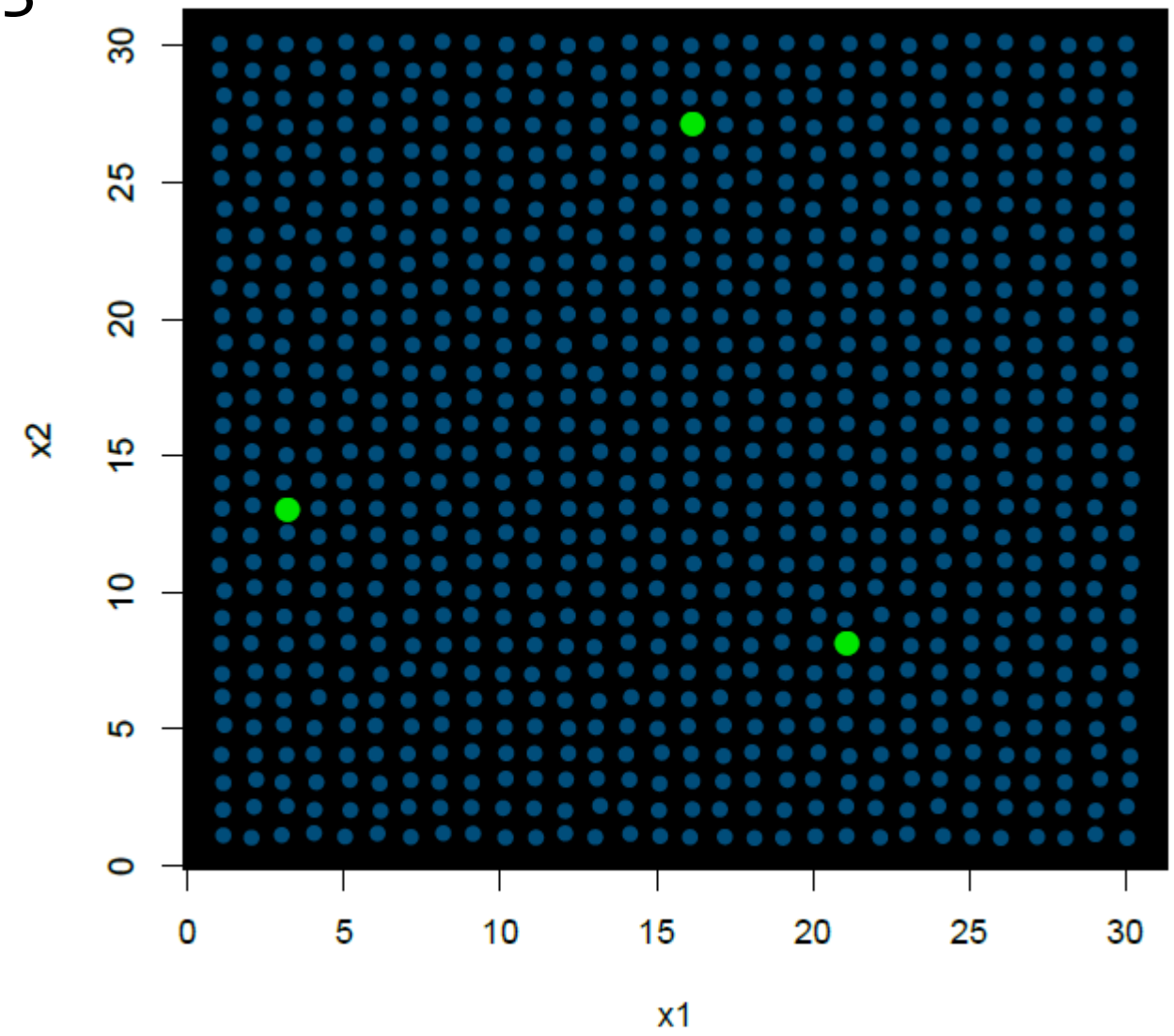
n = 2

contioned Latin Hypercube sampling (cLHS)



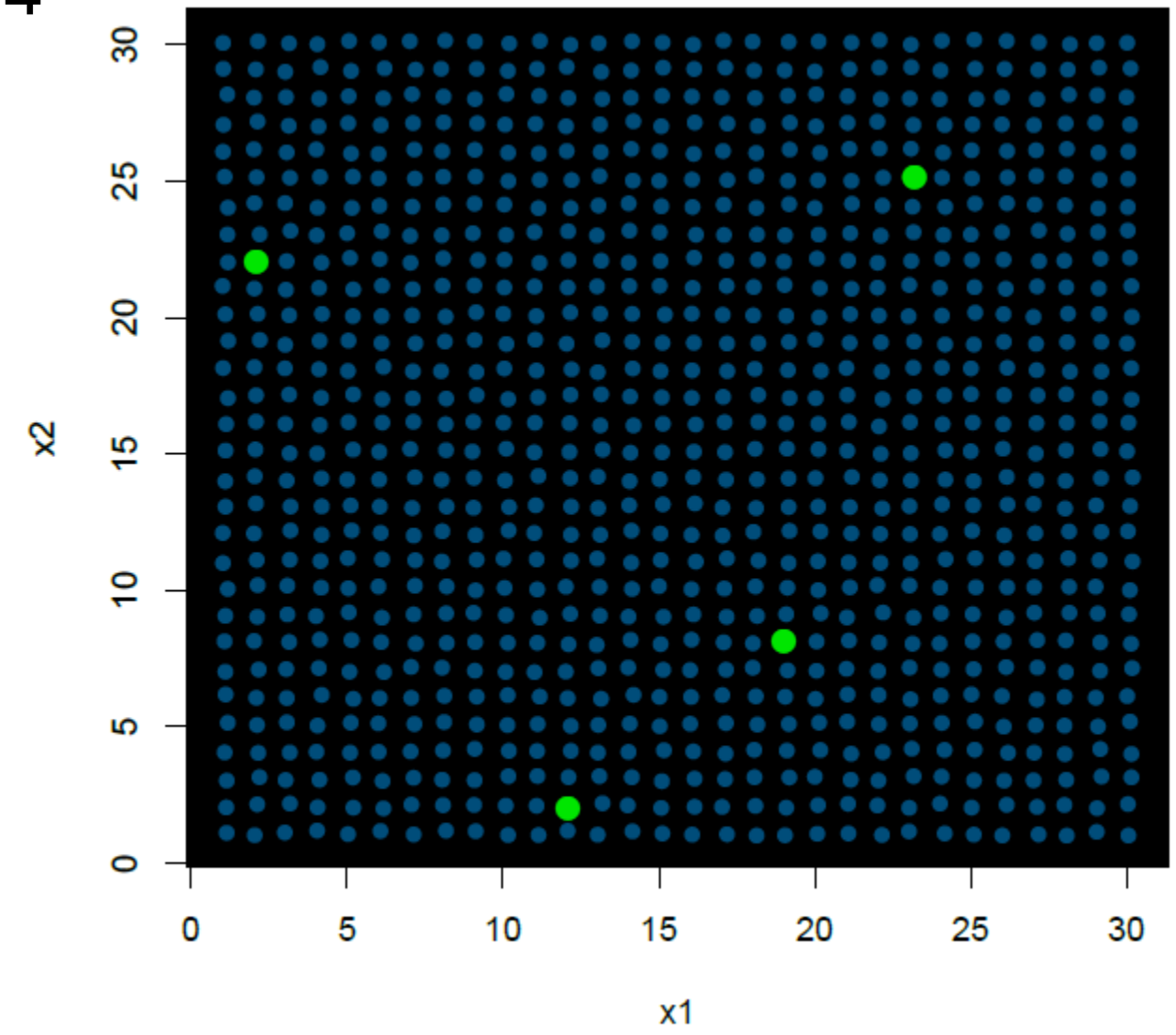
n = 3

contioned Latin Hypercube sampling (cLHS)



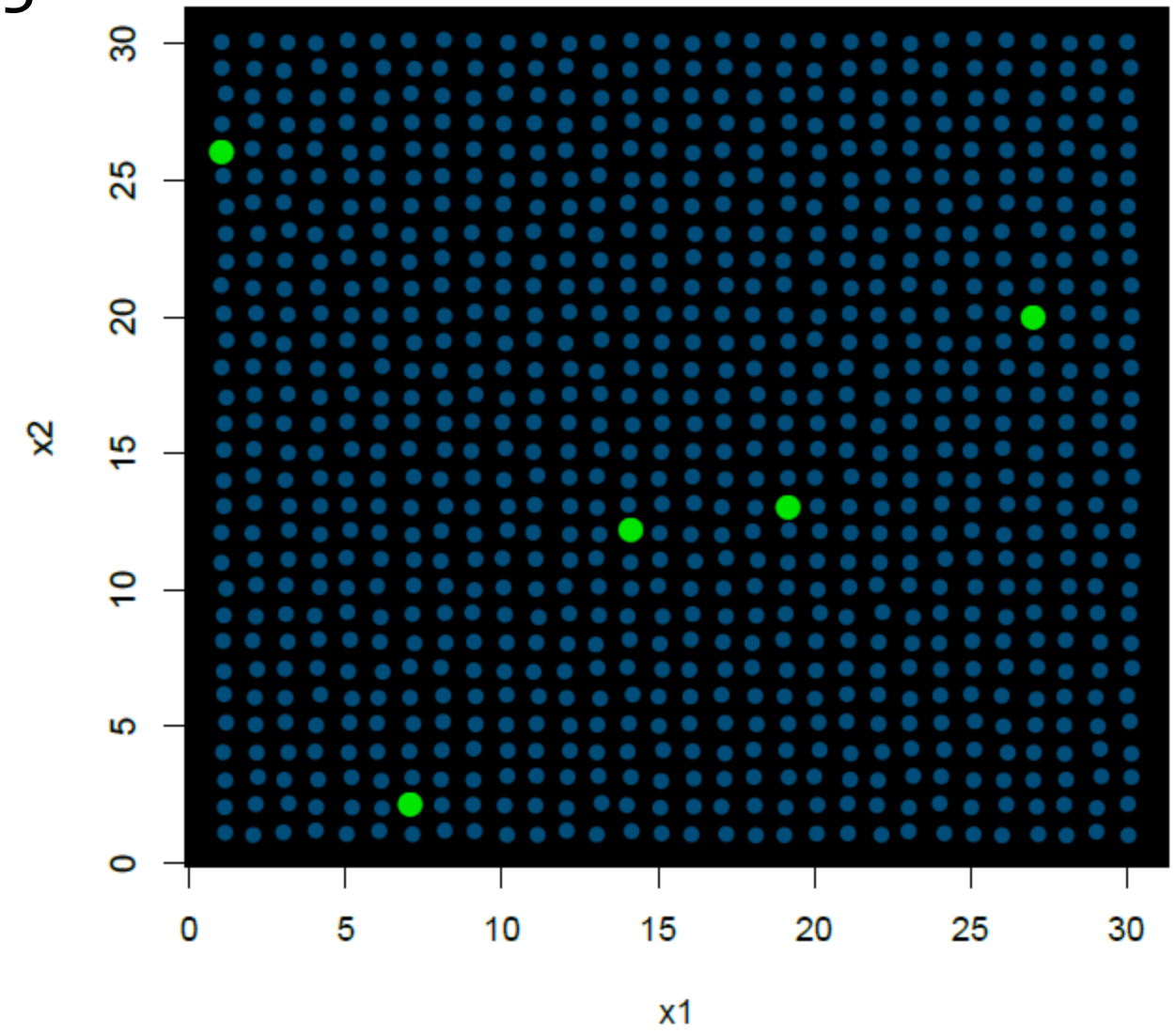
n = 4

contioned Latin Hypercube sampling (cLHS)



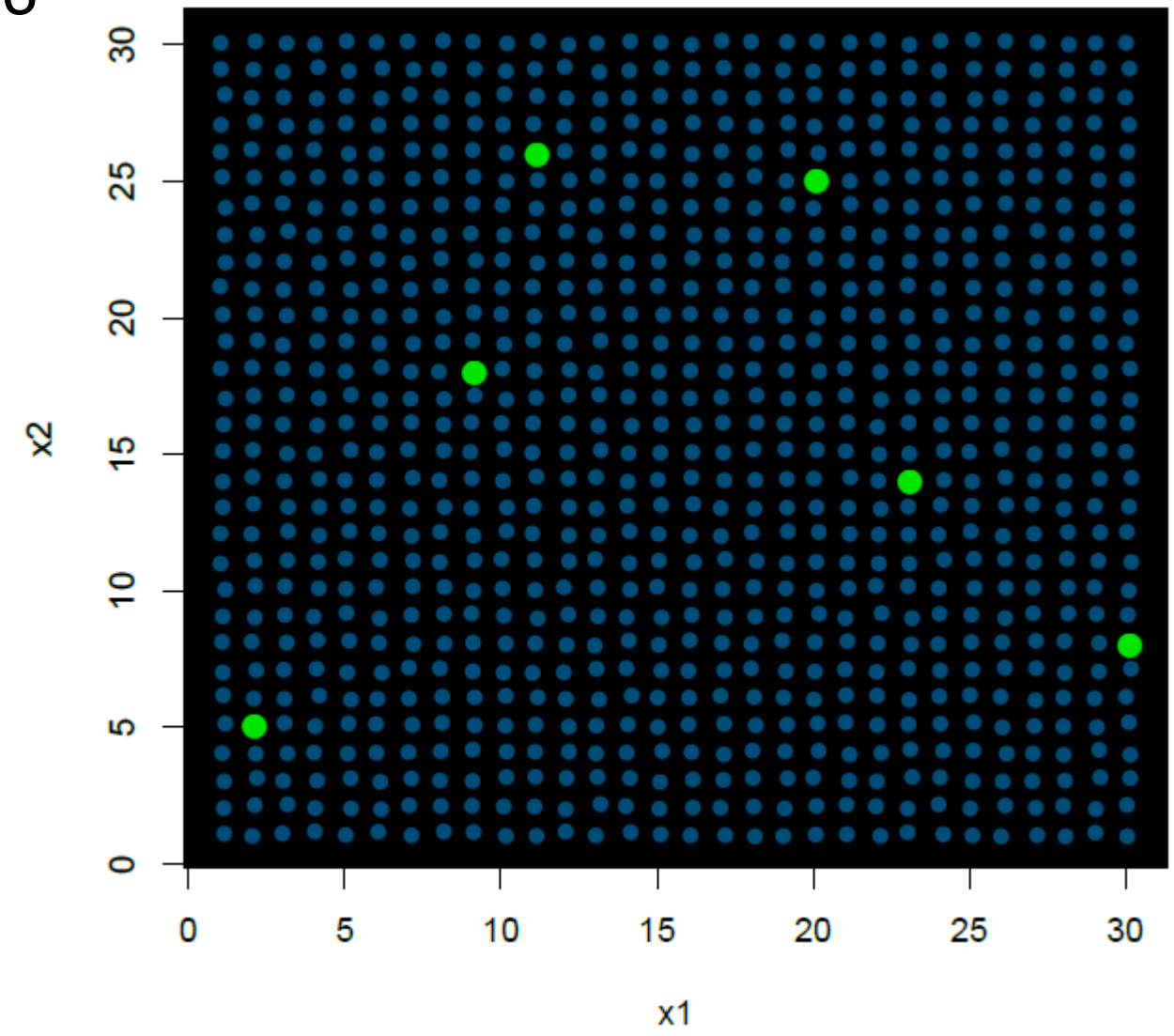
n = 5

contioned Latin Hypercube sampling (cLHS)



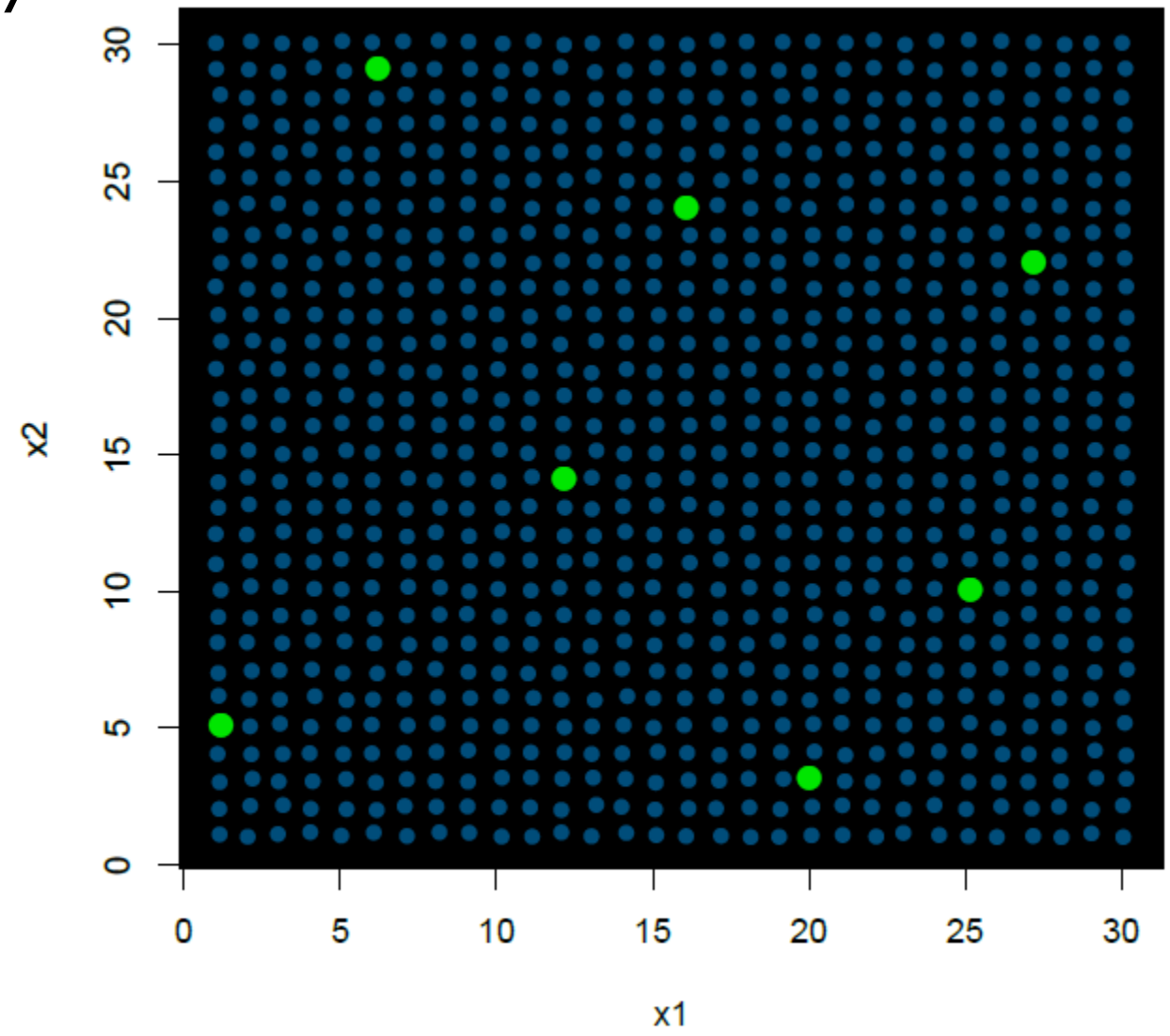
n = 6

contioned Latin Hypercube sampling (cLHS)



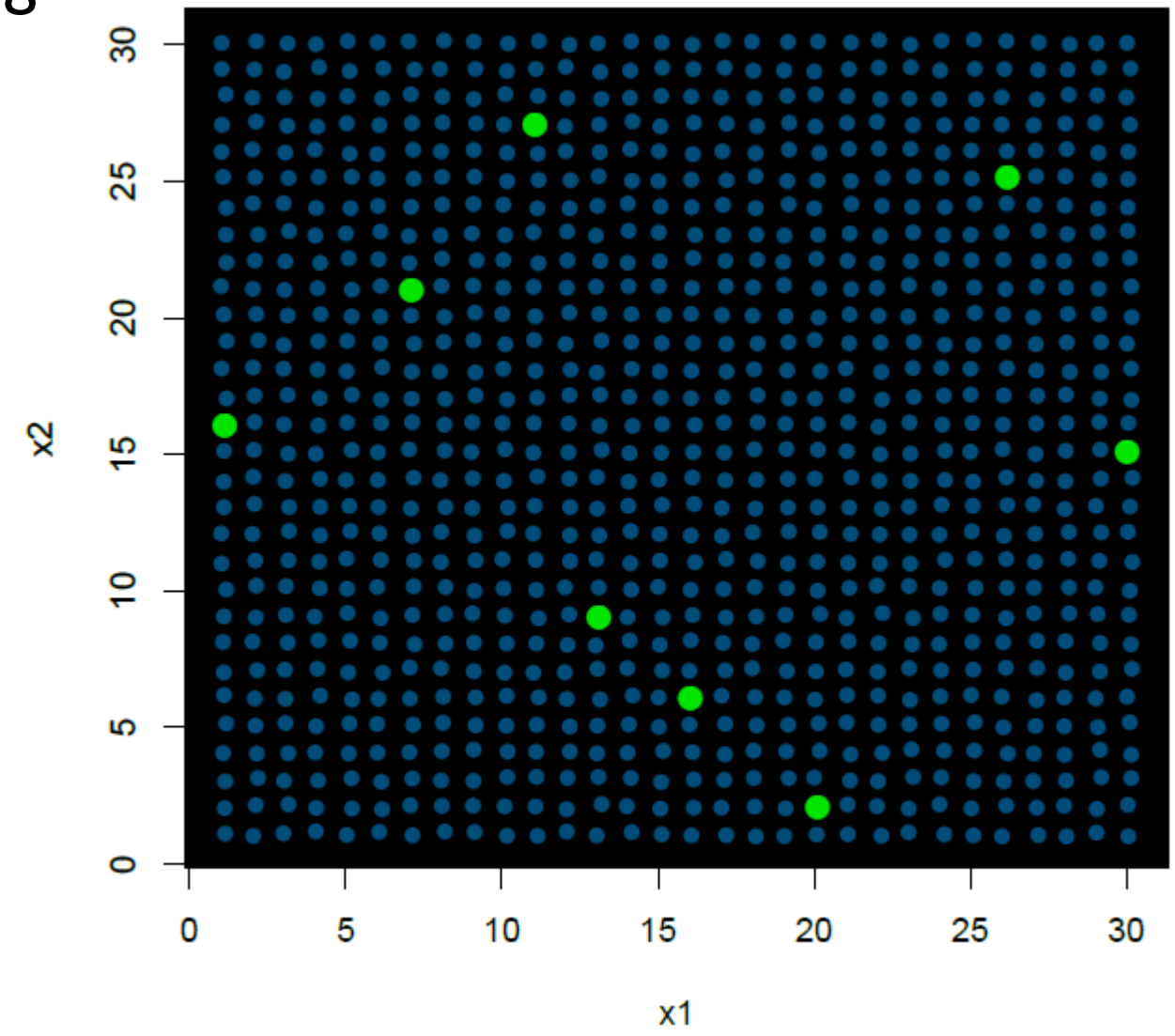
n = 7

contioned Latin Hypercube sampling (cLHS)



