How to properly analyse spectra

Harald Martens, bio-chemometrician

Prof. emerit., DEPARTMENT OF ENGINEERING CYBERNETICS, Norw. U. of Sci. & Technol. NTNU, *Trondheim Norway*

IDLETECHS AS (www.idletechs.com) *Trondheim Norway*

NTNU
Norwegian University of Science and Technology

idletechs

24th HelioSPIR association / French group for chemometrics (GFC), June 13-15 2023 in Montpellier Agropolis

How to properly analyse spectra

My perspective after 50 years: A way to analyze spectra

Harald Martens, bio-chemometrician

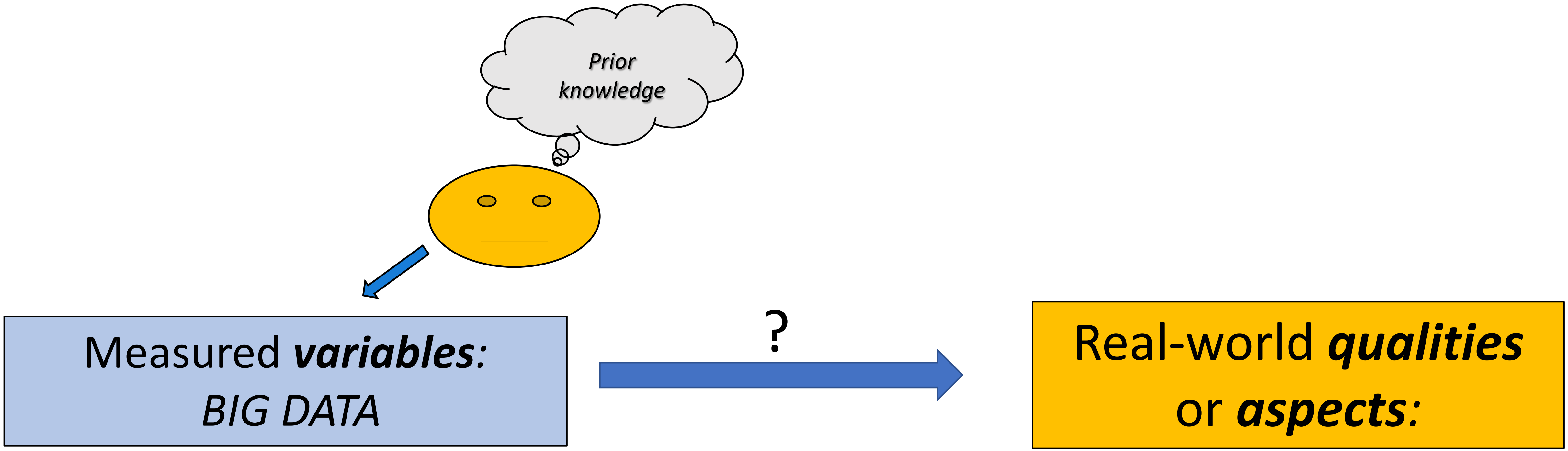
Prof. emerit., DEPARTMENT OF ENGINEERING CYBERNETICS, Norw. U. of Sci. & Technol. NTNU, *Trondheim Norway*

IDLETECHS AS (www.idletechs.com) *Trondheim Norway*

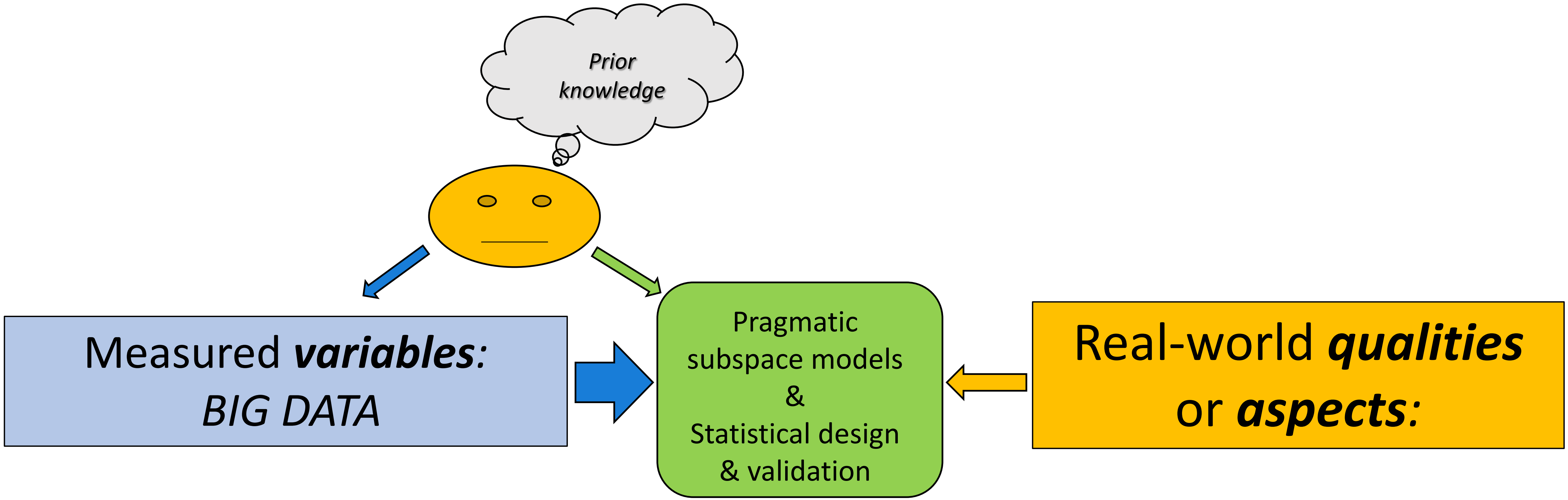
NTNU
Norwegian University of Science and Technology

idletechs

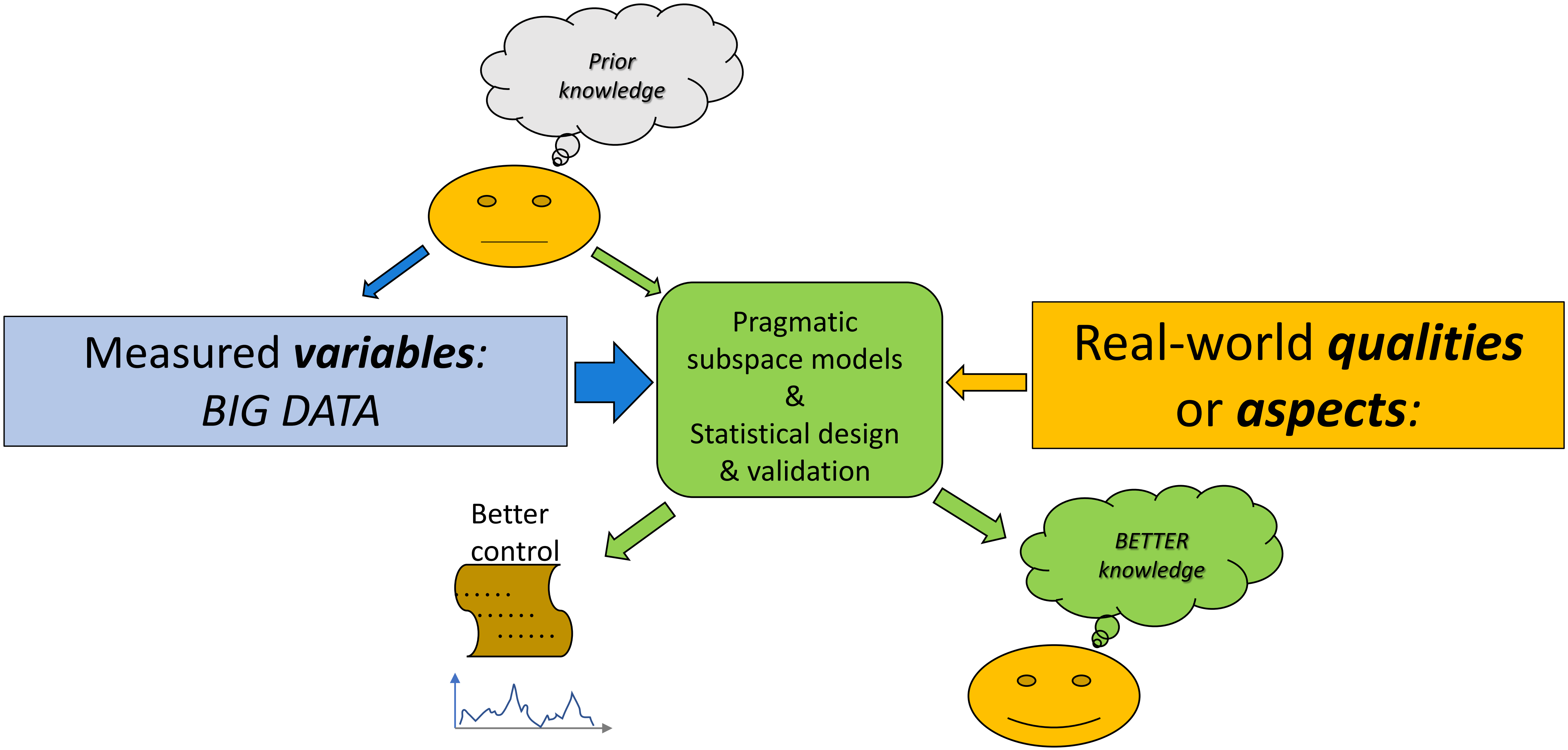
Predicting something from many other things

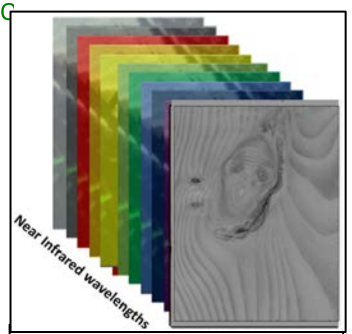


Predicting something from many other things



Predicting something from many other things



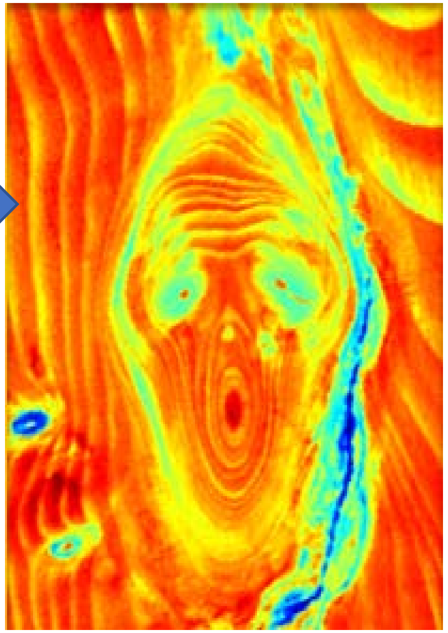


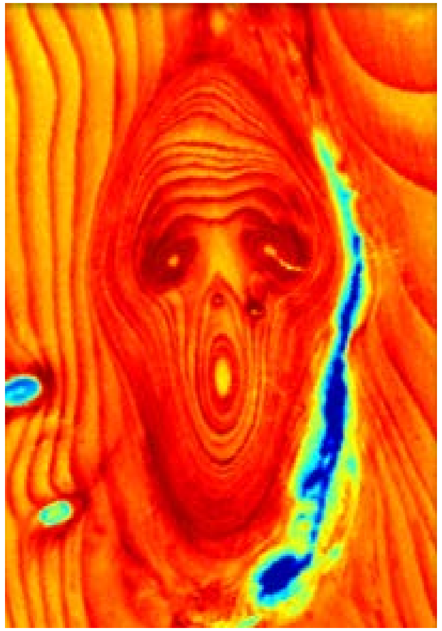
BIG DATA

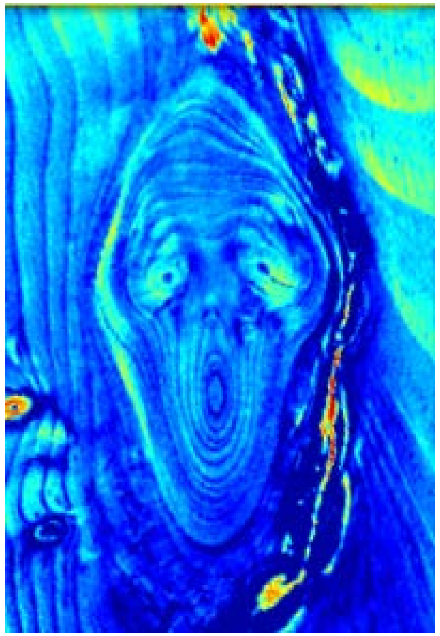
Hybrid Chemometric Subspace Modelling

Hyperspectral image of wood

Multivariate data modelling

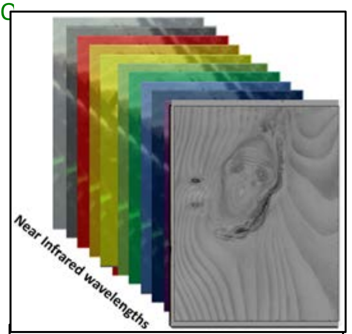






Principal components #1 #2 #3

WHAT IS GOING ON HERE ???

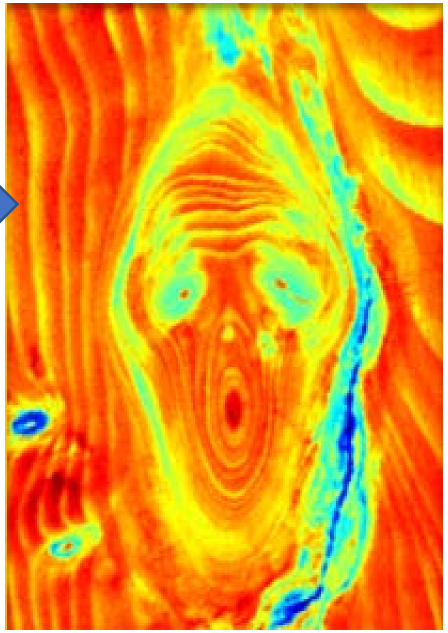
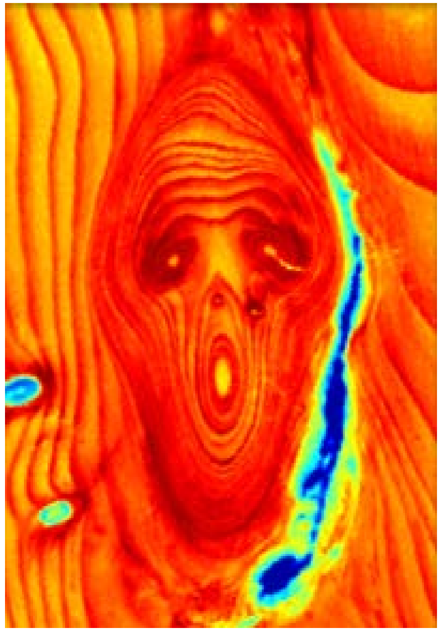
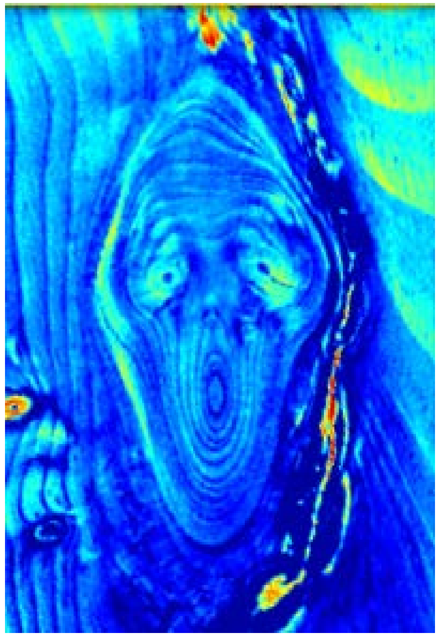


BIG DATA

Hybrid Chemometric Subspace Modelling

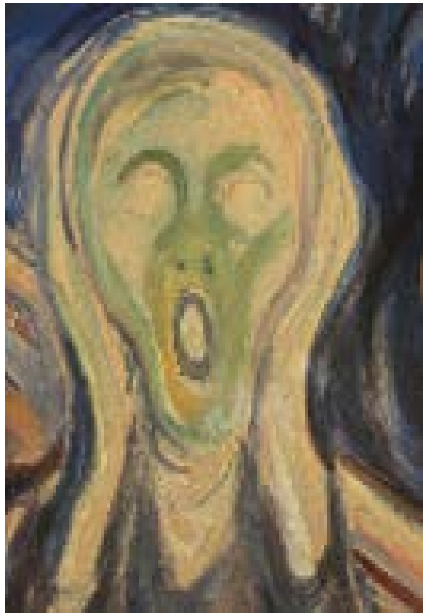
Hyperspectral image of wood

Multivariate data modelling

Principal components #1 #2 #3

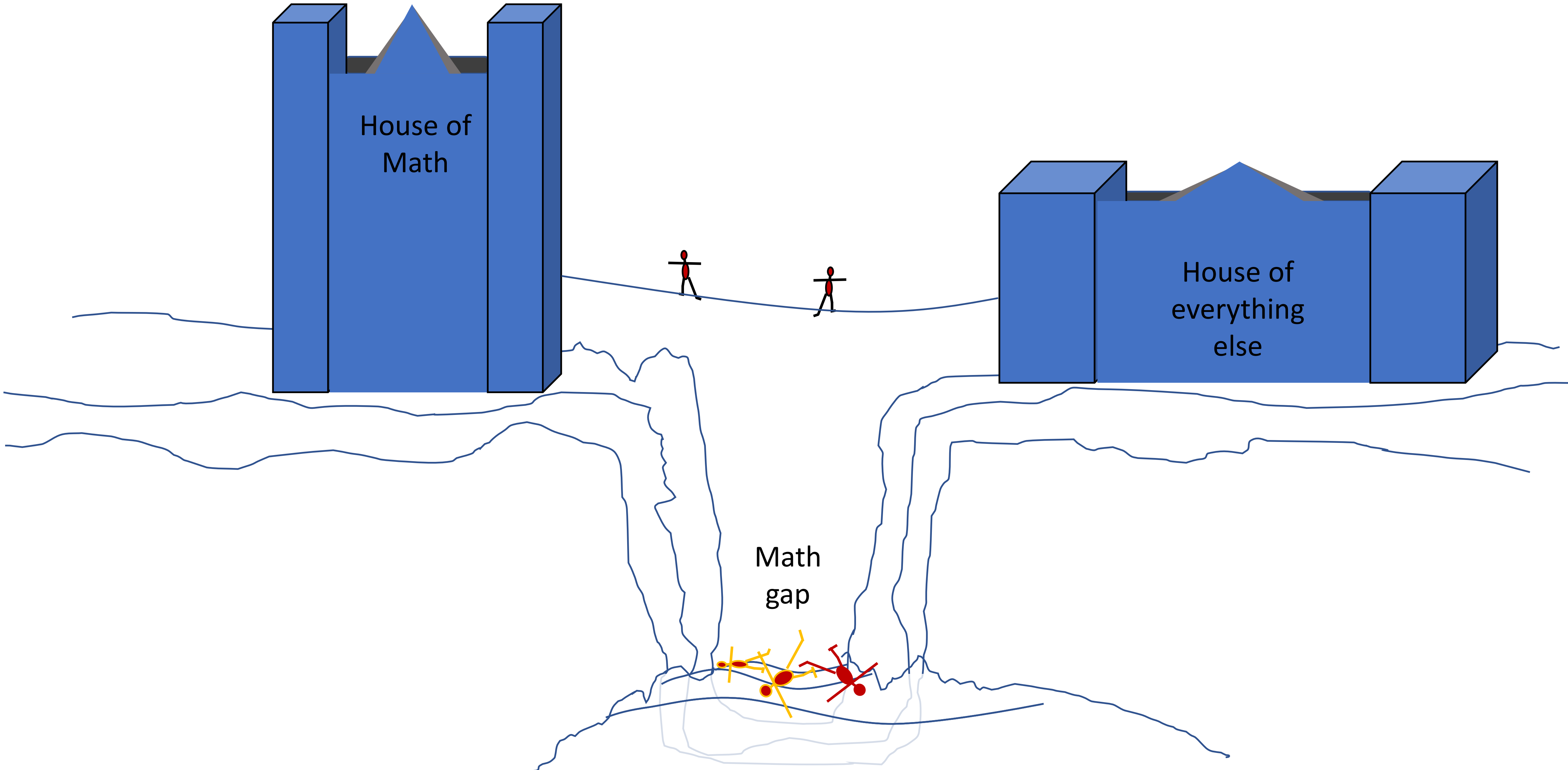
Edvard Munch: SCREAM



WHAT IS GOING ON HERE ???

MATHEMATICAL MODELLING ?

Open



Open

House of Math

House of everything else

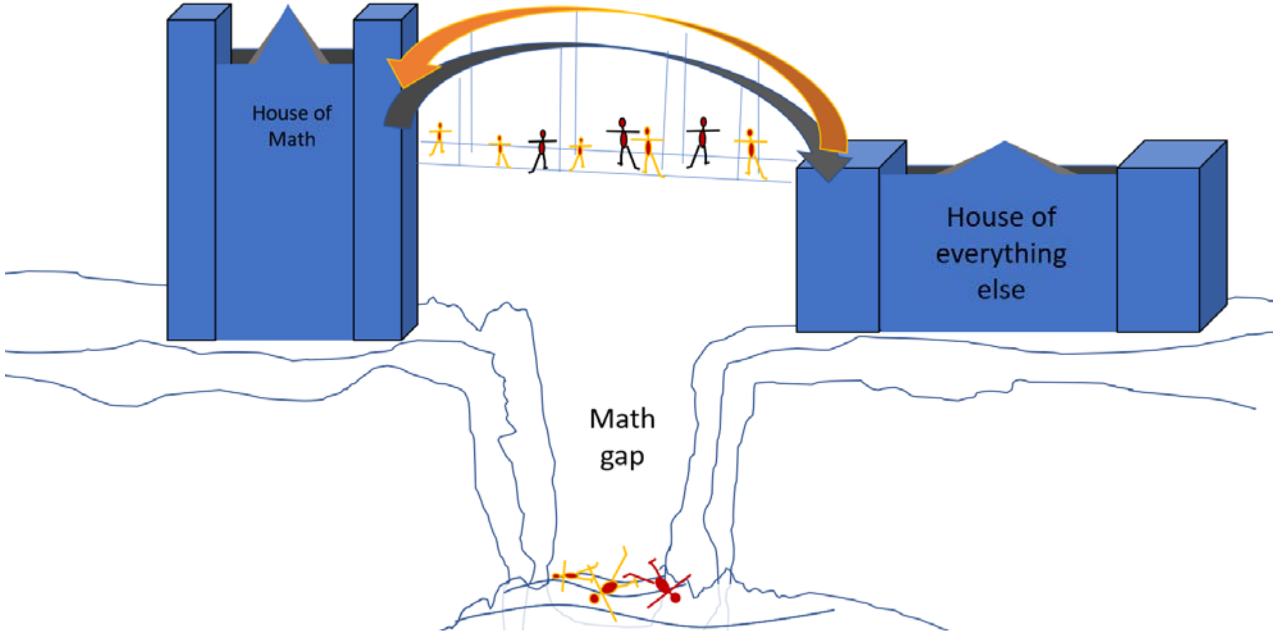
Math gap

Open

House of Math

House of everything else

Math gap



Open

A way to analyze spectra

Science in general :

Euro ⇒

What drives us?