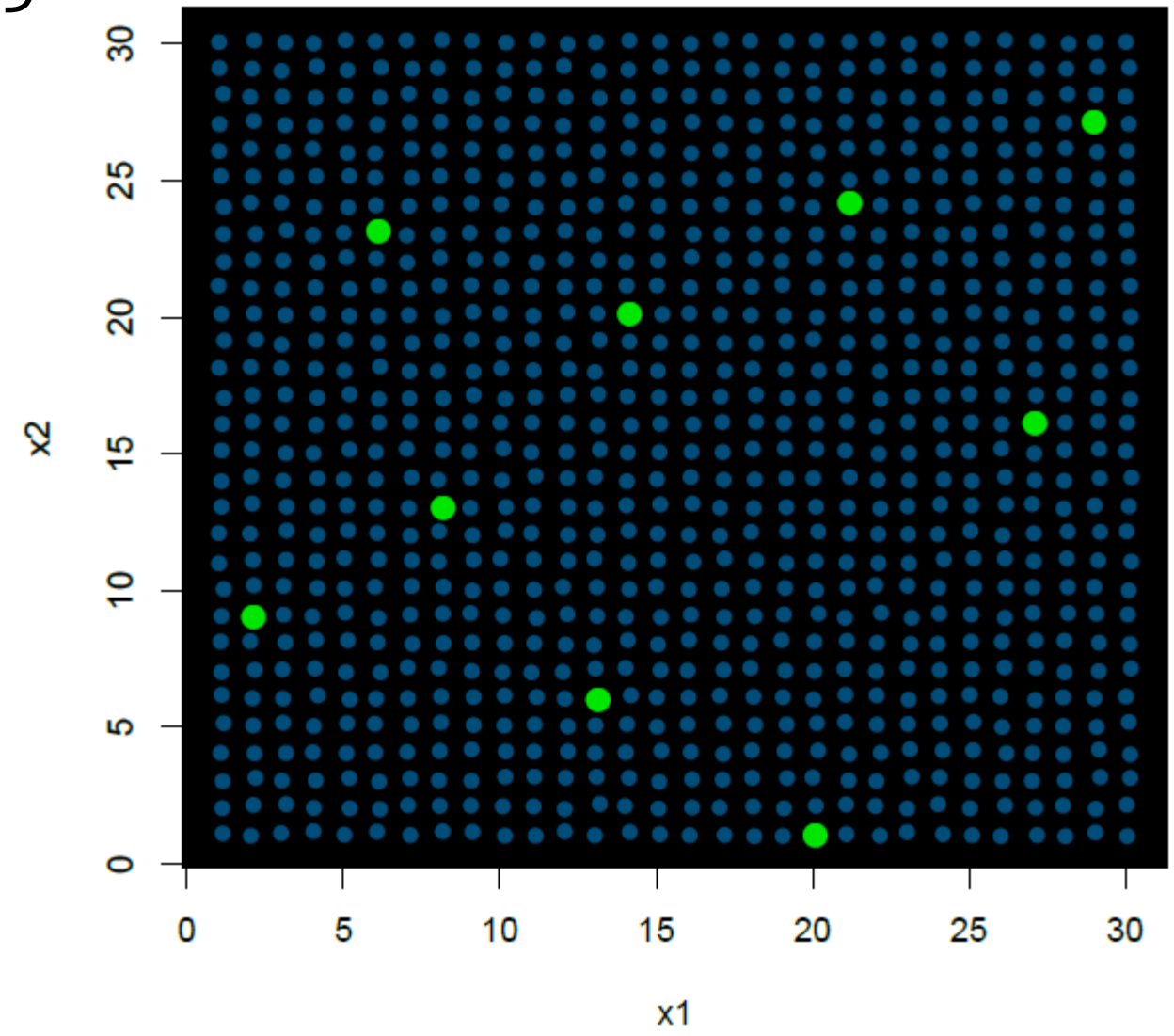
n = 8

contioned Latin Hypercube sampling (cLHS)



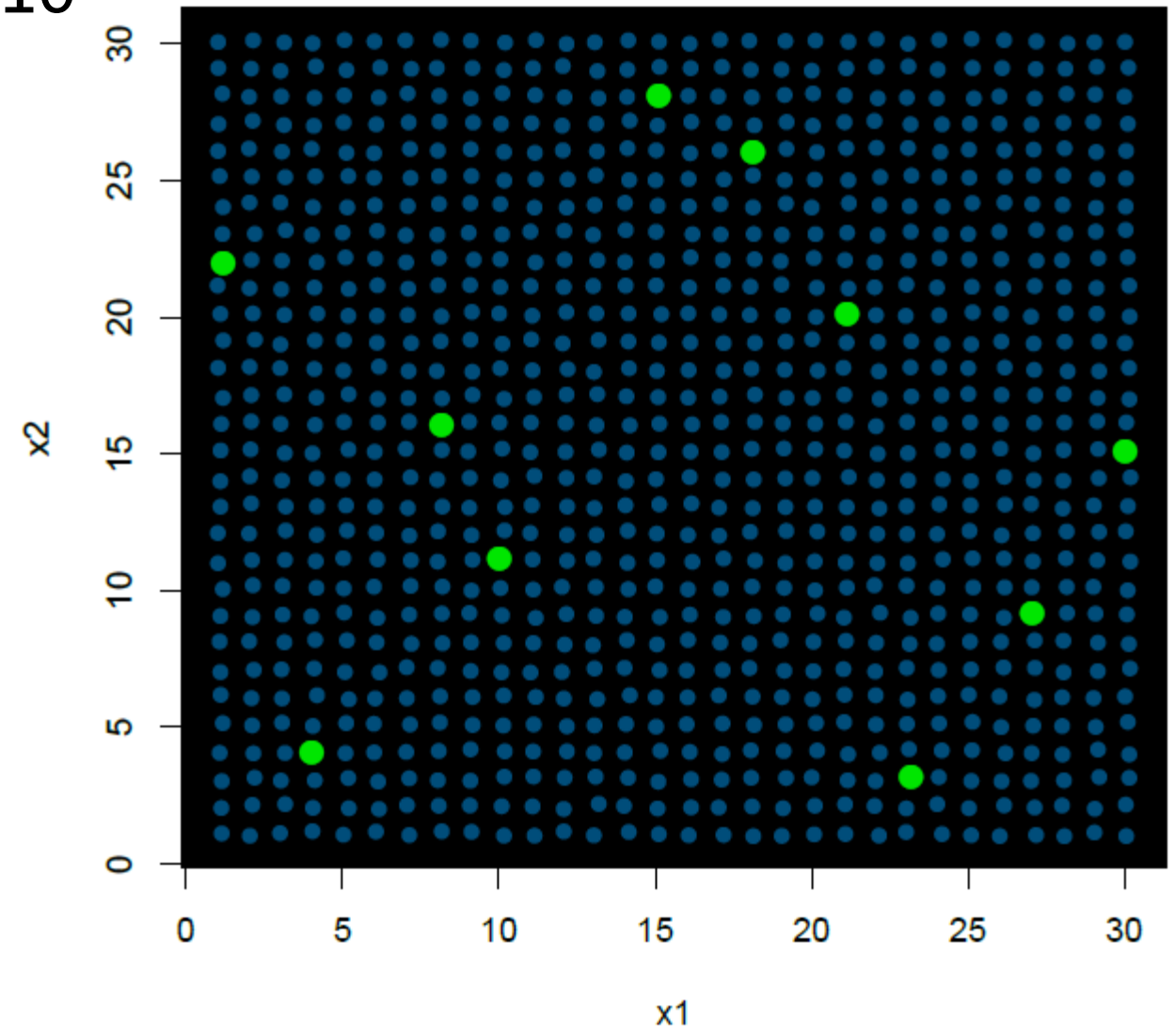
n = 9

contioned Latin Hypercube sampling (cLHS)



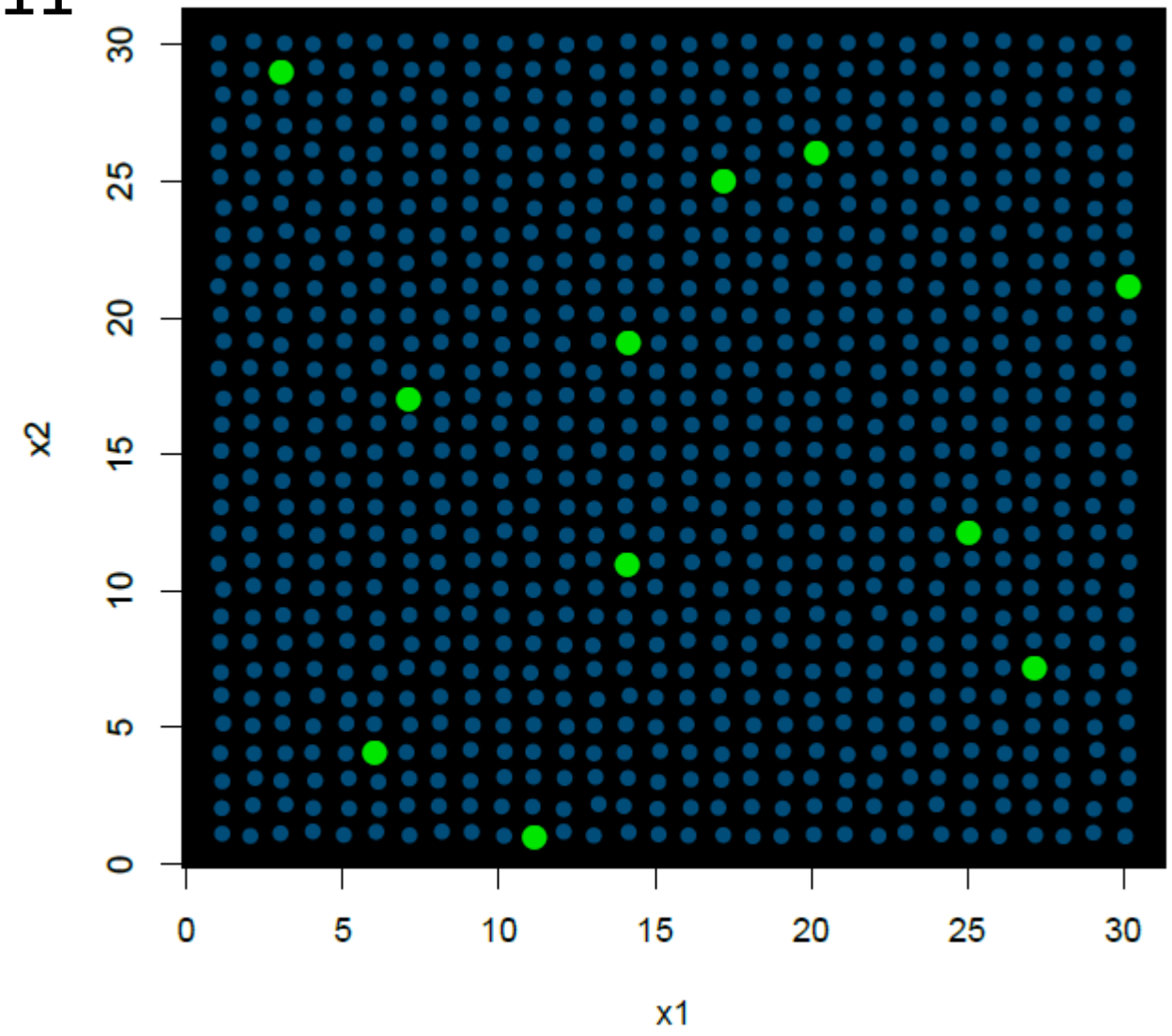
n = 10

contioned Latin Hypercube sampling (cLHS)



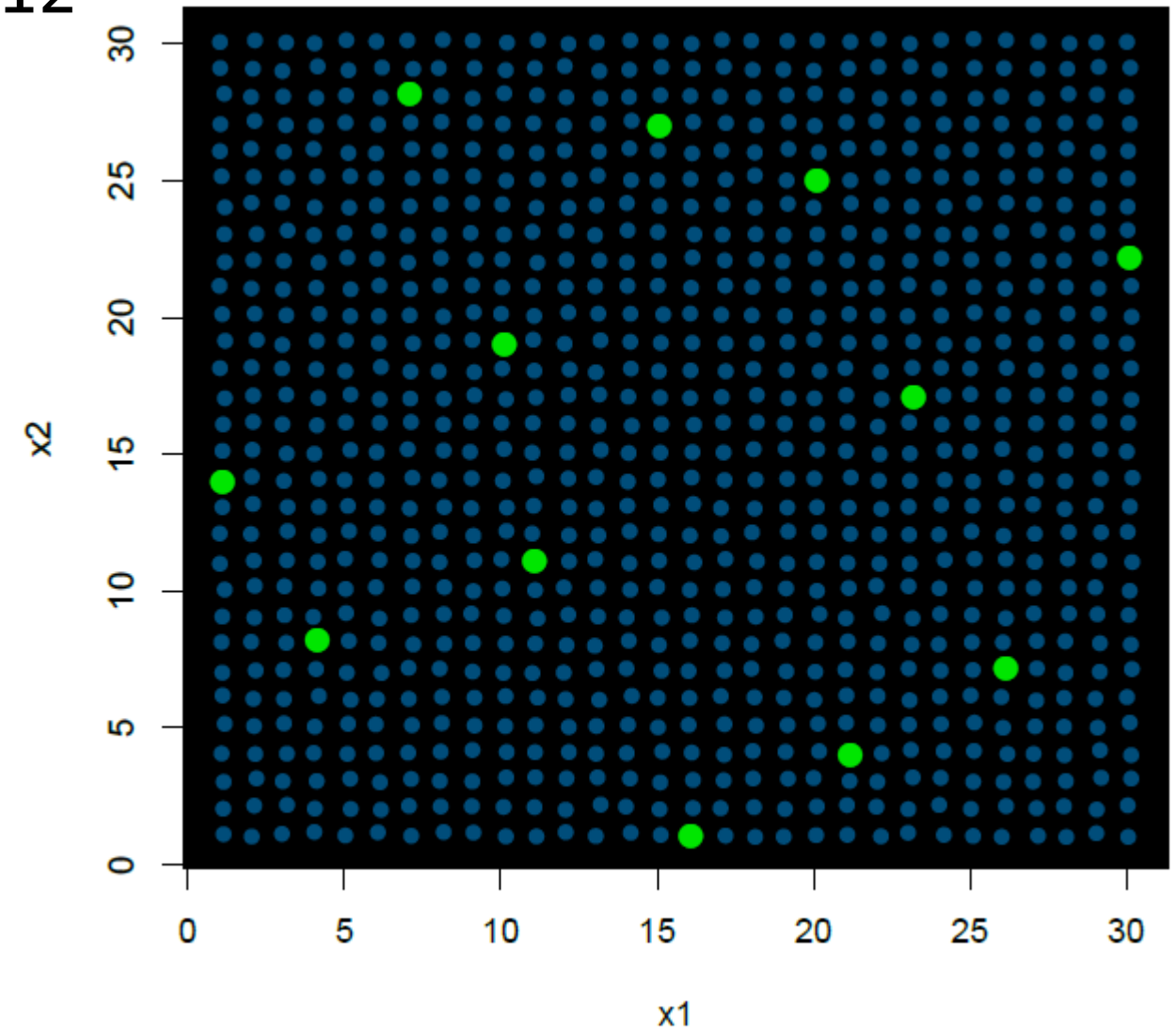
n = 11

contioned Latin Hypercube sampling (cLHS)



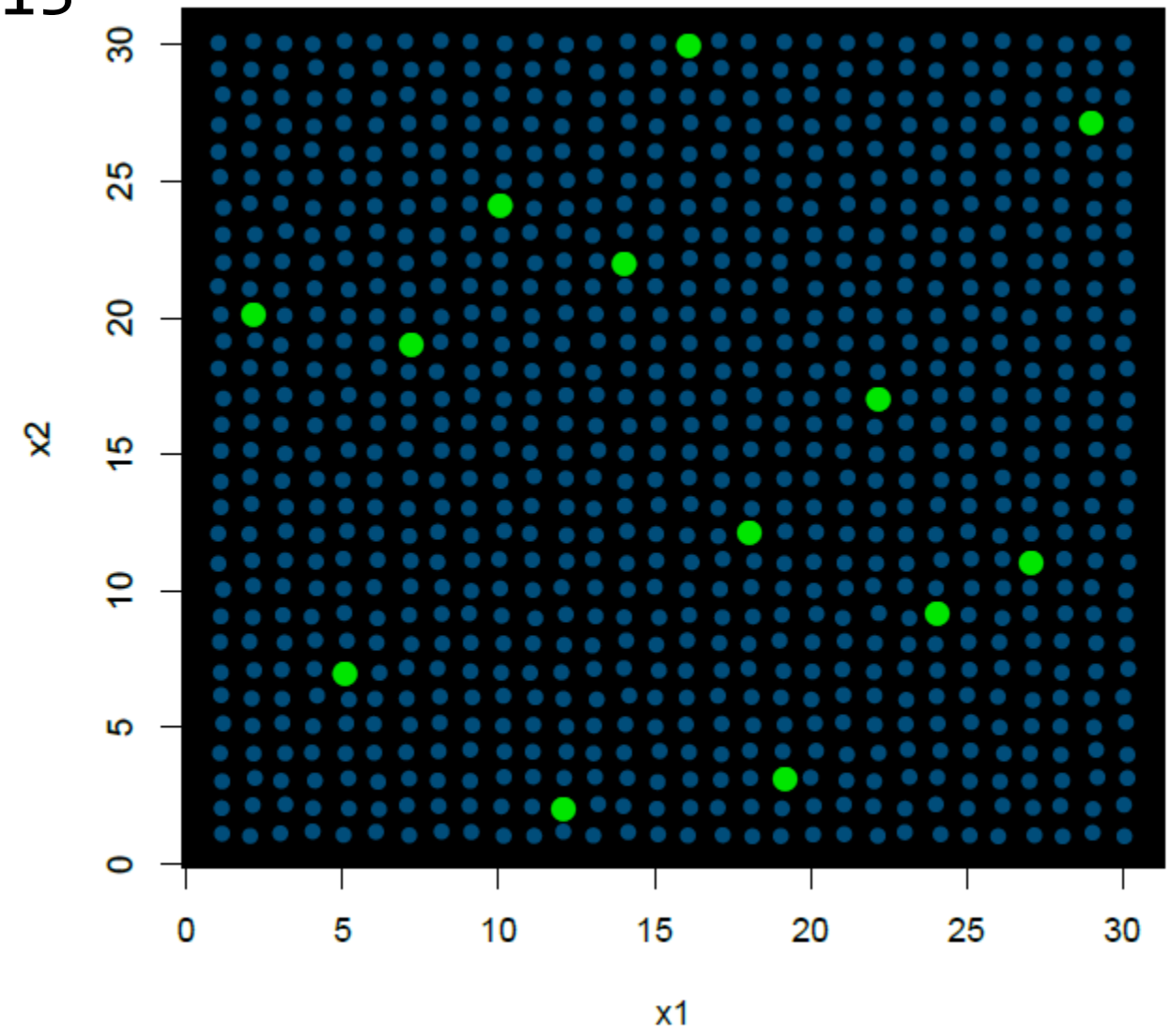
n = 12

contioned Latin Hypercube sampling (cLHS)



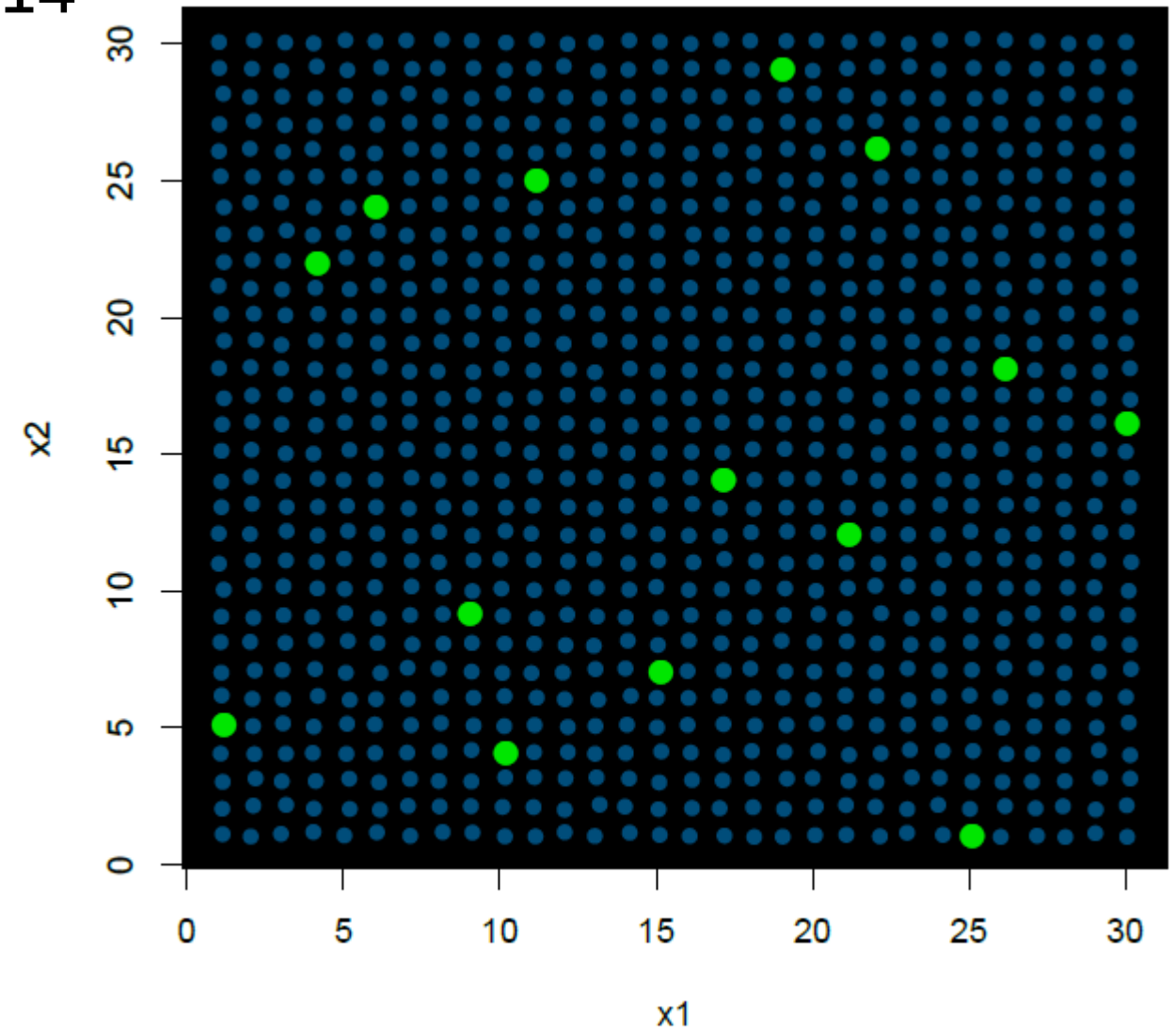
n = 13

contioned Latin Hypercube sampling (cLHS)



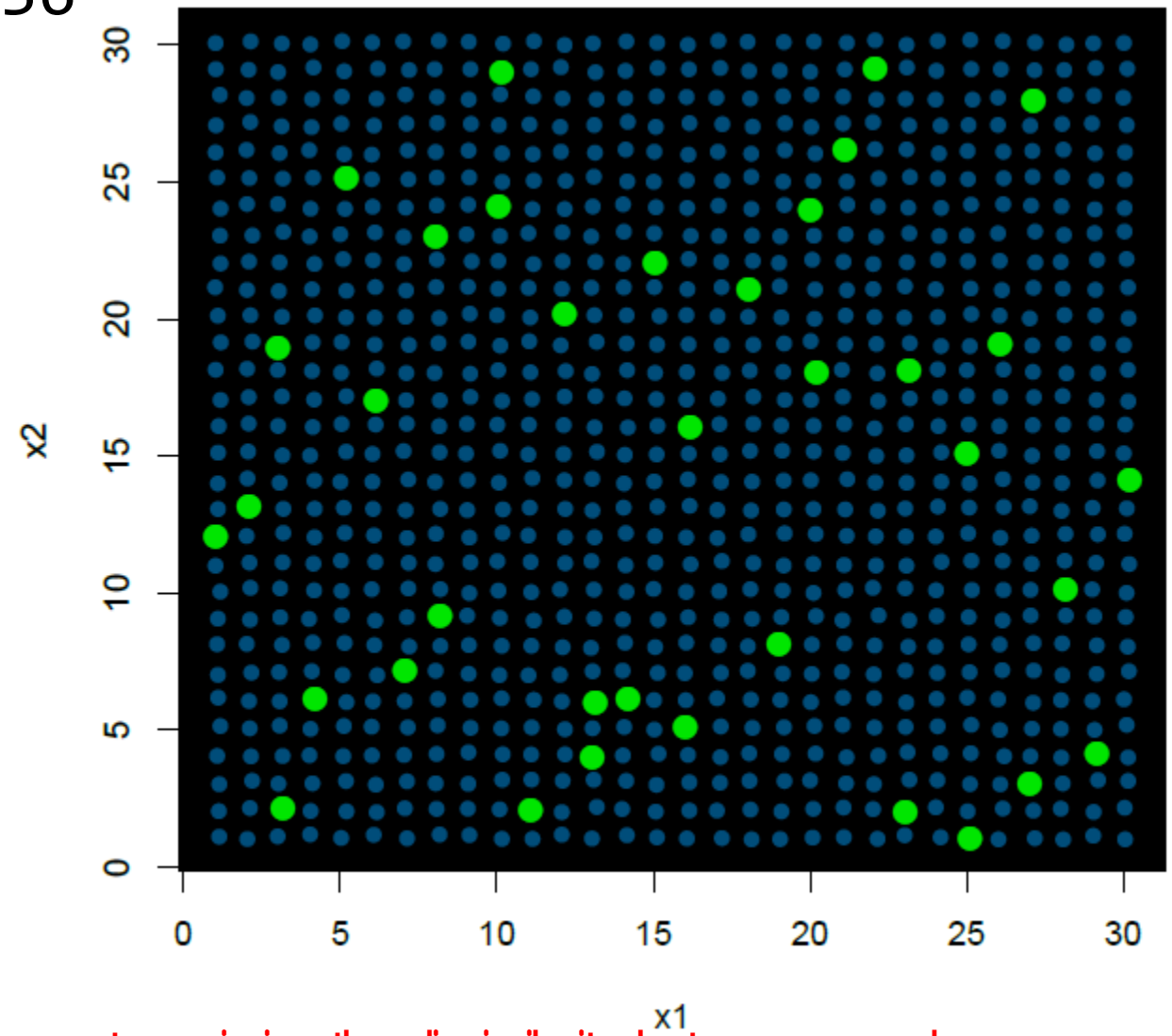
n = 14

contioned Latin Hypercube sampling (cLHS)



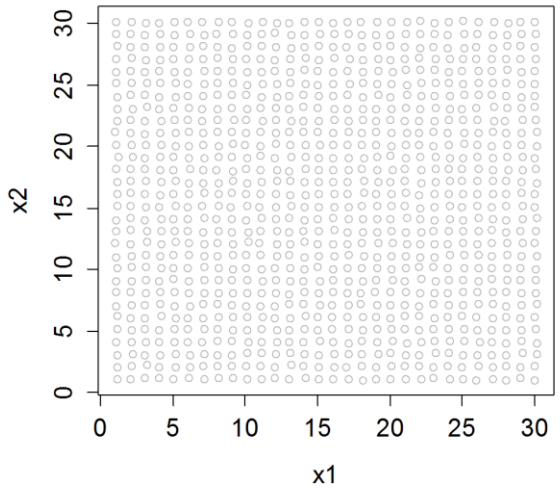
n = 36

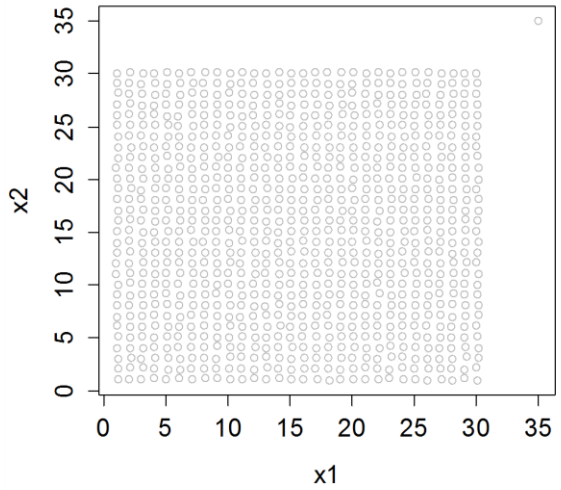
contioned Latin Hypercube sampling (cLHS)



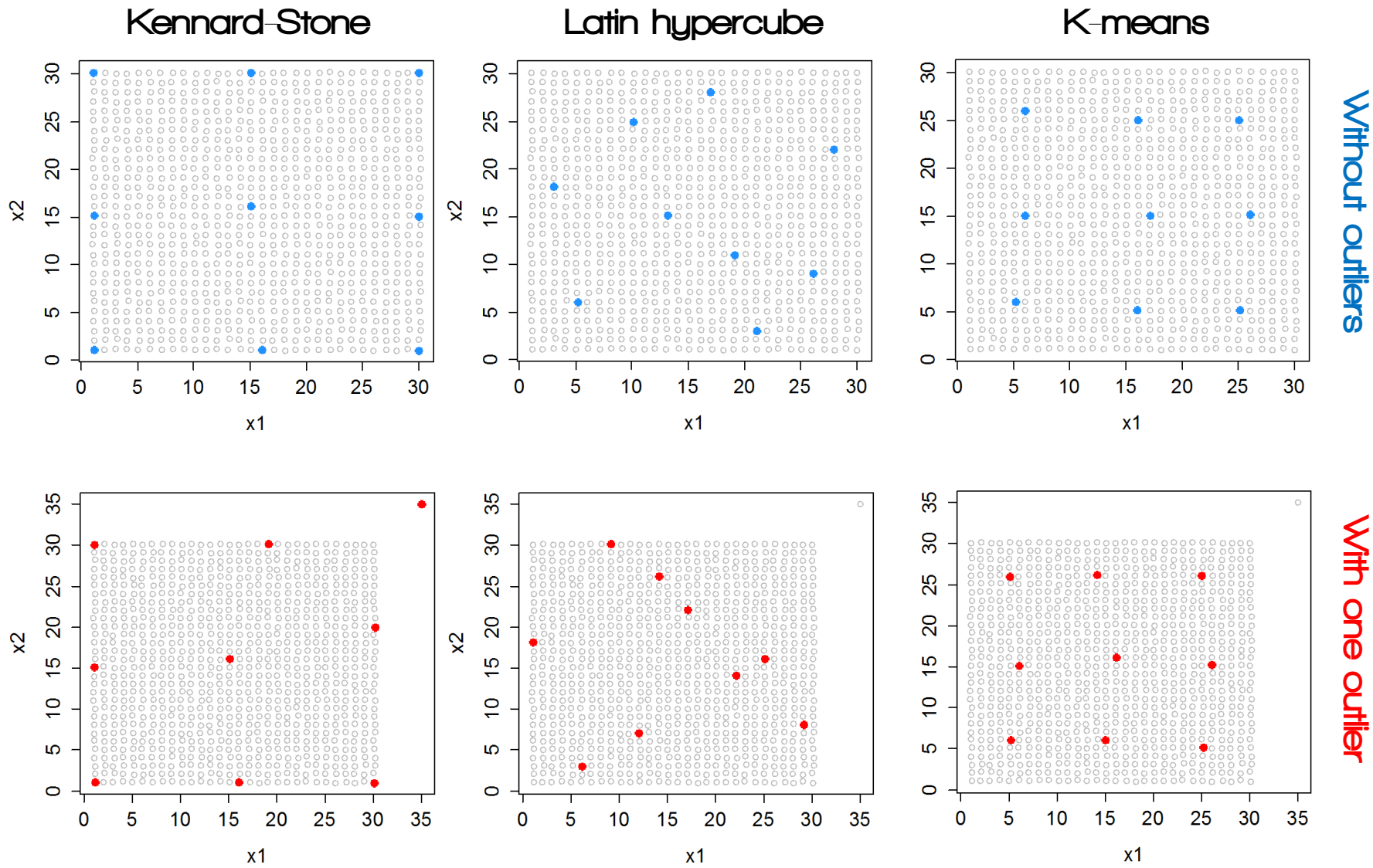
cLHS does not maximize the dissimilarity between samples

What if we have outliers in our data?



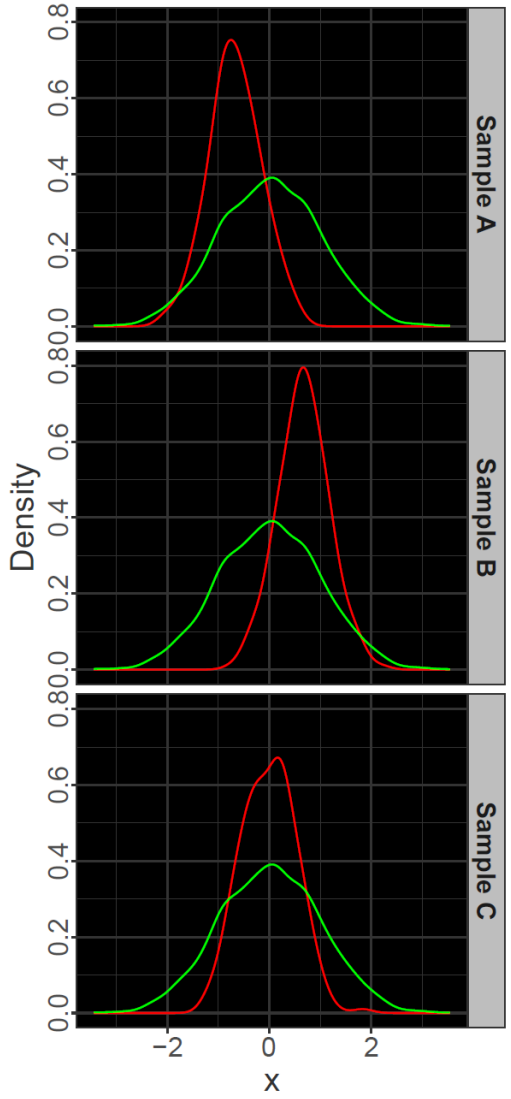


What if we have outliers in our data?



Size of the calibration set

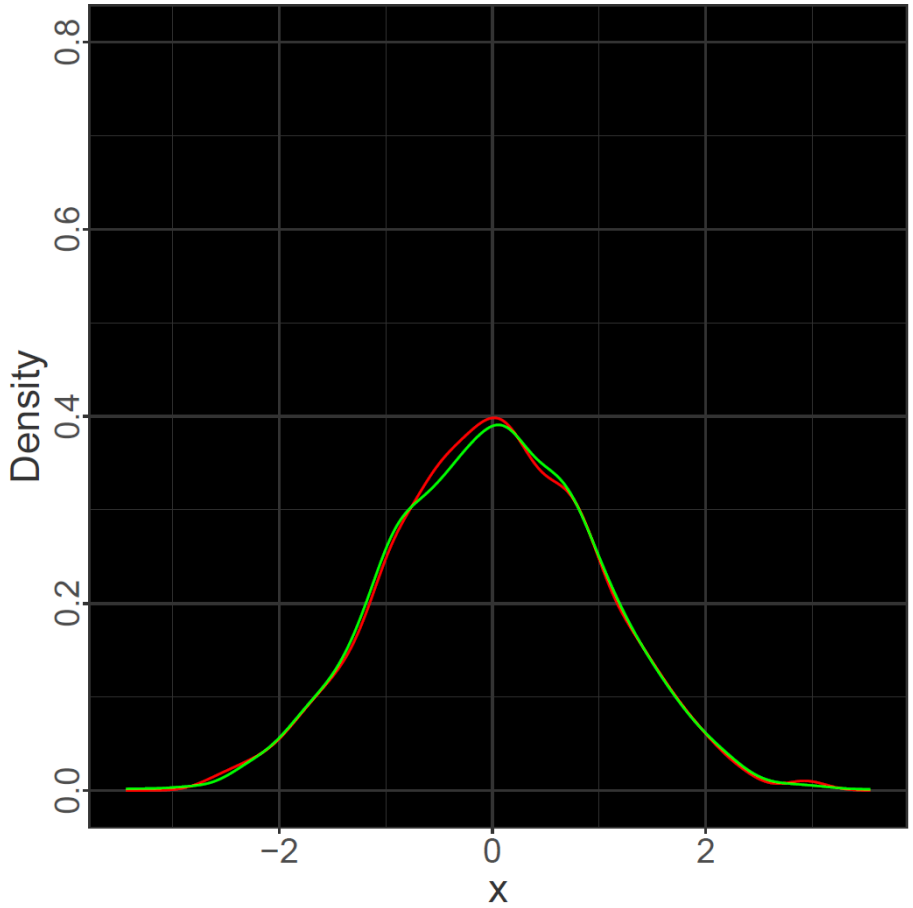
Low sampling efficiency



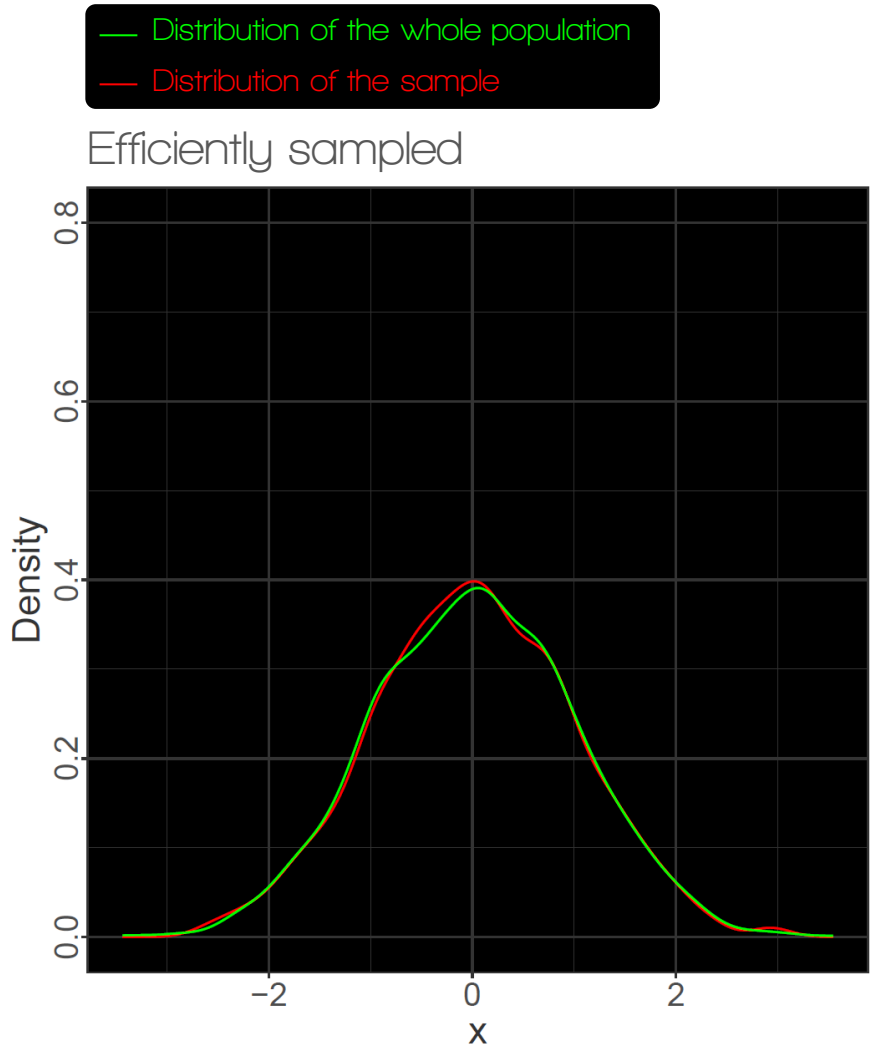
Distribution of the whole population

Distribution of the sample

Efficiently sampled



Size of the calibration set



Comparing the probability density function (pdf) of the population and the pdf of the sample set can be useful for assesing the representativeness of the sample set.

These pdfs can be computed for the PCs of the vis–NIR data.

The mean squared Euclidean distance (msd) can be used as a dissimilarity measure between the pdfs. It can be computed as follows:

$$msd = \frac{1}{k}\sum_{j=1}^{k} d^2\left[P_s(x_j \in cs), P_p(x_j)\right]$$

where:

$$d^2\left[P_s(x_j \in cs), P_p(x_j)\right] = \int_a^b \left[P_p(x_j) - P_s(x_j \in cs)\right]^2 dx_j$$

and where:

cs: a given subset of samples

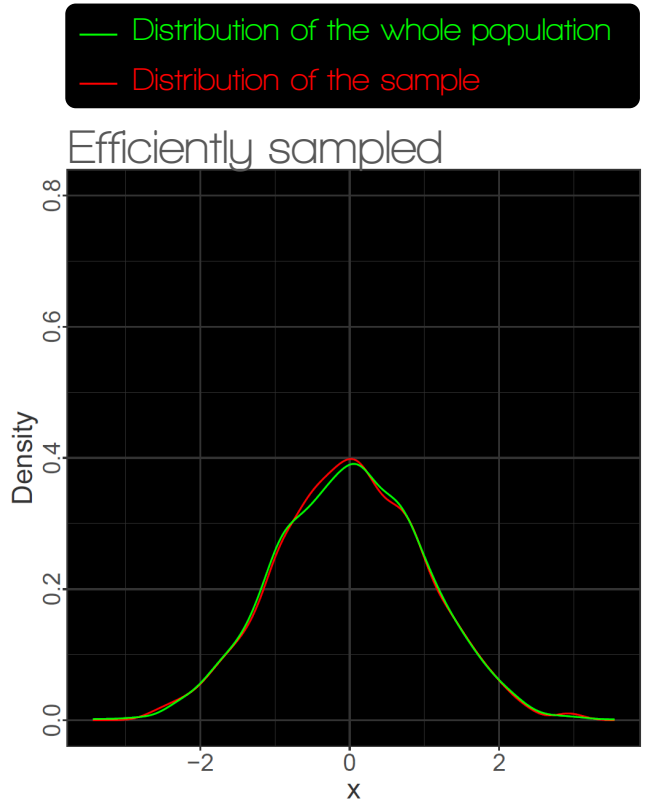
$P_s(x_j \in cs)$: the estimated pdf of the jth PC of cs,

$P_p(x_j)$: is the pdf of the jth PC for the whole population,

a and b: The range of the jth PC,

d^2 : represents the squared Euclidean distance between the two single distributions being compared,

k: the total number of PCs retained,



Example...