I love to do this

This needs to be done

See how good I am

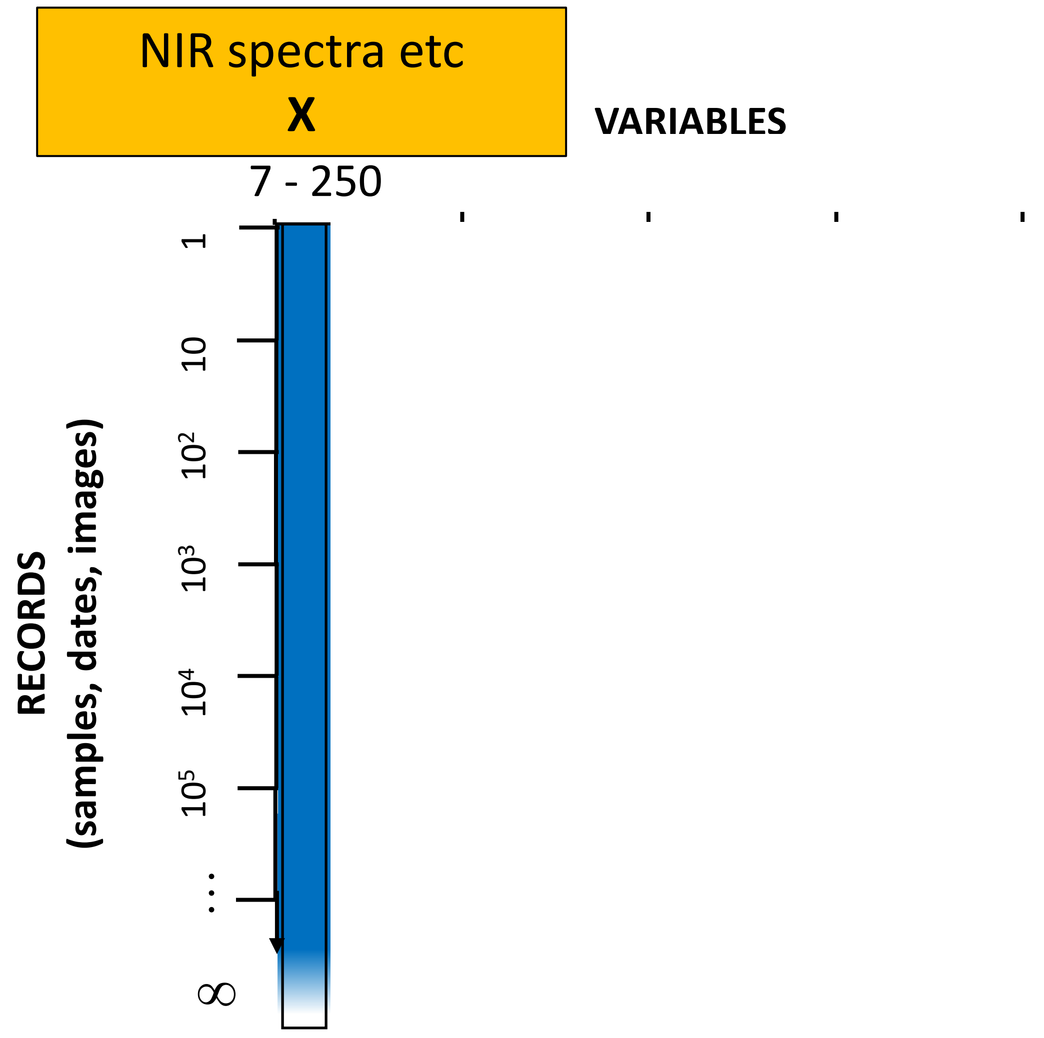
Open

A way to analyze spectra

NIR & Chemometrics:

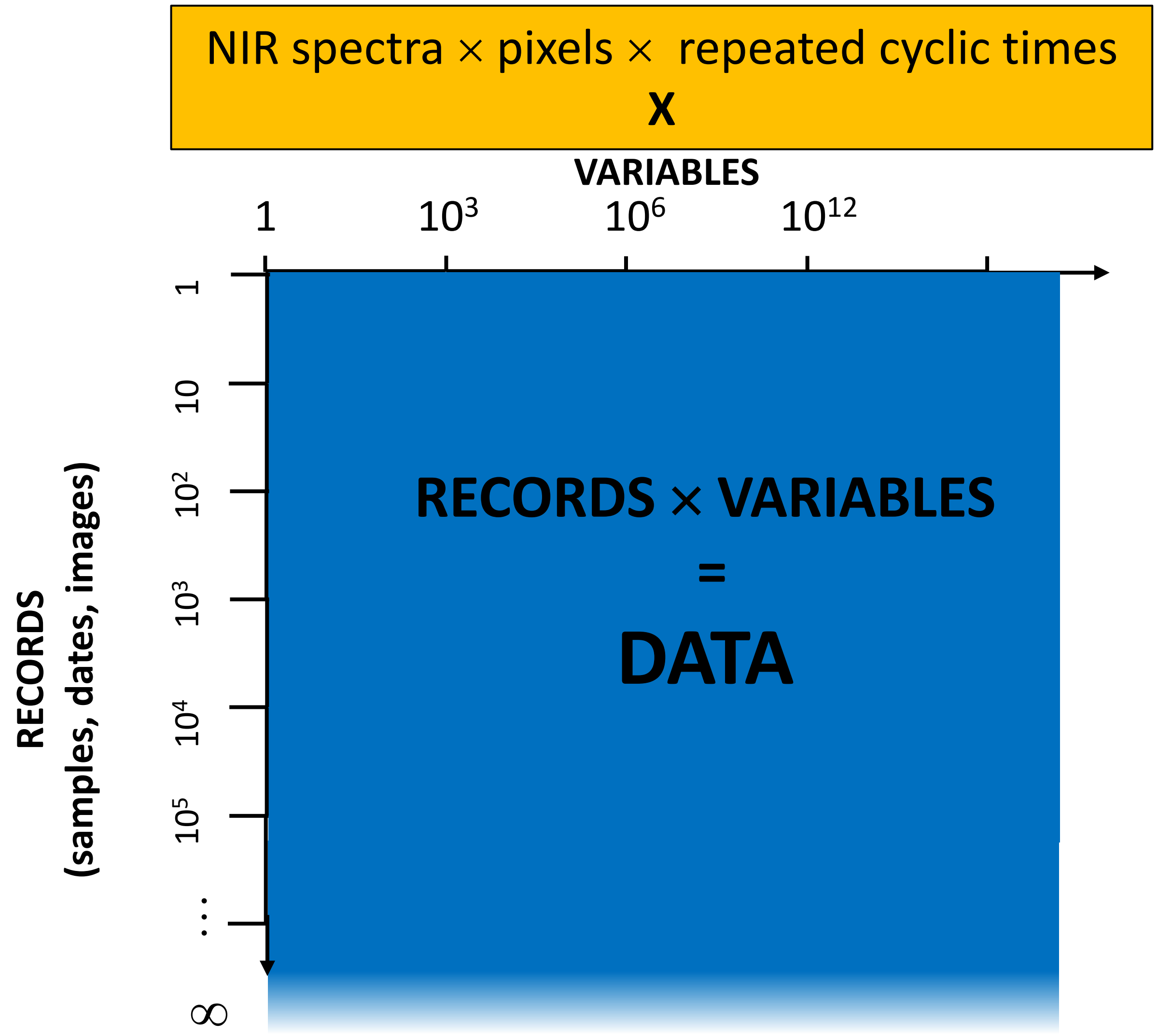
Open

Input **DATA = RECORDS** × **VARIABLES:**

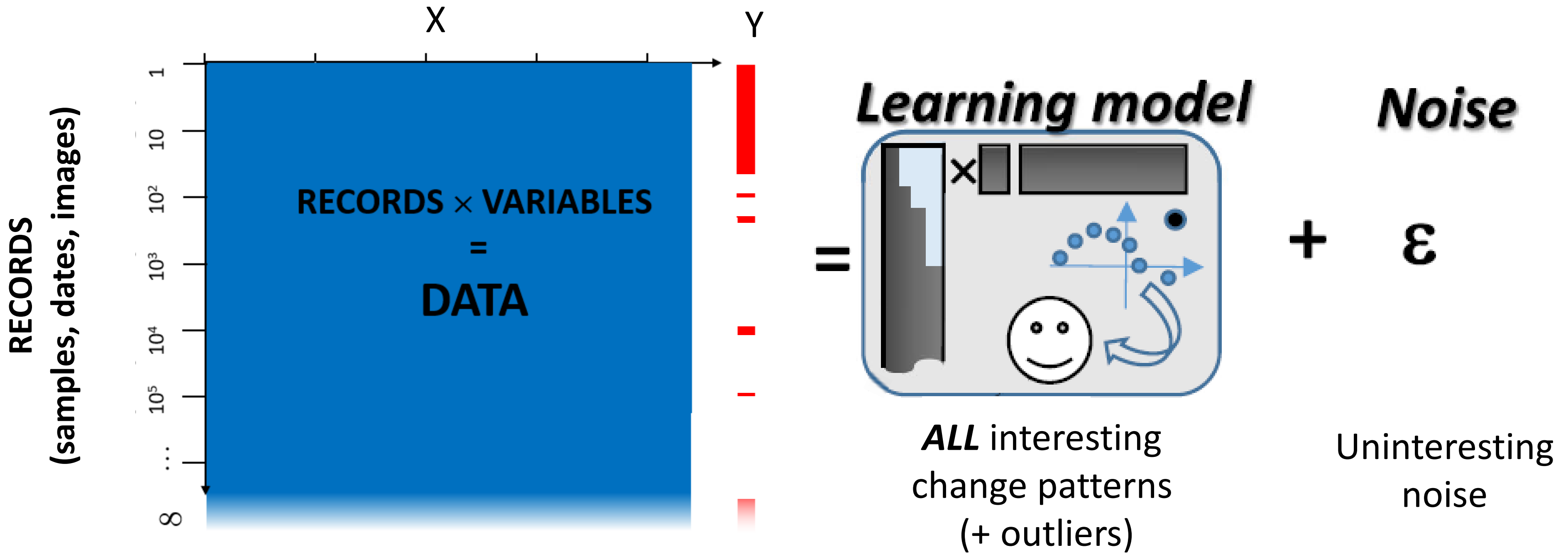


Open

Input **DATA = RECORDS** × **VARIABLES:**



OTFP: Automatic modelling of continuous high-dimensional data streams

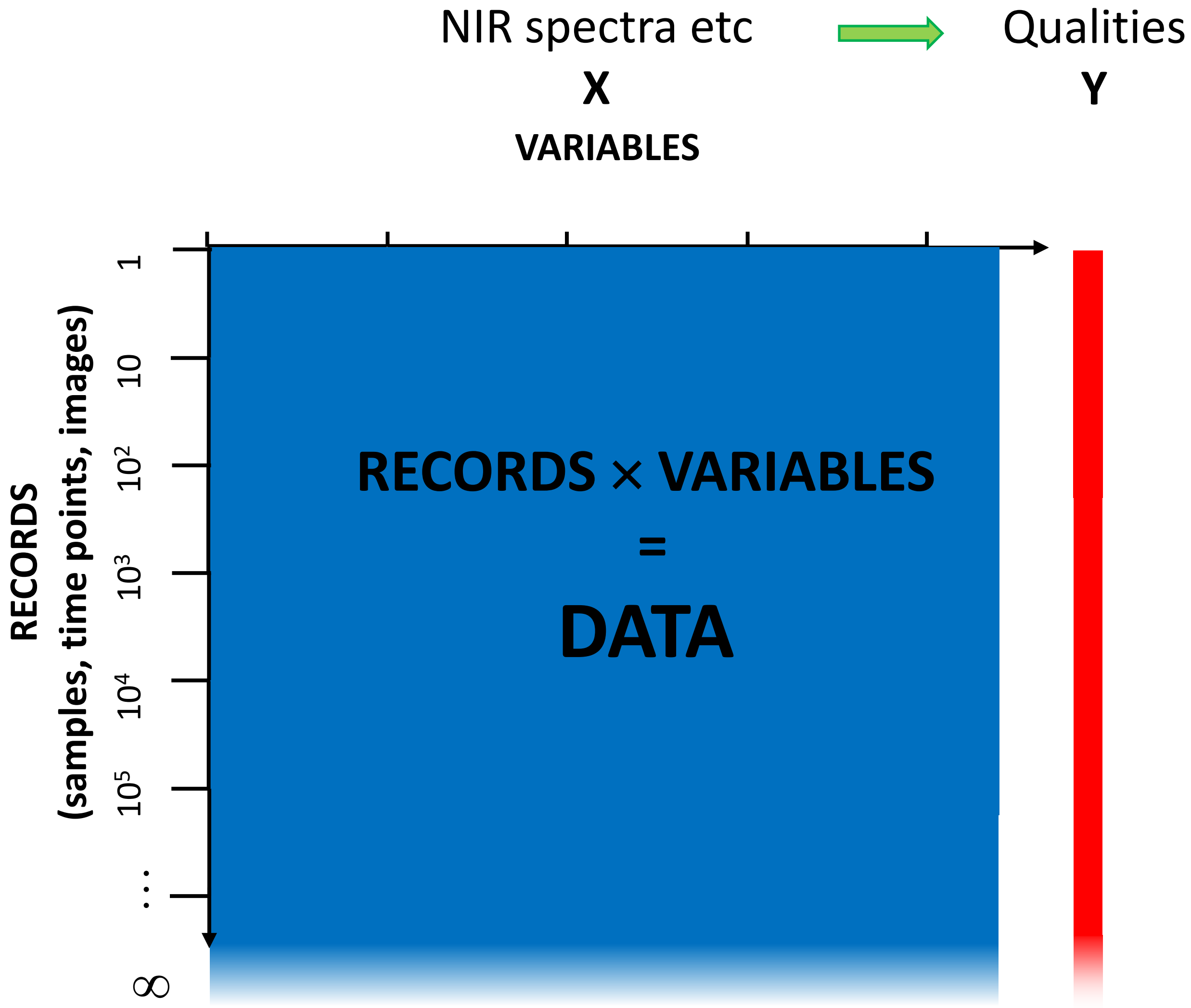


On-The-Fly-Processing software for e.g.
thermal – and hyperspectral video in industry (Vitale et al. 2017)

idletechs

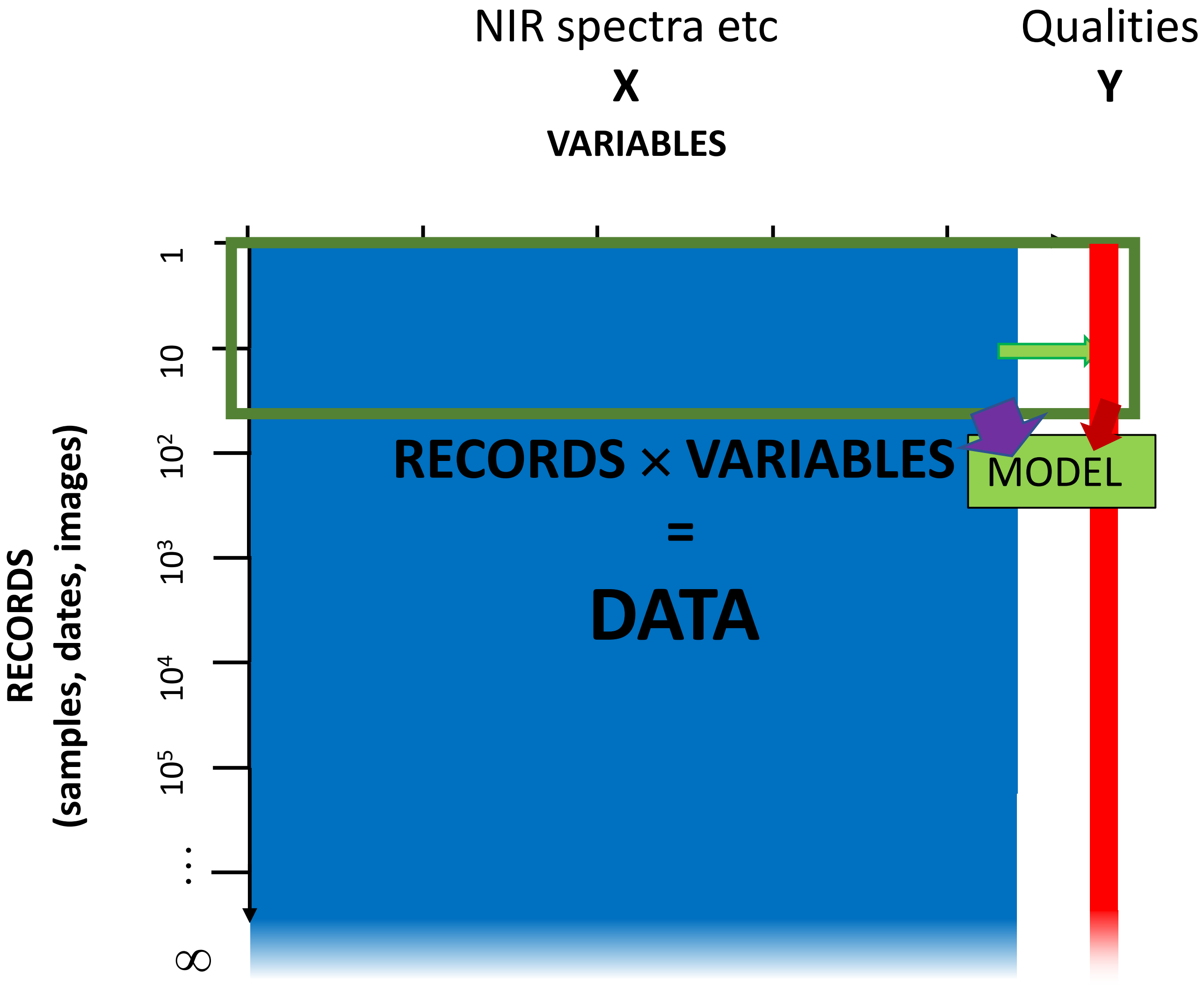
Open

Input **DATA = RECORDS × VARIABLES:**



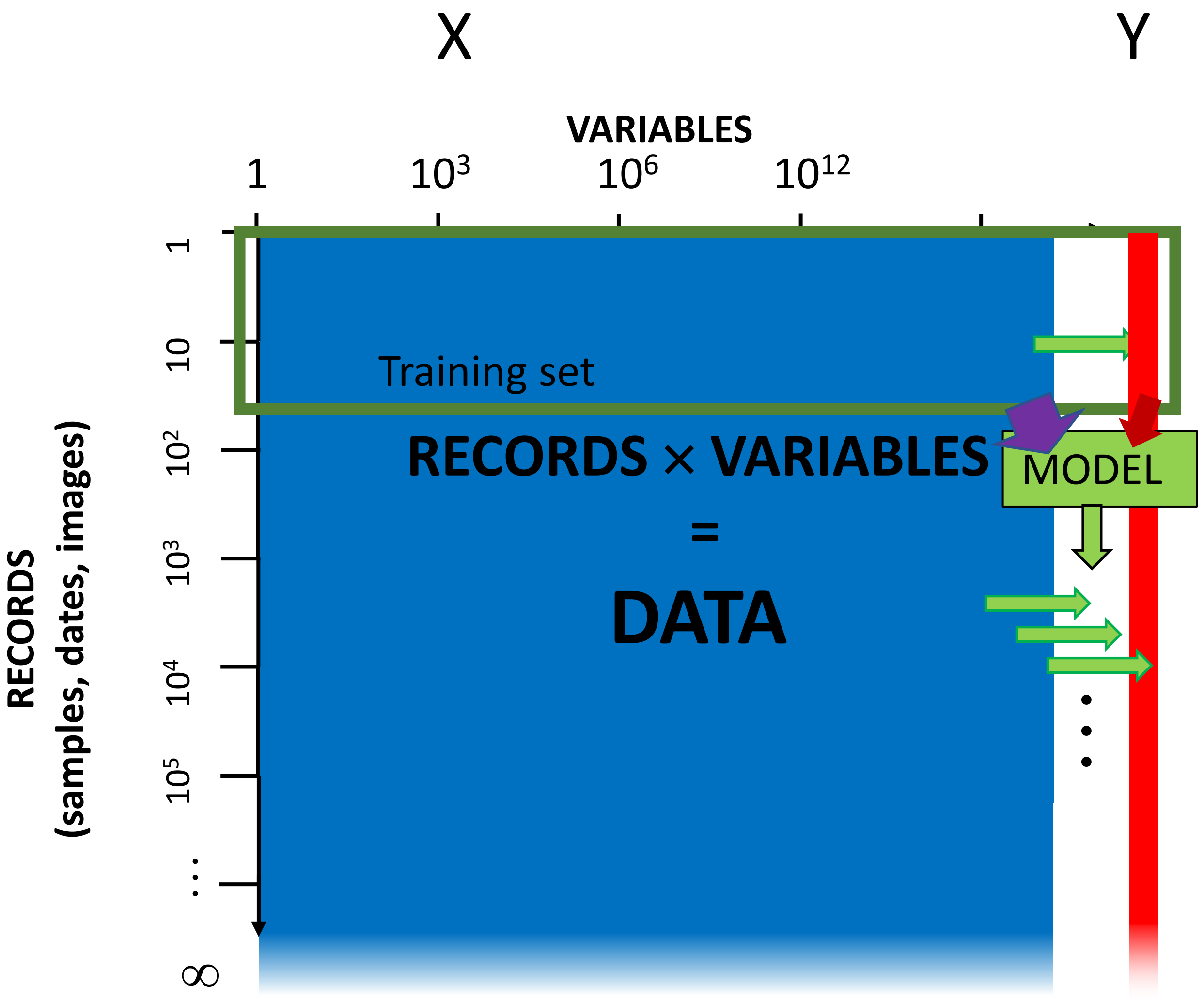
Open

Input **DATA = RECORDS** × **VARIABLES:**



Open

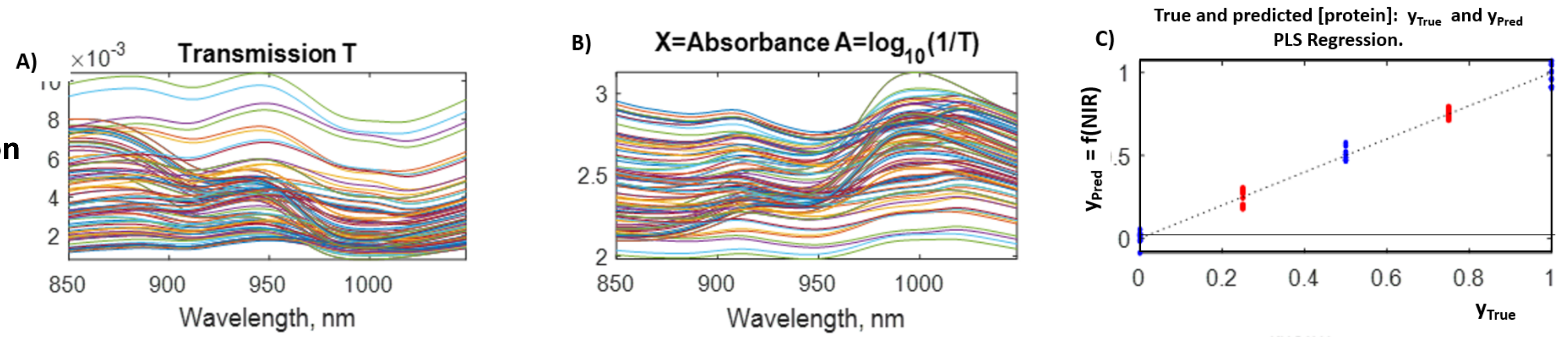
Input **DATA = RECORDS × VARIABLES:**



Open

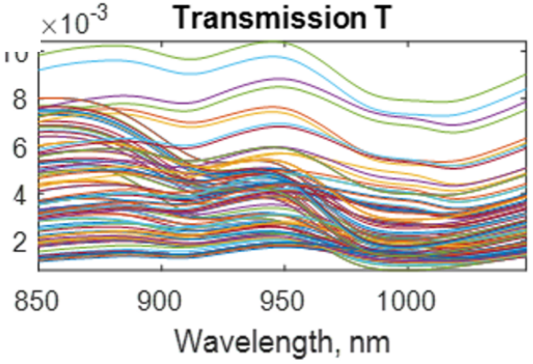
Five different powder mixtures measured by light transmission, each at varying sample - thickness and - compression

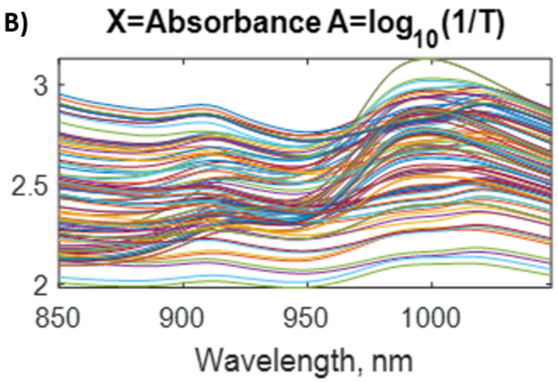
Conventional linearization
+
multivariate calibration (cross-validated PLSR)

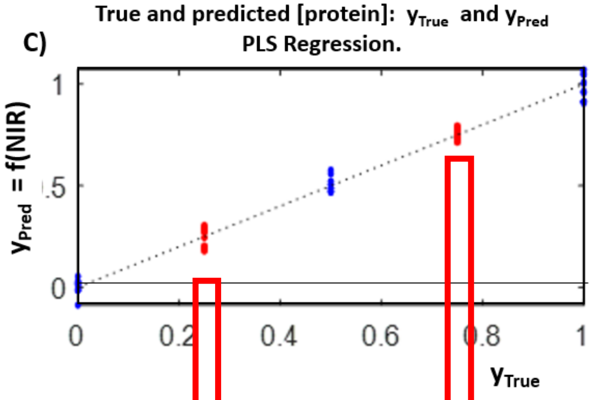


Five different powder mixtures measured by light transmission, each at varying sample - thickness and - compression

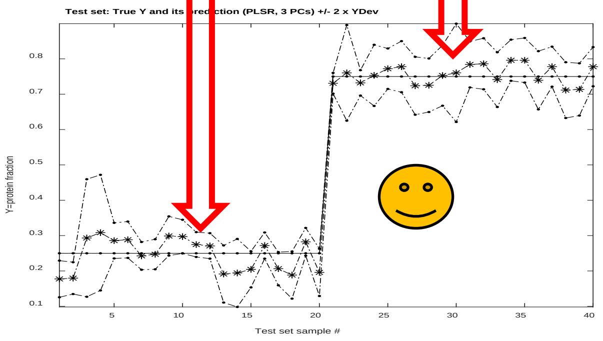
Conventional linearization
+
multivariate calibration
(cross-validated PLSR)

A)


B)


C)


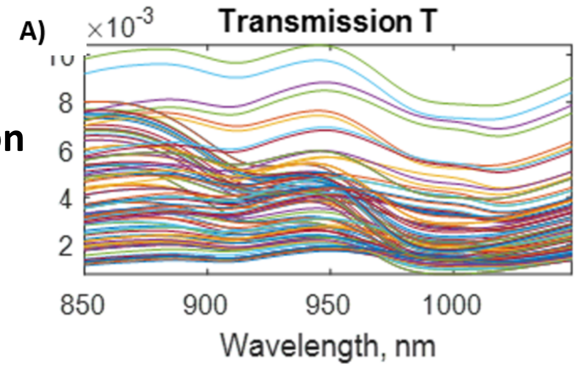
Predicted uncertainty: OK!