Soil vis-NIR library of Europe (n ~ 19.000)

Calibration sampling algorithms

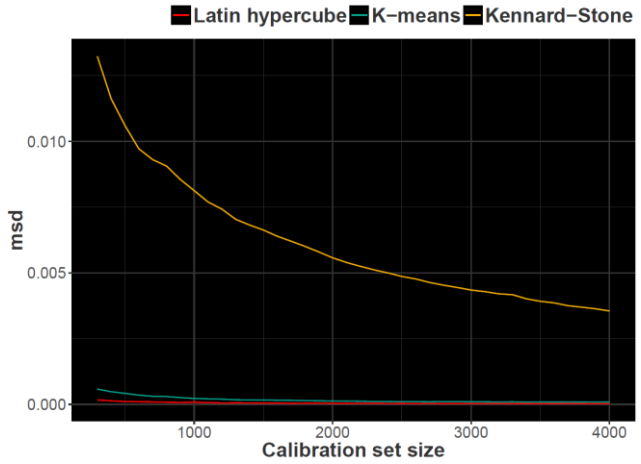
- Kennard-Stone sampling (KSS)
- Conditioned Latin hypercube sampling (cLHS)
- K-means sampling (KMS)

Calibration set size (CSS)

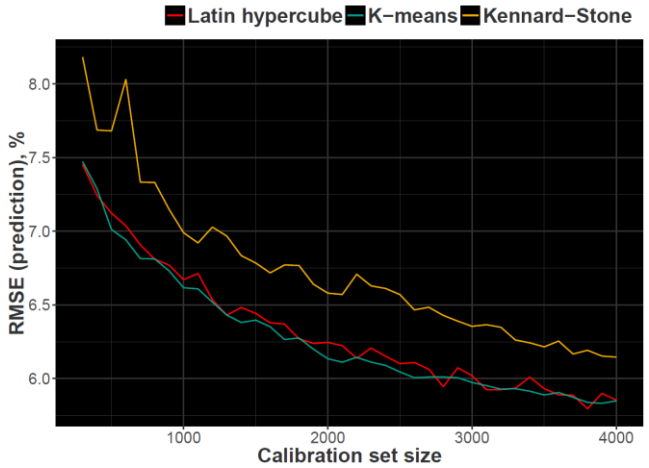
CSS = (300, 200, 400, ..., 4000)

The analyisis is based on the comparison between the probability density function (pdf) of the population and the pdf of the sample set

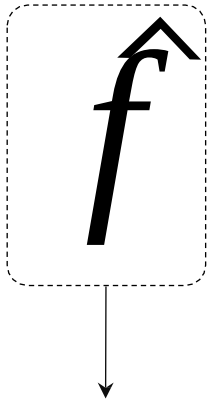
Comparisons of the pdfs



Prediction error



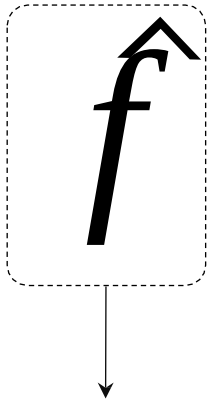
NIR Applications



Function that can be approximated by using chemometrics/machine learning

These functions are device-, product- and property-specific.

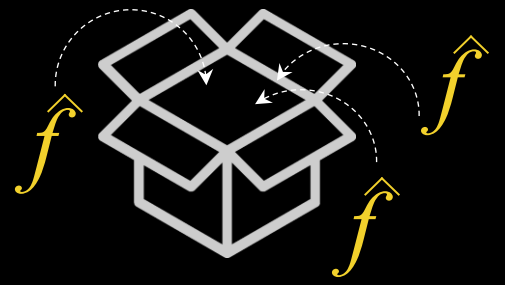
NIR Applications



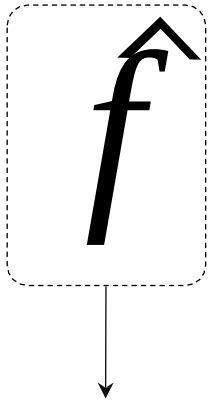
Function that can be approximated by using chemometrics/machine learning

These **functions are device-, product- and property-specific.**

For each product we build and group a set of functions that quantify relevant product properties (e.g. moisture, protein, fat). **This group of functions is called application.**



NIR Applications

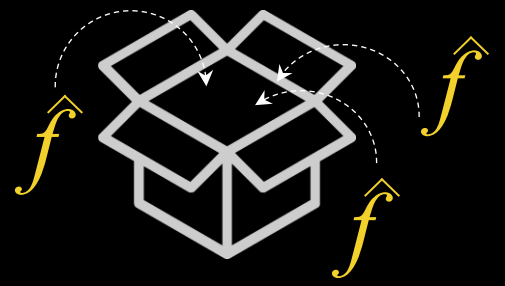


Function that can be approximated by using chemometrics/machine learning

We currently use our own BUCHI software for that!

These **functions are device-, product- and property-specific.**

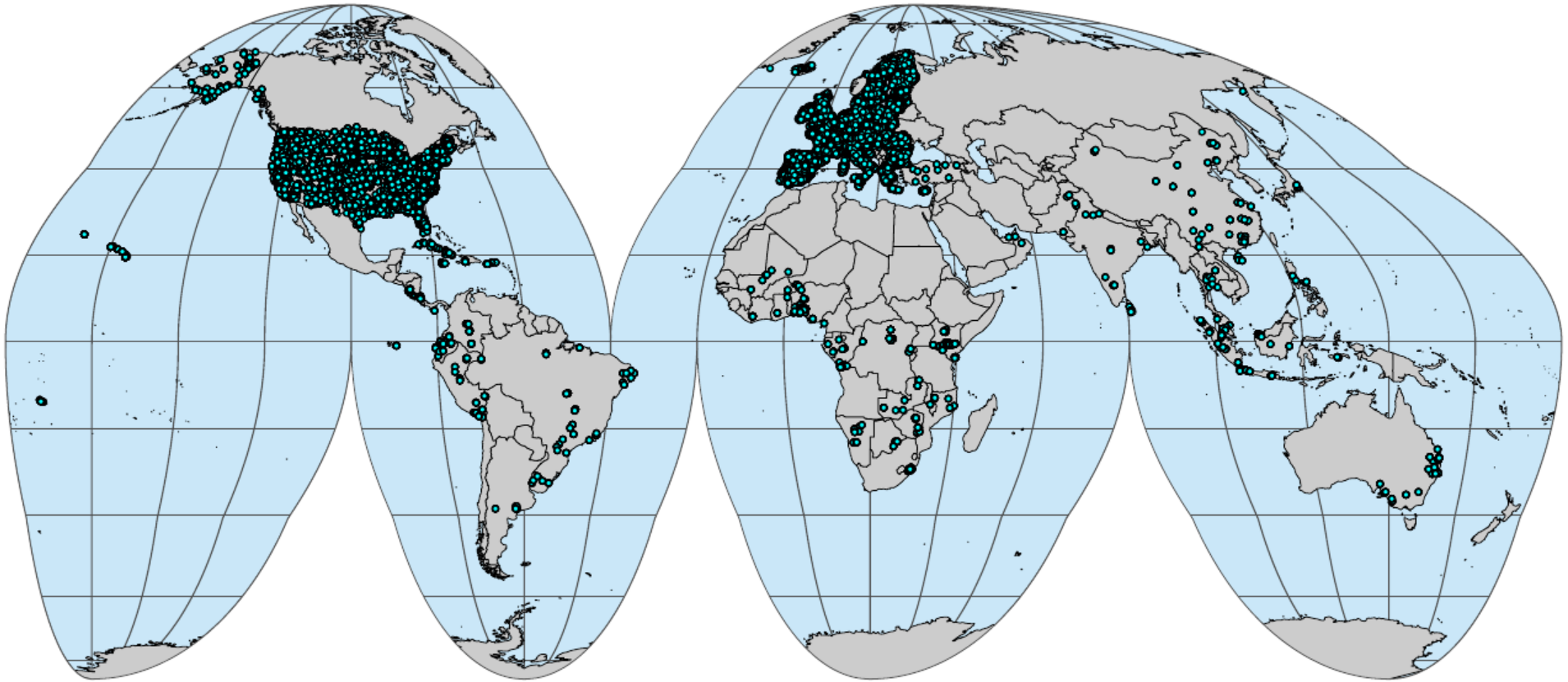
For each product we build and group a set of functions that quantify relevant product properties (e.g. moisture, protein, fat). **This group of functions is called application.**



From a user's perspective, **a NIR device without an application can be perceived as useless.**

How to benefit from spectral libraries to
make soil spectroscopy operational?

Dataset	Coverage	VNIR sample size
ICRAF_ISRIC	World	4438
KSSL	USA	19807
LUCAS	Europe	40818
Total		**65063**



Source: https://soilspectroscopy.org/ossl-updates/

We should aim at reducing entry barrier/implementation costs!!!

Implementation

It can be expensive
- Hardware and software costs
- Cost of reference analyses
- Elevated time investments
- People

Routine analysis

It can make boost the efficiency of your lab:
- Reduces need of conventional methods
- A sample is measured in matter of seconds
- Requires periodic validation and model/method updates

We should aim at reducing entry barrier/implementation costs!!!

Time to routine!

Implementation

It can be expensive
- Hardware and software costs
- Cost of reference analyses
- Elevated time investments
- People

Routine analysis

It can make boost the efficiency of your lab:
- Reduces need of conventional methods
- A sample is measured in matter of seconds
- Requires periodic validation and model/method updates

NIR/IR Spectroscopy – An equation for success

$$\text{Success} = f(\text{hardware, software, data, models, people, automation, ...})$$

Contributions

- Focus on usability of spectral libraries rather than modeling
- Scale up soil spectroscopy
- Make spectroscopy simple and operational for "simple" domains (i.e. fast to implement and service).
- Think globally fit locally*

*Saul, L. K., & Roweis, S. T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of machine learning research*, *4*(Jun), 119-155.

Success = f (hardware, software, data, chemometrics, automation, people, ...)

Success = f (hardware, software, data, chemometrics, automation, people, ...)

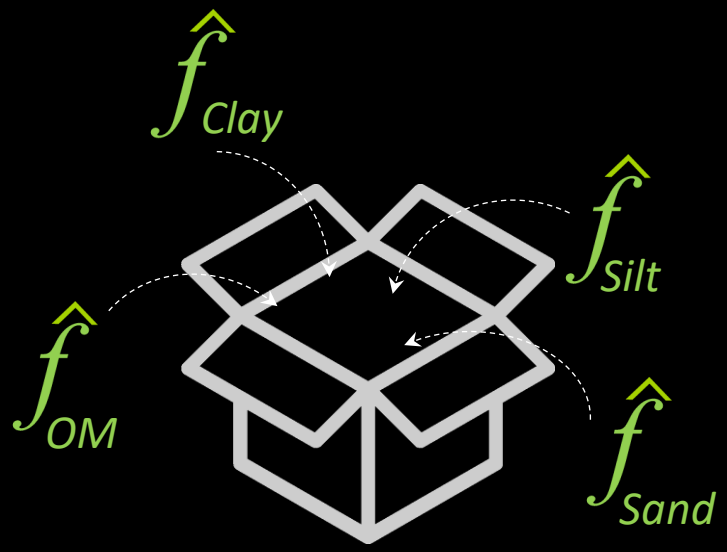
Library search methods

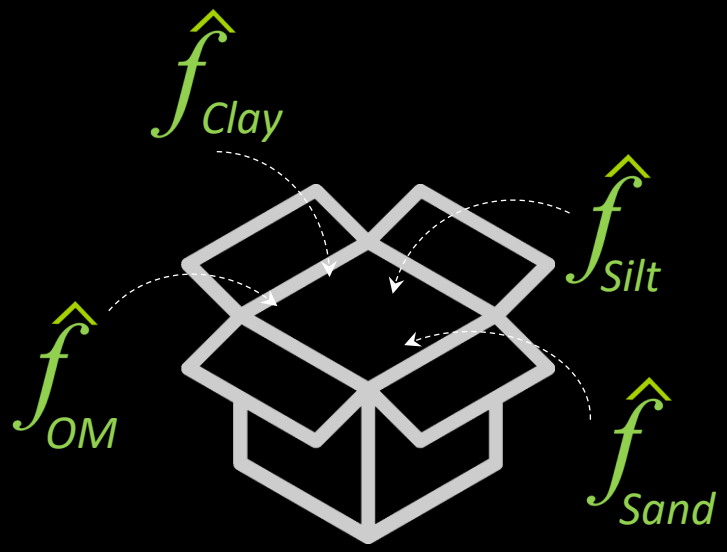
Main users

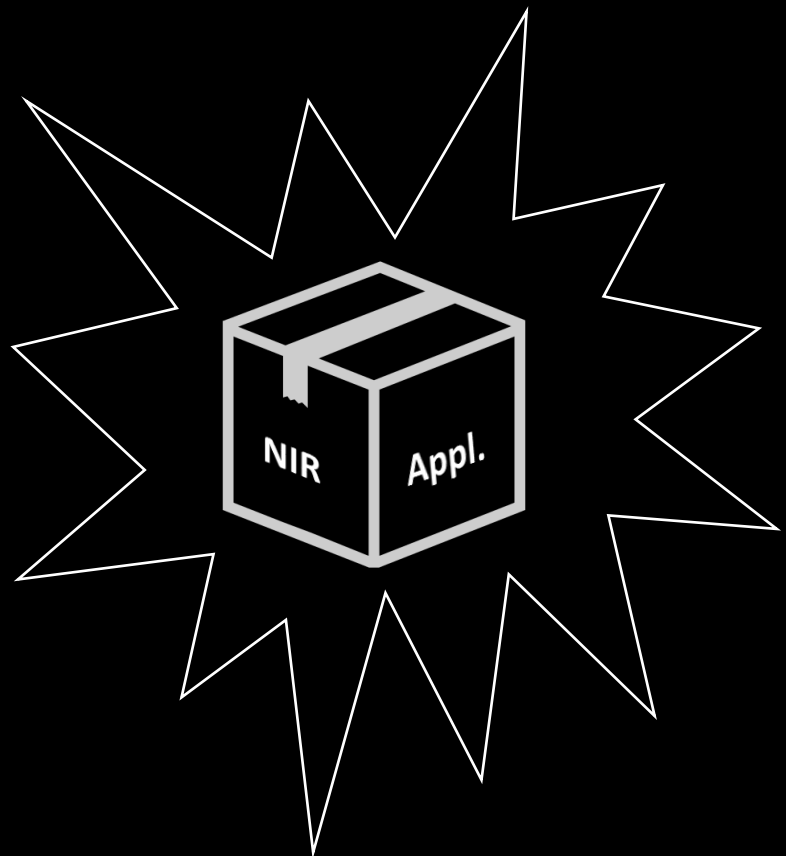
Big farms (soil, plant tissues, fertilizers, final product, etc)
Service labs
Research institutes/Universities

NIR/IR Spectroscopy – An equation for success

$$\text{Success} = f(\text{hardware, software, data, models, people, automation, ...})$$

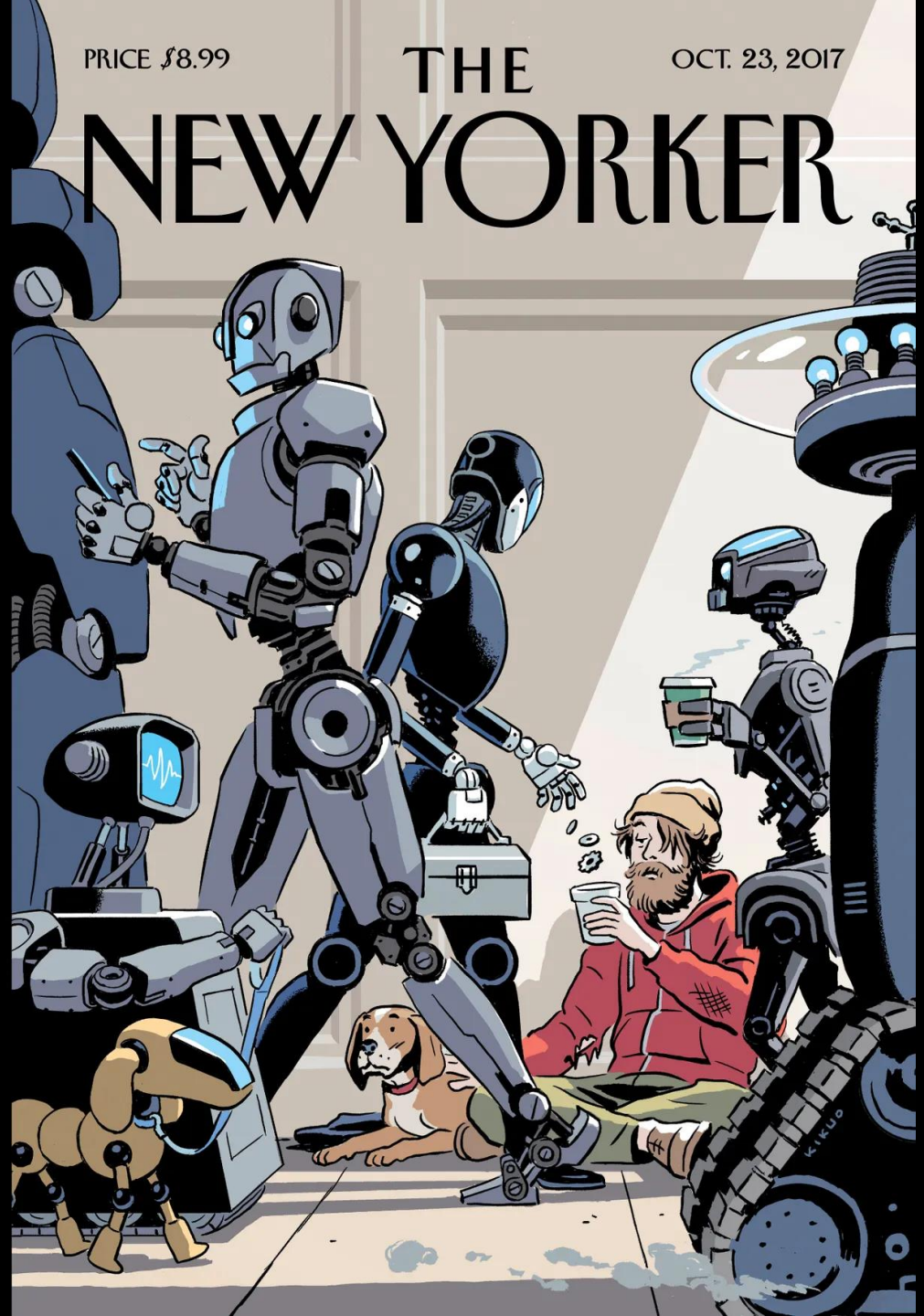

$\hat{f}_{Clay}$
$\hat{f}_{Silt}$
$\hat{f}_{OM}$
$\hat{f}_{Sand}$


$\hat{f}_{Clay}$
$\hat{f}_{Silt}$
$\hat{f}_{OM}$
$\hat{f}_{Sand}$


NIR
Appl.

debunking myths

1. Could soil spectroscopy eventually replace conventional methods?


PRICE $8.99
THE
NEW YORKER
OCT. 23, 2017
K. KUO

Brought to you by: Imperial College London Library

ScienceDirect

Journals & Books

Personalize

Sign out

Purchase PDF

Find at Imperial

Access through another institution

Imperial College London does not subscribe to this content on ScienceDirect.

Article preview

Abstract

Introduction

Section snippets

References (77)

Cited by (322)

ELSEVIER

Advances in Agronomy

Volume 132, 2015, Pages 139-159

Chapter Four - Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring

Marco Nocita [*] [§], Antoine Stevens [§], Bas van Wesemael [§], Matt Aitkenhead [¶], Martin Bachmann [‖], Bernard Barthès [#], Eyal Ben Dor [**], David J. Brown [§§], Michael Clairotte [#], Adam Csorba [¶¶], Pierre Dardenne [‖‖], Jose A.M. Demattê [##], Valerie Genot [†], Cesar Guerrero [***], Maria Knadel [§§§], Luca Montanarella [*], Carole Noon [§], Leonardo Ramirez-Lopez [¶¶¶], Jean Robertson [¶], Hiro Sakai [‖‖‖]... Johanna Wetterlind [§§§§]

Show more

Recommended articles

Occurrence, Detection, and Molecular and Metabolic Characterization of...

Advances in Agronomy, Volume 132, 2015, pp. 1...

Magdalena Frąc, ..., Takashi Yaguchi

Hydrological Aspects of Arsenic Contamination of Groundwater in...

Advances in Agronomy, Volume 132, 2015, pp. 7...

Saugata Datta

Do we really need large spectral libraries for local scale SOC...

Soil and Tillage Research, Volume 155, 2016, p...

César Guerrero, ..., Raphael A. Viscarra Rossel

Show 3 more articles

Brought to you by: Imperial College London Library

ScienceDirect

Journals & Books

Personalize

Sign out

Purchase PDF

Find at Imperial

Access through another institution

Imperial College London does not subscribe to this content on ScienceDirect.

Article preview

Abstract

Introduction

Section snippets

References (77)

Cited by (322)

ELSEVIER

...ces in Ag...

201...

Chapter Four ...troscopy: An Alternative ...ry for Soil Monitori...

Marco Nocita [*] [§], ...ntoi...e Stevens [§], Bas van Wesemael [§], Matt Aitkenhead [¶], Martin Bachmann [‖], Ber...ard Barthès [#], Eyal Ben Dor [**], David J. Brown [§§], Michael Clairotte [#], Adam Csorba [¶¶], Pierre Dardenne [‖‖], Jose A.M. Demattê [##], Valerie Genot [†], Cesar Guerrero [***], Maria Knadel [§§§], Luca Montanarella [*], Carole Noon [§], Leonardo Ramirez-Lopez [¶¶¶], Jean Robertson [¶], Hiro Sakai [‖‖‖]... Johanna Wetterlind [§§§§]

Show more

Recommended articles

Occurrence, Detection, and Molecular and Metabolic Characterization of...

Advances in Agronomy, Volume 132, 2015, pp. 1...

Magdalena Frąc, ..., Takashi Yaguchi

Hydrological Aspects of Arsenic Contamination of Groundwater in...

Advances in Agronomy, Volume 132, 2015, pp. 7...

Saugata Datta

Do we really need large spectral libraries for local scale SOC...

Soil and Tillage Research, Volume 155, 2016, p...