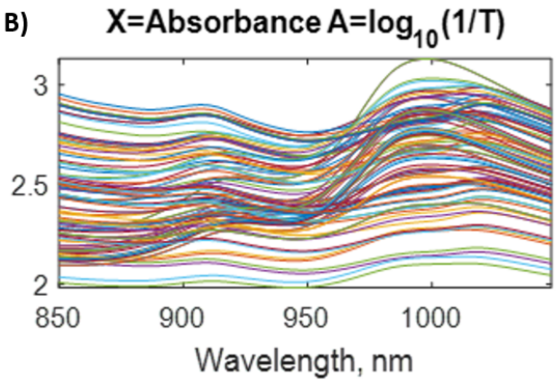


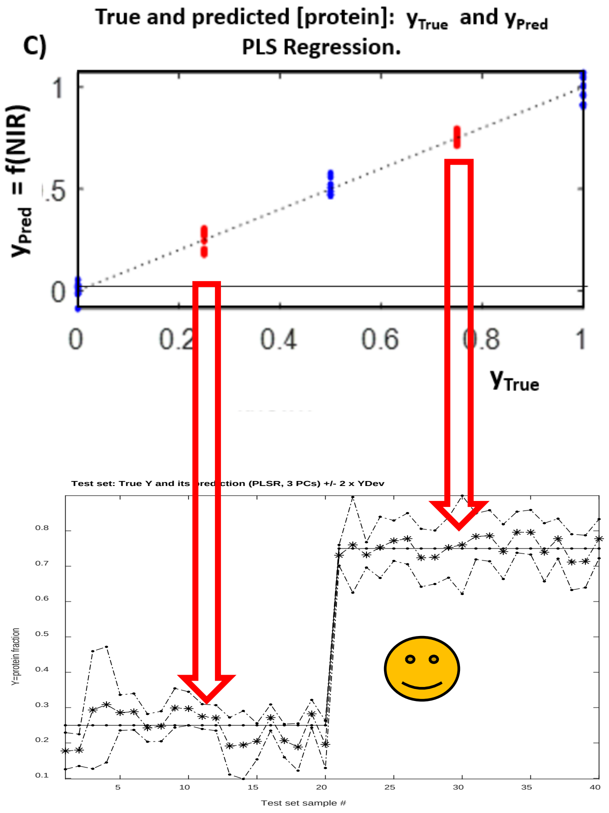
Open

Five different powder mixtures measured by light transmission, each at varying sample - thickness and - compression

Conventional linearization
+
multivariate calibration
(cross-validated PLSR)

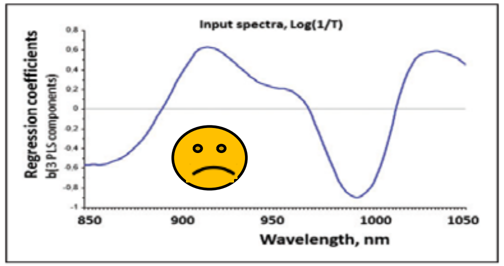






Predicted uncertainty: OK!

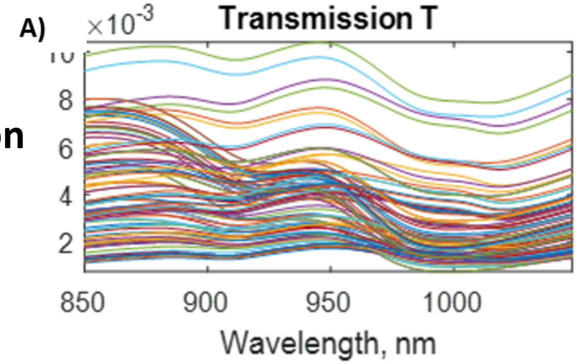
Regression coefficient = «Net Analyte Signal»:

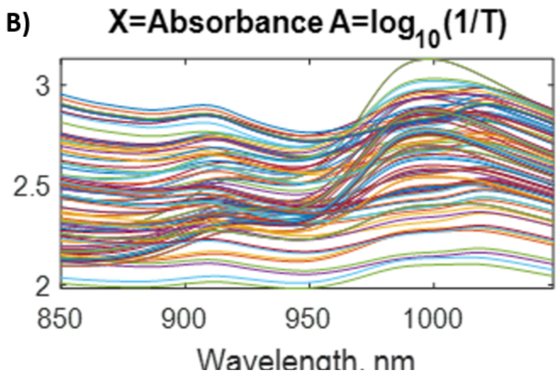


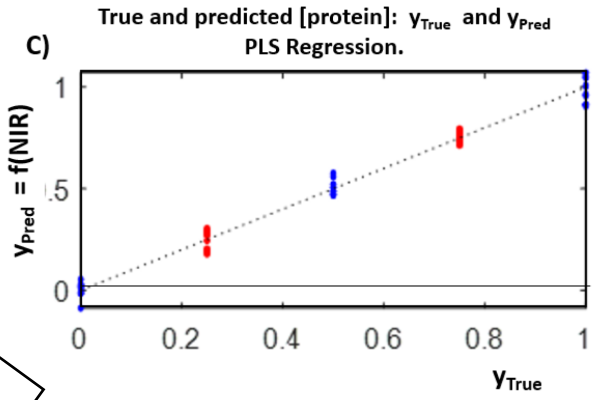
Open

Five different powder mixtures measured by light transmission, each at varying sample - thickness and - compression

Conventional linearization
+
multivariate calibration (cross-validated PLSR)

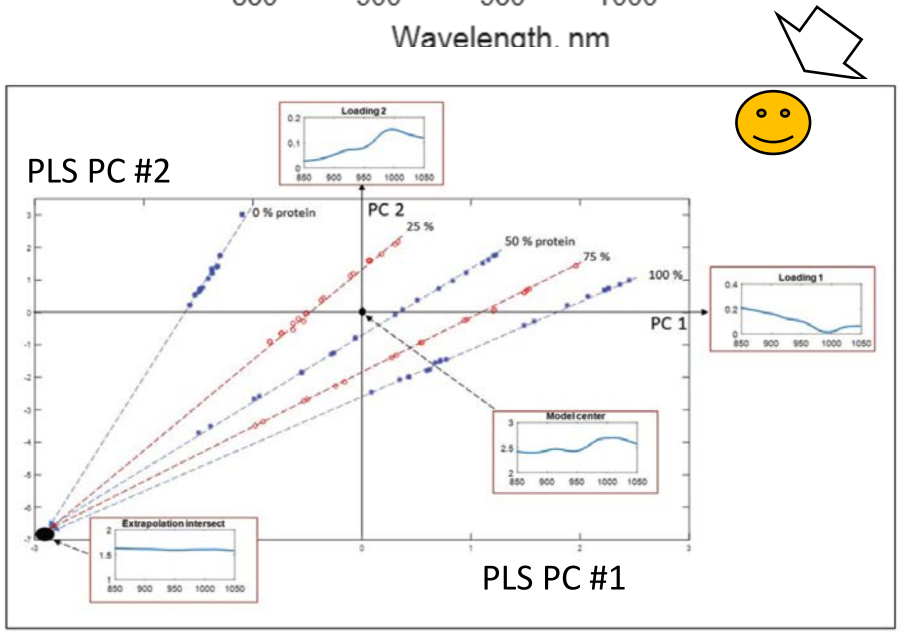






PCA & PLS regression etc :
Low-dimensional subspace !
graphic insight

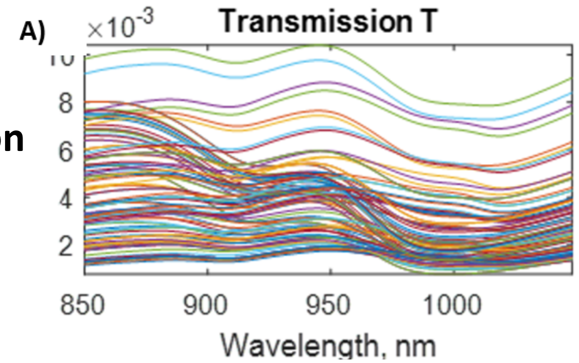
⇒

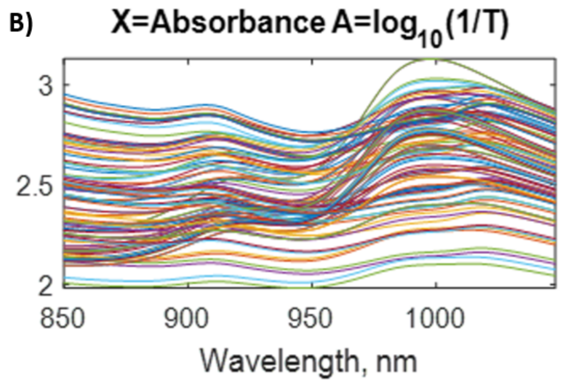


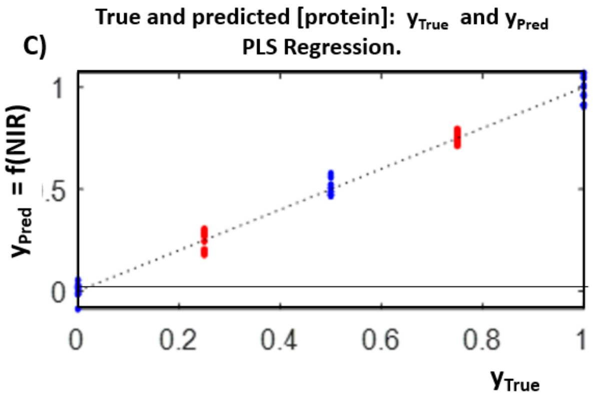
Open

Five different powder mixtures measured by light transmission, each at varying sample - thickness and - compression

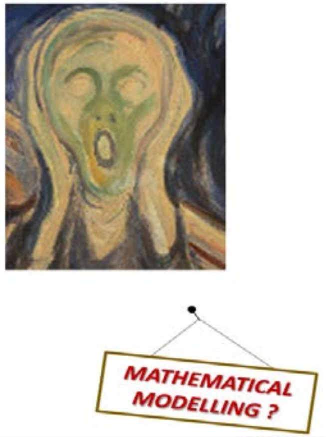
Conventional linearization
+
multivariate calibration
(cross-validated PLSR)



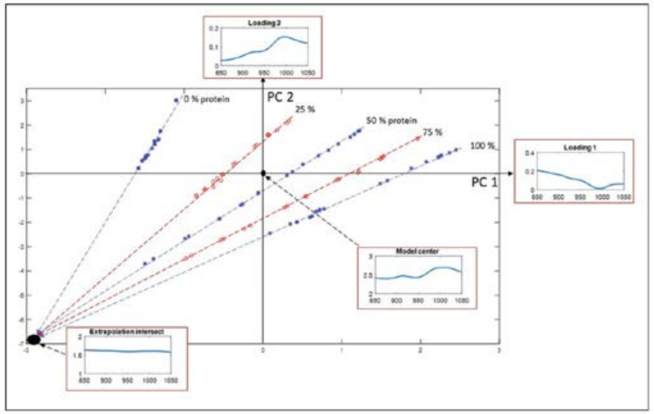




OEMSC linearization
+
multivariate calibration
(cross-validated PLSR)



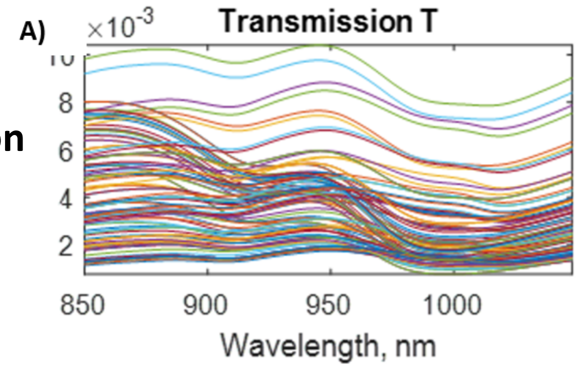
$A \approx B \times C + D$

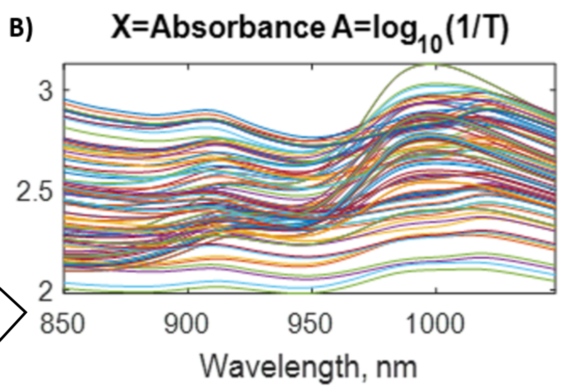


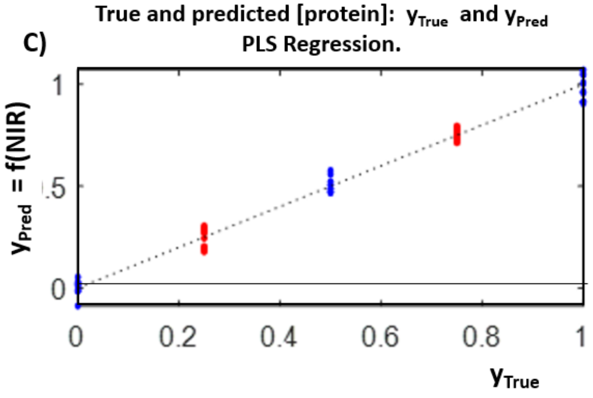
Subspace inspection
(two first PLS PCs)

Five different powder mixtures measured by light transmission, each at varying sample - thickness and - compression

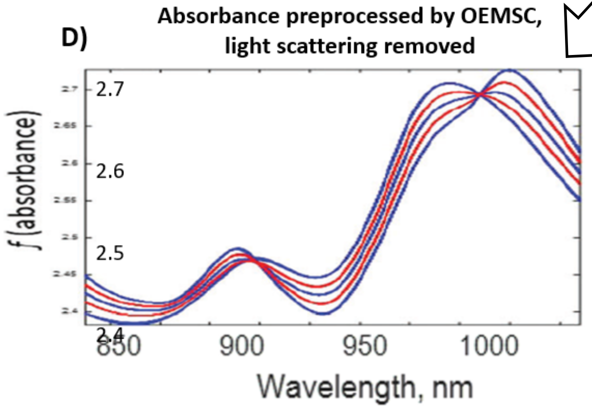
Conventional linearization
+
multivariate calibration
(cross-validated PLSR)



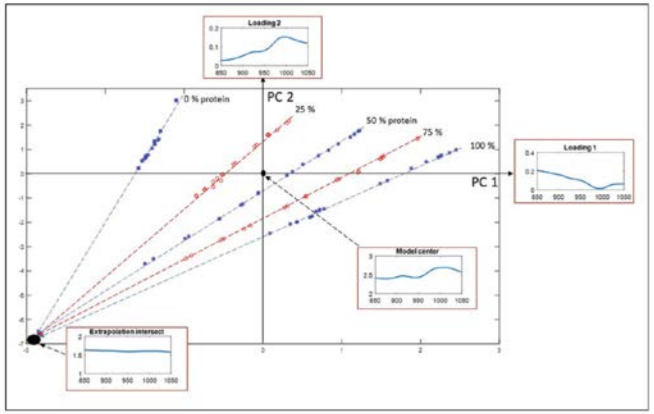




OEMSC linearization
+
multivariate calibration
(cross-validated PLSR)



Retaining only chemical info

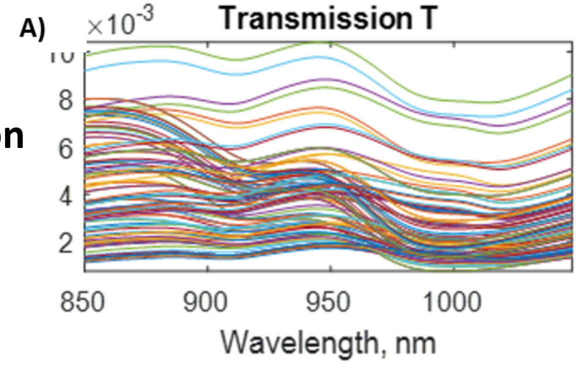


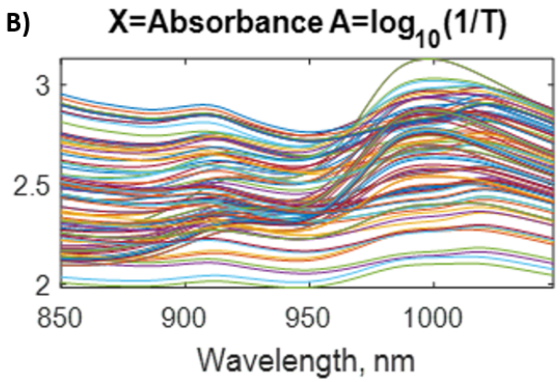
Subspace inspection
(two first PLS PCs)

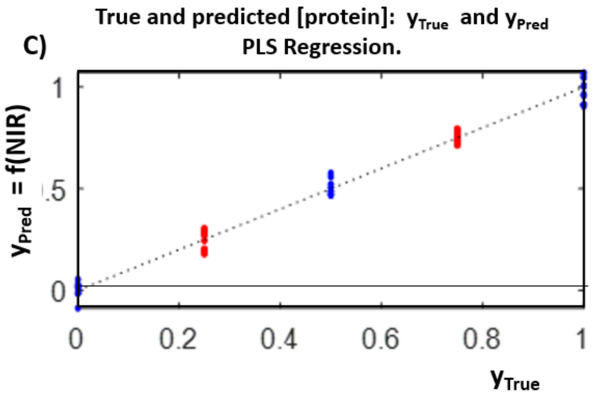
Open

Five different powder mixtures measured by light transmission, each at varying sample - thickness and - compression

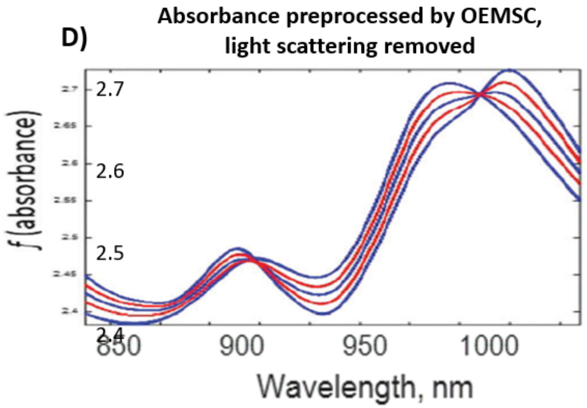
Conventional linearization
+
multivariate calibration (cross-validated PLSR)

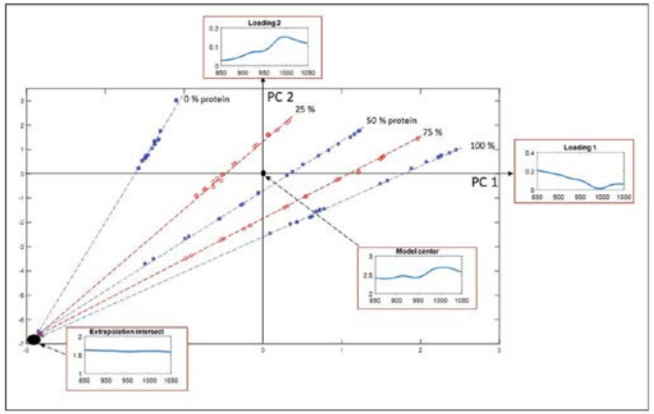


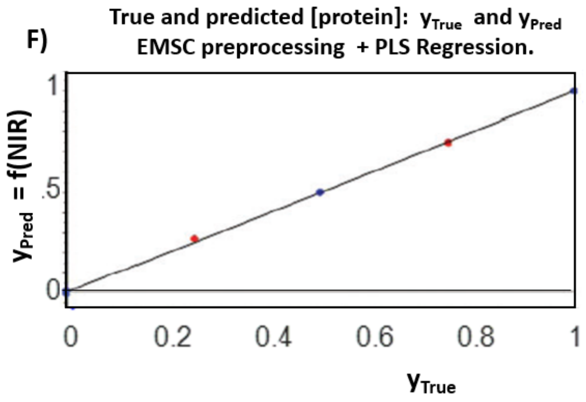




OEMSC linearization
+
multivariate calibration (cross-validated PLSR)



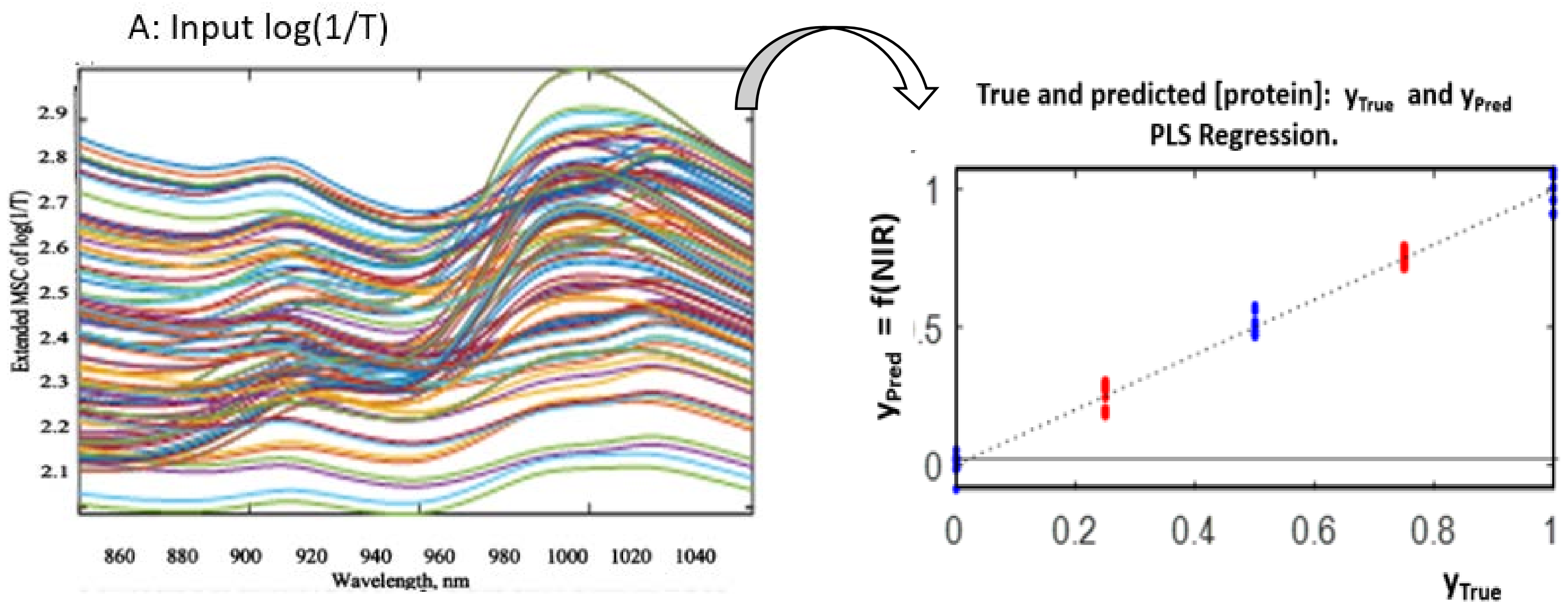




Retaining only chemical info

Subspace inspection (two first PLS PCs)

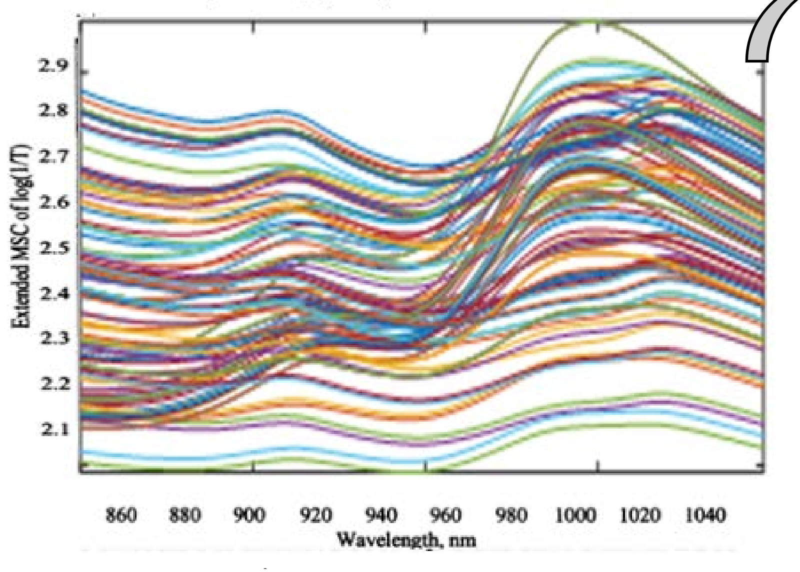
Open

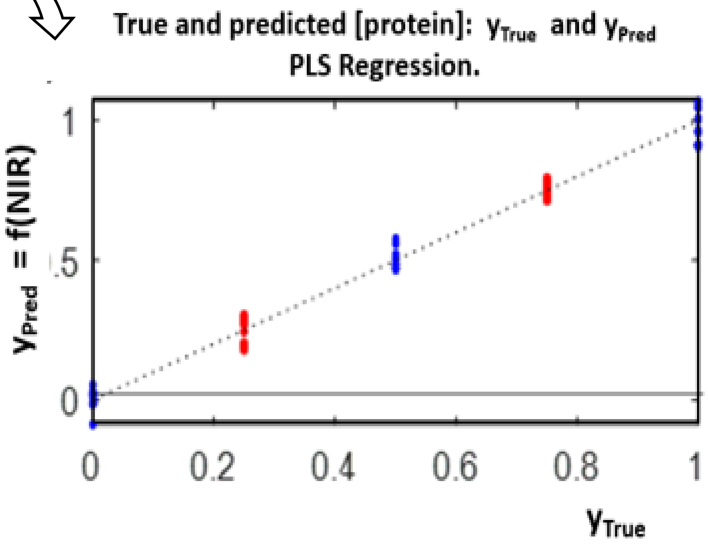


A: Input log(1/T)
Extended MSC of log(1/T)
2.9
2.8
2.7
2.6
2.5
2.4
2.3
2.2
2.1
860
880
900
920
940
960
980
1000
1020
1040
Wavelength, nm
True and predicted [protein]: y_{True} and y_{Pred}
PLS Regression.
y_{Pred} = f(NIR)
1
.5
0
0
0.2
0.4
0.6
0.8
1
y_{True}

Open

A: Input log(1/T)





EMSC: Extended Multiplicative Signal Correction

Simple linear model, using high-school algebra:

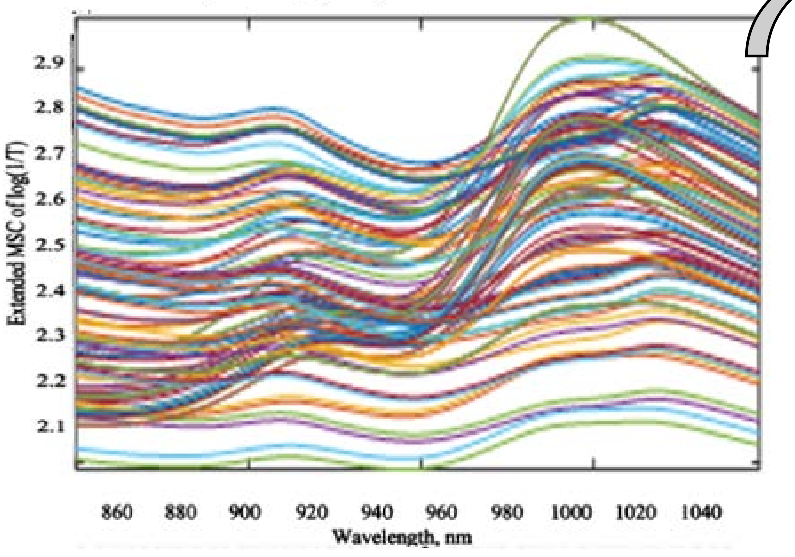
X=NIR absorbance log(1/T) or log(1/R)

$X \approx A \times B + C$: Find B and C, then $X_{corrected}=(X-C)/B$

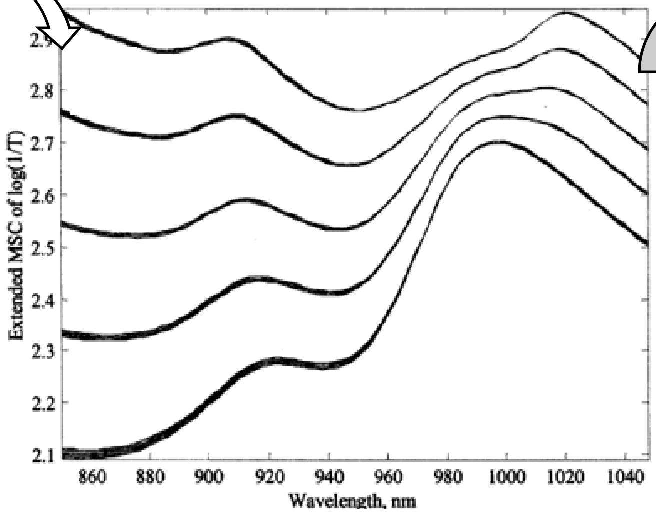
A=Spectral knowledge

Open

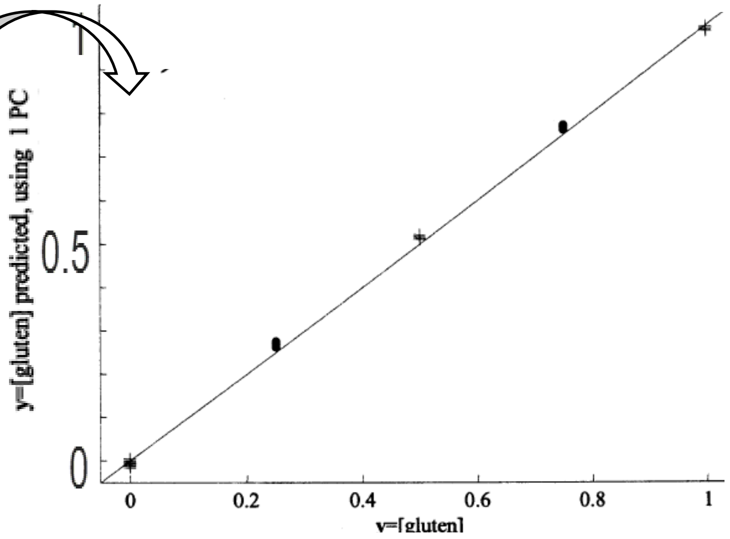
A: Input log(1/T)



B: EMSC w/ **known** constituent spectra



C: EMSC w/ **known** constituent spectra



EMSC: Extended Multiplicative Signal Correction

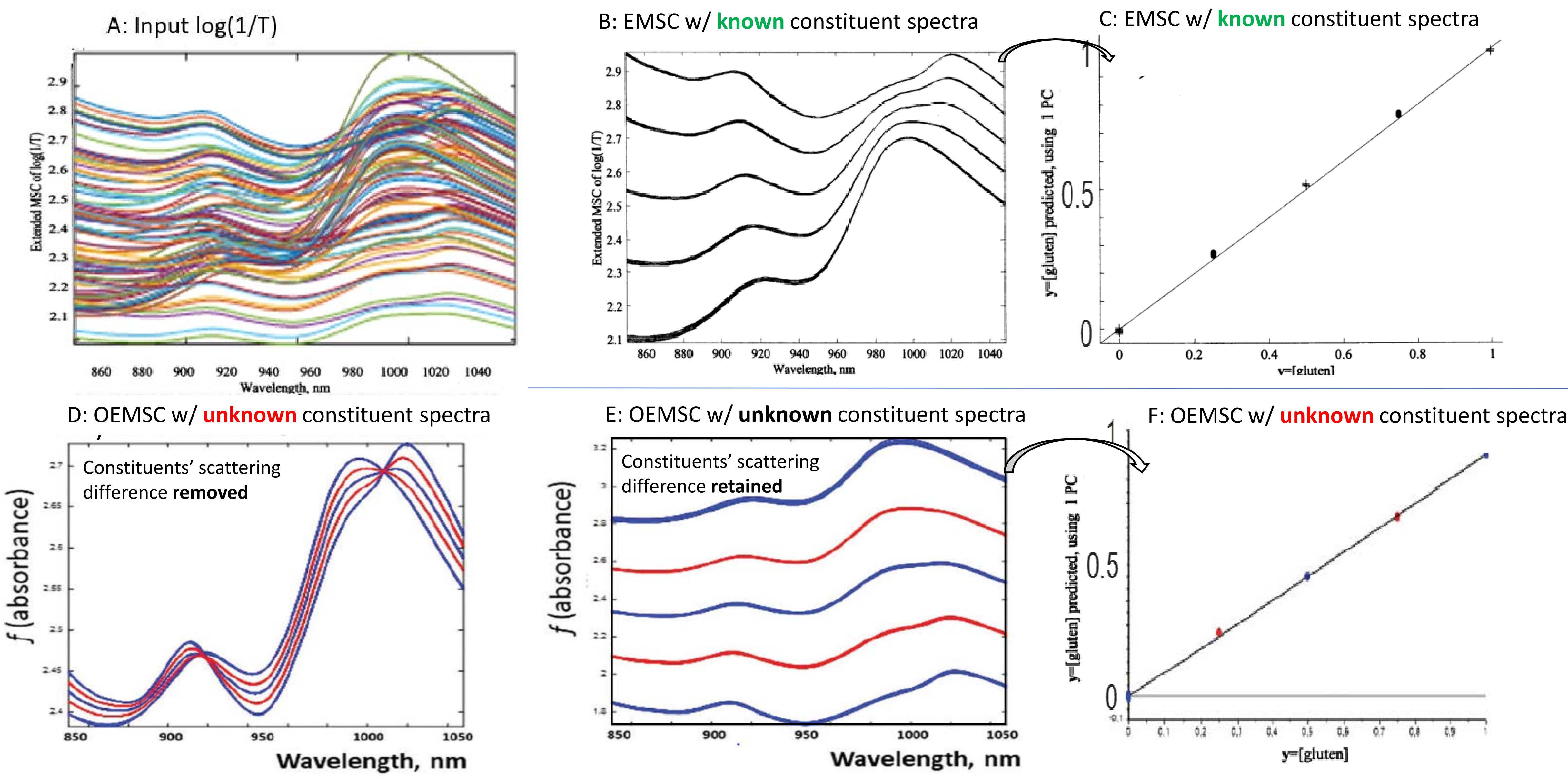
Simple linear model, using high-school algebra:

X=NIR absorbance log(1/T) or log(1/R)

$X \approx A \times B + C$: Find B and C, then $X_{corrected}=(X-C)/B$

A=Spectral knowledge

Open

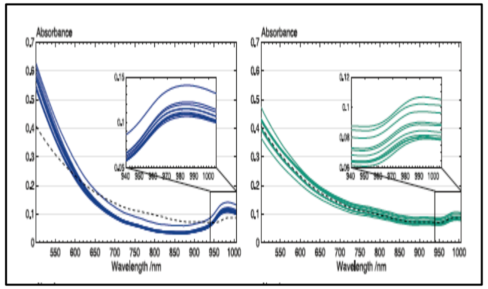


A: Input log(1/T)
Extended MSC of log(1/T)
Wavelength, nm
B: EMSC w/ known constituent spectra
Extended MSC of log(1/T)
Wavelength, nm
C: EMSC w/ known constituent spectra
y=[gluten] predicted, using 1 PC
y=[gluten]
D: OEMSC w/ unknown constituent spectra
Constituents' scattering difference removed
f (absorbance)
Wavelength, nm
E: OEMSC w/ unknown constituent spectra
Constituents' scattering difference retained
f (absorbance)
Wavelength, nm
F: OEMSC w/ unknown constituent spectra
y=[gluten] predicted, using 1 PC
y=[gluten]

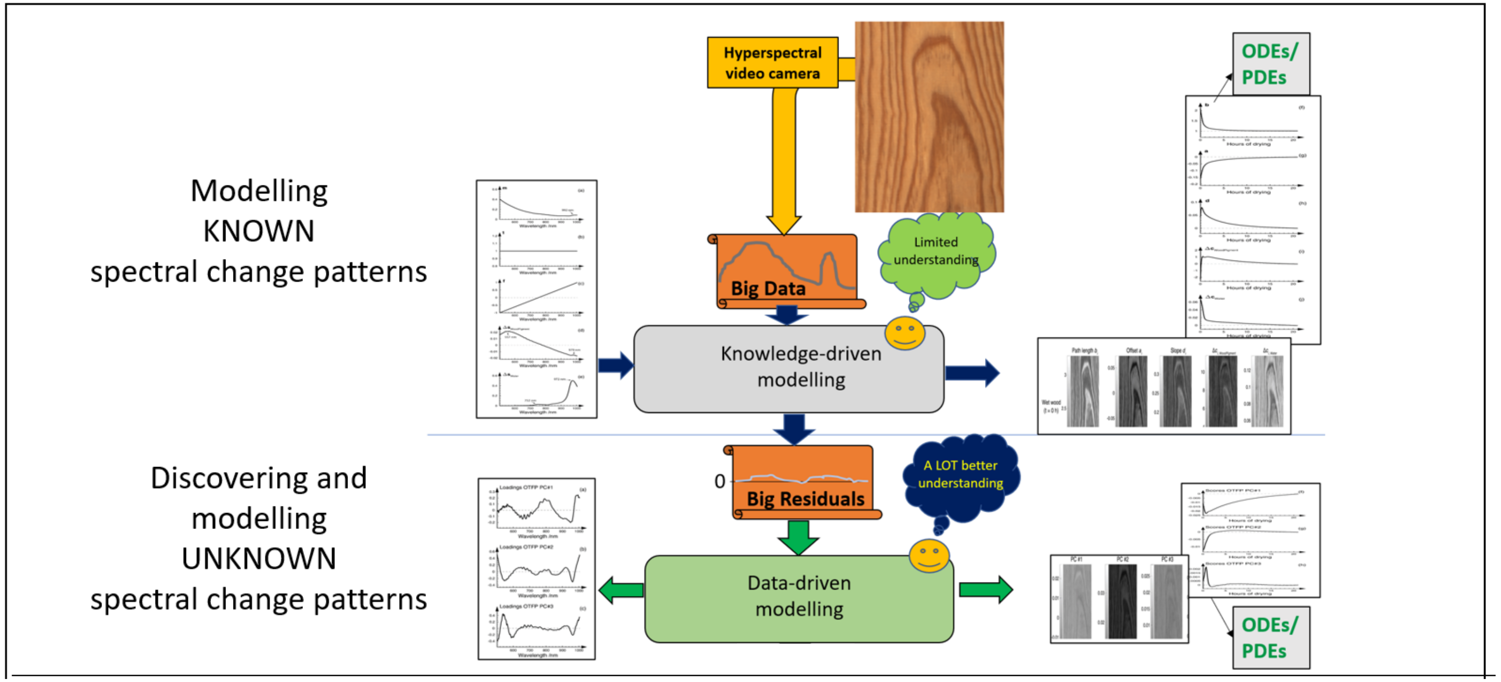
Big Data: Hyperspectral «video»

A single piece of drying wood:
>350 000 000 VNIR reflectance spectra,
≈200 channels each, measured at 150 consecutive times

VNIR;
400-1000 nm



EMSC modelling KNOWN and UNKNOWN physics & chemistry:

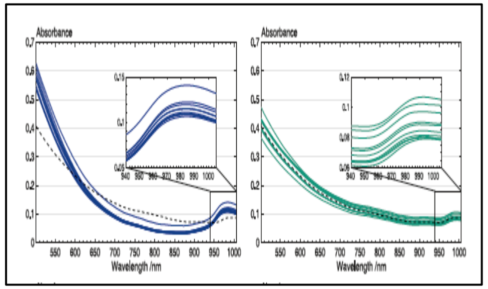


Open

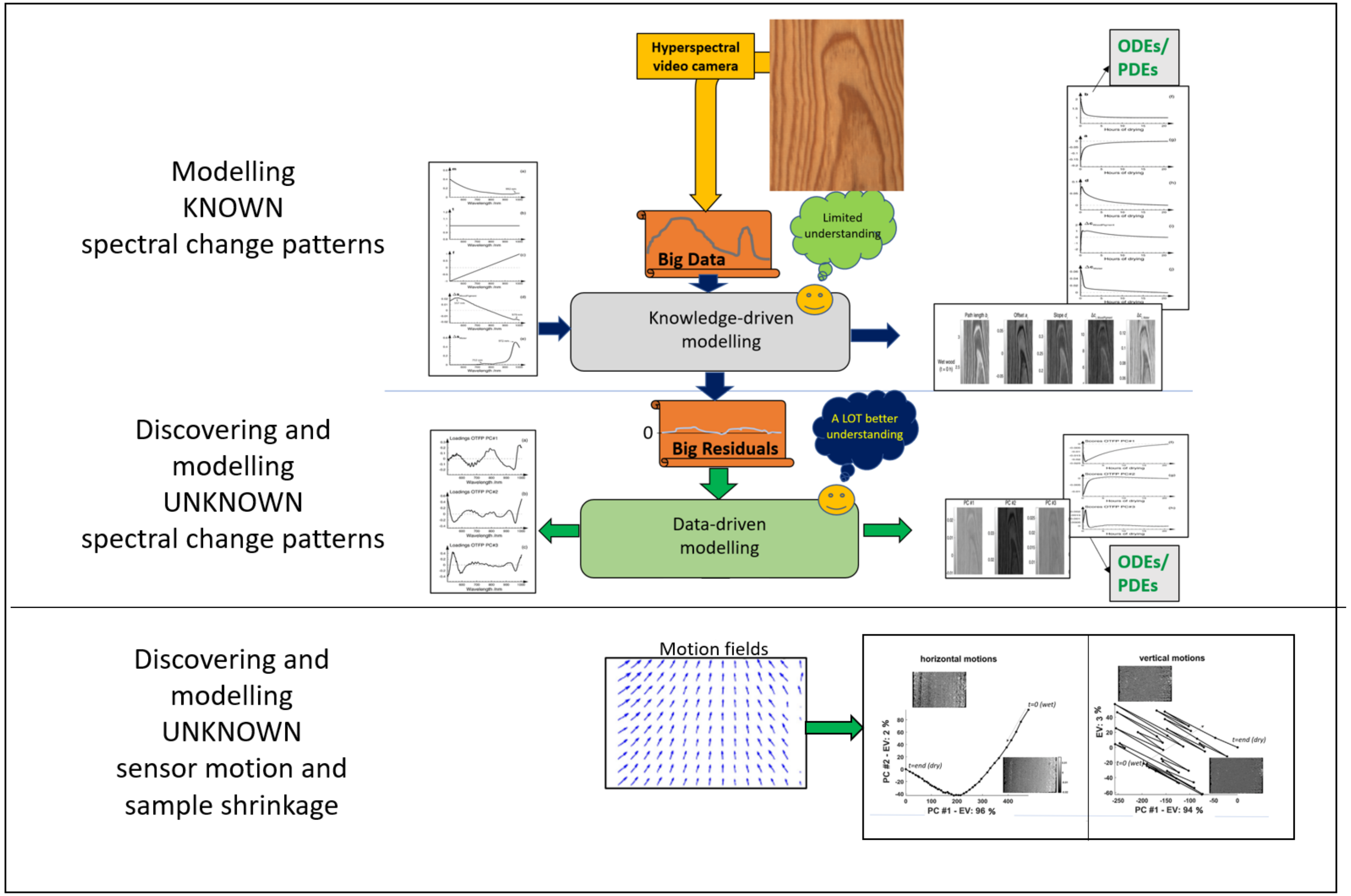
Big Data: Hyperspectral «video»

A single piece of drying wood:
>350 000 000 VNIR reflectance spectra,
≈200 channels each, measured at 150 consecutive times

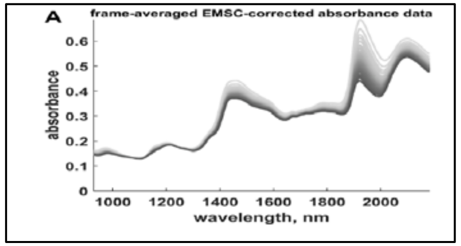
VNIR:
400-1000 nm



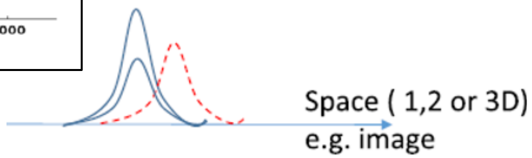
EMSC modelling KNOWN and UNKNOWN physics & chemistry:



SWIR:
900-2500 nm

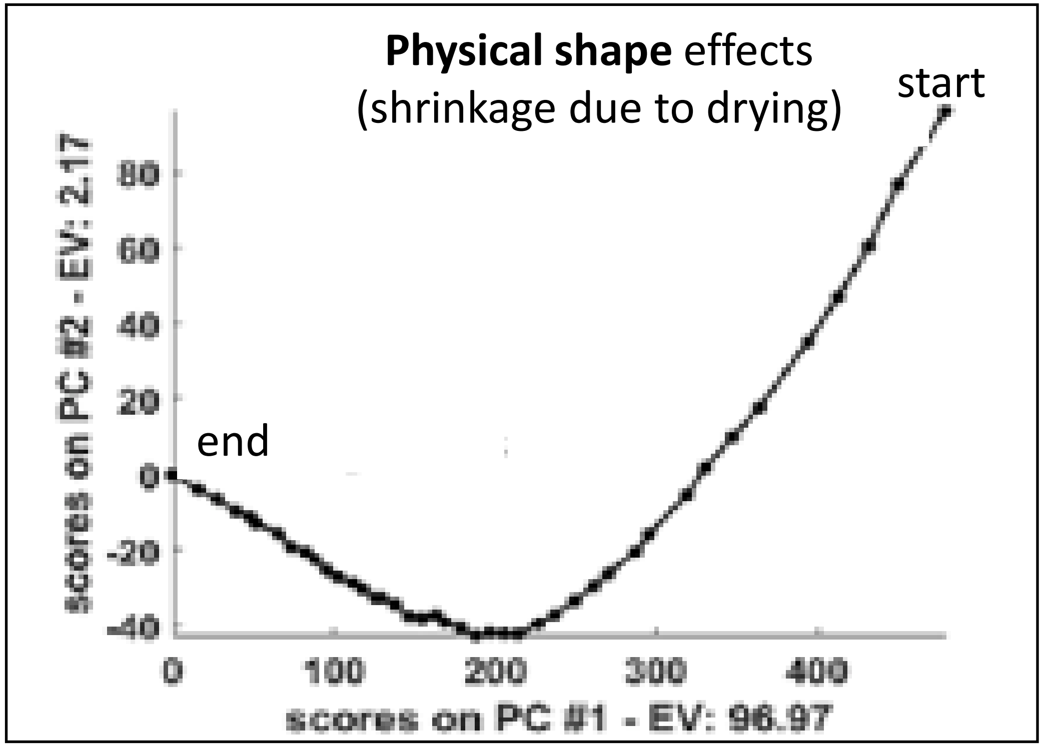


Two-domain IDLE modelling:

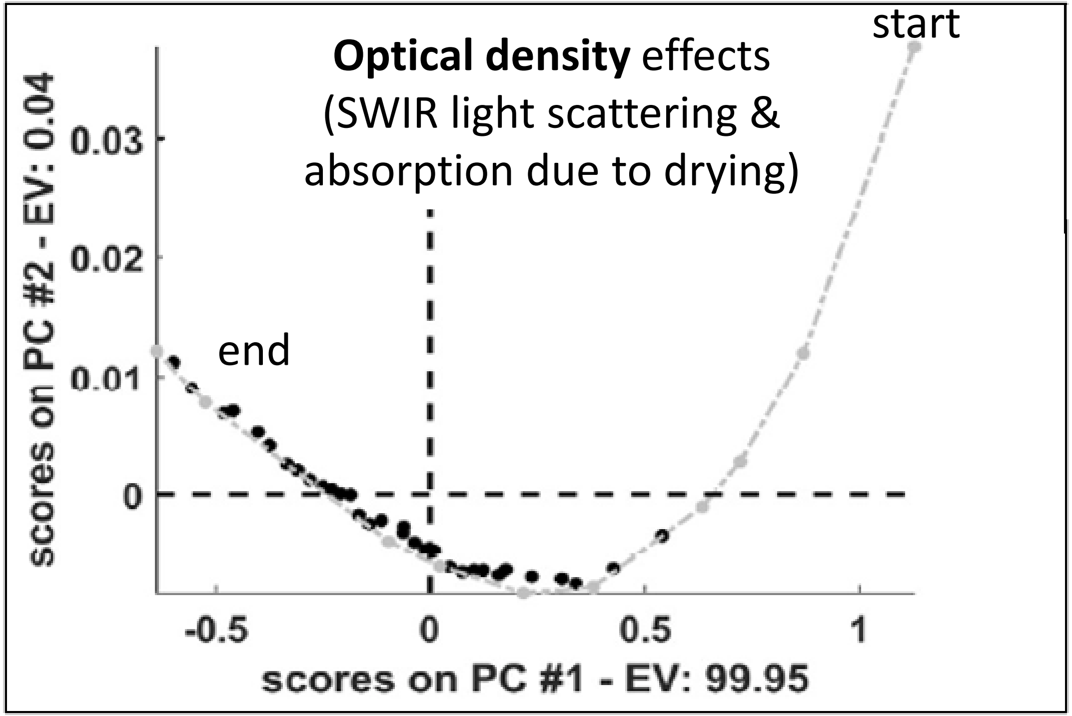


Drying wood in SWIR (900-2500 nm)

Similar two phase-kinetics for physical shrinkage and chemical composition change



≈

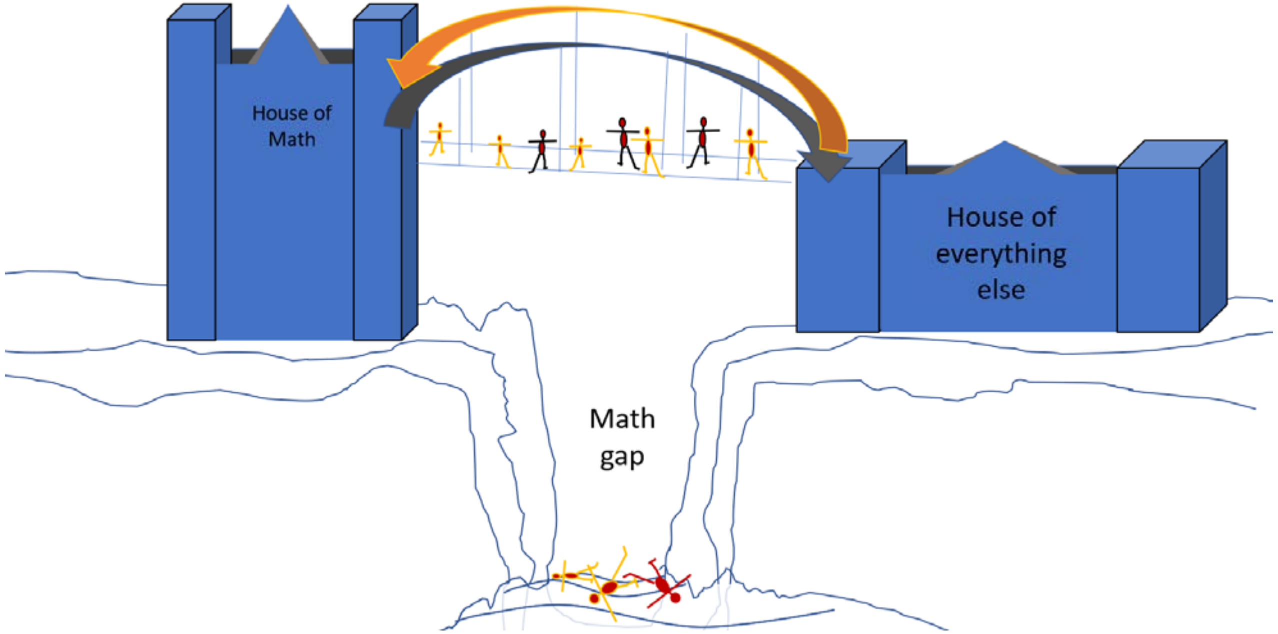


Open

House of Math

House of everything else

Math gap



NTNU
Norwegian University of
Science and Technology

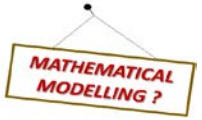
idletechs

Thank you!

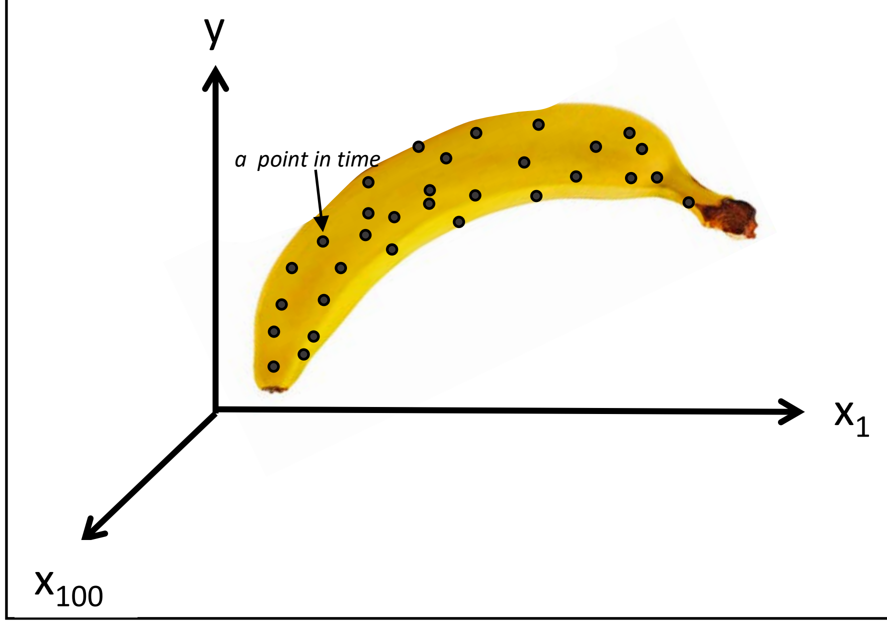
harald.martens@ntnu.no

Open

PLSR etc uses a linear method, but can often handle non-linear responses automatically

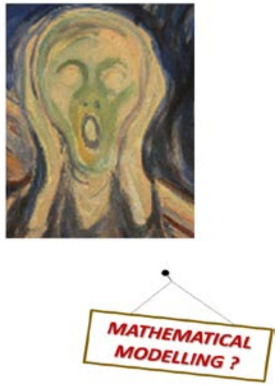


Many data points in a high-dimensional space
e.g. 100 wavelengths (y, x_1, x_2,… , x_{100}),
happen to form a banana-shaped cloud :

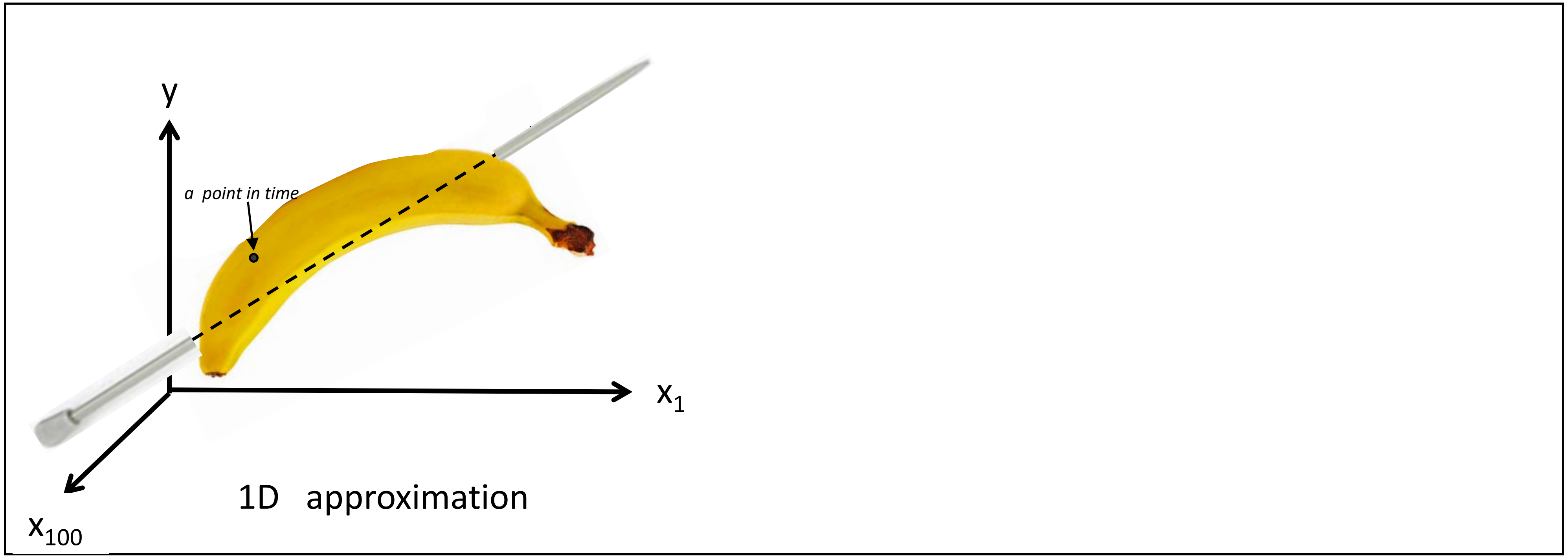


Open

PLSR etc uses a linear method, but can often handle non-linear responses automatically

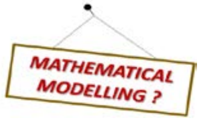


Many data points in a high-dimensional space
e.g. 100 wavelengths (y, x_1, x_2,… , x_{100}),
happen to form a banana-shaped cloud :

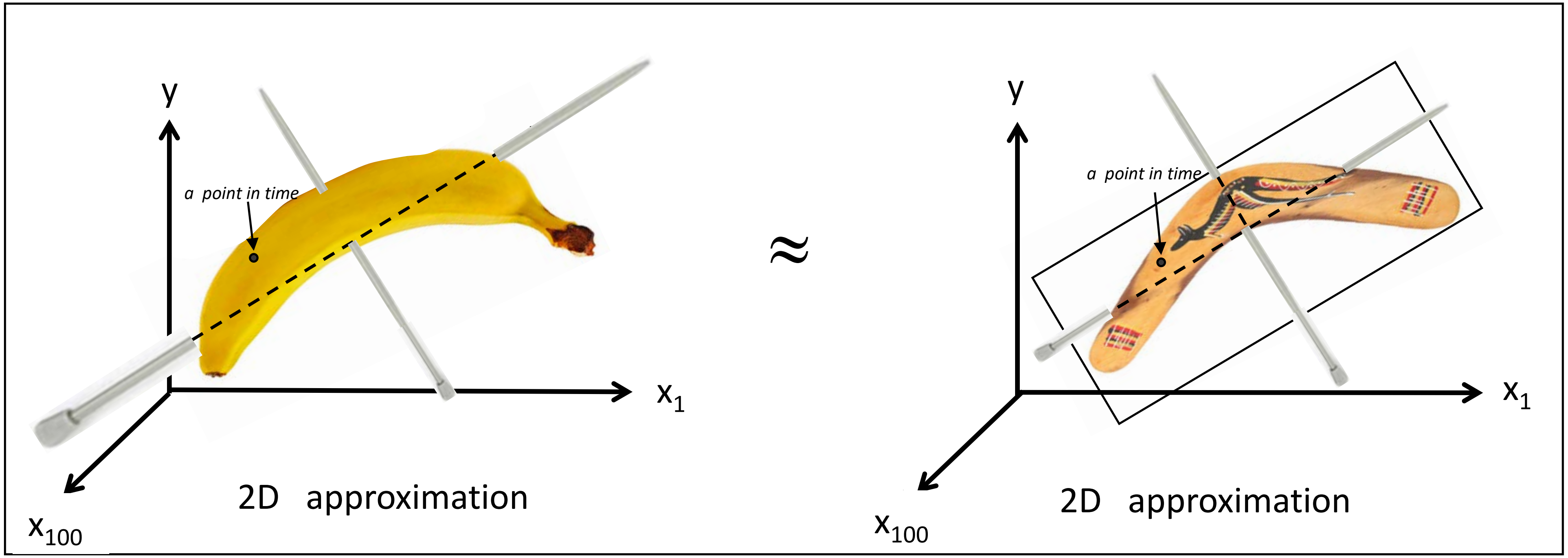


Open

PLSR etc uses a linear method, but can often handle non-linear responses automatically

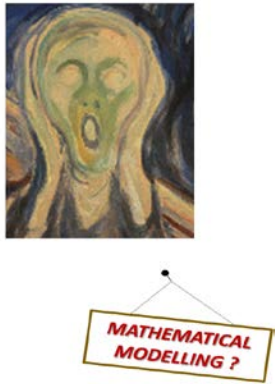


Many data points in a high-dimensional space
e.g. 100 wavelengths (y, x_1, x_2,… , x_{100}),
happen to form a banana-shaped cloud :

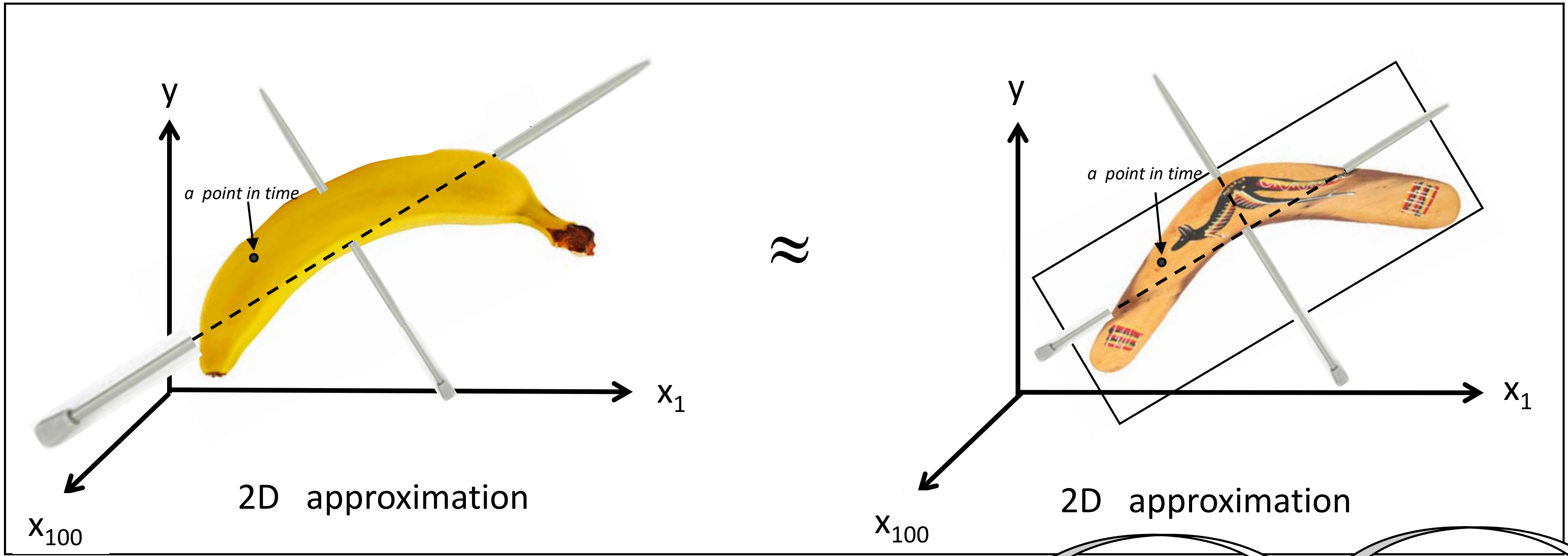


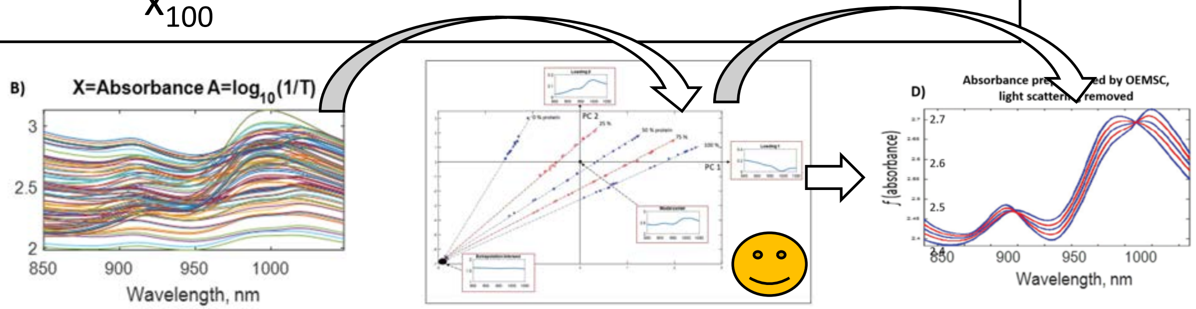
Open

PLSR etc uses a linear method, but can often handle non-linear responses automatically



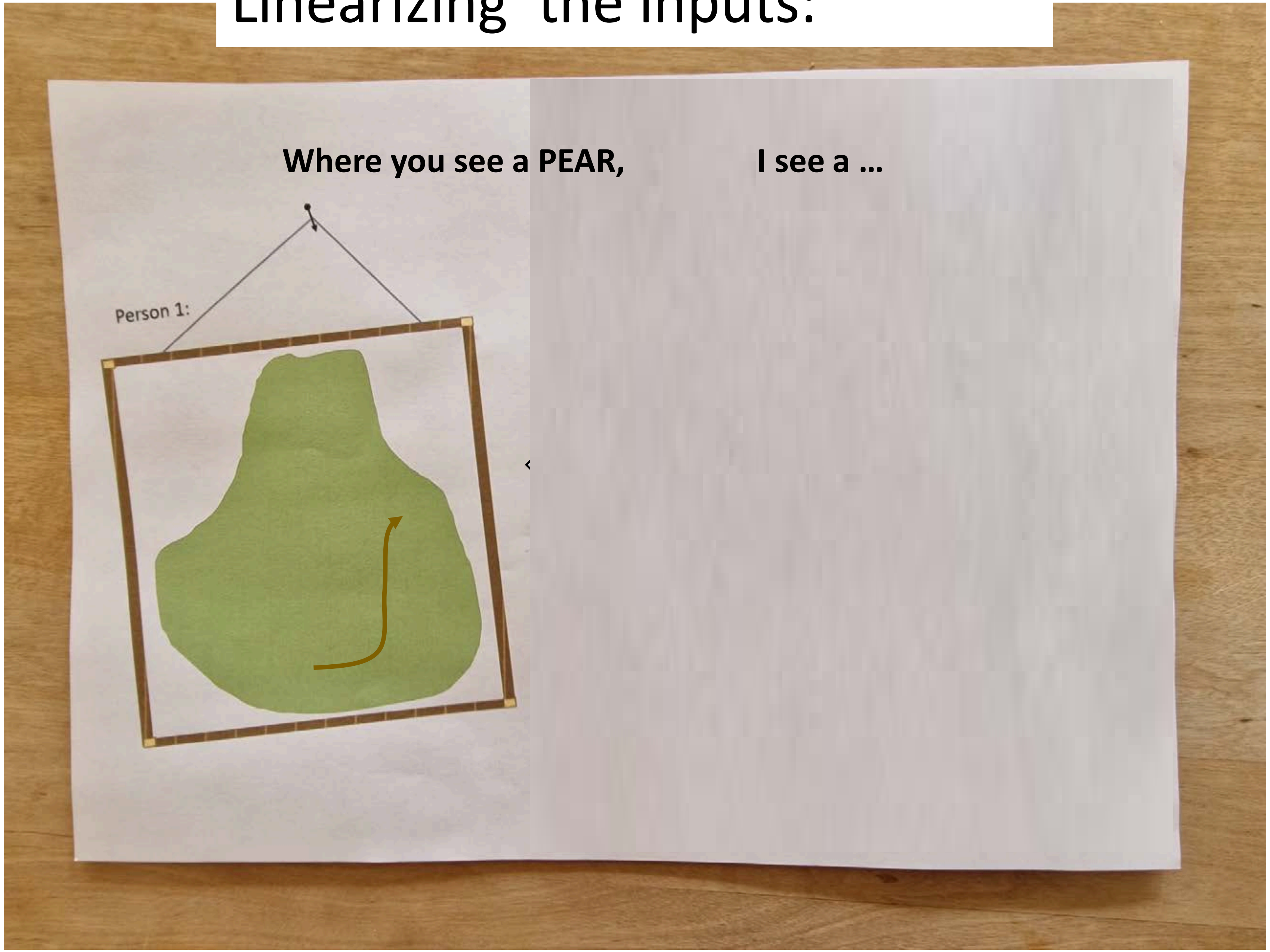
Many data points in a high-dimensional space
e.g. 100 wavelengths ($y, x_1, x_2, \ldots, x_{100}$),
happen to form a banana-shaped cloud :