César Guerrero, ..., Raphael A. Viscarra Rossel

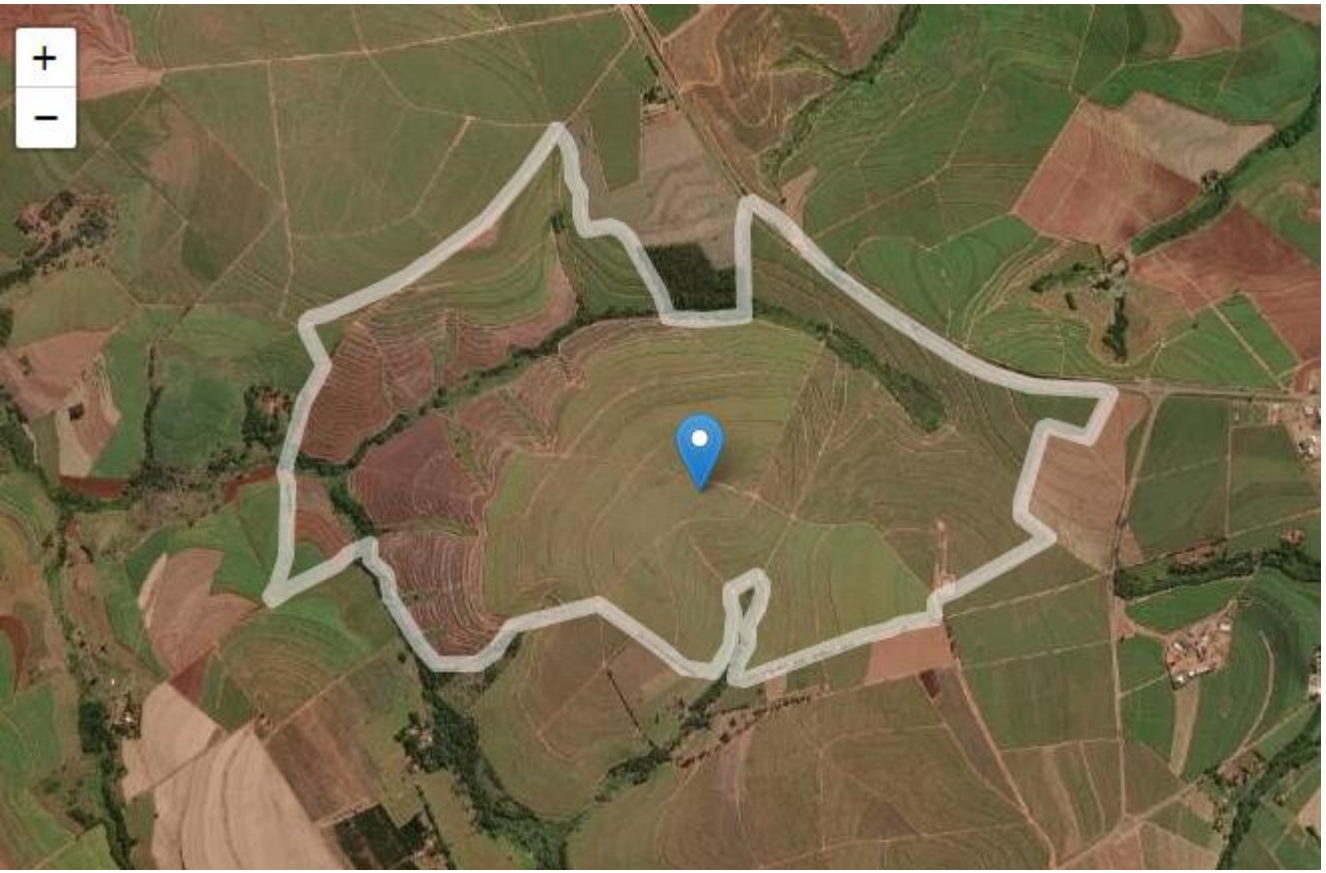
Show 3 more articles

Example: NIR spectroscopy for soil mapping and management

Ramirez-Lopez, Leonardo, et al. "Robust soil mapping at the farm scale with vis–NIR spectroscopy." *European Journal of Soil Science* 70.2 (2019): 378-393.

Example:
473 ha in the municipality of Barra Bonita (Brazil)

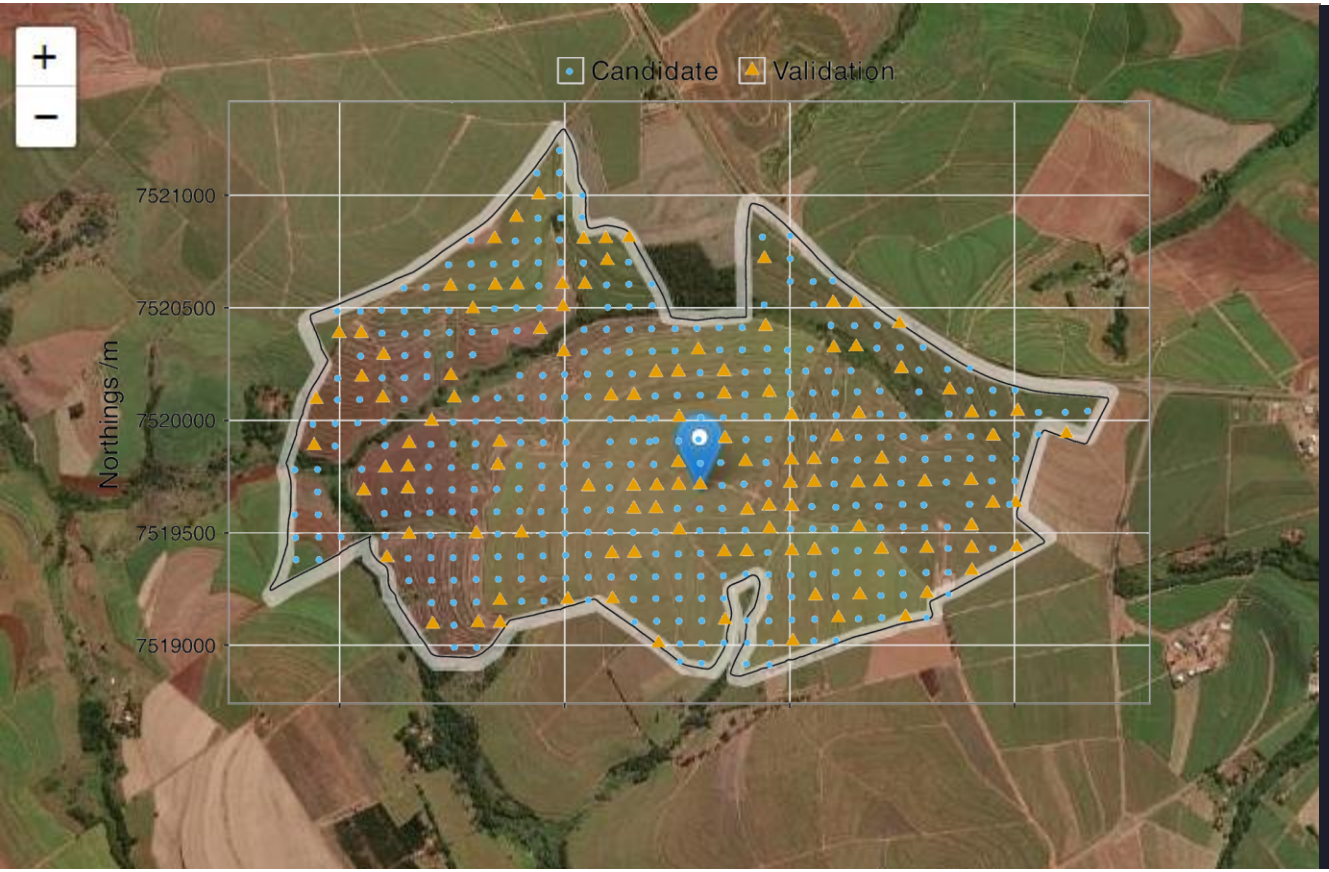
Sampling at two depths (A: 0 – 20 cm; B: 80 – 100 cm)
Total samples: 910

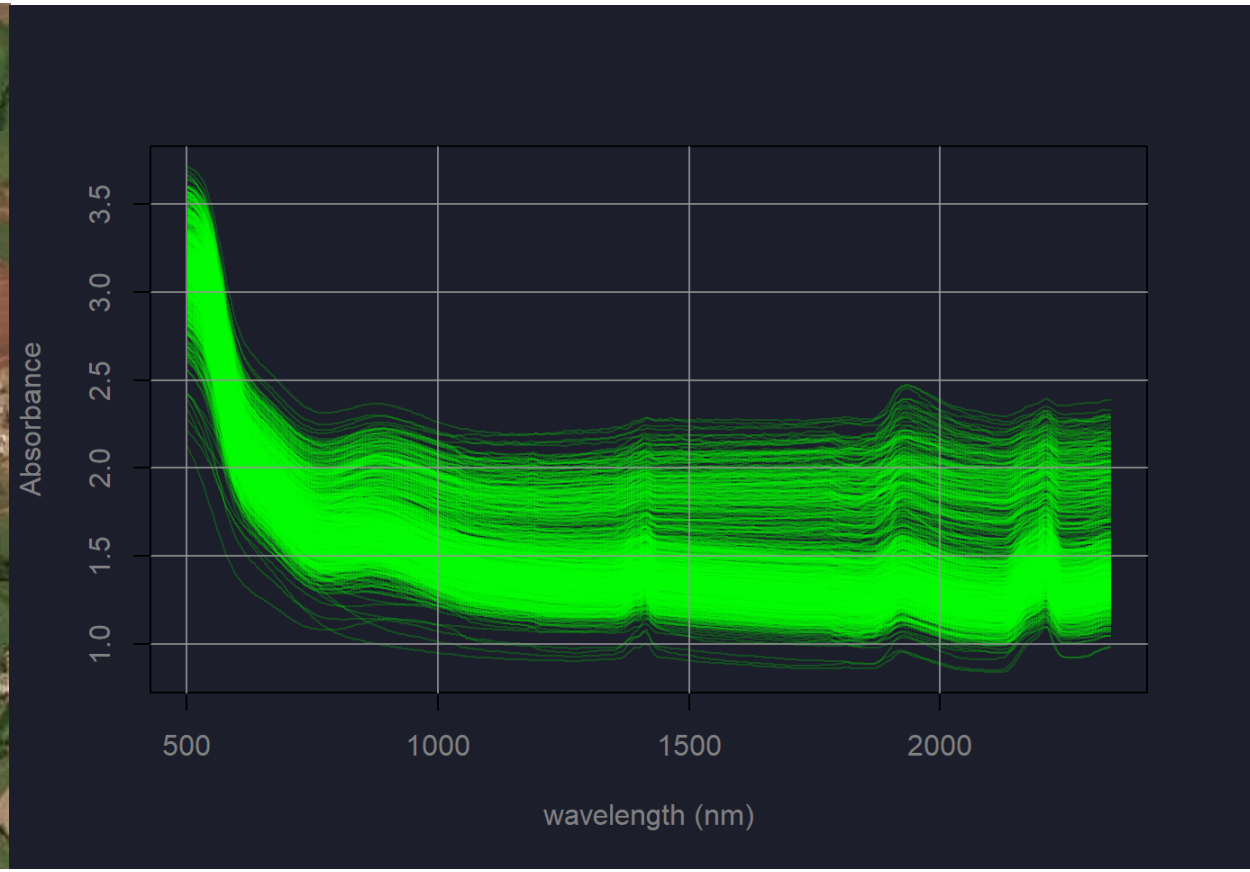


Example:
473 ha in the municipality of Barra Bonita (Brazil)

Sampling at two depths (A: 0 – 20 cm; B: 80 – 100 cm)
Total samples: 910

Sampling at two depths (A: 0 – 20 cm; B: 80 – 100 cm)
Total samples: 910

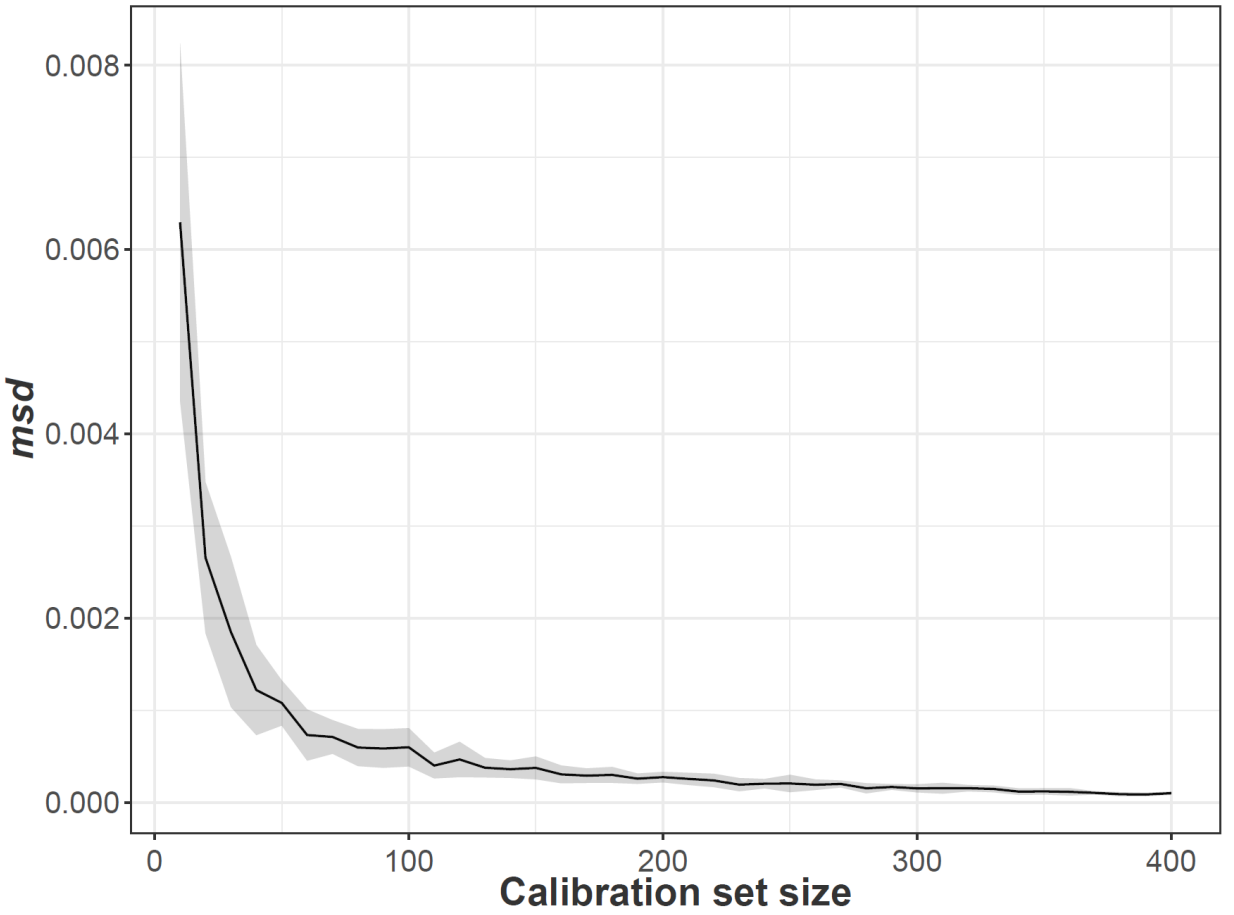


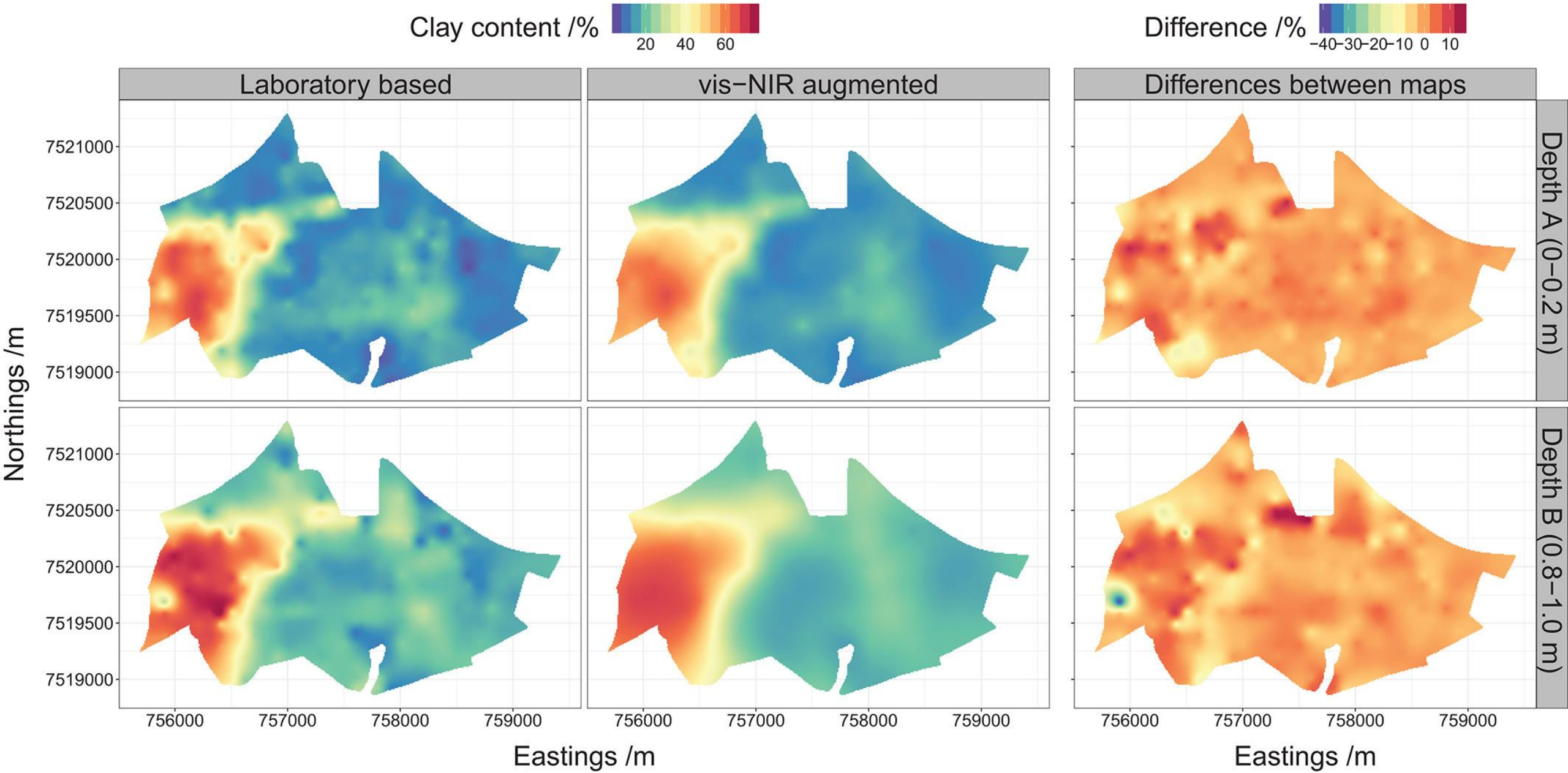


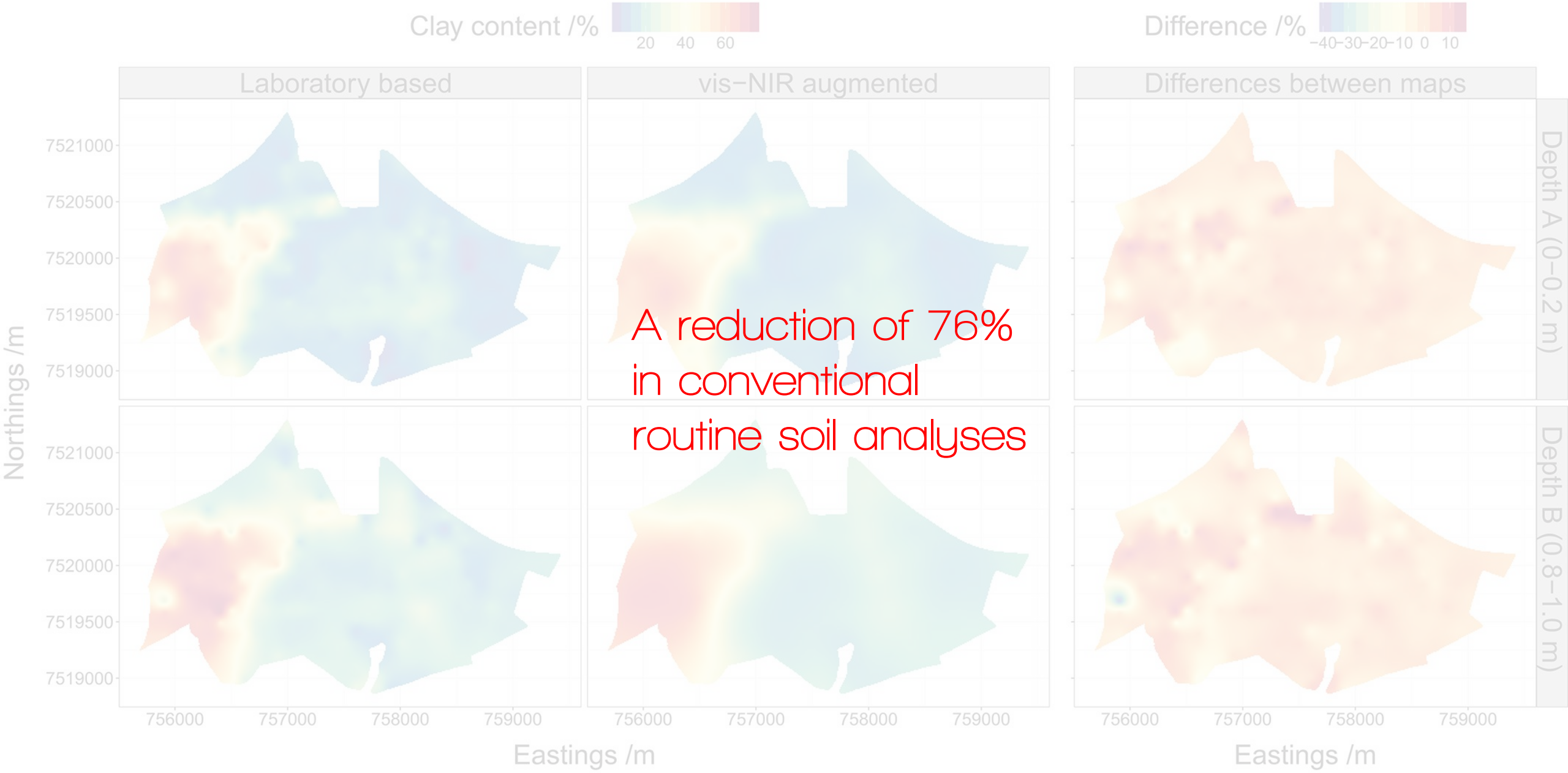
Example:
473 ha in the municipality of Barra Bonita (Brazil)

Sampling at two depths (A: 0 – 20 cm; B: 80 – 100 cm)
Total samples: 910

Optimal calibration size: 180 samples




Clay content /%
20
40
60
Difference /%
−40
−30
−20
−10
0
10
Laboratory based
vis−NIR augmented
Differences between maps
Depth A (0−0.2 m)
Depth B (0.8−1.0 m)
Northings /m
7521000
7520500
7520000
7519500
7519000
756000
757000
758000
759000
Eastings /m


Clay content /%
20 40 60
Difference /%
−40 −30 −20 −10 0 10
Laboratory based
vis−NIR augmented
Differences between maps
Depth A (0−0.2 m)
Depth B (0.8−1.0 m)
Northings /m
7521000
7520500
7520000
7519500
7519000
756000
757000
758000
759000
Eastings /m
A reduction of 76%
in conventional
routine soil analyses

Scaling soil spectroscopy

Short term

- Learn and adjust

Scaling soil spectroscopy

Short term

- Learn and adjust

Mid term

- Move to modest amount of samples processed with NIR/IR

Scaling soil spectroscopy

Short term

Learn and adjust

Mid term

Move to modest amount of samples processed with NIR/IR

Long term

Make it fully operational for your properties and move to a significant amount of samples processed with NIR/IR (get all your investments back)

Soil spectroscopy complements conventional methods (to different extents)

2. Is soil spectroscopy cheap and fast?

Mmm

Mmmmmmmmmmmmmm

Mmmmmmmmmmmmmm,
Yeah BUT

Mmmmmmmmmmmmmm,
Yeah BUT not really

Mmmmmmmmmmmmmm,
Yeah BUT not really

Implementation

Routine analysis

Mmmmmmmmmmmmmm,
Yeah BUT not really

Implementation

It can be expensive
- Hardware and software costs
- Cost of reference analyses
- Elevated time investments
- People

Routine analysis

Mmmmmmmmmmmmmm,
Yeah BUT not really

Implementation

It can be expensive
- Hardware and software costs
- Cost of reference analyses
- Elevated time investments
- People

Routine analysis

It can make boost the efficiency of your lab:
- Reduces need of conventional methods
- A sample is measured in matter of seconds
- Requires periodic validation and model/method updates

3. Can all important soil properties be quantified?

Let's ask a friend who has read TONS of scientific literature

Sound scientific evidence
[consistently showing good results]

Under research

Sound scientific evidence
[consistently showing good results]

Examples:
- Particle size
- Organic C
- N

Under research

Where do you want go?