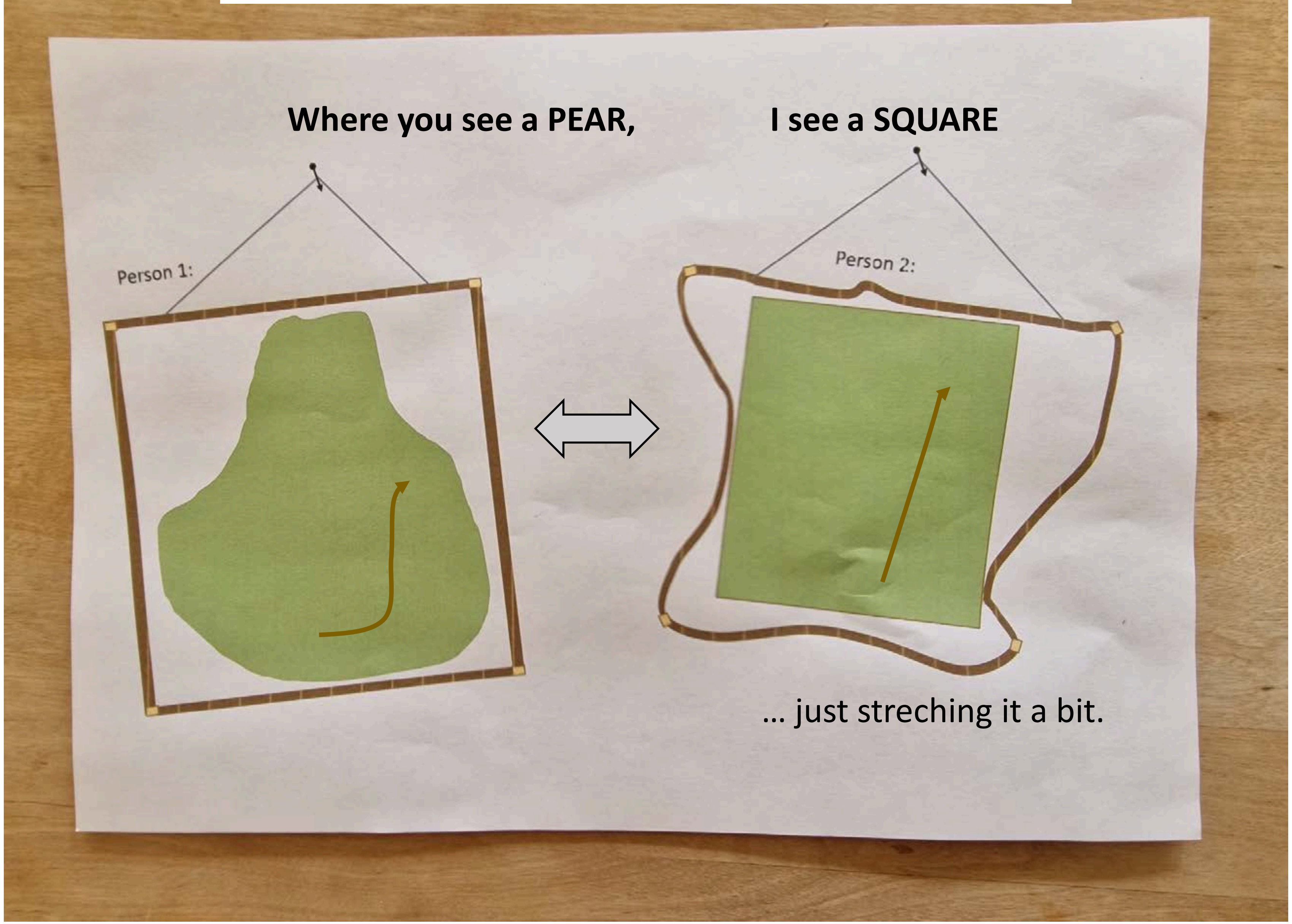
Open

Linearizing the inputs:



Open

Linearizing the inputs:



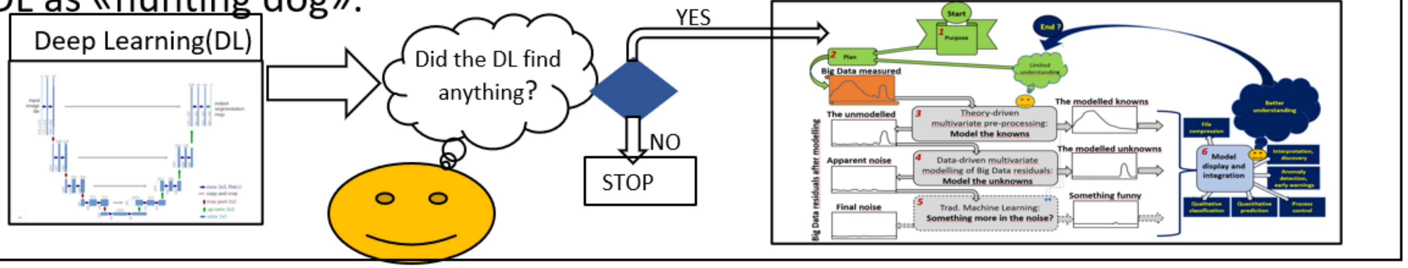
Open

Open

Open

Ways to combine Deep Learning (DL) and Hybrid Chemometrics (HC)

1. DL as «hunting dog»:



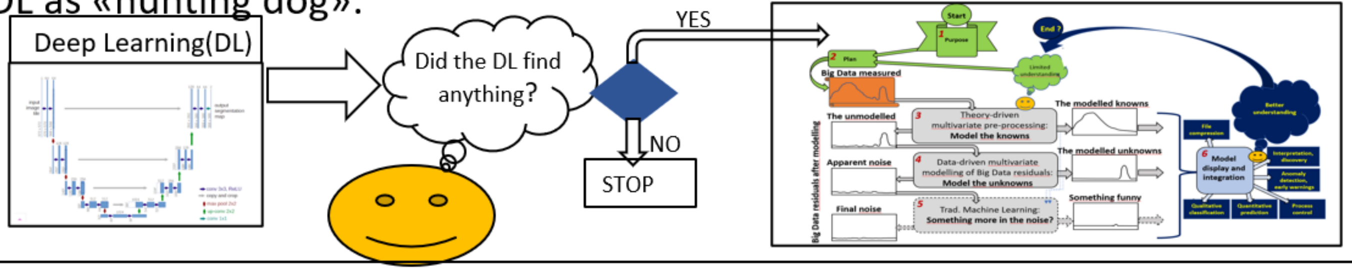
If DL did not find anyting, then find a better project !

Open

Open

Ways to combine Deep Learning (DL) and Hybrid Chemometrics (HC)

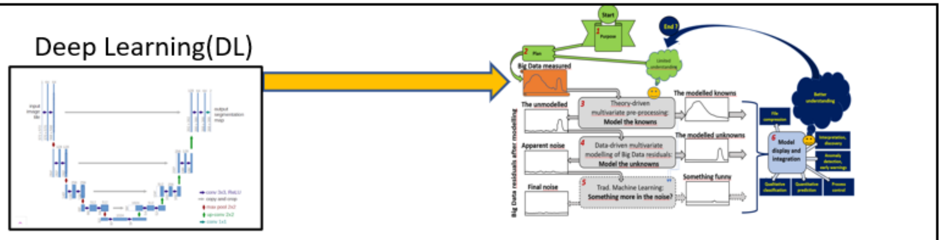
1. DL as «hunting dog»:



If DL did not find anyting, then find a better project !

Find patterns in the hidden nodes of DL

2. DL «hidden node» scores as data input:

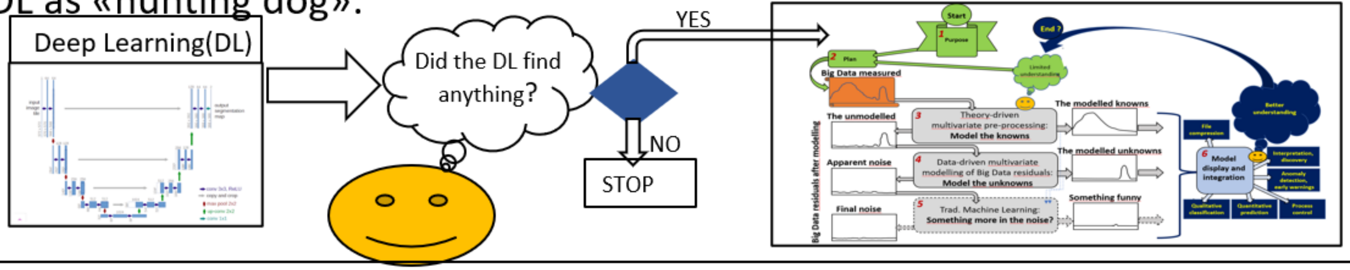


Open

Open

Ways to combine Deep Learning (DL) and Hybrid Chemometrics (HC)

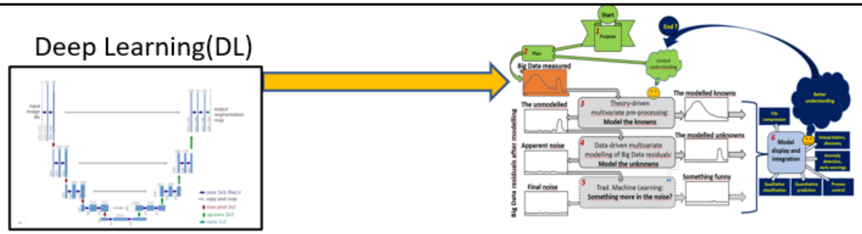
1. DL as «hunting dog»:



If DL did not find anyting, then find a better project !

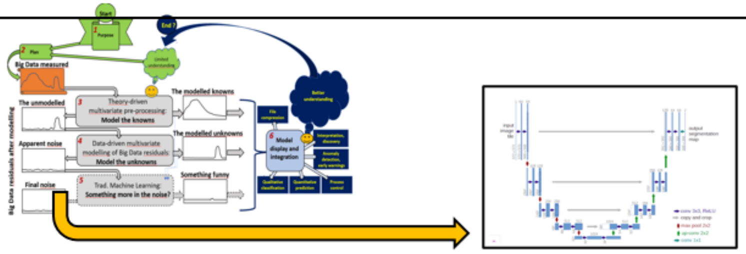
Find patterns in the hidden nodes of DL

2. DL «hidden node» scores as data input:



Look for «funny» patterns in PCA residuals

3. DL for clean-up:
Look for structures in residuals after PCA

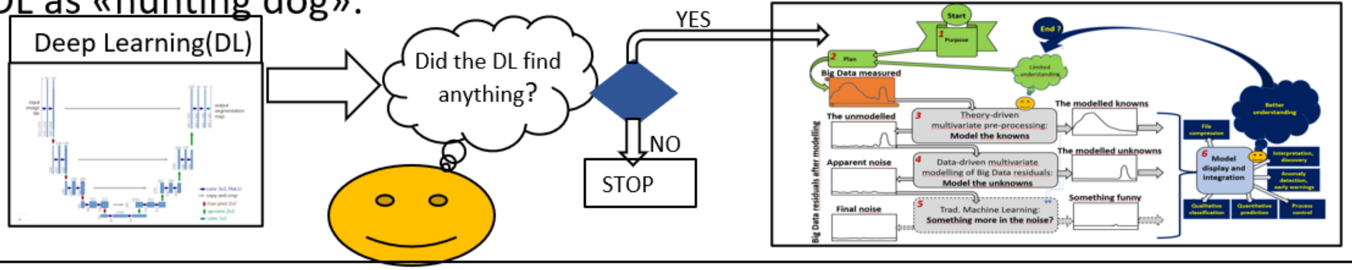


Open

Open

Ways to combine Deep Learning (DL) and Hybrid Chemometrics (HC)

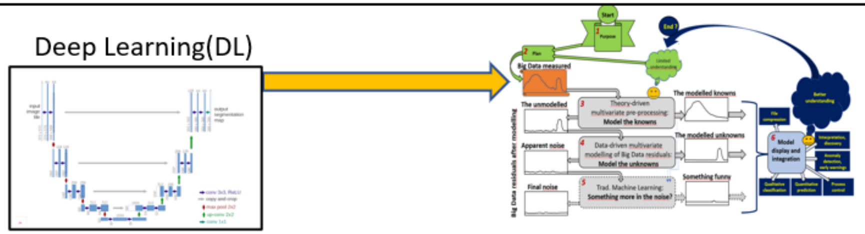
1. DL as «hunting dog»:



If DL did not find anyting, then find a better project !

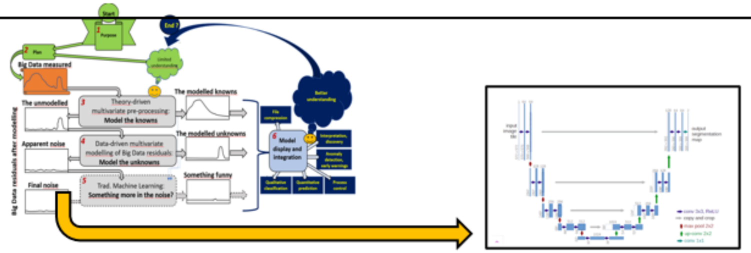
Find patterns in the hidden nodes of DL

2. DL «hidden node» scores as data input:



Look for «funny» patterns in PCA residuals

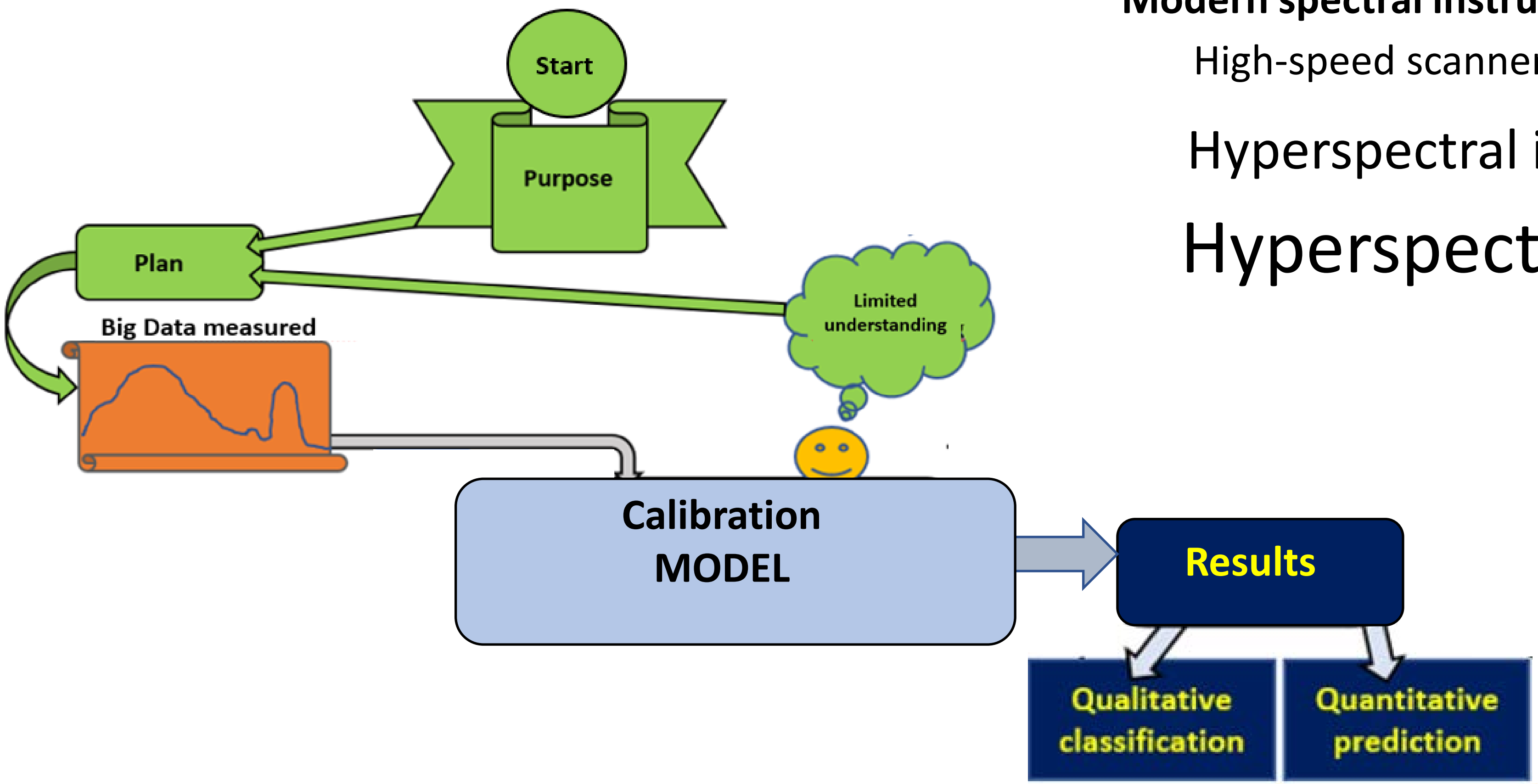
3. DL for clean-up:
Look for structures in residuals after PCA



Look for Higher-order complexities in modelled subspace

4. DL as post-processing:
Look for more complex structures in the combined output scores and residual statistics from EMSC & PCA

BIG DATA in Science and Technology (S&T)



Modern spectral instruments generate Big Data:

High-speed scanners

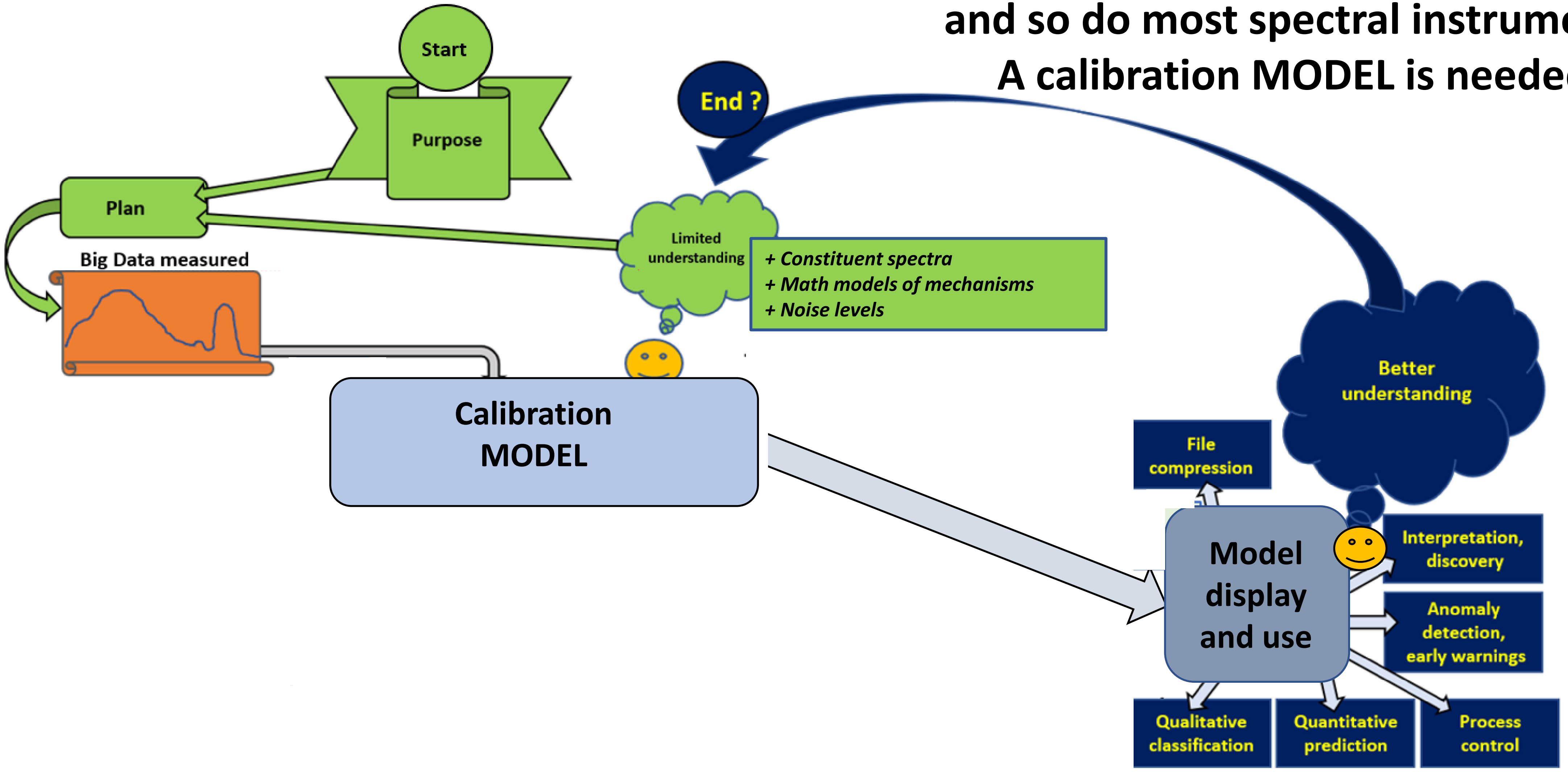
Hyperspectral imaging

Hyperspectral video

Open

Laws of nature, and other common causes

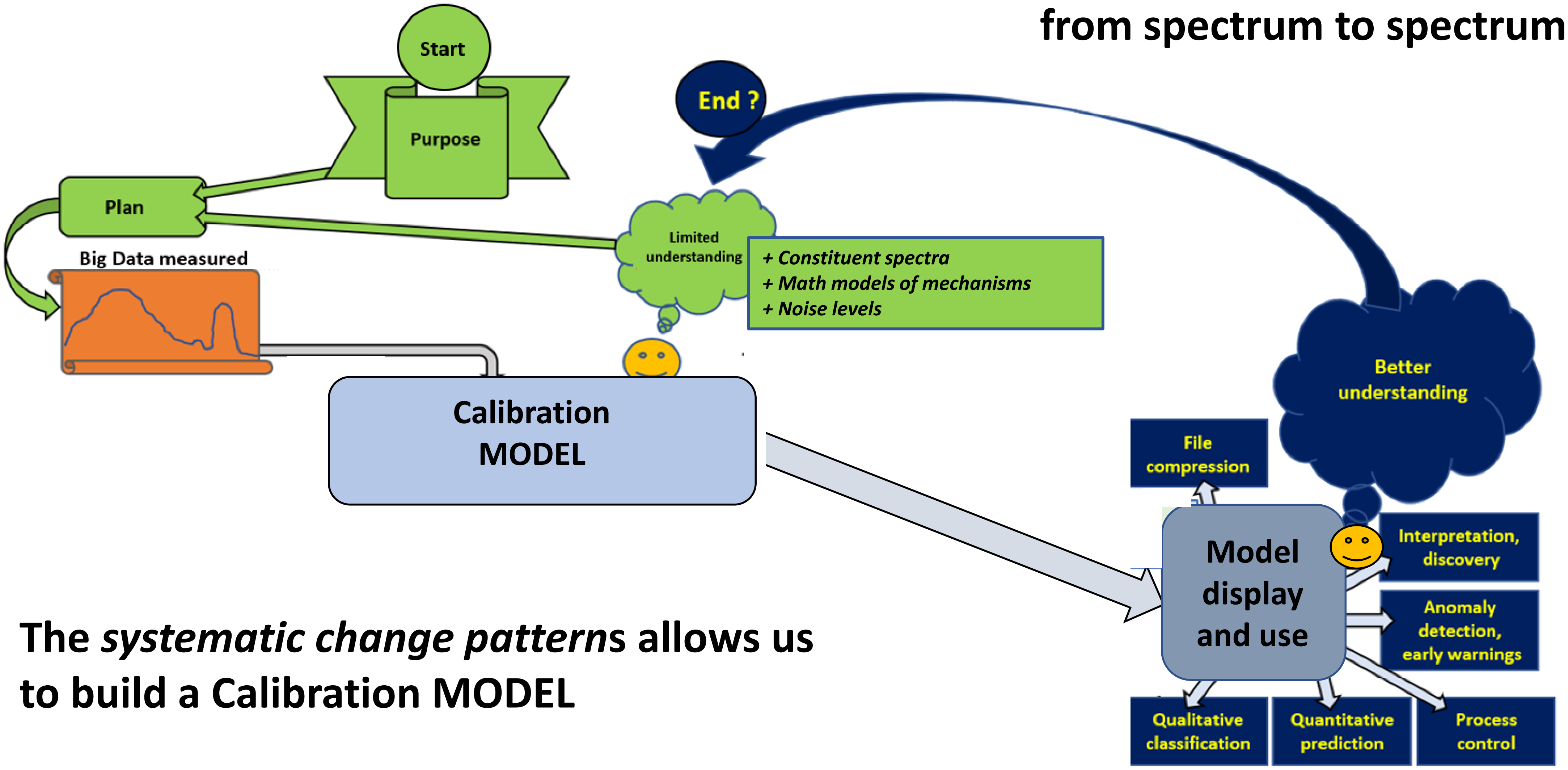
Most technical scenes and samples change in systematic ways according to laws of nature, and so do most spectral instruments. A calibration MODEL is needed.