Sound scientific evidence
[consistently showing good results]

Under research

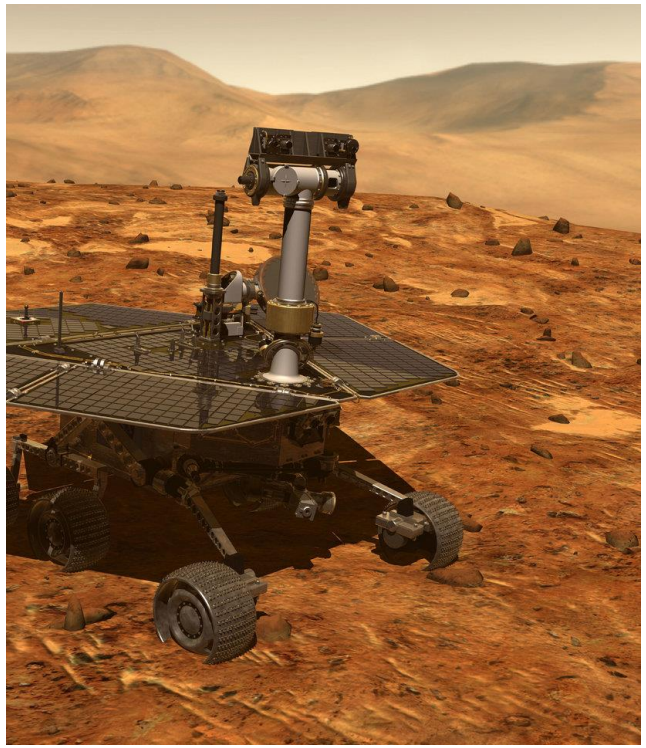
4. Are global/universal soil spectroscopy models feasible?

Svalbard, Norway

Svalbard, Norway

Uganda

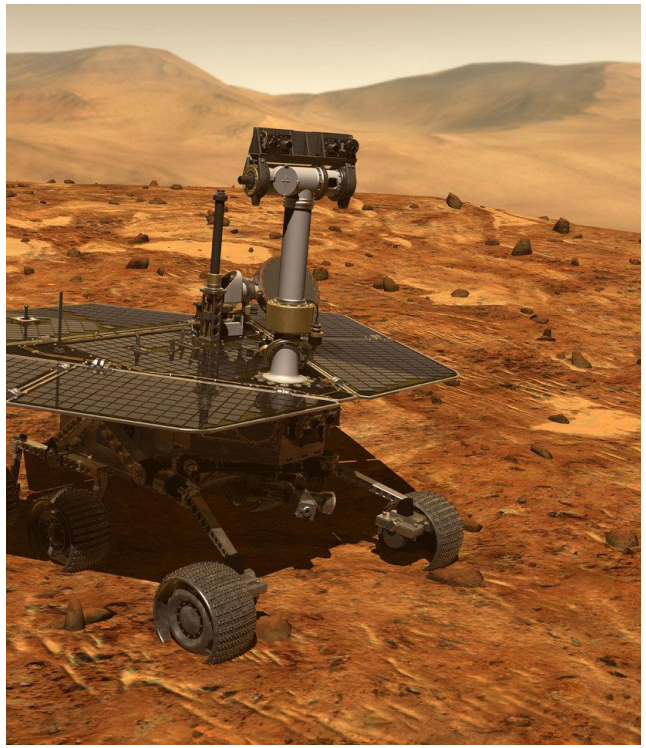
Mars



Svalbard, Norway

Uganda

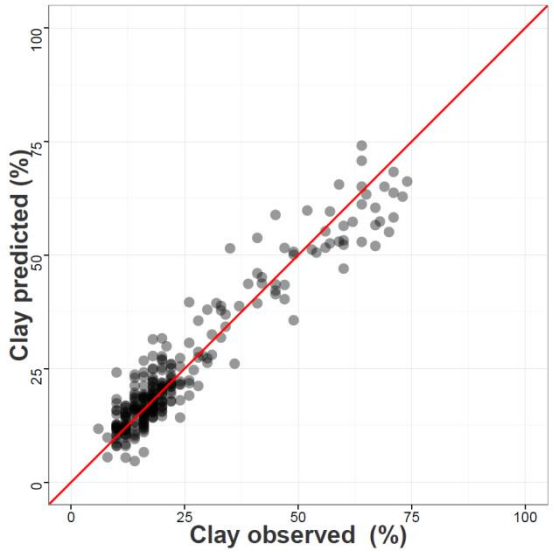
Mars



➔ Different soils might require different empirical models

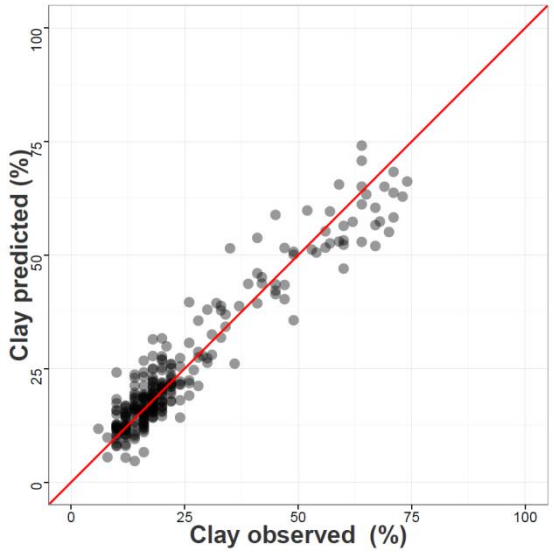
Field scale

R^2 = 0.91; RMSE=5.02%



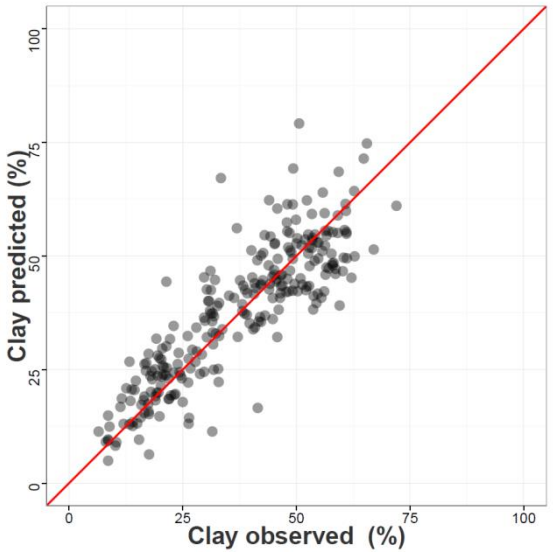
Field scale

R^2 = 0.91; RMSE=5.02%



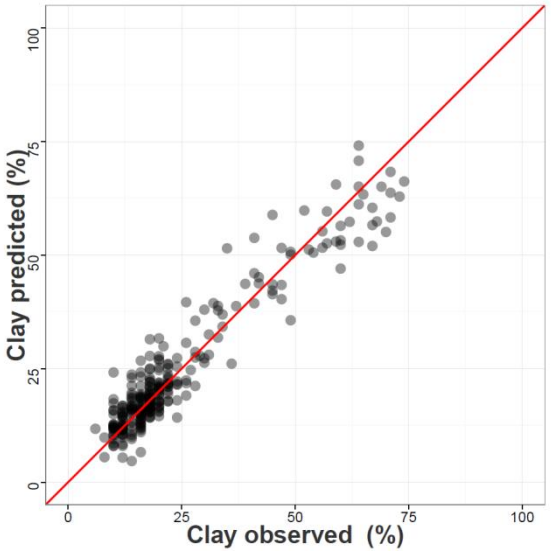
Regional scale

R^2 = 0.75; RMSE=8.03%



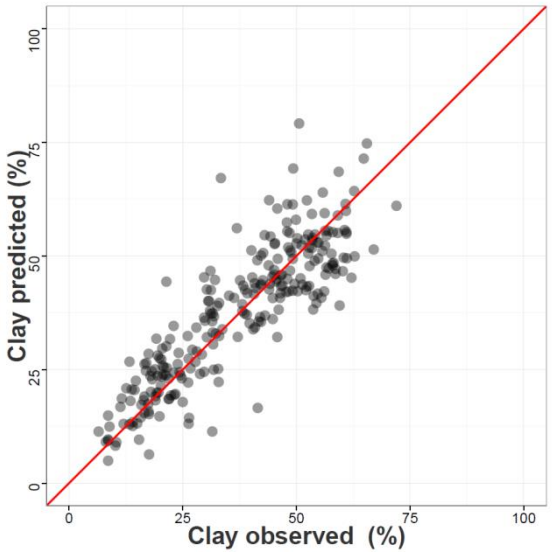
Field scale

R^2 = 0.91; RMSE=5.02%



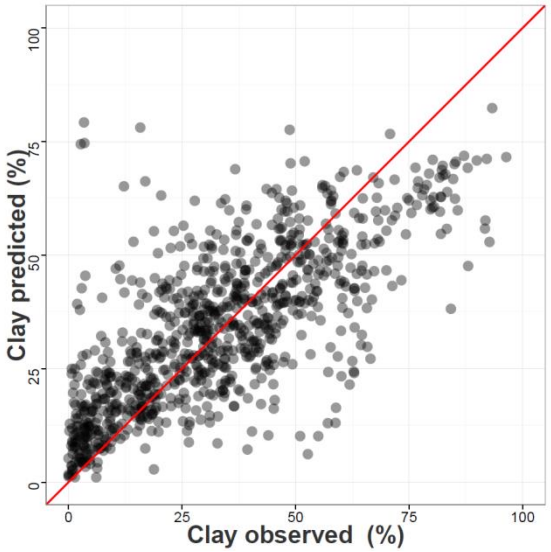
Regional scale

R^2 = 0.75; RMSE=8.03%



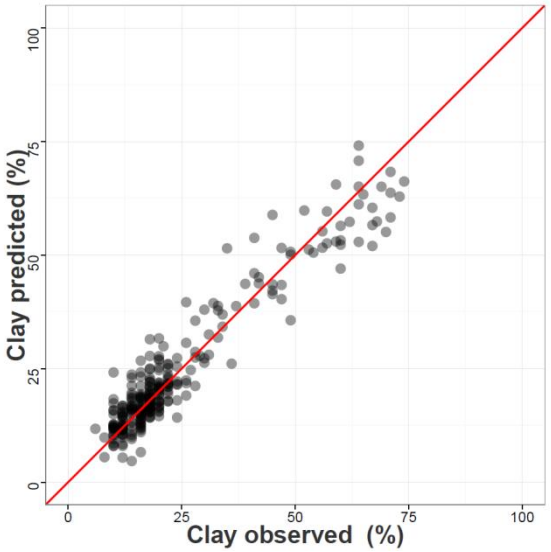
Global scale

R^2 = 0.50; RMSE=15.53%



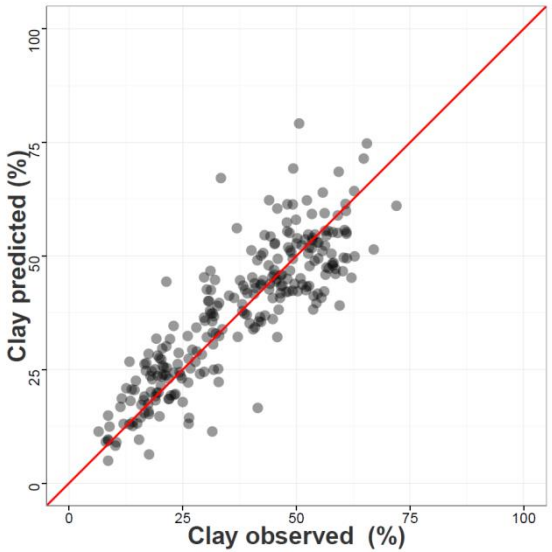
Field scale

R^2 = 0.91; RMSE=5.02%



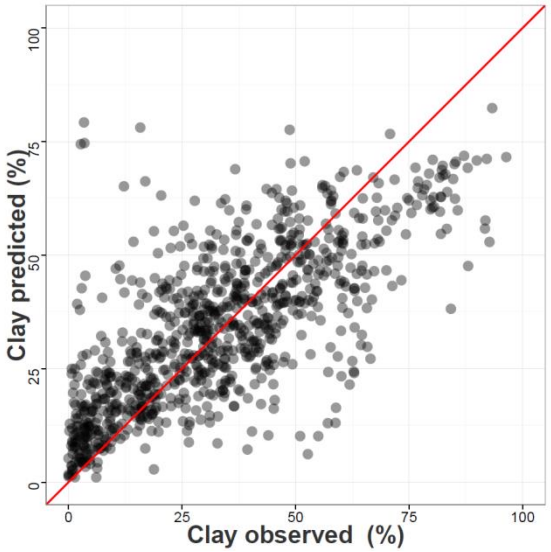
Regional scale

R^2 = 0.75; RMSE=8.03%



Global scale

R^2 = 0.50; RMSE=15.53%



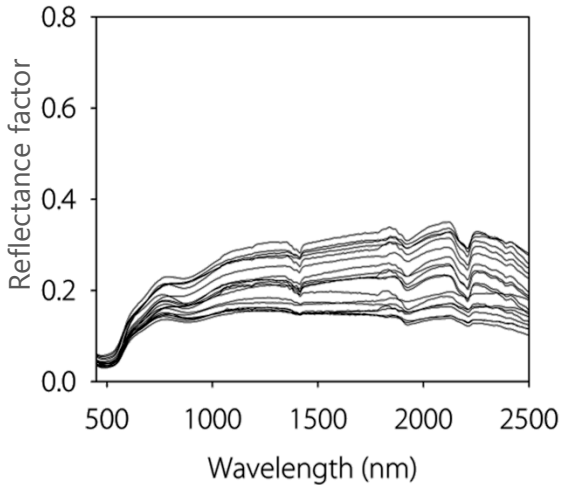
−

Data complexity

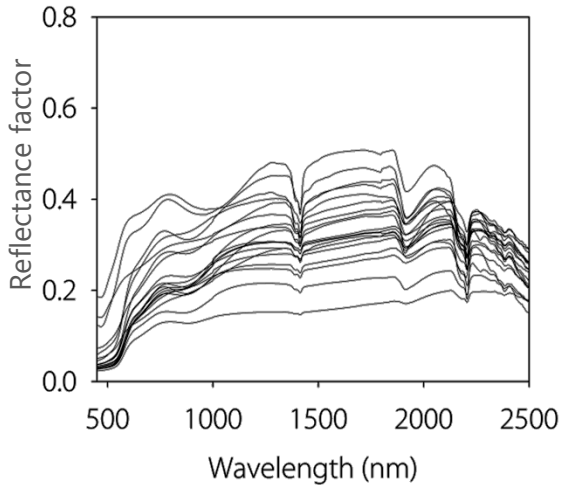
+

Degradation of the accuracy

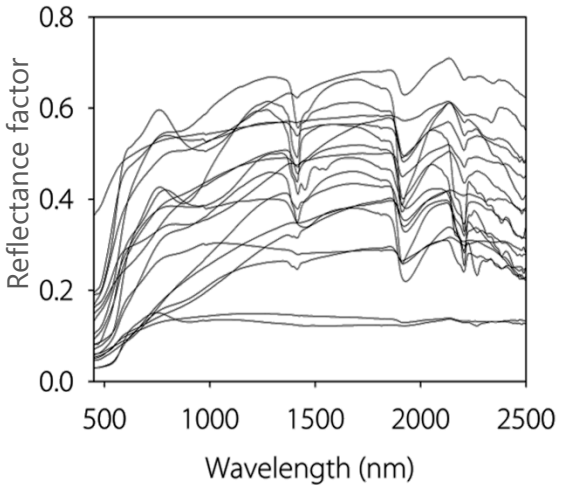
Field scale



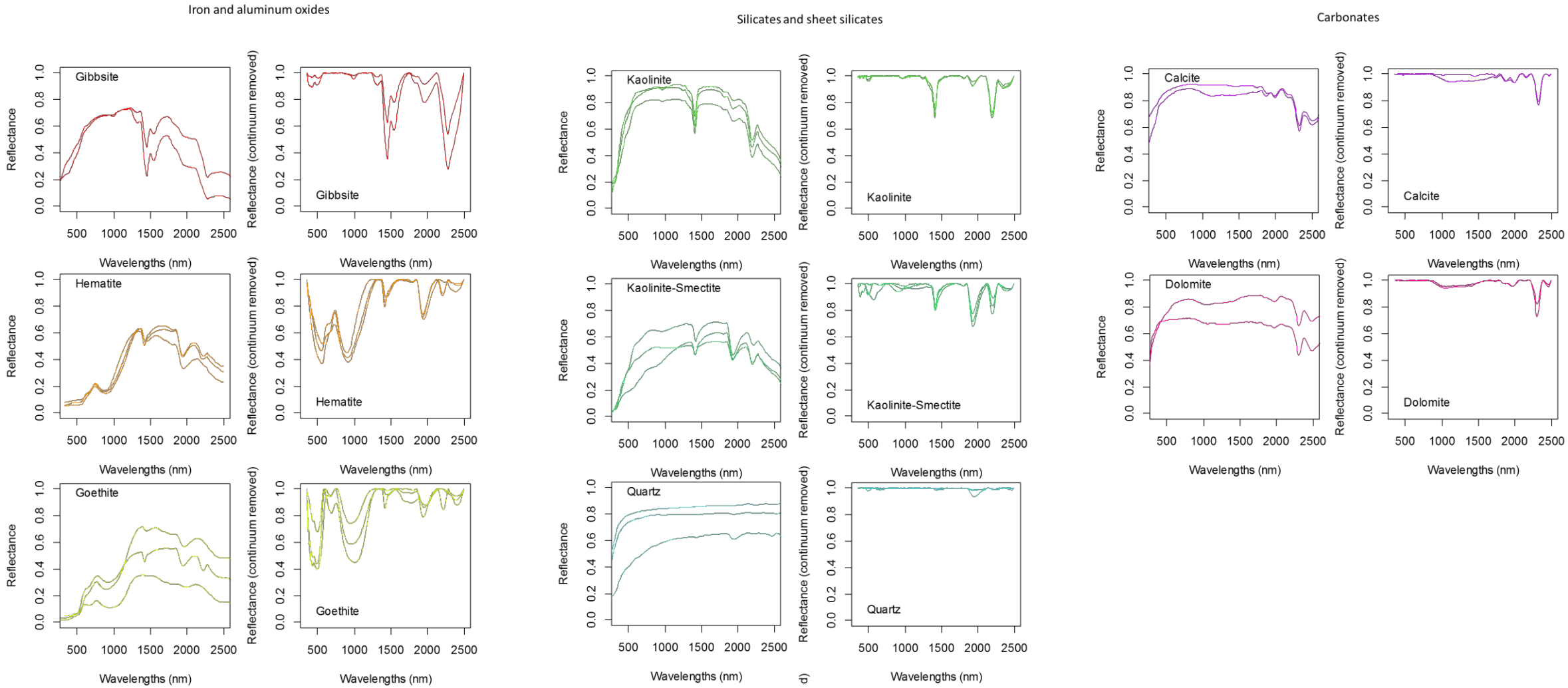
Regional scale

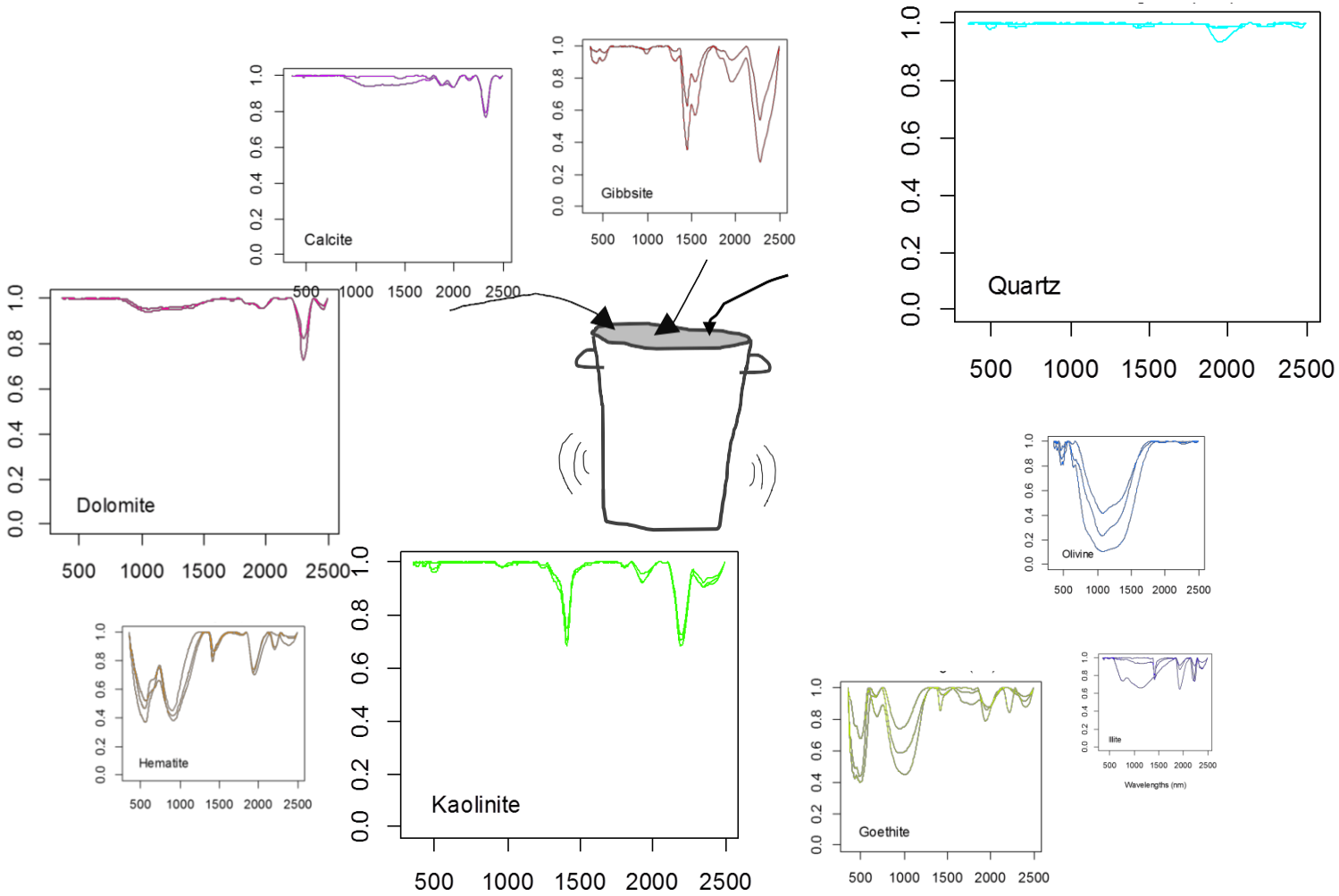


Global scale



20 spectra sampled at random


Iron and aluminum oxides
Silicates and sheet silicates
Carbonates
Gibbsite
Hematite
Goethite
Kaolinite
Kaolinite-Smectite
Quartz
Calcite
Dolomite
Reflectance
Reflectance (continuum removed)
Wavelengths (nm)
0.0
0.2
0.4
0.6
0.8
1.0
500
1000
1500
2000
2500

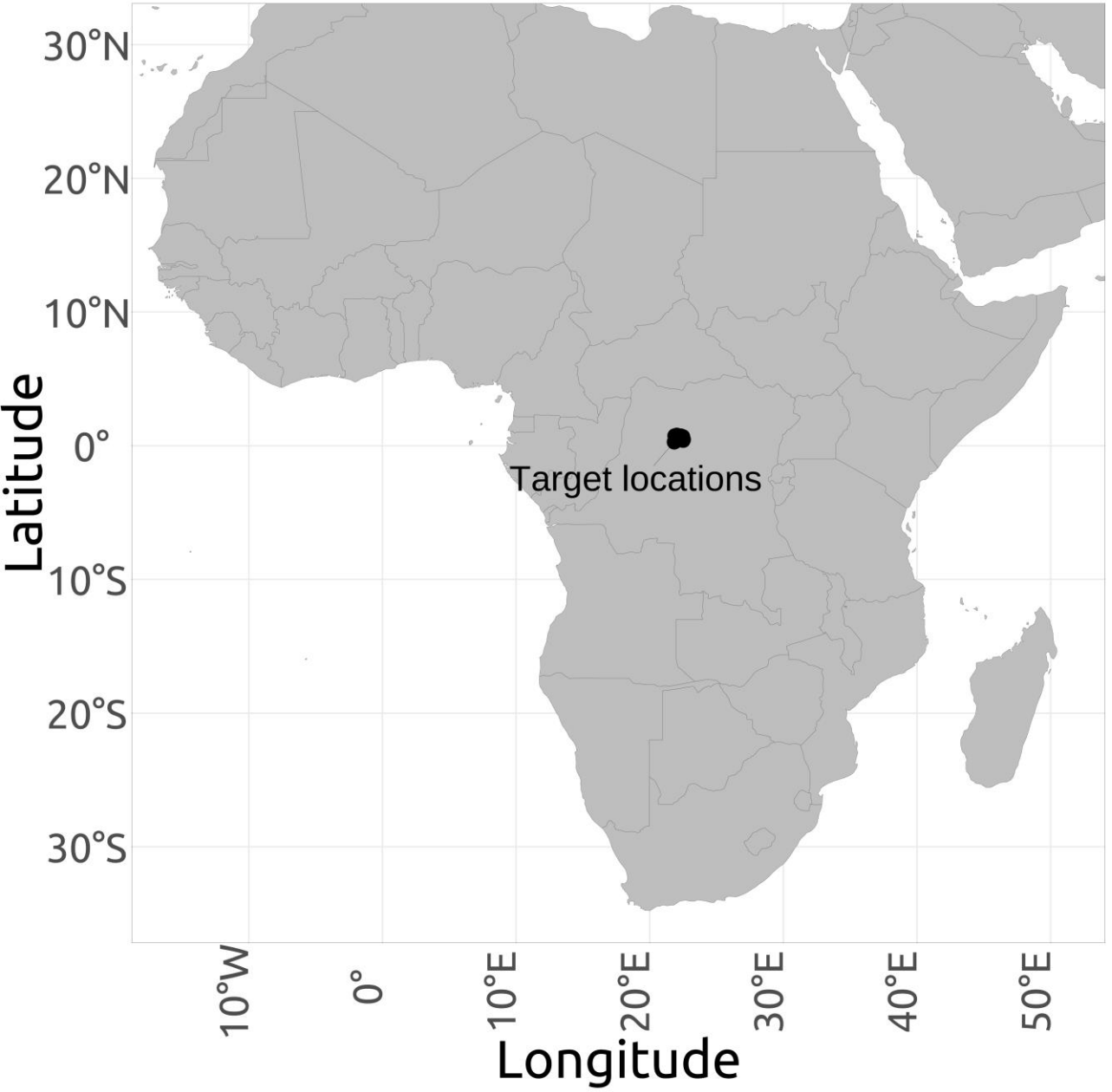

Calcite
Gibbsite
Quartz
Dolomite
Olivine
Hematite
Kaolinite
Goethite
Illite
Wavelengths (nm)

Think globally, fit locally*!

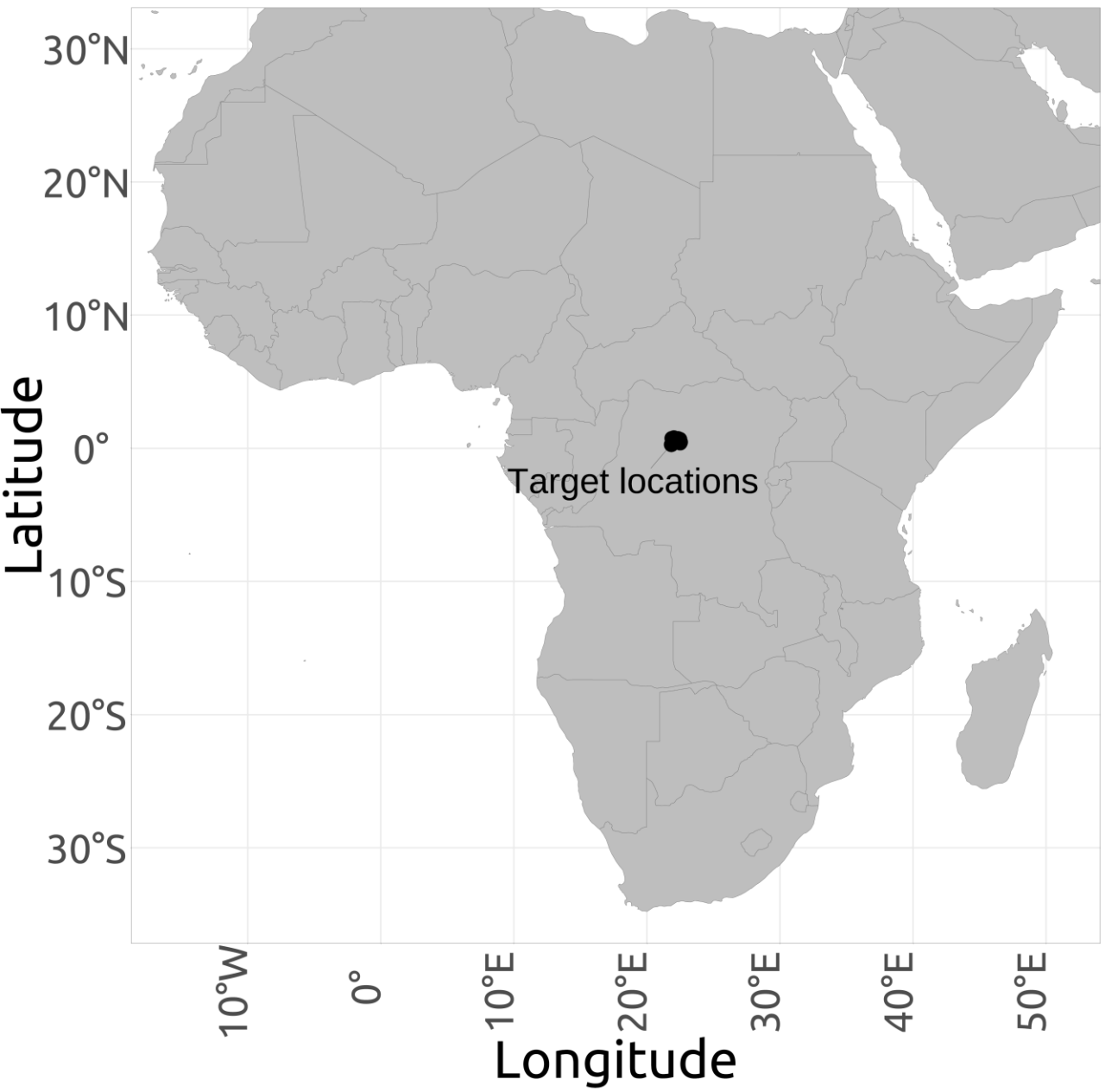
*Saul, L. K., & Roweis, S. T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of machine learning research*, *4*(Jun), 119-155.

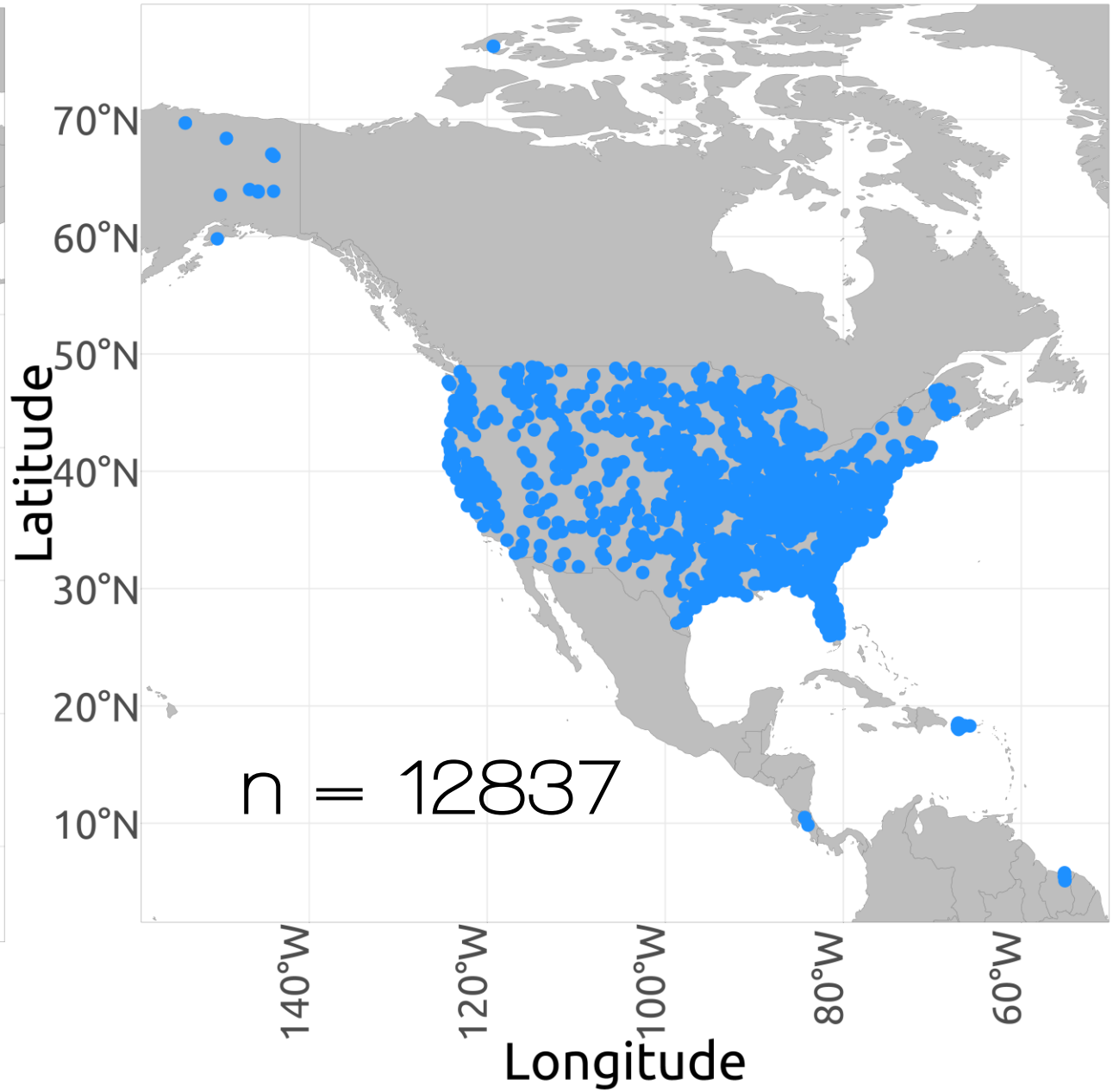
Example

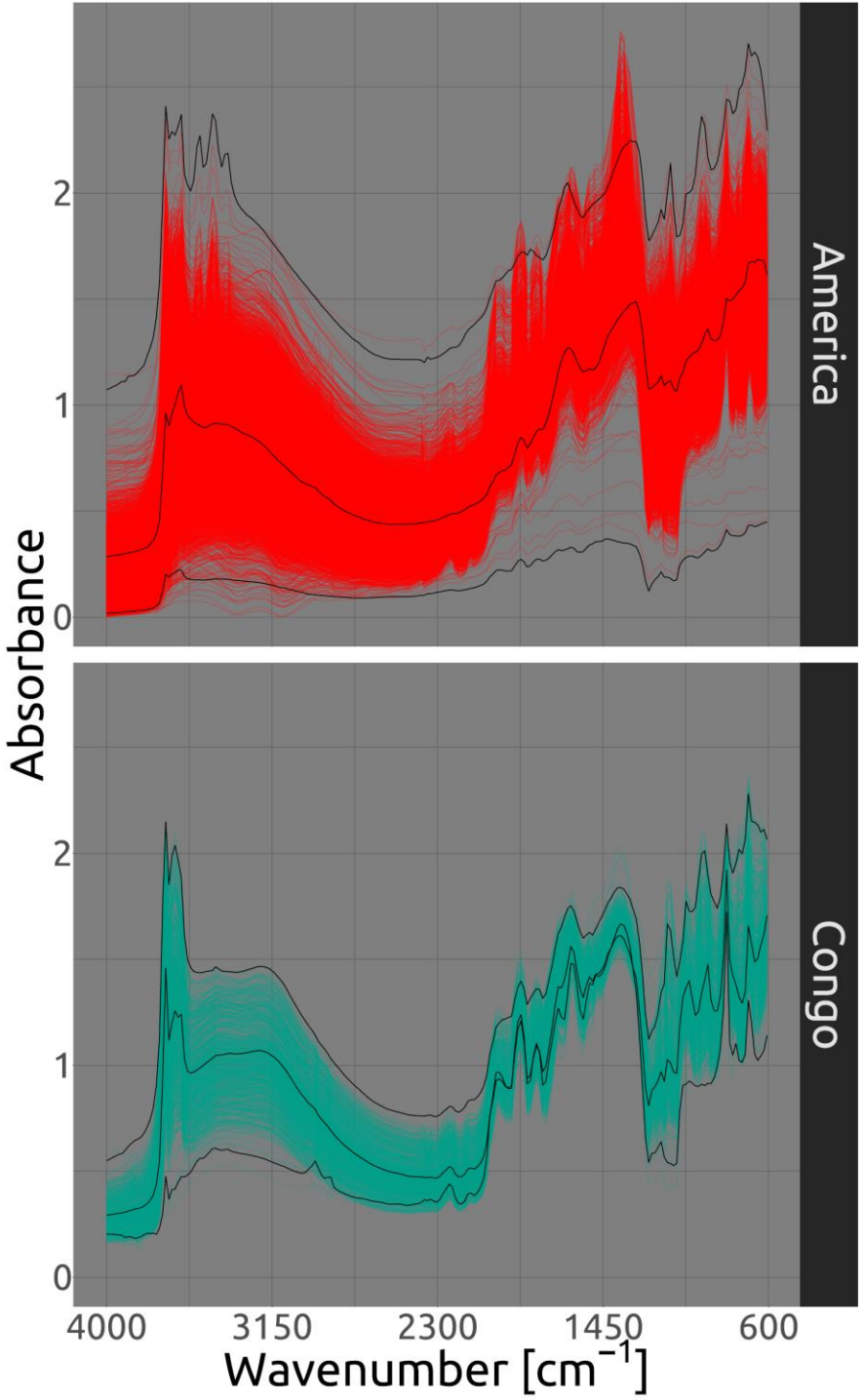
Total Carbon in soils


30°N
20°N
10°N
0°
10°S
20°S
30°S
Latitude
Target locations
10°W
0°
10°E
20°E
30°E
40°E
50°E
Longitude

n: 729


30°N
20°N
10°N
0°
10°S
20°S
30°S
Latitude
Target locations
10°W
0°
10°E
20°E
30°E
40°E
50°E
Longitude


70°N
60°N
50°N
40°N
30°N
20°N
10°N
Latitude
n = 12837
140°W
120°W
100°W
80°W
60°W
Longitude



12837 samples

729 samples

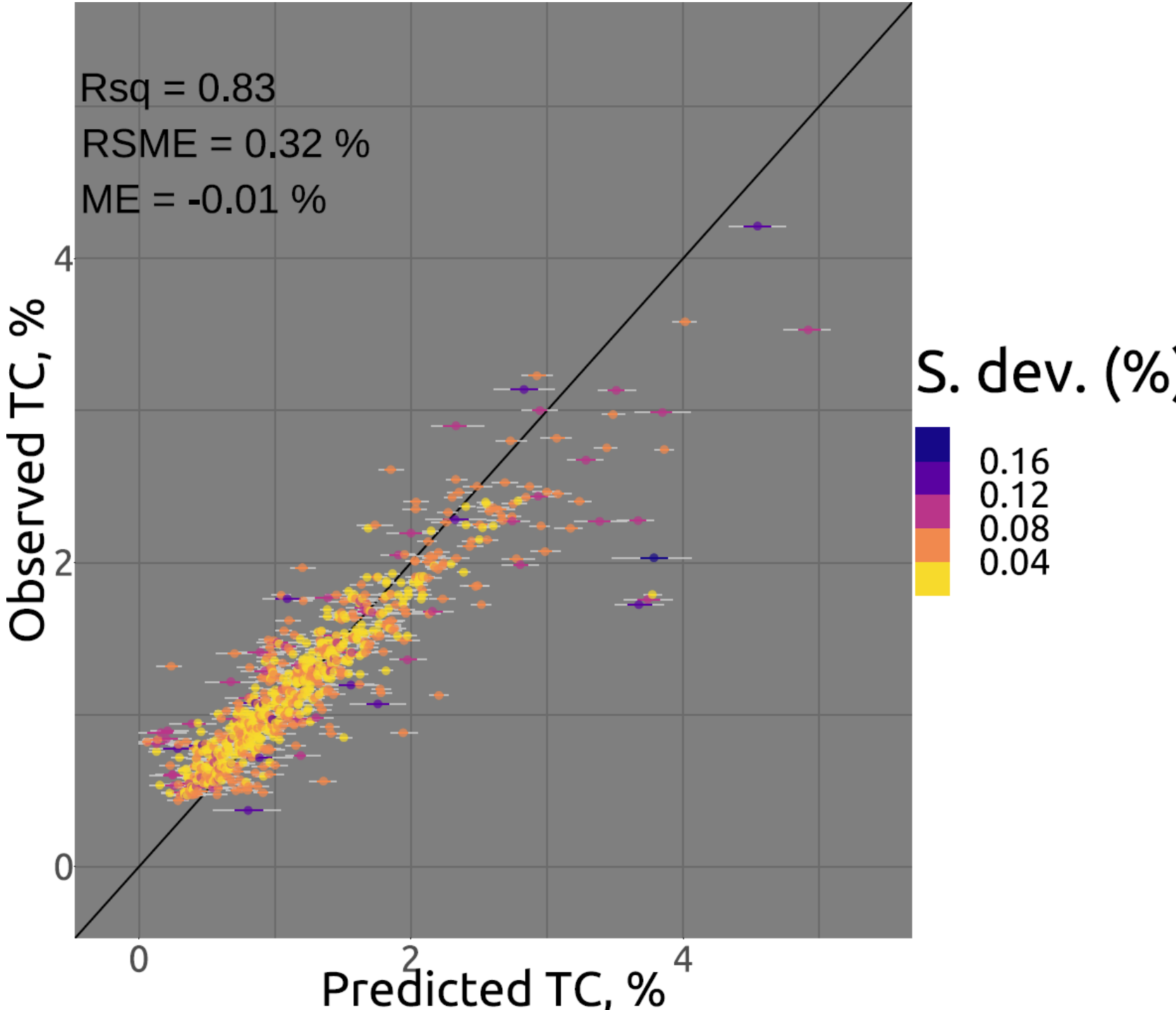
Genetically derived model

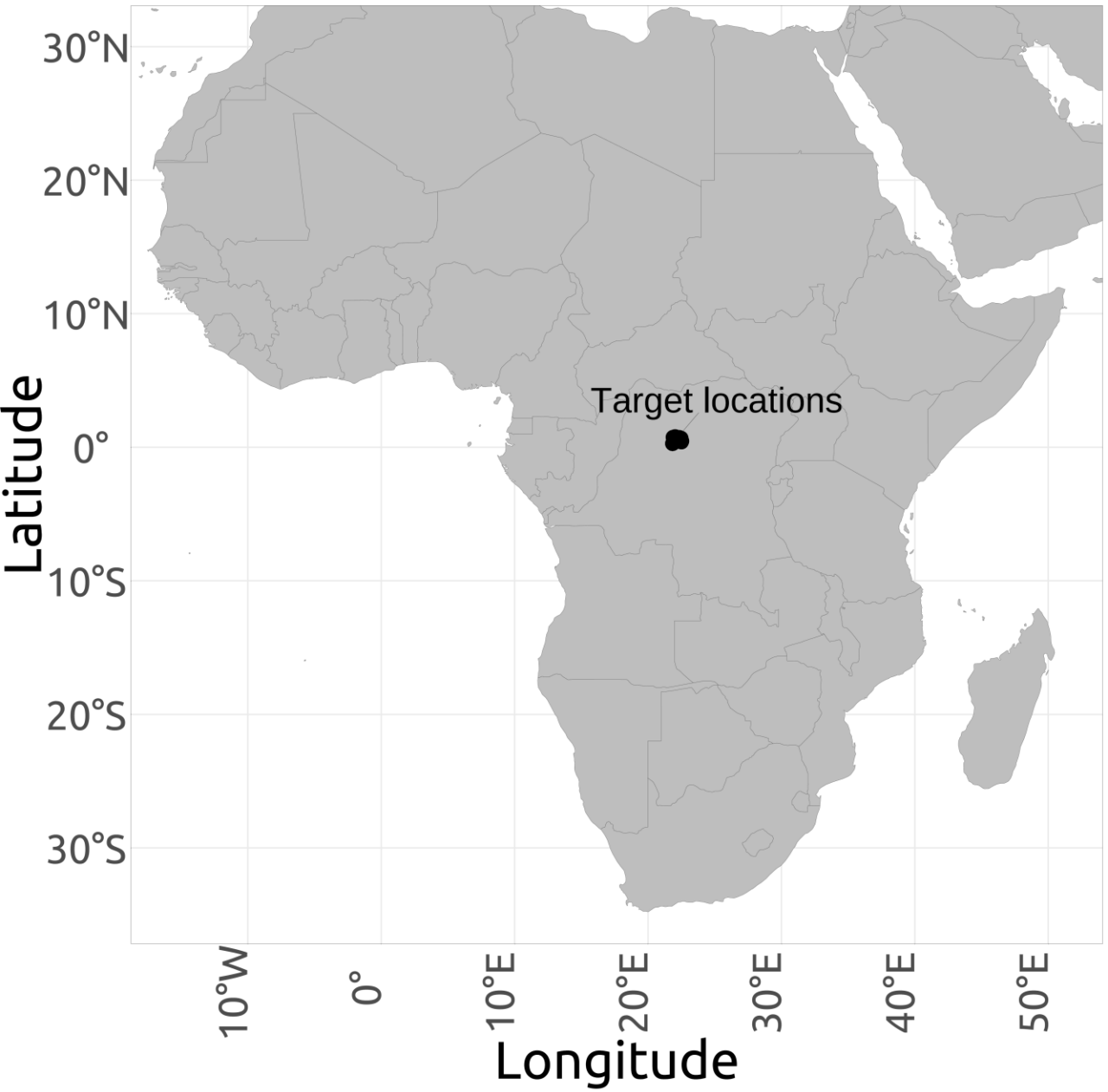
Weakness scores:

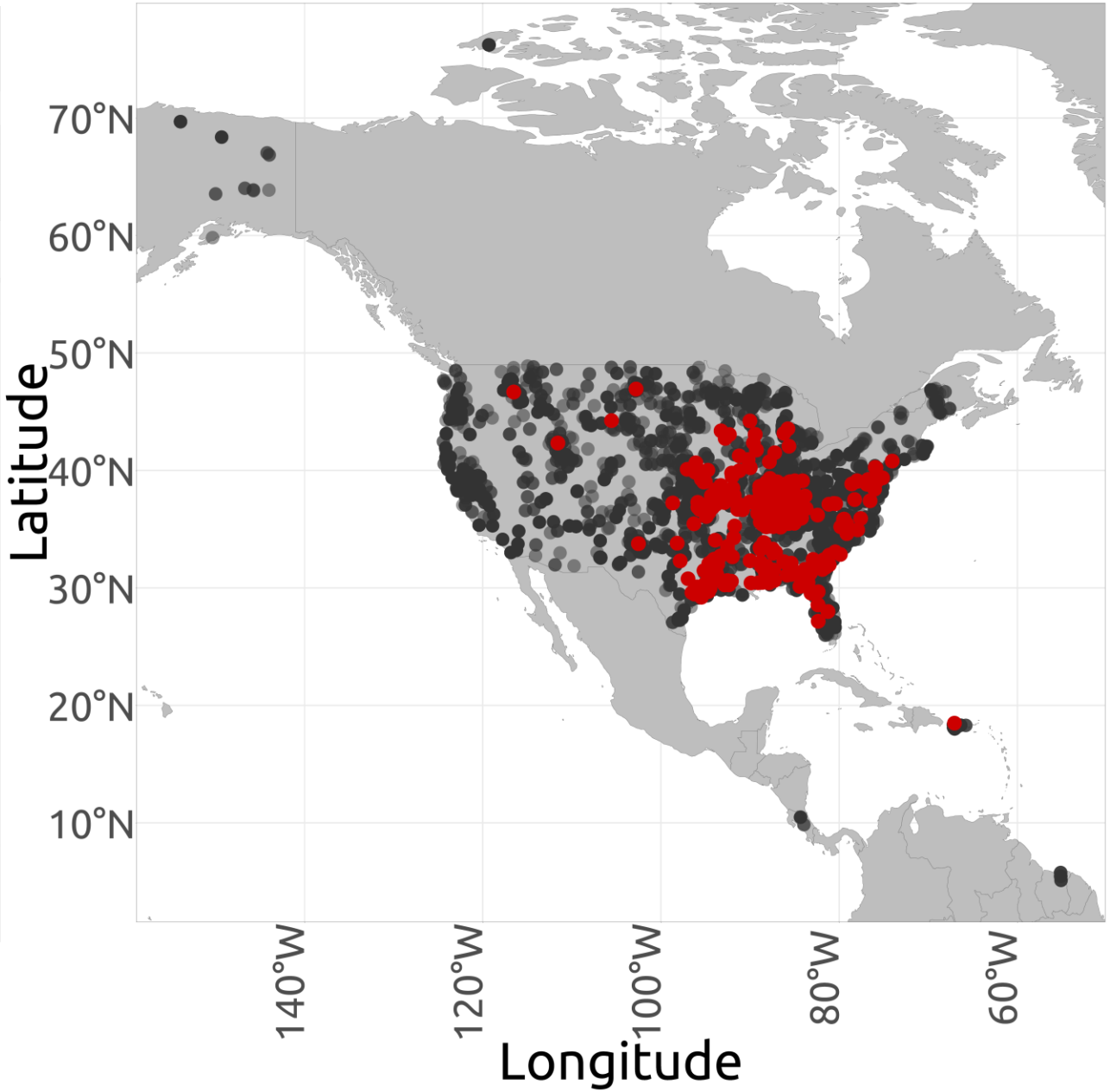
- Spectral reconstruction
- Spectral similarity
- **Response accuracy**

↓

20 samples with response values



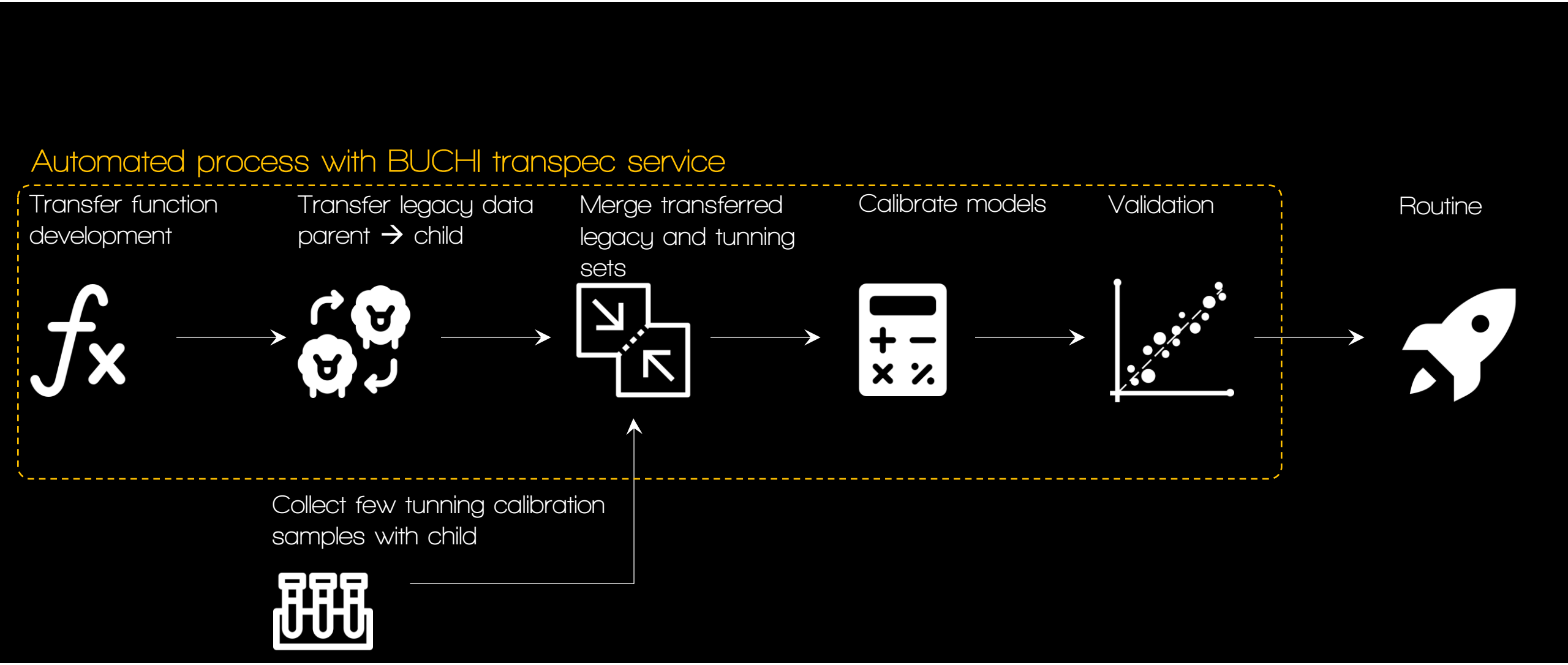

Target locations
30°N
20°N
10°N
0°
10°S
20°S
30°S
Latitude
10°W
0°
10°E
20°E
30°E
40°E
50°E
Longitude


70°N
60°N
50°N
40°N
30°N
20°N
10°N
Latitude
140°W
120°W
100°W
80°W
60°W
Longitude

Think globally, fit locally*!

*Saul, L. K., & Roweis, S. T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of machine learning research*, *4*(Jun), 119-155.

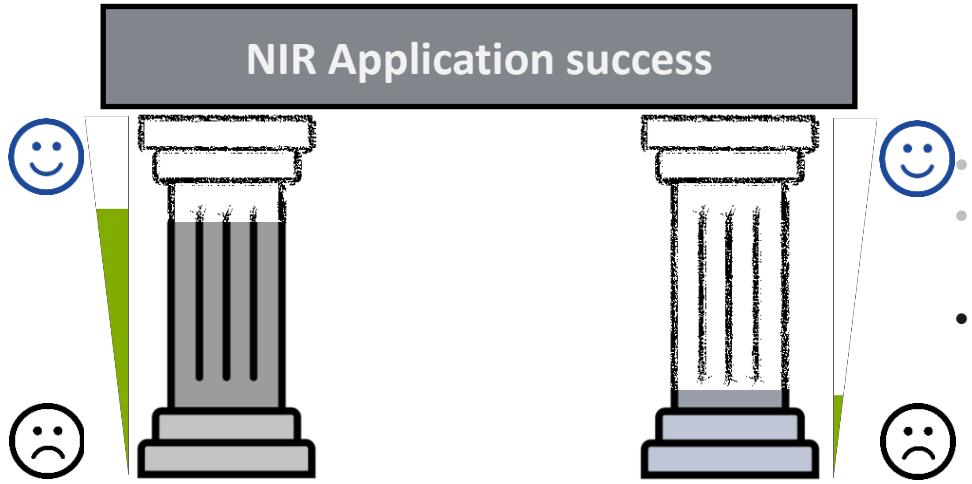
Our generalization



AI: the **what** and the why?

Is AI the holy grail for success?

How much cement do we have to build the pillars of an excellent NIR application system?



- Little automation
- Manually-engineered optmizations
- **Potential of data is underexploited**

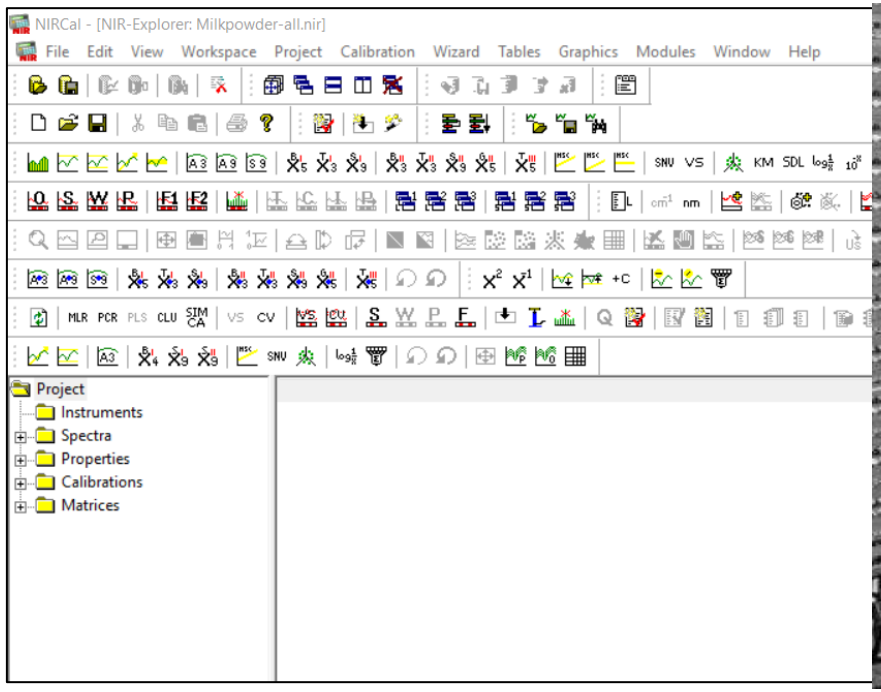
Data

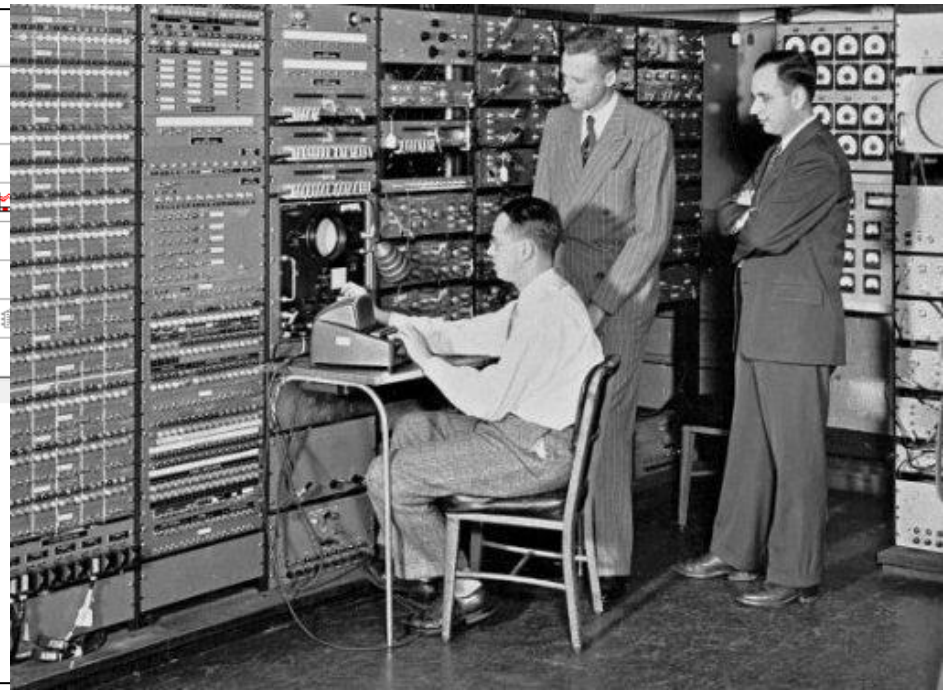
Representative -
Low noise -
Storage -

AI

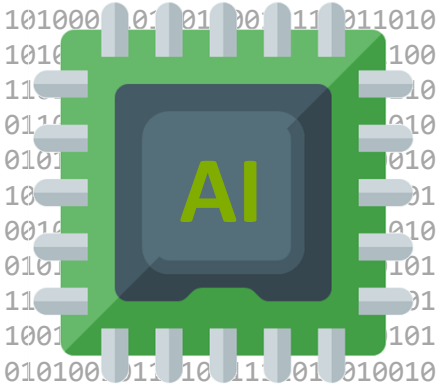
- Precise, accurate
- Highly automated
- Computationally efficient
- Safe
- Interpetable

From current situation… (BUCHI NIRCal sofware)





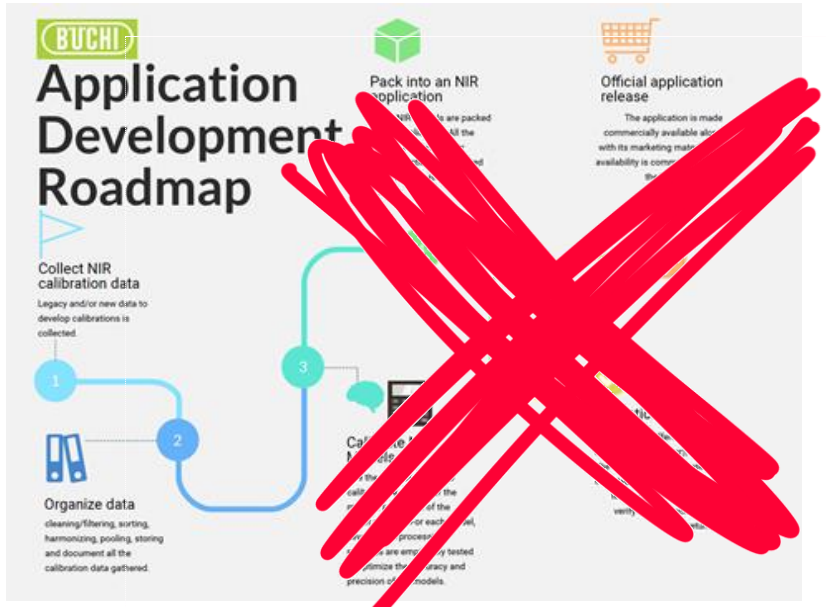
To a sound NETCAL Solution…



AI: the what and the **why?**

We need to remove all this heavy calibration burden because:

- **It is expensive**



AI: the what and the **why?**

We need to remove all this heavy calibration burden because:

- **It is expensive**
- **Sales force should focus on sales** (not on learning and dealing with chemometrics)

AI: the what and the **why?**

We need to remove all this heavy calibration burden because:

- **It is expensive**
- **Sales force should focus on sales** (not on learning and dealing with chemometrics)
- **For customers,** many time-consuming iterations to get their applications ready **create frustration**.

AI: the what and the **why?**

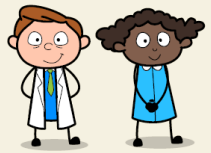
We need to remove all this heavy calibration burden because:

- **It is expensive**
- **Sales force should focus on sales** (not on learning and dealing with chemometrics)
- **For customers,** many time-consuming iterations to get their applications ready **create frustration**.
- **Chemometrics** is a complex subject which **requires time to be learned/taught**

Food and Agriculture Organization of the United Nations

GLOSOLAN is a Global Soil Laboratory Network which aims to harmonize soil analysis methods and data so that soil information is comparable and interpretable across laboratories, countries and regions.

Good practices on purchasing and operating laboratory equipment



This document aims to provide laboratory staff and managers with guidance on **what to do** and **what not to do** when purchasing laboratory equipment or receiving it as a donation.

The document is divided into three sections to provide users with as much support as possible. It also includes guidance on good practices in managing consumables and hazardous substances.

GLOBAL SOIL PARTNERSHIP

For service labs...

https://www.fao.org/global-soil-partnership/glosolan-old/equipment/en/

Proposed Machine Learning Framework

3 Machine Learning (ML) API

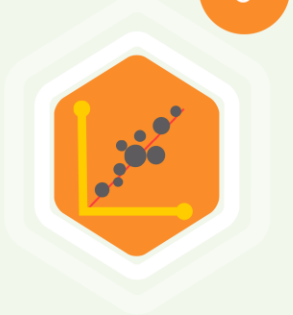
BUCHI ML algorithms are used to build optimal predictive models with selected BUCHI legacy training data that optimally match the user samples. If user data contains reference/response values, this samples are included in the models.

4 Download Application

Download the application developed by ML.

5 Validate

User checks if the performance is reach his expectations.
If not, go to step 1.



1 Sampling

Customer measures few samples which are representative of its product variability. Reference property values are not mandatory. Embedded BUCHI ML-based sampling tools guide this process.

2 Upload data

User uploads all the data collected from their product.

Contributions

- Understanding of the standardization concepts developed by competitors
- Build tools an internal standardization service
- Better understanding of the uncertainty sources and their impact on application performance
- Development of our own standardization concept
- Close cooperation with NIR assembly
- Development of standard operating procedures
- Implementation of incoming inspection of white references