Open

Spectra from common causes show patterns

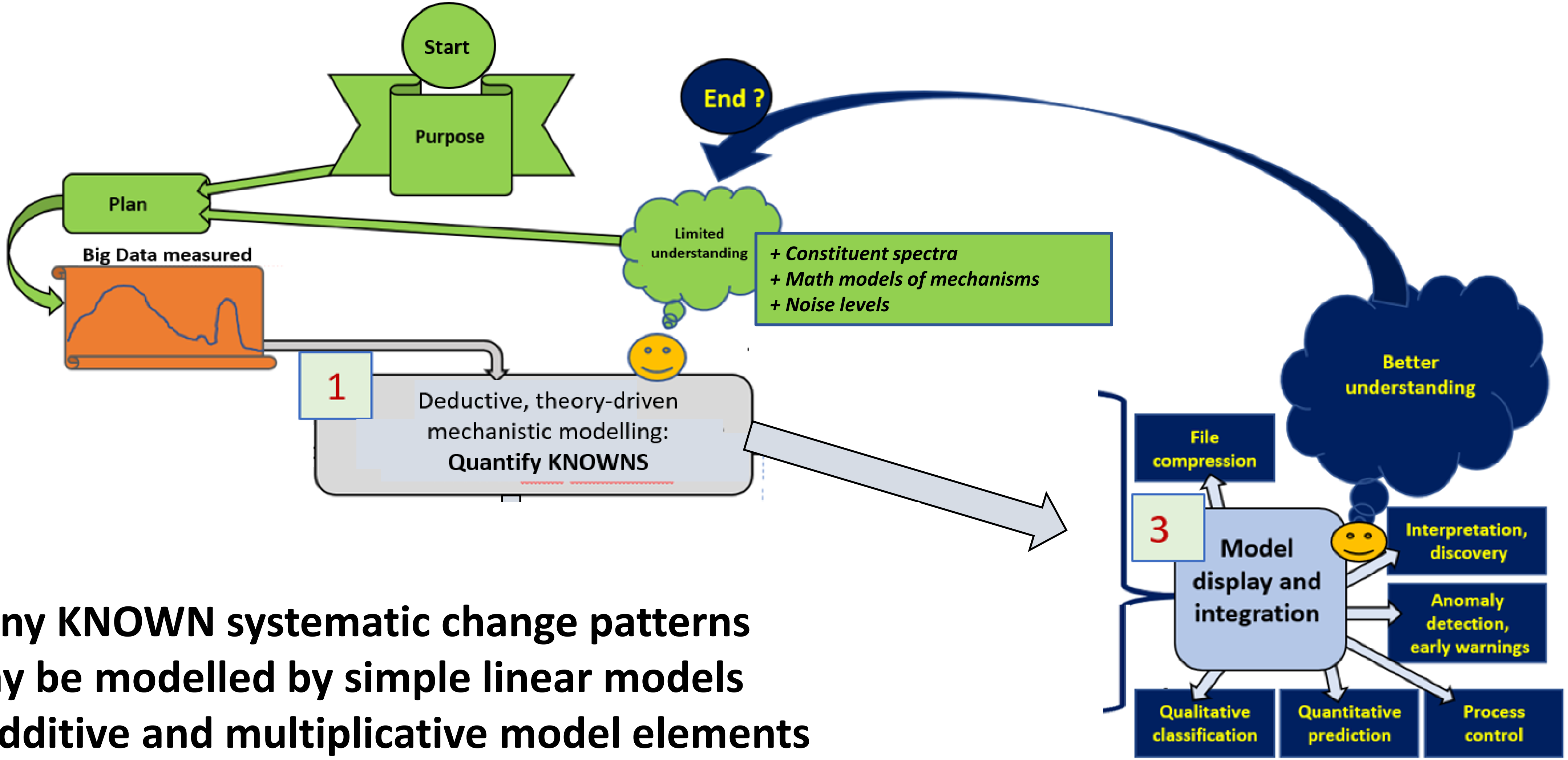
Common causes generate *systematic change patterns* from spectrum to spectrum



The *systematic change patterns* allows us to build a Calibration MODEL

Some causes are expected

Many KNOWN causes give NICE, *systematic change patterns*

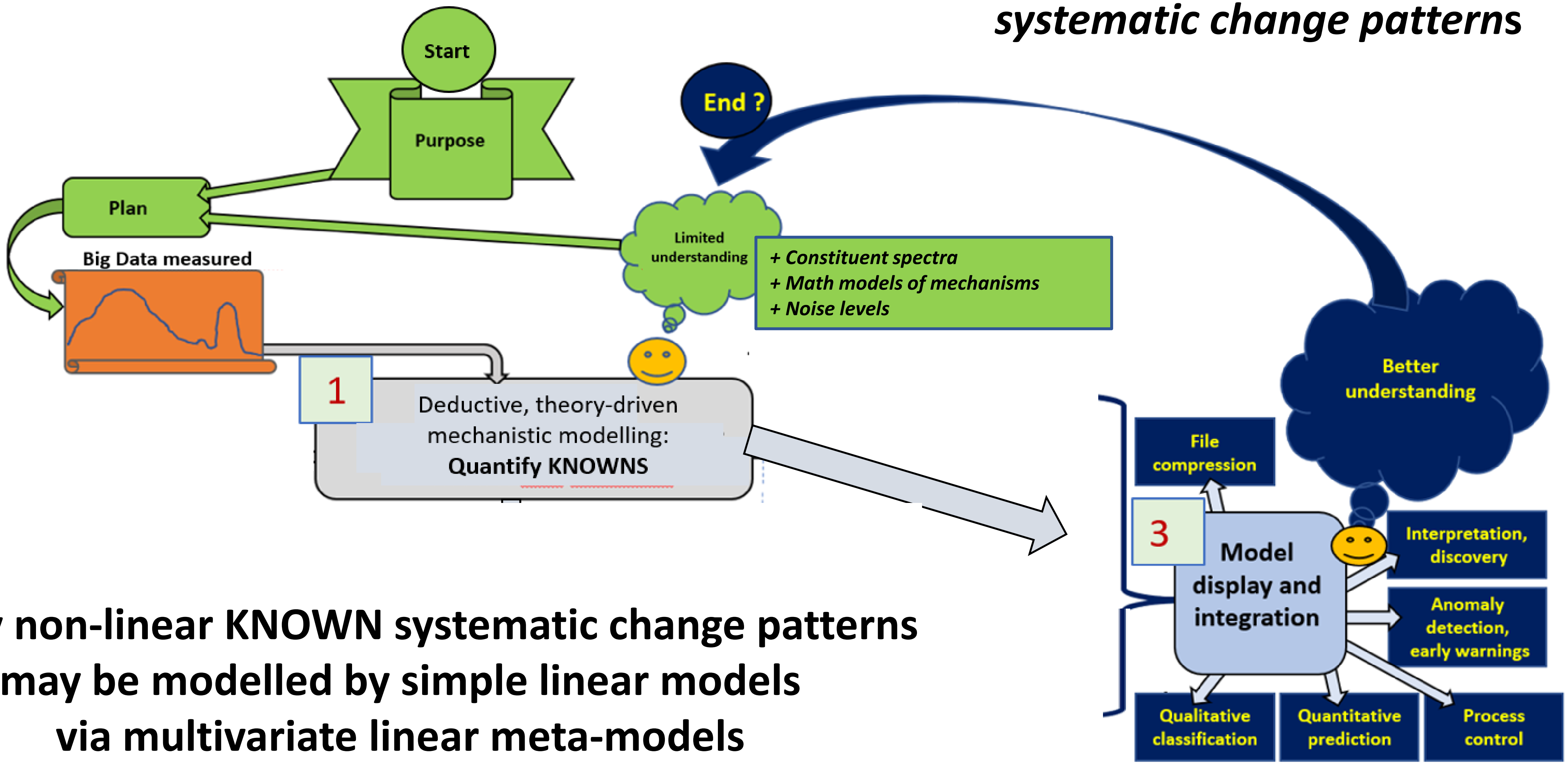


Many KNOWN systematic change patterns may be modelled by simple linear models with additive and multiplicative model elements

Open

Some causes are expected

Some KNOWN causes give NOT SO NICE, but still *systematic change patterns*

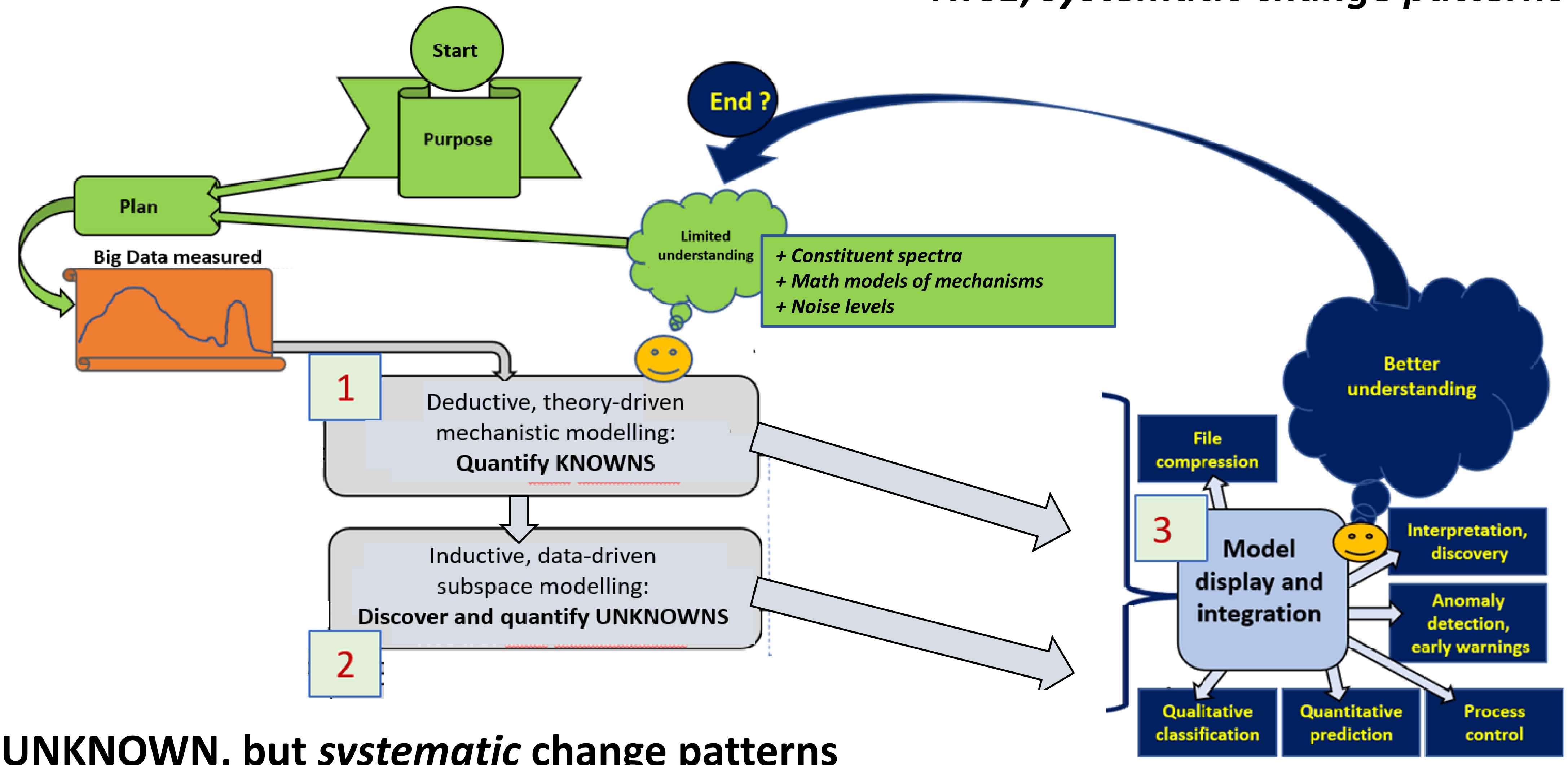


Highly non-linear KNOWN systematic change patterns may be modelled by simple linear models via multivariate linear meta-models based on computer simulation studies

Open

Other causes are unexpected

But UNKNOWN causes can still give NICE, *systematic change patterns*

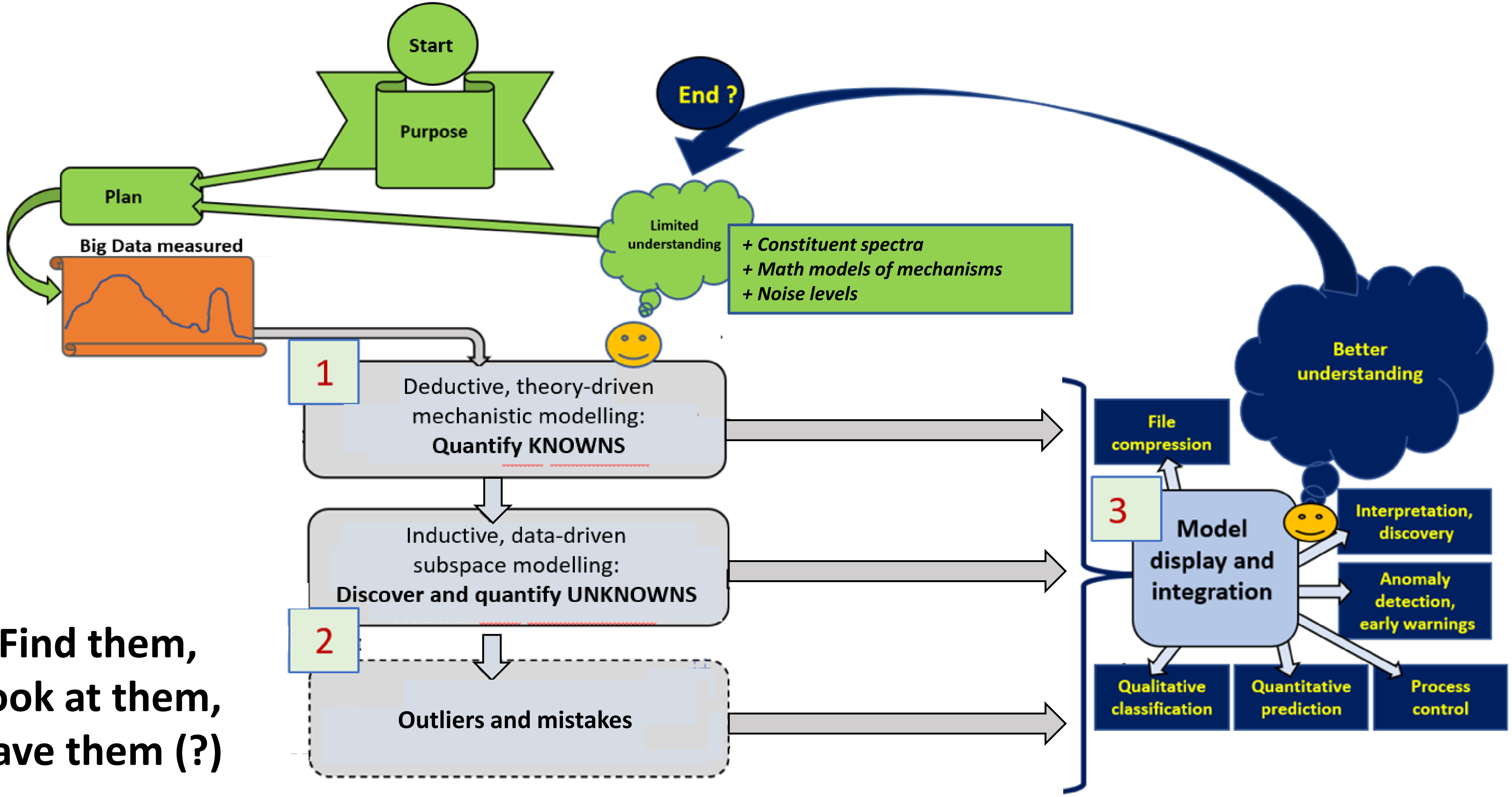


Many UNKNOWN, but *systematic* change patterns may be modelled by purely additive elements

Open

Outliers and irrelevant anomalies

UNEXPECTED OUTLIERS etc
may have to be handled

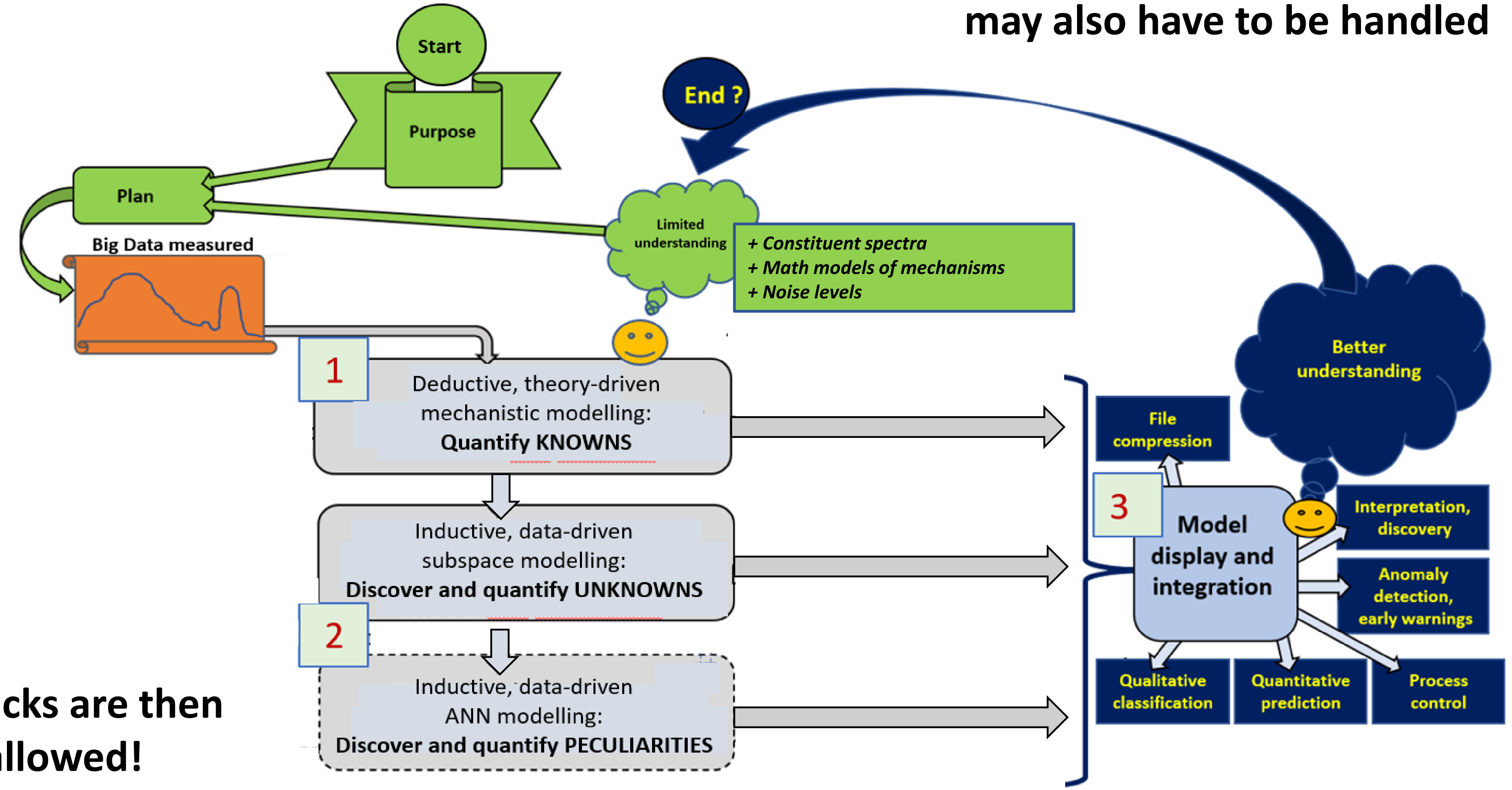


Find them, look at them, save them (?)

Open

Very strange behaviours

UNEXPECTED, non-systematic PECULIARITIES
may also have to be handled

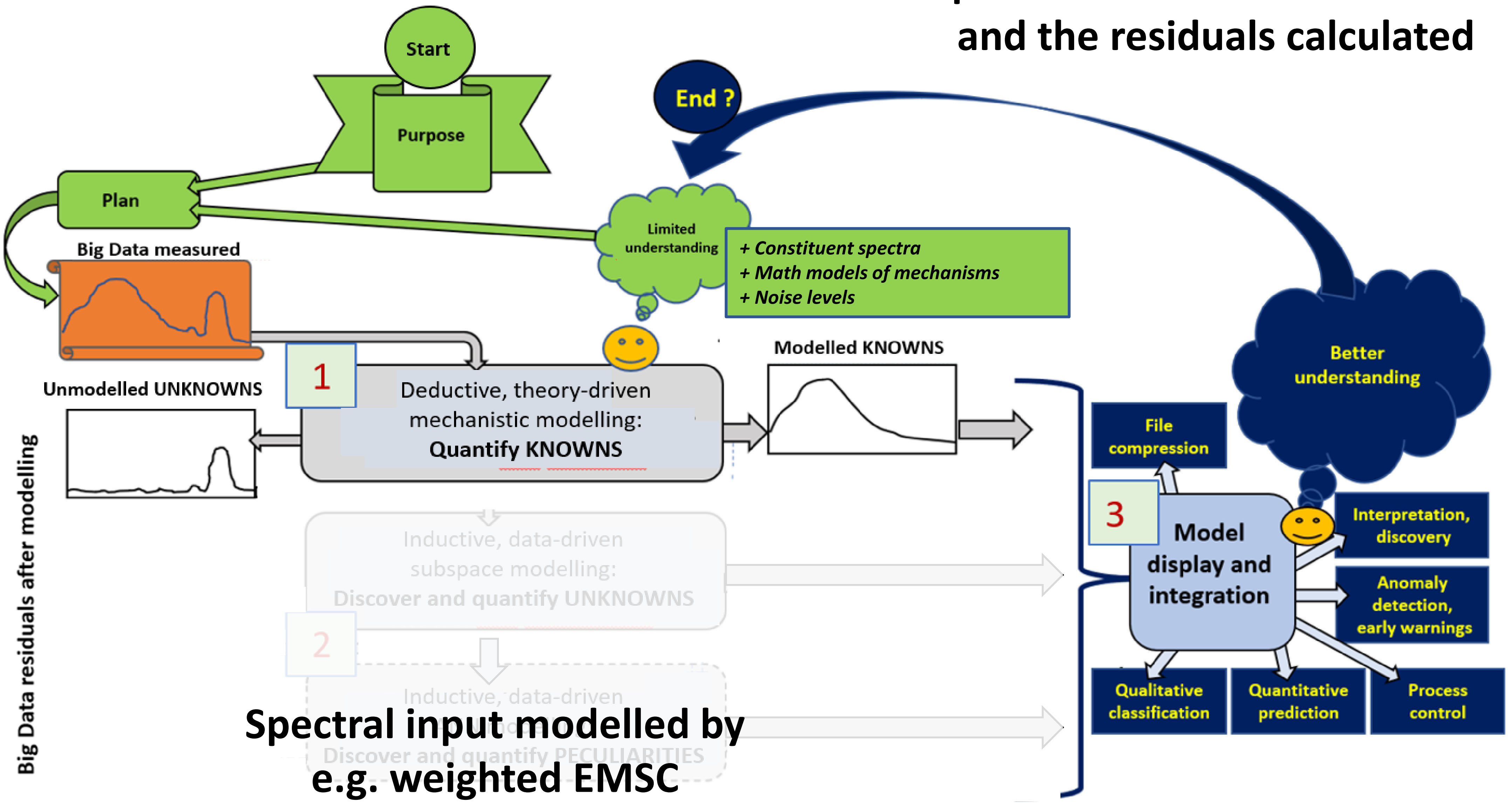


All tricks are then allowed!

Open

Modelling KNOWN patterns

KNOWN
systematic change patterns
are quantified in multivariate model,
and the residuals calculated

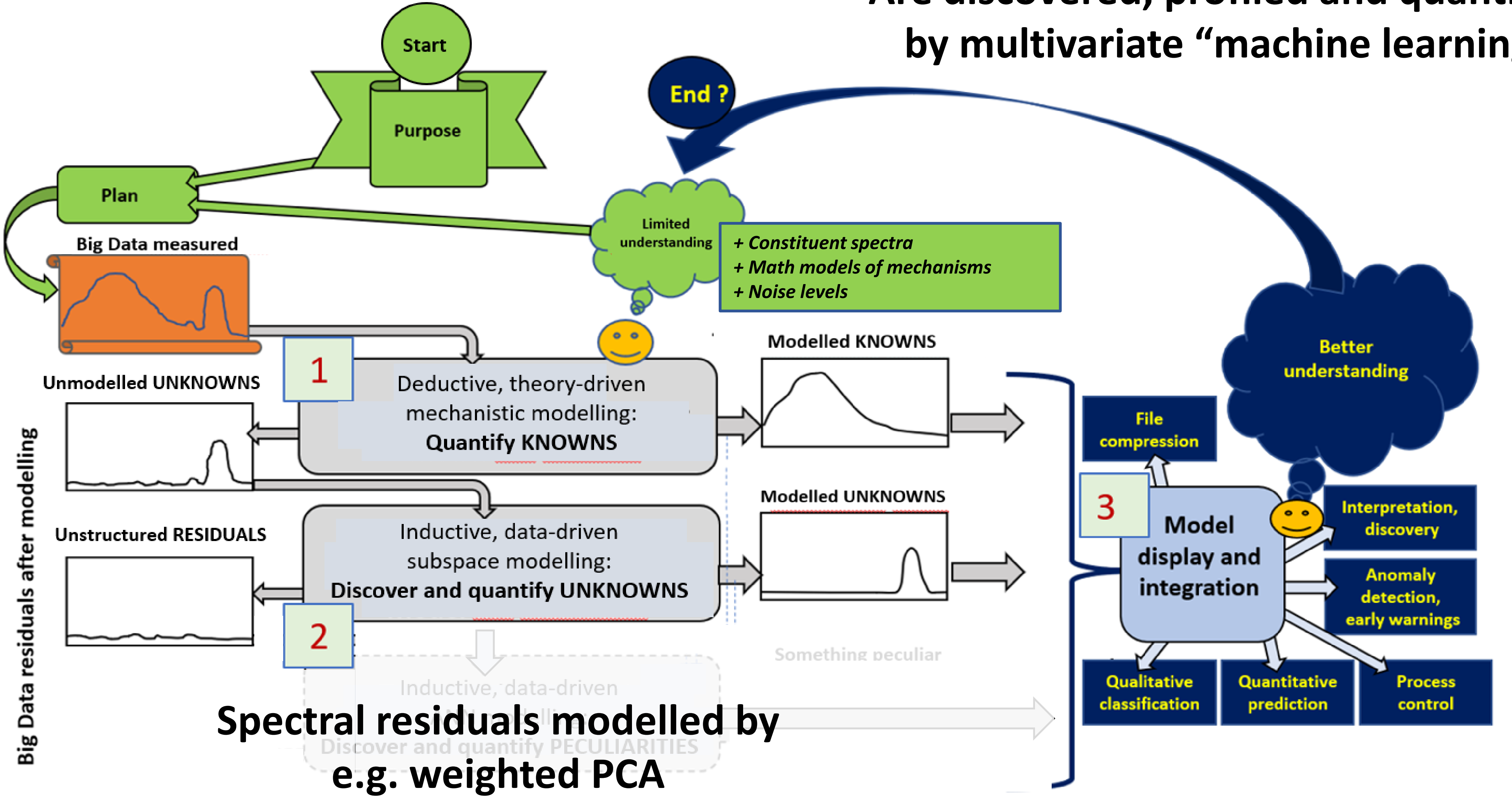


Spectral input modelled by
e.g. weighted EMSC

Open

Modelling UNKNOWN patterns

The UNKNOWN, but *systematic* change patterns Are discovered, profiled and quantified by multivariate "machine learning"



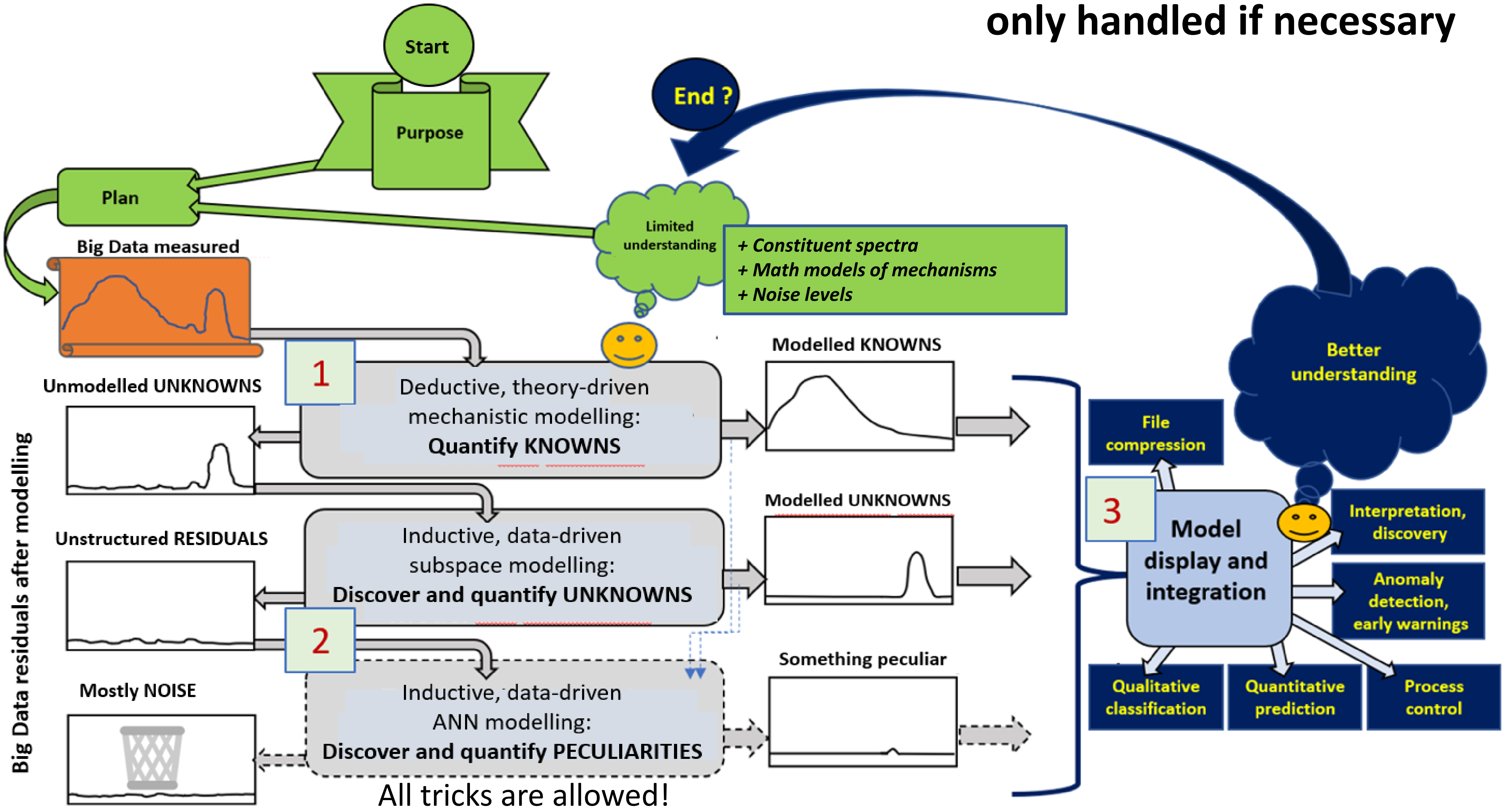
Spectral residuals modelled by e.g. weighted PCA

Open

Modelling UNKNOWN patterns

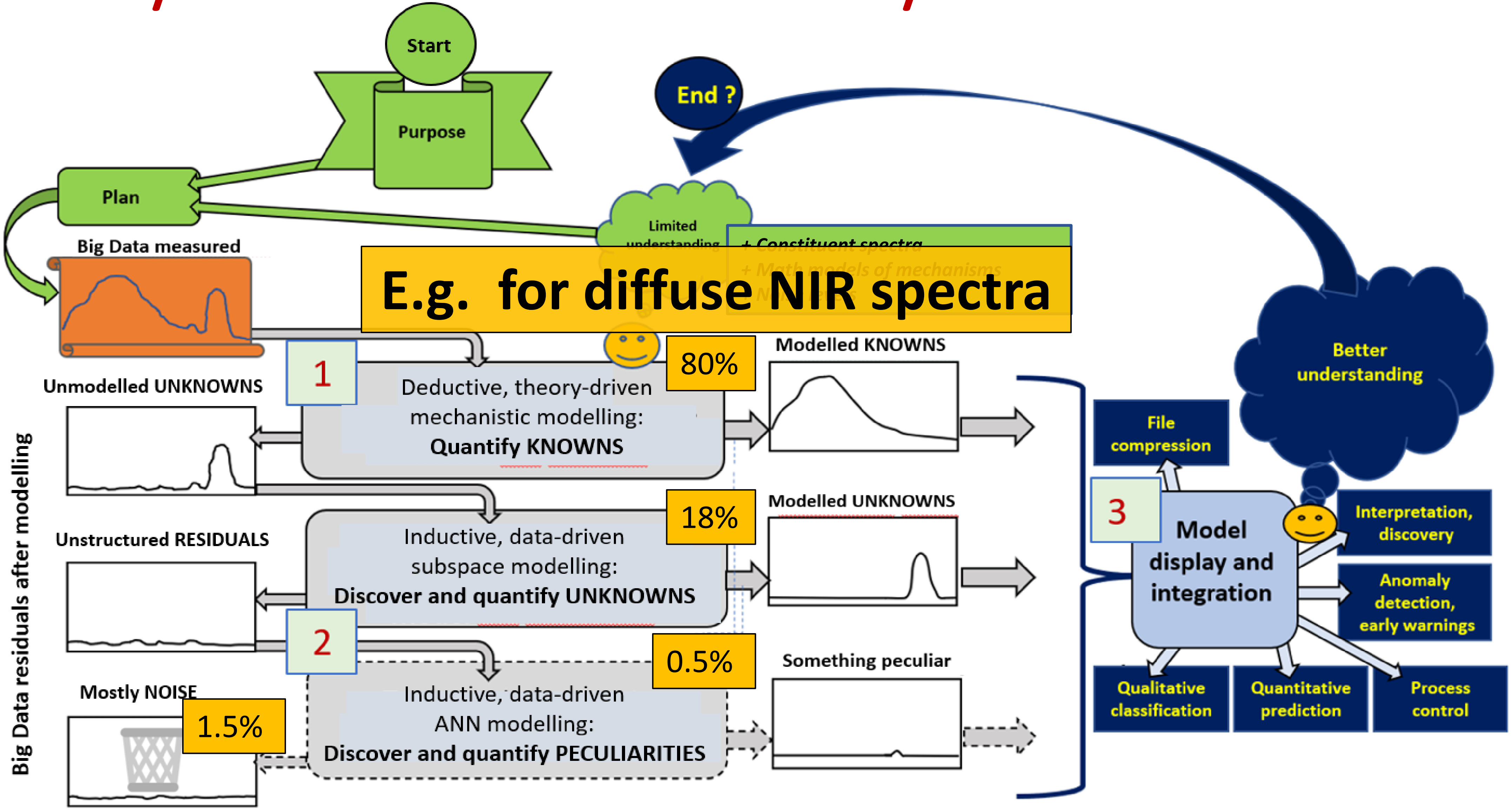
UNEXPECTED, non-systematic PECULIARITIES

only handled if necessary



Open

Variances explained in a representative set of samples



Open

Ontology: position and intensity variation in time, space and properties

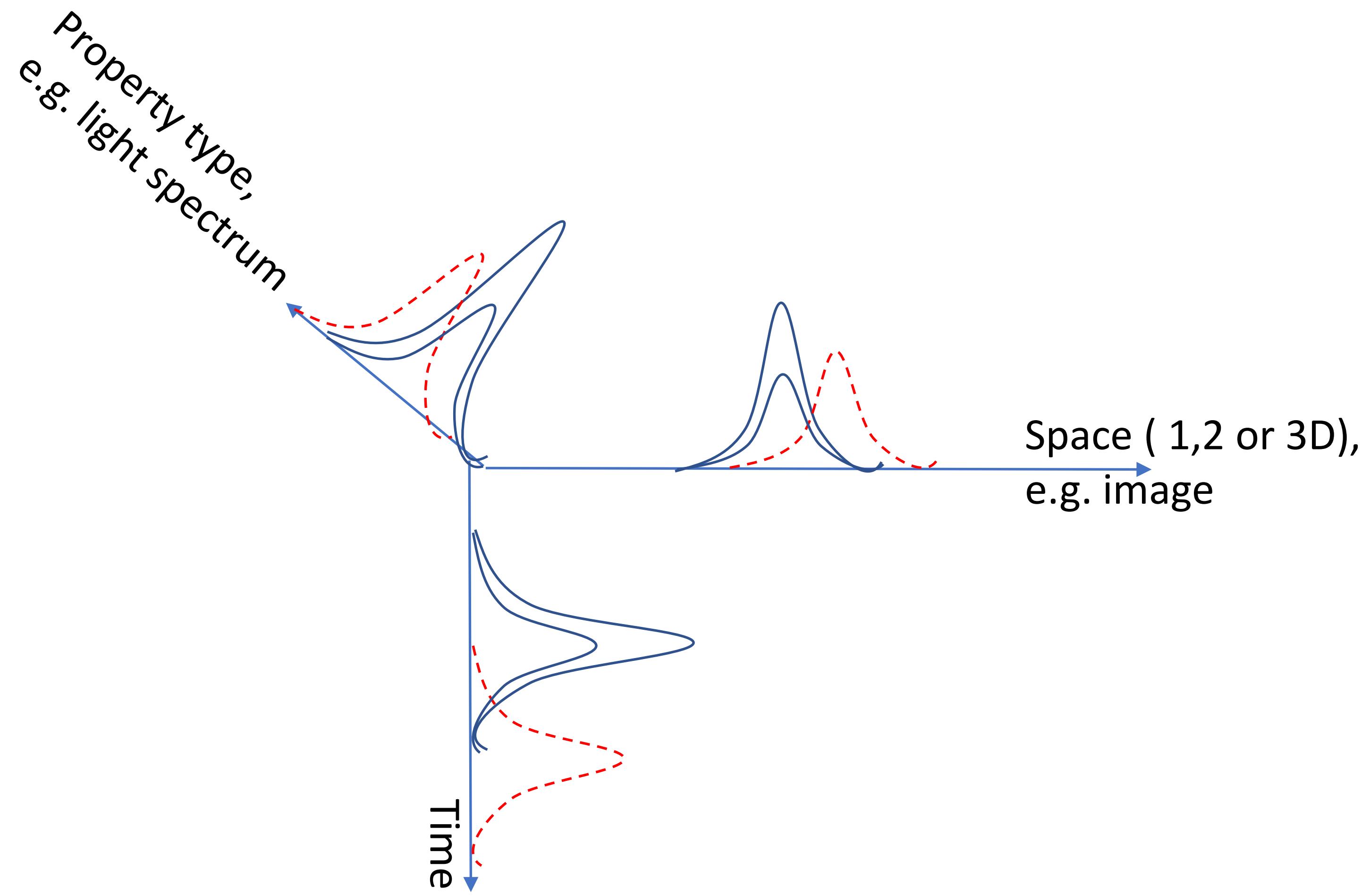
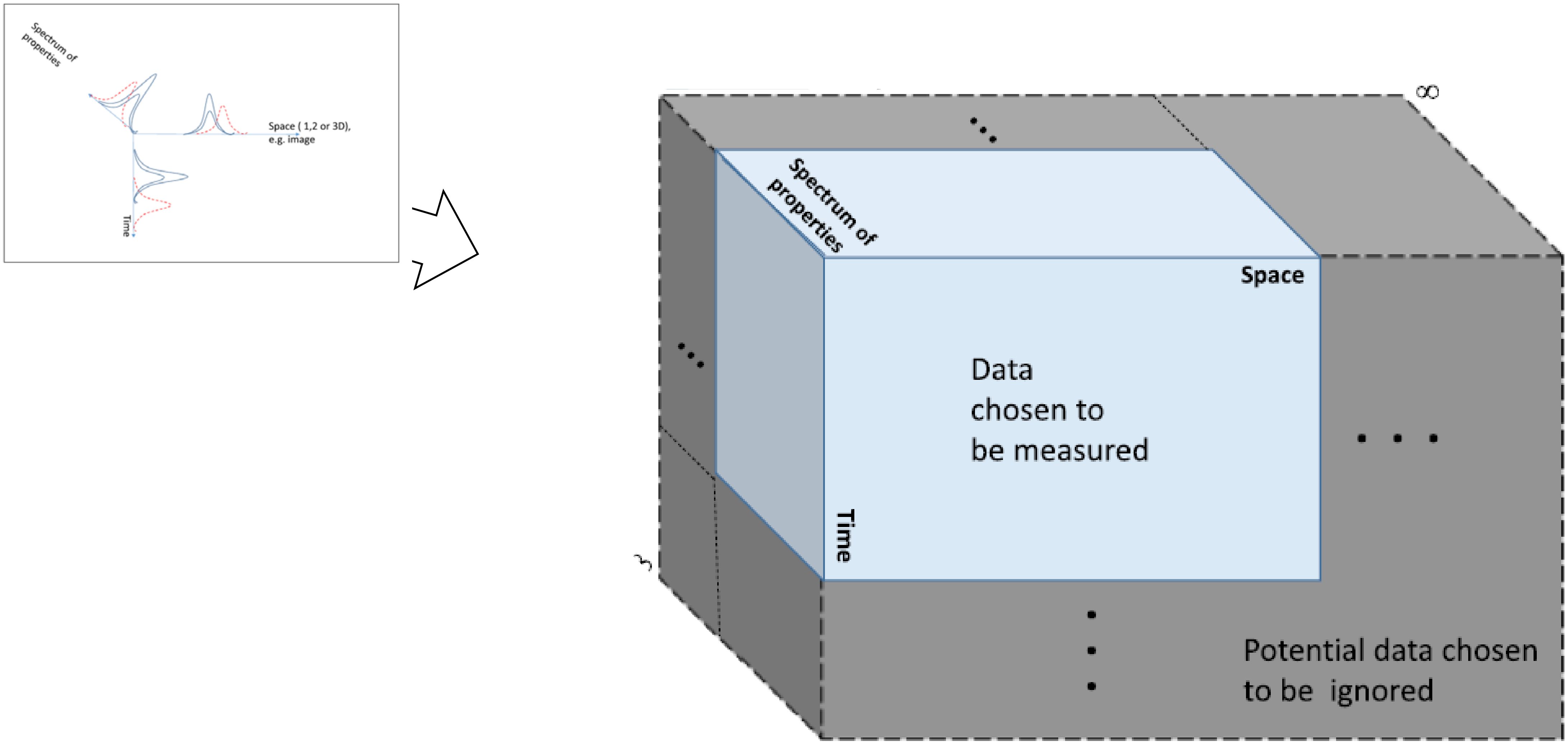


Figure Ontology

Open

Which DATA are measured?



Open

Epistemology: measure position and intensity variation in time, space and properties, and extract interpretable essence by data modelling

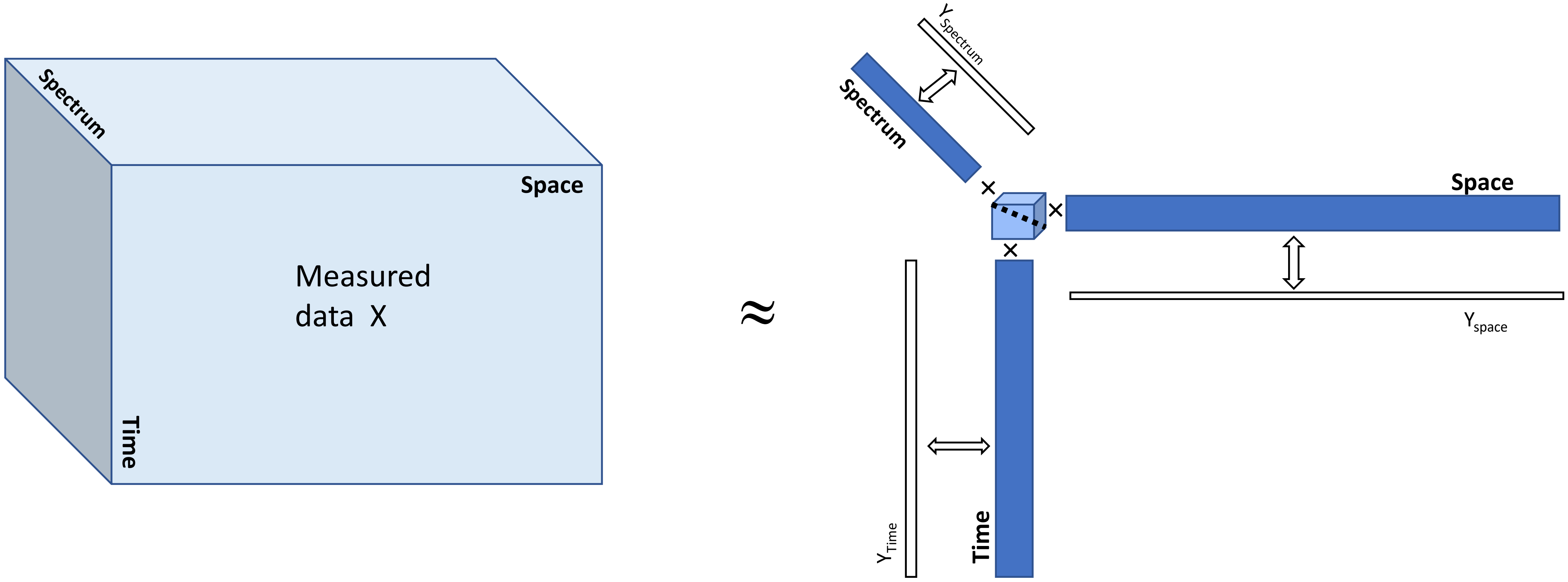
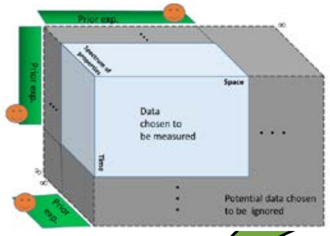


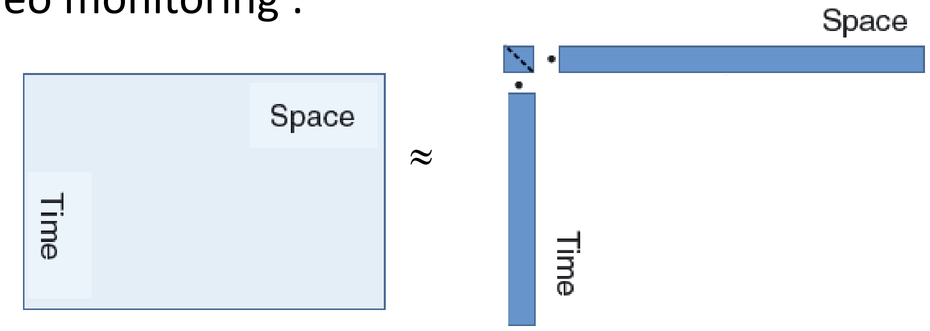
Figure N-linear

Subspace autoencoder, examples:

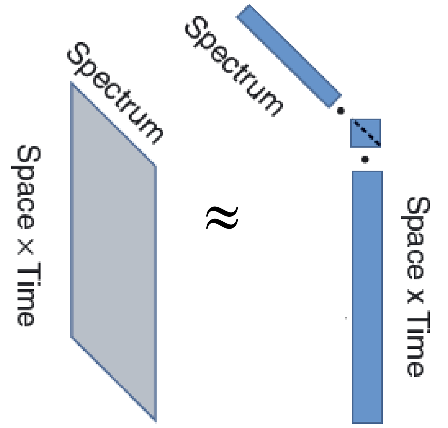


Pragmatic subspace models
&
Statistical design, validatio, graphics

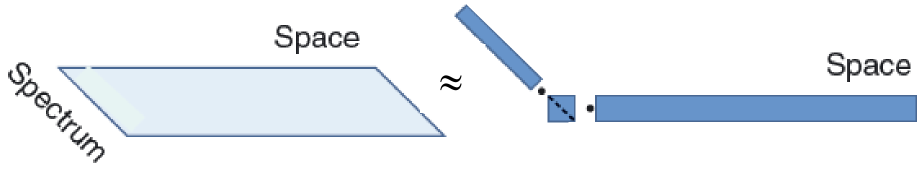
Process video monitoring :



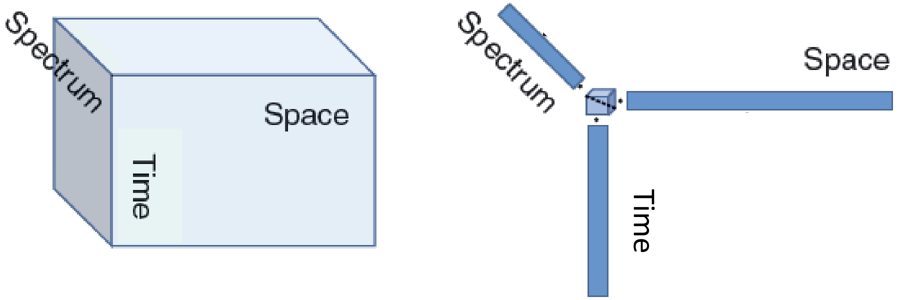
Hyperspectral video monitoring :



Hyperspectral imaging :



Hyperspectral video monitoring :



Subspace regressions, examples: