Black Box Machine Learning may be *overkill* and *risky*. **When possible, use Knowledge-based, Knowledge-generating Machine Learning instead**

Harald Martens, bio-chemetrician

1. Senior consultant, Idletechs AS

https://idletechs.com/, E-mail: harald.martens@idletechs.com

2. Prof. emerit. Big Data Cybernetics,

Dept. Engineering Cybernetics, Norwegian U. of Sci.& Technol. NTNU,

Trondheim Norway. https://www.ntnu.edu/employees/harald.martens

Main conclusion:

- Spectroscopy of intact samples in NIR e.g. needs multivariate calibration.
- Black Box ANN-based «Machine learning» from AI may be tempting
- It may work, but is not cost-effective , and does not make you smarter.
- Better alternative: **Physics-informed, hybrid chemometrics:**
  = « Knowledge-based and knowledge-generating machine learning»

Content

- **My 50 years in multivariate data modelling**
- **Importance of good data modelling**
- **What we do, and can do for *you*, in Idletechs AS**
- **Future applications and challenges**

When possible, use Knowledge-based, Knowledge-generating Machine Learning:

My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

Chemometrics etc:
Human-interpretable
"machine learning"

1970-1990: **NIR**

for foods and feeds, pharma, petrochem.:

High-speed real-world measurements

The Unscrambler: Multivariate calibration for better selectivity

When possible, use Knowledge-based, Knowledge-generating Machine Learning:

My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

Chemometrics etc:
Human-interpretable
"machine learning"

1970-1990: **NIR**
for foods and feeds, pharma, petrochem.:
High-speed real-world measurements
The Unscrambler: Multivariate calibration for better selectivity

1990-2024: **Technical BIG DATA**
for Industry, health, telecom., environment, space:
Hyperspectral & Thermal imaging and –video
Idletechs: Efficient learning algorithms
Data compression and interpretation
Professional industry implementations

When possible, use Knowledge-based, Knowledge-generating Machine Learning:

My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

Classical mechanistic models:
Poweful, but rigid, oversimplified?

Chemometrics etc:
Human-interpretable
"machine learning"

1970-1990: **NIR**
for foods and feeds, pharma, petrochem.:
High-speed real-world measurements
The Unscrambler: Multivariate calibration for better selectivity

1990-2024: **Technical BIG DATA**
for Industry, health, telecom., environment, space:
Hyperspectral & Thermal imaging and –video
Idletechs: Efficient learning algorithms
Data compression and interpretation
Professional industry implementations

When possible, use Knowledge-based, Knowledge-generating Machine Learning:

My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

Classical mechanistic models:
Poweful, but rigid, oversimplified?

⟺

Chemometrics etc:
Human-interpretable
"machine learning"

⟺

"Modern" black box AI:
Powerful, but ignorant overkill?

1970-1990: **NIR**
for foods and feeds, pharma, petrochem.:
High-speed real-world measurements
The Unscrambler: Multivariate calibration for better selectivity

1990-2024: **Technical BIG DATA**
for Industry, health, telecom., environment, space:
Hyperspectral & Thermal imaging and –video
Idletechs: Efficient learning algorithms
Data compression and interpretation
Professional industry implementations

Importance of good data modelling:

$\Rightarrow$ Good predictions, classification

Importance of good data modelling:

$\Rightarrow$ Good predictions, classification
& Uncertainty-estimates and Anomaly warnings

Importance of good data modelling:

$\Rightarrow$ Good predictions, classification
& Uncertainty-estimates and Anomaly warnings

$\Rightarrow$ Better understanding $\Rightarrow$ Safer use and New opportunities

Technical BIG DATA from modern instruments:

idletechs

- Necessary info from Technical BIG DATA :

See THAT it works, HOW it works and WHY it works!

Technical BIG DATA from modern instruments:

- Necessary info from Technical BIG DATA :

See THAT it works, HO Black Box rks and WHY Black Box ks!

Conventional AI = Black Box

idletechs

Technical BIG DATA from modern instruments:

- Necessary info from Technical BIG DATA :

See THAT it works, HOW it works and WH[redacted]rks!

XAI ?

XAI= eXplainable AI

Technical BIG DATA from modern instruments:

idletechs

- Necessary info from Technical BIG DATA :

See THAT it works, HOW it works and WHY it works!

«Understandable AI»

idletechs

What we do in Idletechs AS

Interpretation and use of Technical BIG DATA for industry, environmental etc.:

- Data modelling
- Data visualization and warnings
- Software for
  - data input , modelling, display , prediction, classification, outliers, control
    - & compression, storage & retrieval

idletechs

What we do in Idletechs AS

- Software: White label- or stand-alone software. All standard protocols

Idletechs AS has two customer types:

What we do in Idletechs AS

- **Spectrometeres:** Vis/NIR, Raman, IR, MS, Chromatogr., InSAR, …
- **Imagers:** Thermal video, VNIR, SWIR, Raman, Xray, MRI, …

  - Modelling:

    - Identify, separate and quantify different spectral changes in e.g. VNIR/SWIR/RGB:

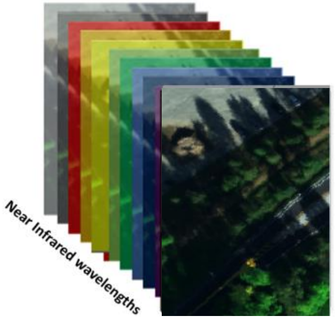
• **light absorptions**	chemical composition
• **light scattering** / effective optical path length	particle size and –density
• **Specular surface effects**	stray light, surface “mirrors”
• **Illumination**	spectral deshadowing

⇒ Innovative semi-mechanistic modelling of real-world light measurements

What we can do for your spectral measurements

- Increase the **relevance and interpretability**

- **Example 1:** Remove ***shadows*** from hyperspectral images



Spectral de-shadowing

idletechs

Hyper-spectral VNIR camera

Technical BIG DATA

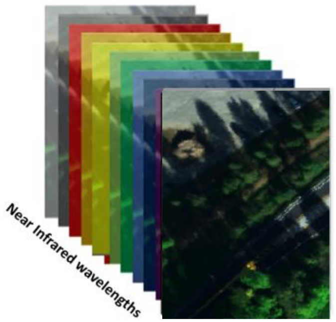
Hybrid, interpretable machine learning

From HSI-camera (shown as RGB)

(Norsk Elektro-optikk/Terratec)

idletechs

Spectral de-shadowing



Hyper-spectral VNIR camera

Technical BIG DATA

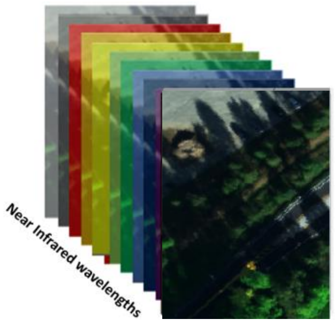
Hybrid, interpretable machine learning

From HSI-camera (shown as RGB)

(Norsk Elektro-optikk/Terratec)

After spectral de-shadowing

(Idletechs)



Spectral de-shadowing

idletechs

Hyper-spectral VNIR camera

Technical BIG DATA

Hybrid, interpretable machine learning

From HSI-camera (shown as RGB)

(Norsk Elektro-optikk/Terratec)

After spectral de-shadowing

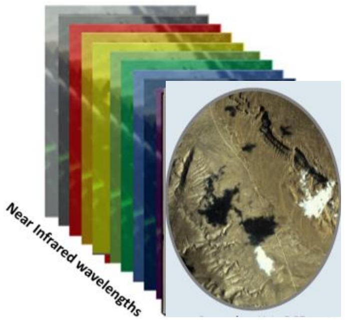
(Idletechs)

How?

Input
The spectral difference between **yellow sun and blue sky.**

Then: Simplified EMSC
Multivariate linear model-based pre-processing of log(1/R) spectrum for each pixel.

May be optimized in different ways



Spectral de-shadowing

NTNU
Norwegian University of Science and Technology

idletechs

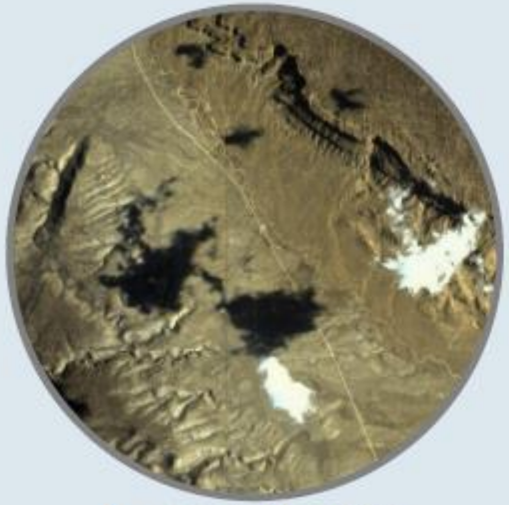
Hyper-spectral VNIR camera

Technical BIG DATA

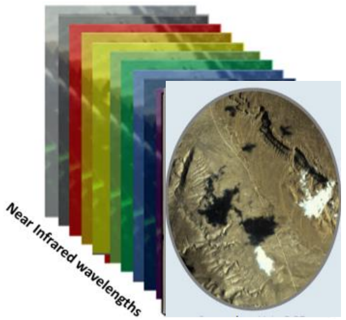
Hybrid, interpretable machine learning

Earth Observing-1

Data from the Hyperion instrument onboard the EO-1 Satellite. Data contains 200 bands in the VIS-NIR region. Clouds were the main source of shadows in this dataset.



Input data Y, in RGB



Spectral de-shadowing

NTNU
Norwegian University of Science and Technology

idletechs

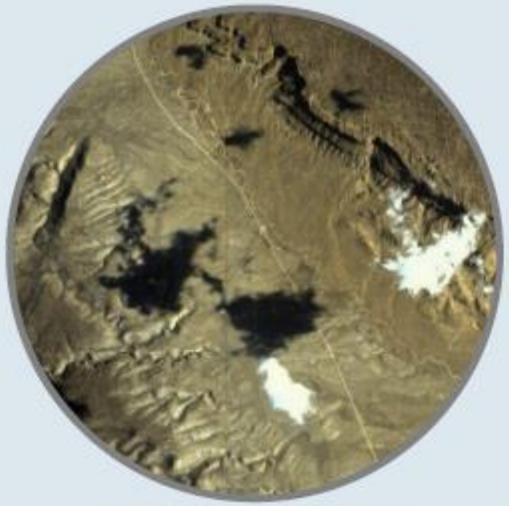
Hyper-spectral VNIR camera

Technical BIG DATA

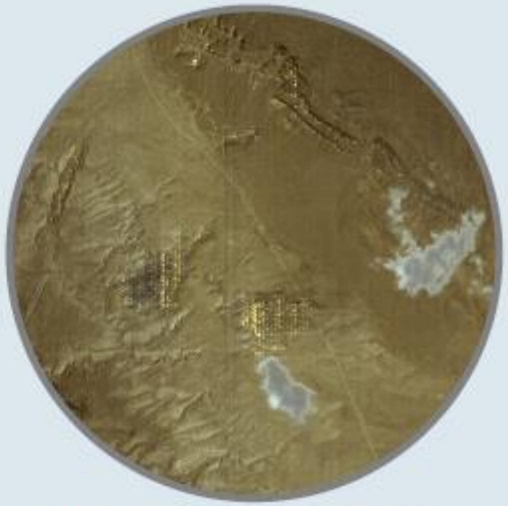
Hybrid, interpretable machine learning

Earth Observing-1

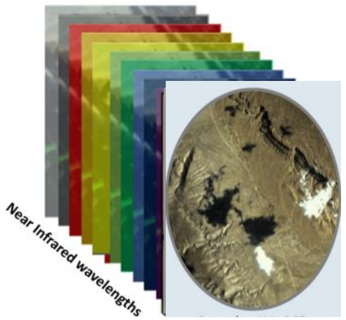
Data from the Hyperion instrument onboard the EO-1 Satellite. Data contains 200 bands in the VIS-NIR region. Clouds were the main source of shadows in this dataset.



Input data Y, in RGB



Deshadowed image, in RGB



Spectral de-shadowing

NTNU
Norwegian University of Science and Technology

idletechs

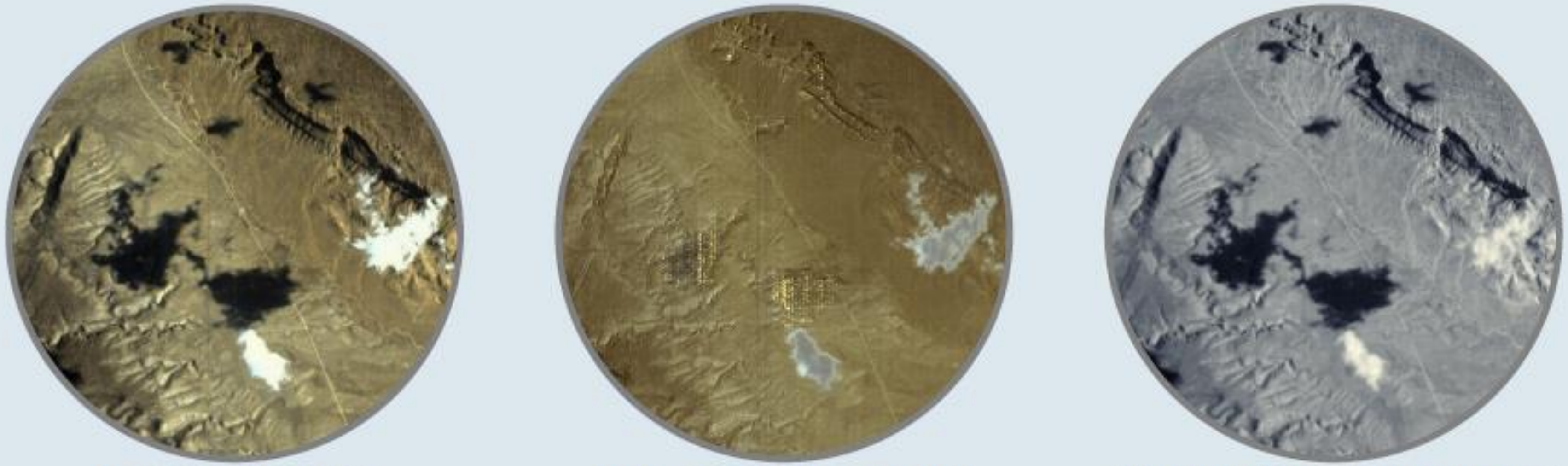
Hyper-spectral VNIR camera

Technical BIG DATA

Hybrid, interpretable machine learning

Earth Observing-1

Data from the Hyperion instrument onboard the EO-1 Satellite. Data contains 200 bands in the VIS-NIR region. Clouds were the main source of shadows in this dataset.



Input data Y, in RGB | Deshadowed image, in RGB | "Shadow" (illumination change)

idletechs

What we can do for your spectral measurements

- Increase the **selectivity and linearity**

Example 2: Separate **«chemical» light absorptions** from **«physical» light scattering**

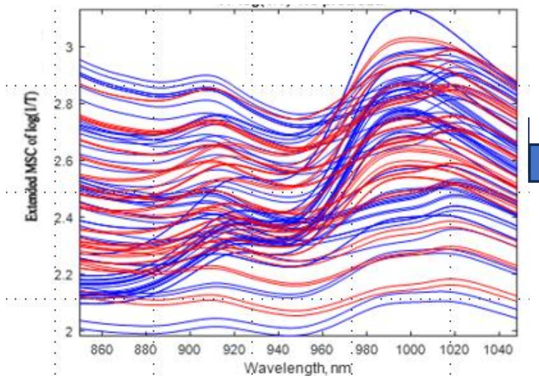
Beer's law, reliable **Lambert's law, unreliable ?**

Model-based preprocessing of NIR spectra:

Separation of «chemical» light scattering and «physical» light absorption by EMSC and OEMSC

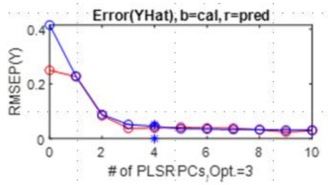
Didactic example:
Mixtures of two powders

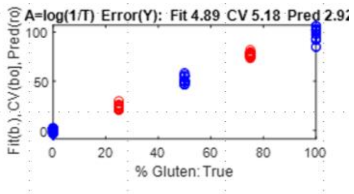
Measured NIR Absorbances, 50 samples



Conventional multivariate calibration based on PLSR_

Works,
but not well enough.



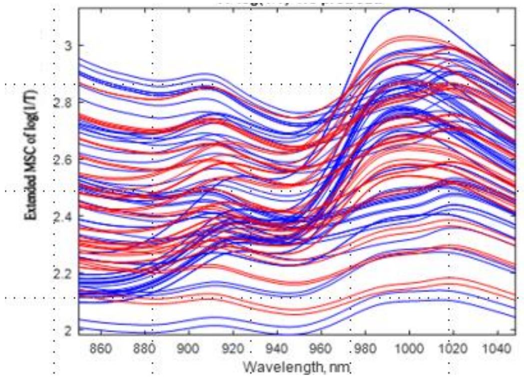


Model-based preprocessing of NIR spectra:

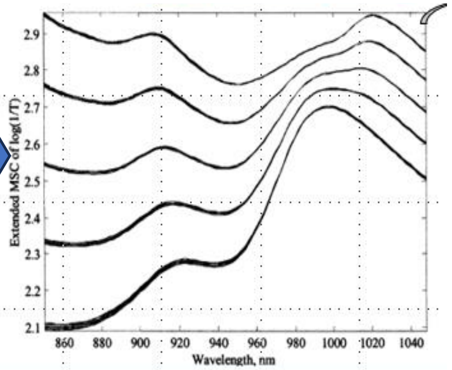
Separation of «chemical» light scattering and «physical» light absorption by EMSC and OEMSC

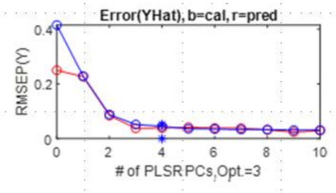
Didactic example:
Mixtures of two powders

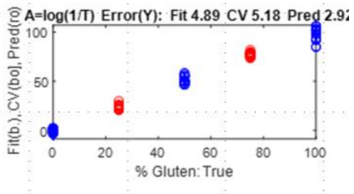
Measured NIR Absorbances, 50 samples



When the constitutents' spectra are KNOWN: Same spectra after EMSC preprocessing





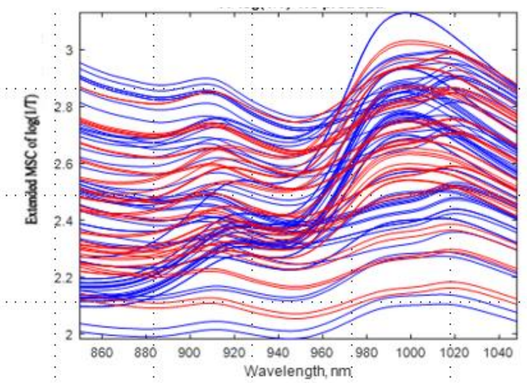


Model-based preprocessing of NIR spectra:

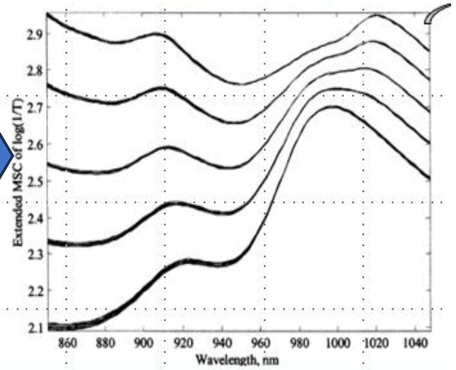
Separation of «chemical» light scattering and «physical» light absorption by EMSC and OEMSC

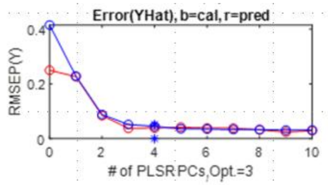
Didactic example:
Mixtures of two powders

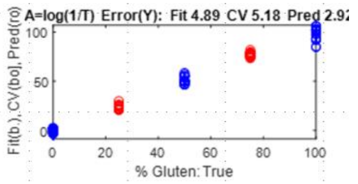
Measured NIR Absorbances, 50 samples



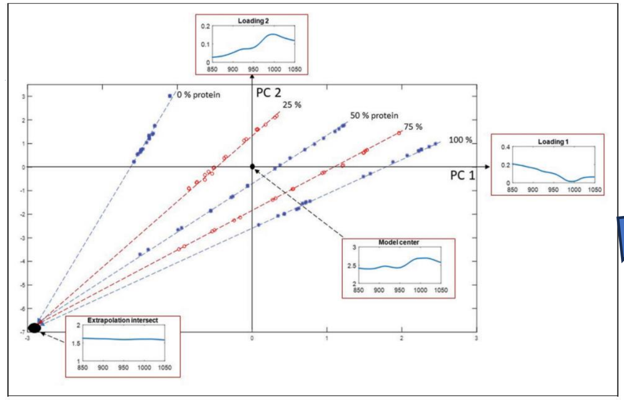
When the constitutents' spectra are KNOWN: Same spectra after EMSC preprocessing





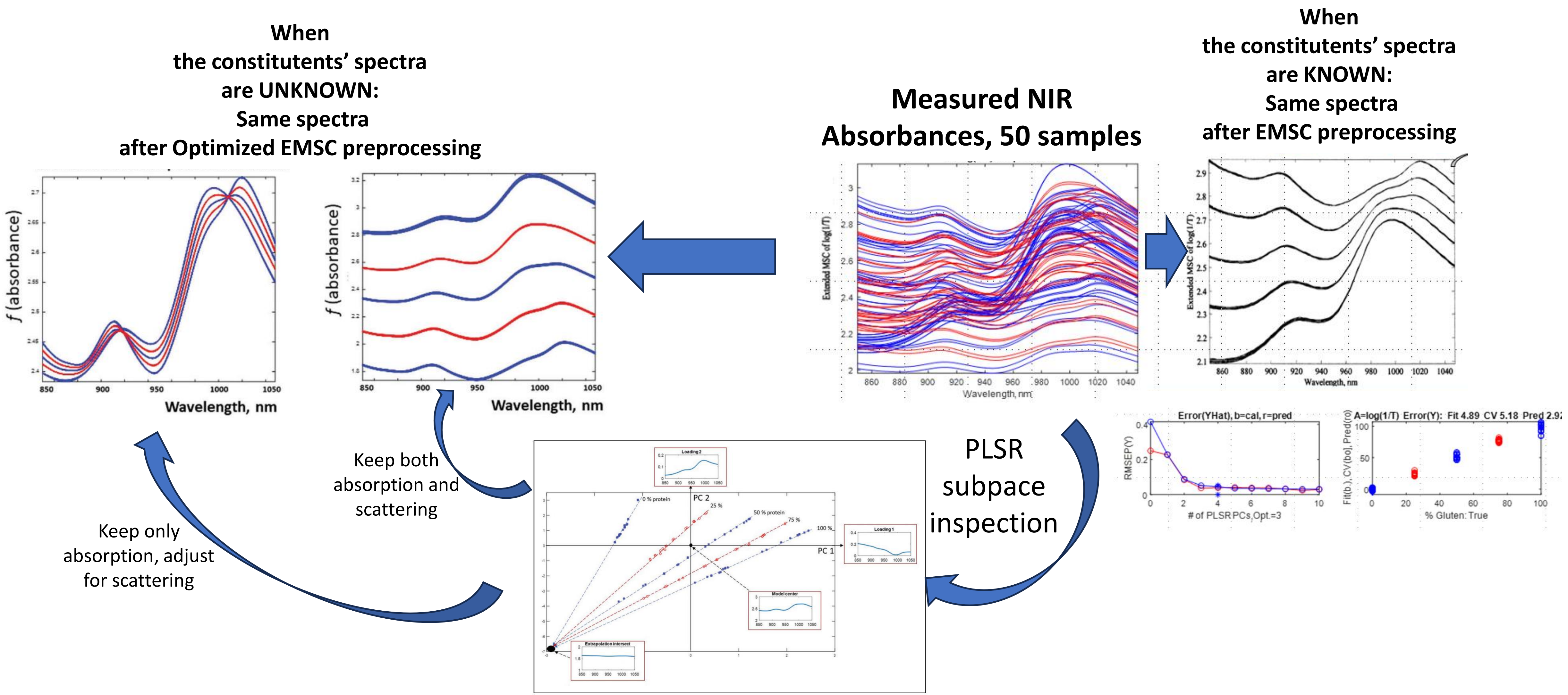


PLSR subpace inspection



Model-based preprocessing of NIR spectra:

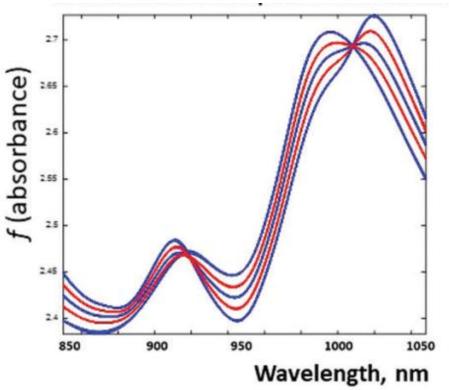
Separation of «chemical» light scattering and «physical» light absorption by EMSC and OEMSC

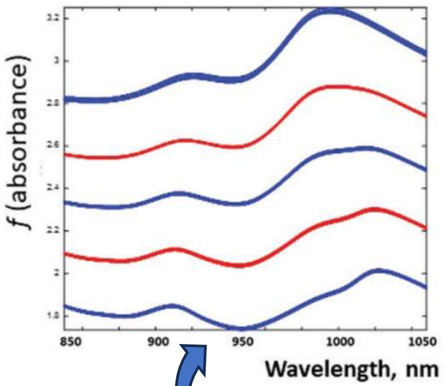


Model-based preprocessing of NIR spectra:

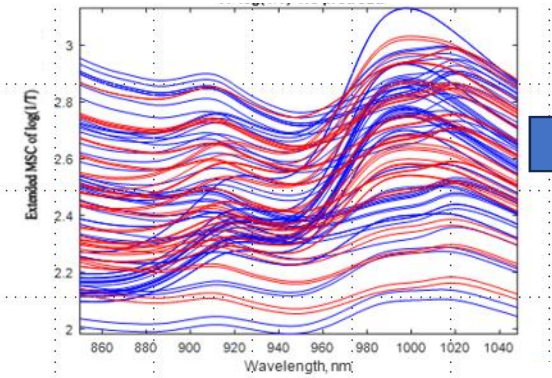
Separation of «chemical» light scattering and «physical» light absorption by EMSC and OEMSC

When
the constitutents' spectra
are UNKNOWN:
Same spectra
after Optimized EMSC preprocessing

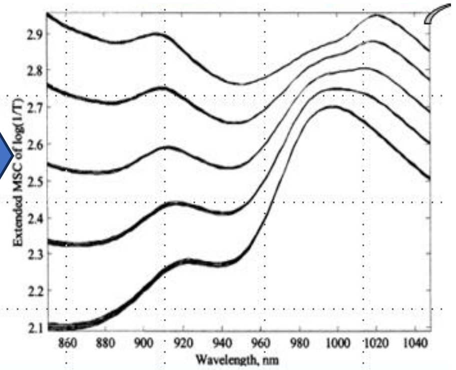




Measured NIR
Absorbances, 50 samples

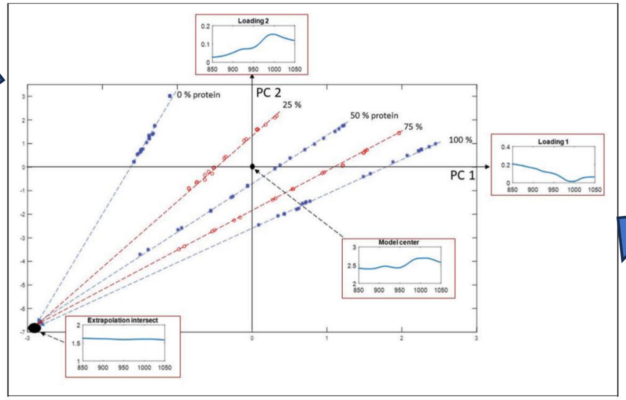


When
the constitutents' spectra
are KNOWN:
Same spectra
after EMSC preprocessing

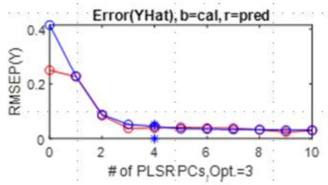


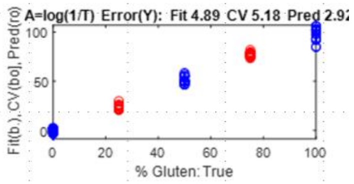
Keep only absorption, adjust for scattering

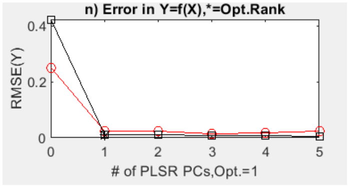
Keep both absorption and scattering

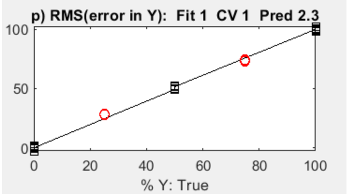


PLSR subpace inspection









idletechs

What we can do for your spectral measurements

- Separate effects of **motions** from effects of **intensity** changes.
  - IDLE modelling I=D(L)+E: **Intensity**=**D**isplacement of **L**ocal structure + **E**rror

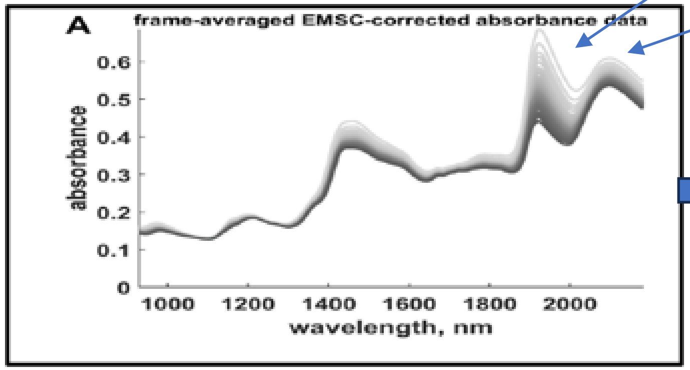
What we can do for your spectral measurements

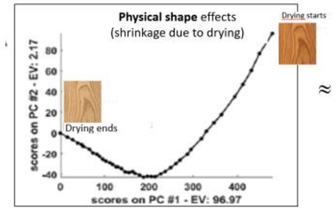
- Separate effects of **motions** from effects of **intensity** changes.
  - IDLE modelling I=D(L)+E: **Intensity**=**D**isplacement of **L**ocal structure + **E**rror

Example 3: NIR HSI of Drying wood

Drying wood: ⇒ Δ Chemistry (water) , Δ Physics (scattering),

SWIR: 900-2500 nm





idletechs

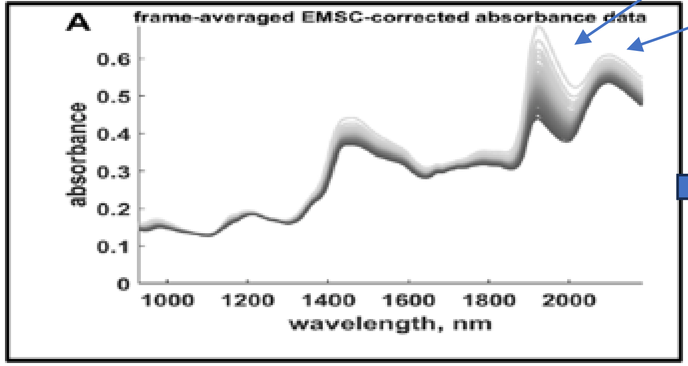
What we can do for your spectral measurements

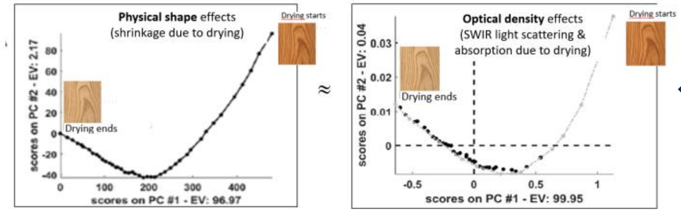
- Separate effects of **motions** from effects of **intensity** changes.
  - IDLE modelling I=D(L)+E: **I**ntensity=**D**isplacement of **L**ocal structure + **E**rror

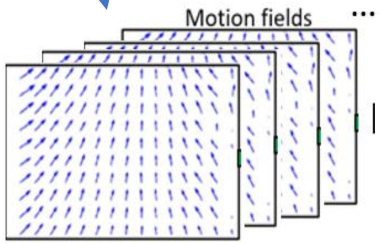
Example 3: NIR HSI of Drying wood

Drying wood: ⇒ Δ Chemistry (water) , Δ Physics (scattering), Δ shape (shrinking),

SWIR: 900-2500 nm







Similar kinetics path!

idletechs

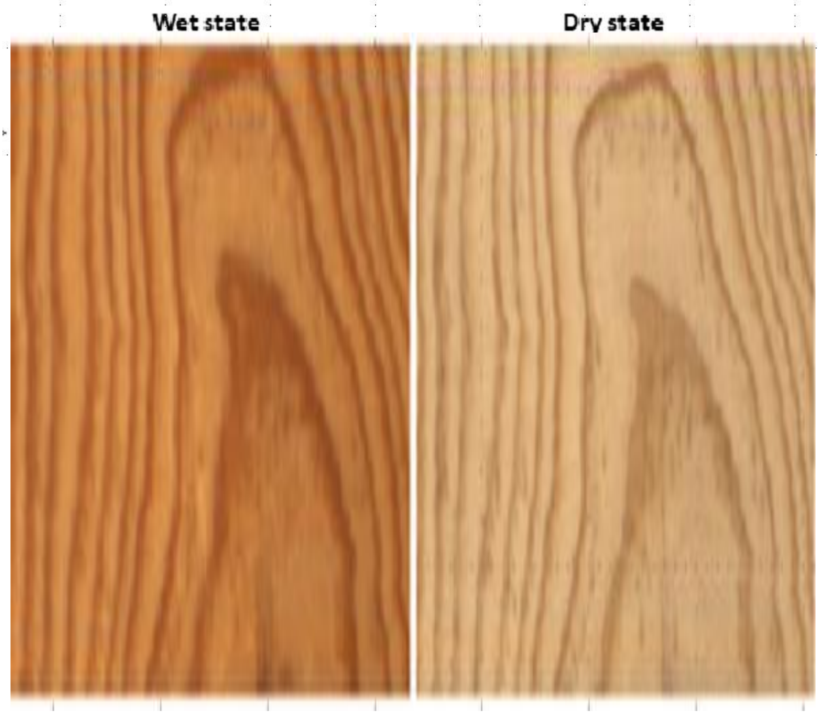
What we can do for your spectral measurements

- **Compress** Technical BIG DATA without loss of relevant information

Example 4: **Technical Big Data: Hyperspectral «video»**

Norwegian University of Life Sciences

idletechs



A single piece of drying wood:

>350 000 000 VNIR reflectance spectra, 200 channels each:

Idletechs' physics-informed machine learning

Only 8 «change patterns»: $\Rightarrow$ 99.8% NIRS variance explained

Example 4: **Technical Big Data: Hyperspectral «video»**

A single piece of drying wood:
>350 000 000 VNIR reflectance spectra, 200 channels each:

Idletechs' physics-informed machine learning

Only 8 «change patterns»: ⇒ 99.8% NIRS variance explained

+ IDLE based motion compensation ⇒ much higher compression

Future applications and challenges

idletechs

- **Modelling Technical BIG DATA will require more:**
  - **Greeen AI**
    - Low energy use for data transmission and storage, model calibration and use
    - Hybrid subspace machine learning better than ANN based deep learning
  - **Humane AI**
    - Someone must always know THAT, HOW and WHY the instrument works
    - Use prior knowledge (scientific models & measurements, human experience)
    - Display results to people
    - Stimulate people to generate new knowledge
  - **Safe AI**
    - Use methods that reveal anomalies
      - Extreme levels of known variation types
      - New variation types
      - Outliers
      - Instrument problems
    - Always provide uncertainty estimates, outlier warnings and model-selfcritique

idletechs

What we do in Idletechs AS

- Interpretation and use of Technical BIG DATA for industry, environmental etc.

    - Software: White label- or stand-alone software. All standard protocols

Idletechs AS has two customer types:

**Instrument MAKERS
who need
methods and software to
calibrate and employ their
Technical BIG DATA**

⟷

**Idletechs AS:
Methods,
software,
Consulting
wrt
Technical BIG DATA**

⟷

**Instrument USERS
who need
methods and softwareto
utilize their
Technical BIG DATA**

We seek **partners and collaborators**
in our further development of **methods and software**
to convert **multichannel spectra and images**
into **relevant and reliable information**

idletechs

- For food, agriculture, environmental:
  - **Low-end instruments**, e.g. for smart-phones
  - **High-end instruments**, e.g. from satellites drones, vehicles, in processses

Also for industry in general, medicine, space, defense

Acknowledgements:

Ingunn Burud & Petter Stefansson, Norwegian University of Life Sciences NMBU, Ås, Norway:
Raffaele Vitale, U. Lille, France

When possible, use Knowledge-based, Knowledge-generating Machine Learning:

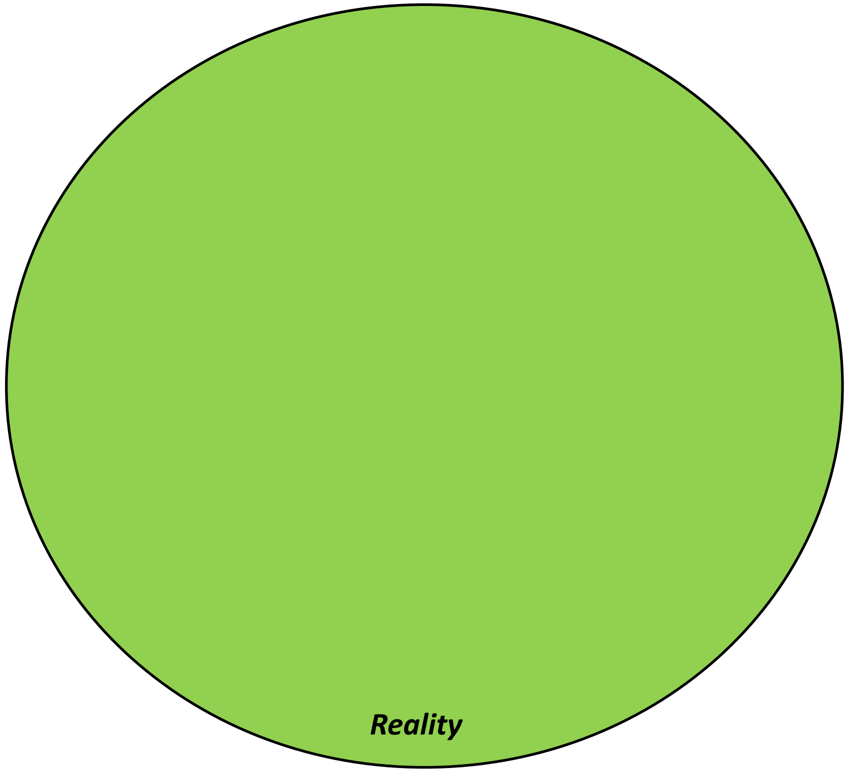
My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

Classical mechanistic models:
Poweful, but rigid, oversimplified?

Chemometrics etc:
Human-interpretable
"machine learning"

"Modern" black box AI:
Powerful, but ignorant overkill?



When possible, use Knowledge-based, Knowledge-generating Machine Learning:

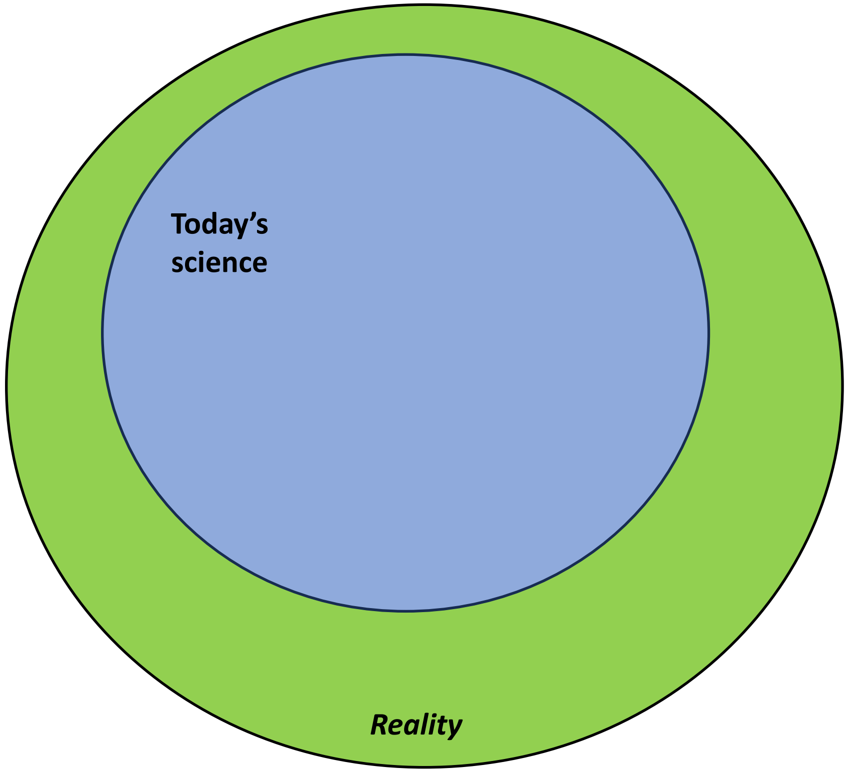
My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

Classical mechanistic models:
Poweful, but rigid, oversimplified?

Chemometrics etc:
Human-interpretable
"machine learning"

"Modern" black box AI:
Powerful, but ignorant overkill?



When possible, use Knowledge-based, Knowledge-generating Machine Learning:

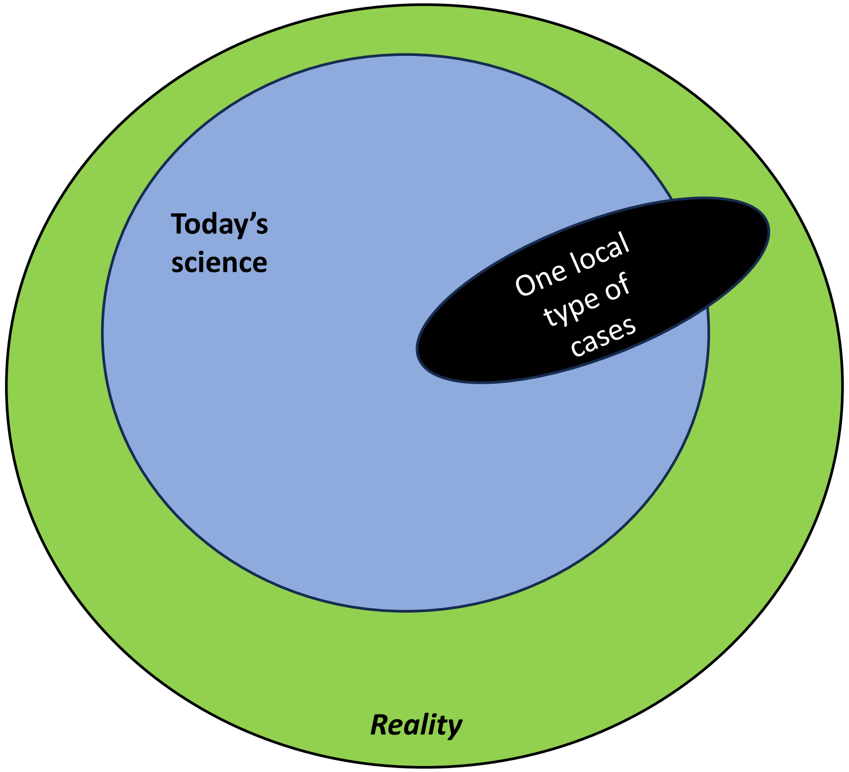
My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

Classical mechanistic models:
Poweful, but rigid, oversimplified?

Chemometrics etc:
Human-interpretable
"machine learning"

"Modern" black box AI:
Powerful, but ignorant overkill?



When possible, use Knowledge-based, Knowledge-generating Machine Learning:

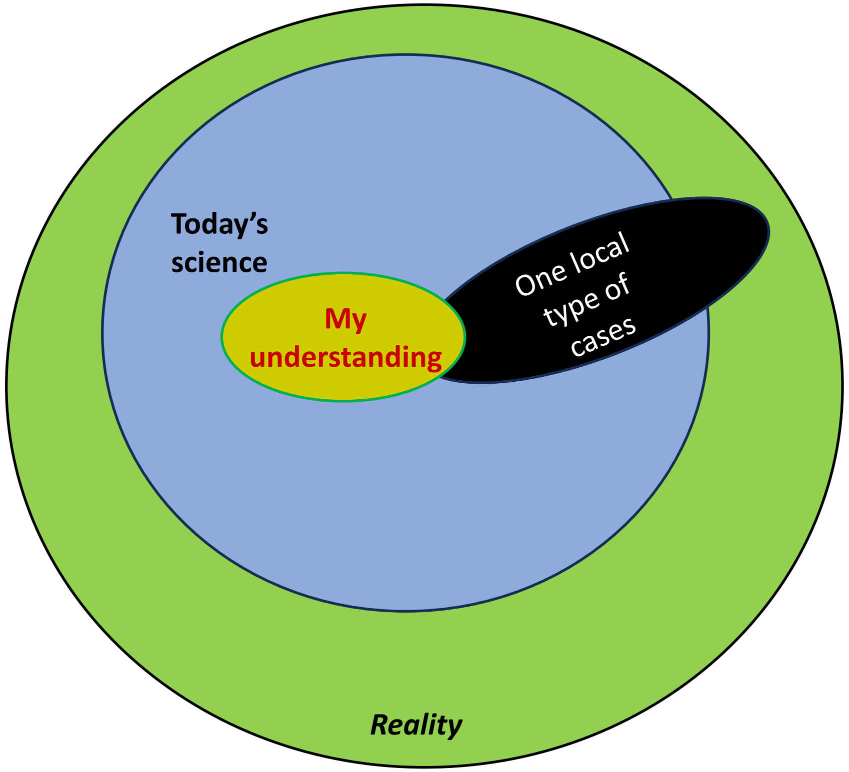
My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

Classical mechanistic models:
Poweful, but rigid, oversimplified?

Chemometrics etc:
Human-interpretable
"machine learning"

"Modern" black box AI:
Powerful, but ignorant overkill?



When possible, use Knowledge-based, Knowledge-generating Machine Learning:

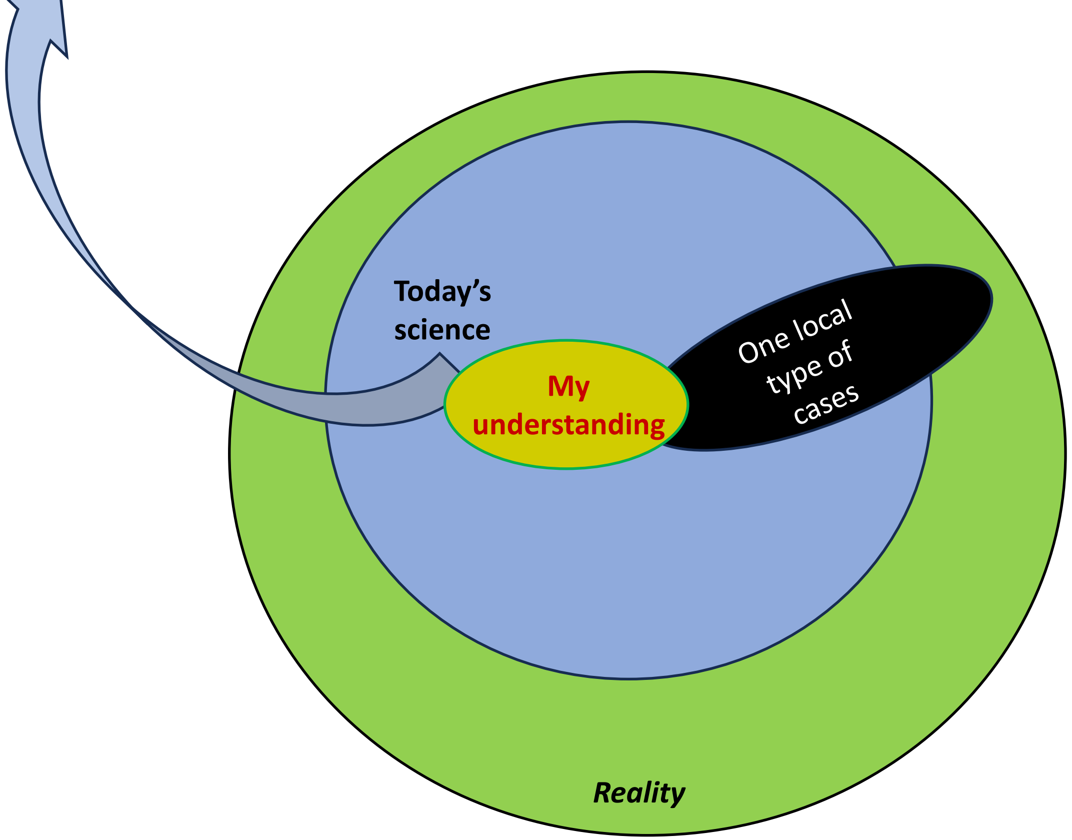
My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

Classical mechanistic models:
Poweful, but rigid, oversimplified?

Chemometrics etc:
Human-interpretable
"machine learning"

"Modern" black box AI:
Powerful, but ignorant overkill?



When possible, use Knowledge-based, Knowledge-generating Machine Learning:

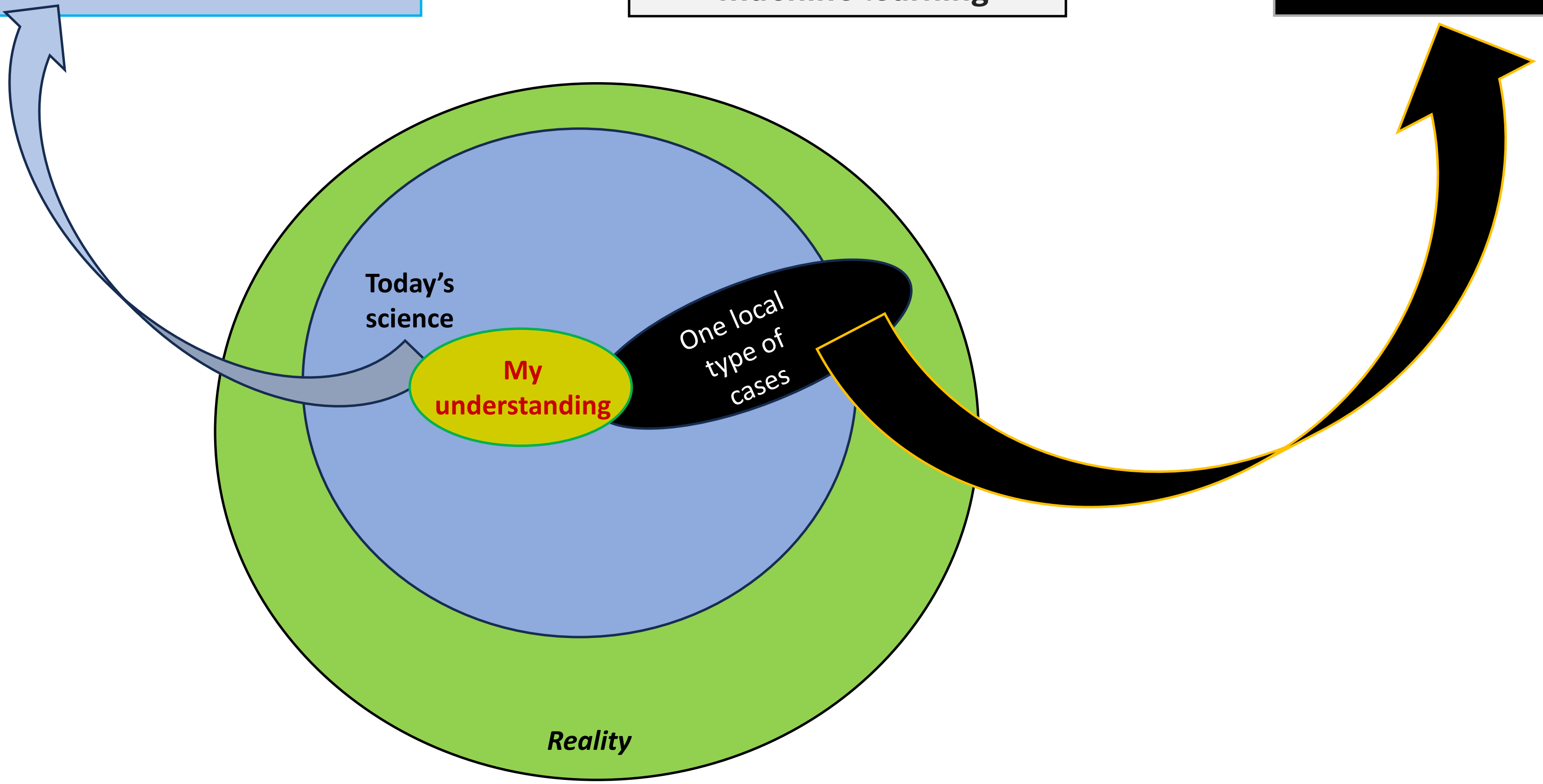
My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

Classical mechanistic models:
Poweful, but rigid, oversimplified?

Chemometrics etc:
Human-interpretable
"machine learning"

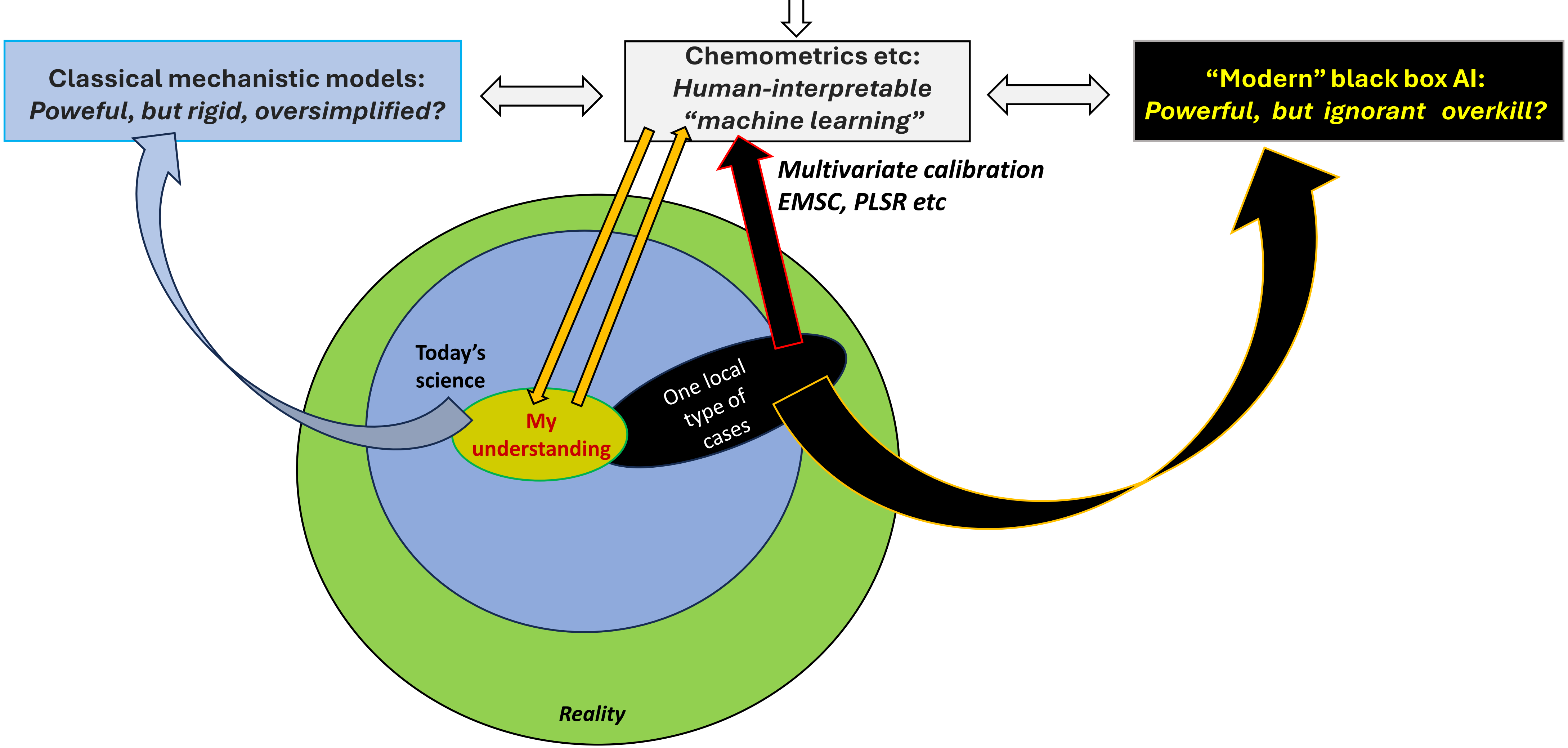
"Modern" black box AI:
Powerful, but ignorant overkill?



When possible, use Knowledge-based, Knowledge-generating Machine Learning:

My 50 years in multivariate data modelling

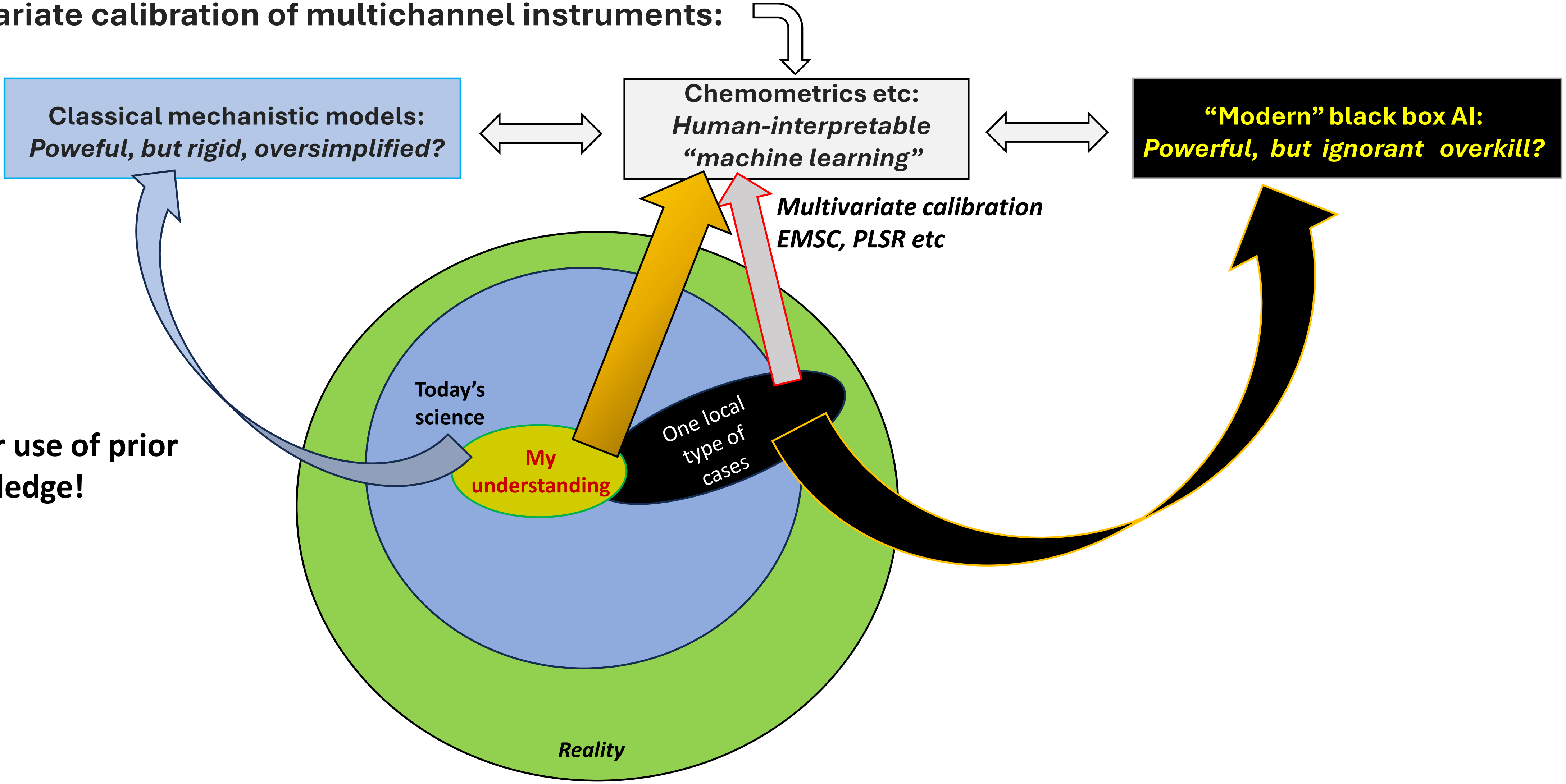
Multivariate calibration of multichannel instruments:



When possible, use Knowledge-based, Knowledge-generating Machine Learning:

My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:

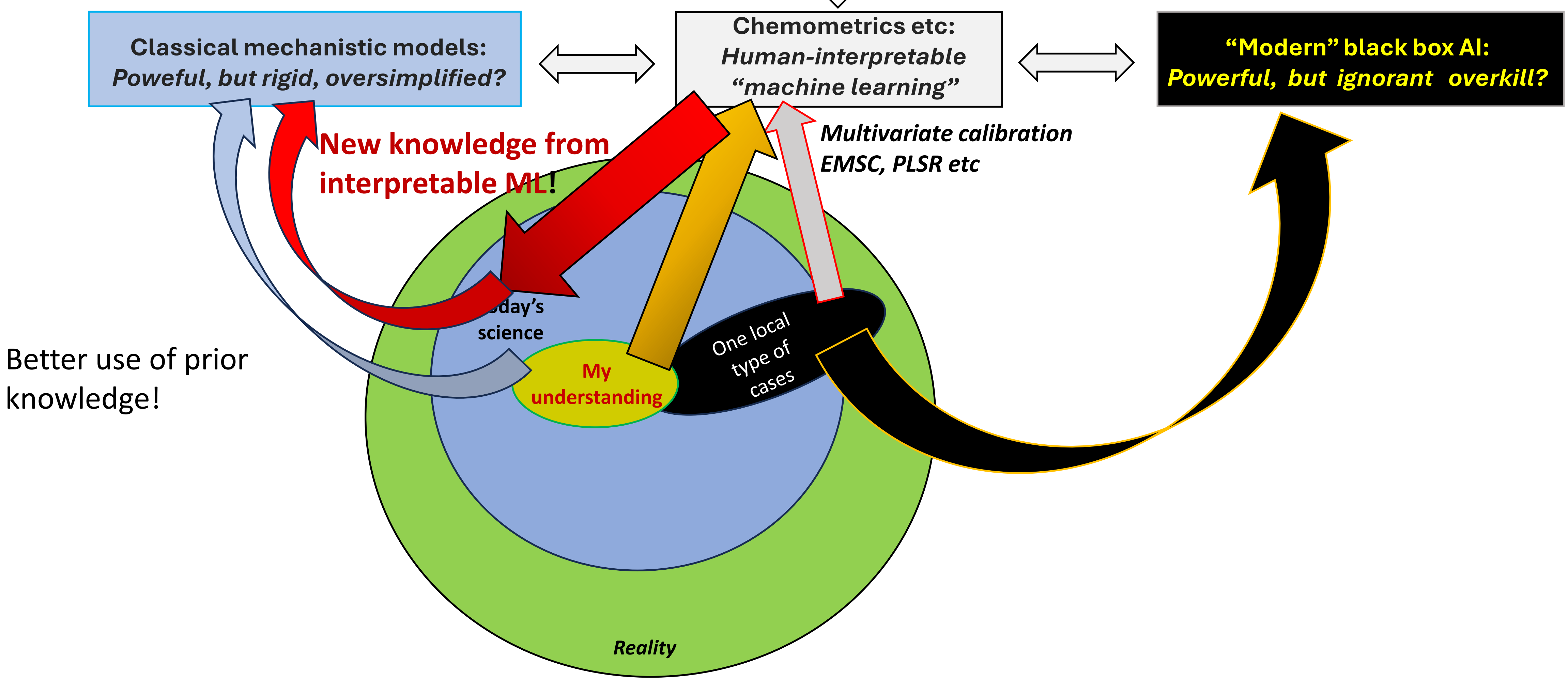


Better use of prior knowledge!

When possible, use Knowledge-based, Knowledge-generating Machine Learning:

My 50 years in multivariate data modelling

Multivariate calibration of multichannel instruments:



Remember:

A *problem* has a *cause*.

Therefore a ***selectivity problem*** may point to a new *opportunity*

Remember:

A *problem* has a *cause*.

Therefore a ***selectivity problem*** may point to a new *opportunity*

Remember:

A *problem* has a *cause*.

Therefore a ***selectivity problem*** may point to a new *opportunity*

Motto 1: Try to see why

Remember:

A *problem* has a *cause*.

Therefore a ***selectivity problem*** may point to a new *opportunity*

Motto 1: Try to see why

Motto 2: It is better to be approximately right than precisely wrong

Remember:

A *problem* has a *cause*.

Therefore a ***selectivity problem*** may point
to a new *opportunity*

Motto 1: Try to see why

**Motto 2: It is better to be approximately right
than precisely wrong**

**Mottos 3 and 4:
No interpretation without proper validation!**

No prediction without attempted interpretation

Thank you

For possible discussions:

Compress Technical BIG DATA without loss of relevant information

idletechs

- Model- based compression and reconstruction:
  - **Measurements** = "**Systematic**" variation patterns (compressed) + "**Random**" noise + **Anomalies**
  - Simpler **transmission & storage** of Technical BIG DATA
  - Measurements are **interpretable in their compressed form**
  - Simpler **reconstruction & visualization** of the relevant information

    (If needed: Lossless reconstruction – but at a bit-rate price: "Random noise" cannot be compressed well.)

Future applications and challenges

(40 % time). 8 min= 4 figs

- Measure for a better world:
  - Science was right about ozone-layer, bio-diversity and global warming
  - Science was wrong microplastic and about water structure (?)
  - We need more measurements, for
    - better products and
    - better understanding

- Max. food production and food quality, min. environment problems:
  - Massive use of low-end scanners and imagers
  - A different way to teach math and statistics to users: “ like music” ?

- Multi-channel scanners and imagers: More cost-effective measurements.
  - Technical Big Data needs data modelling.

Future applications and challenges

(40 % time). 8 min= 4 figs

- For multi-channel scanners and imagers:
  - Good instrument design. But remember “math is cheaper than physics”
  - Better methods and software for “machine learning” (multivariate calibration)
    - Non-linearity challenges (but no need for Artificial Neural Net):
      - Curvatures (e.g. “banana” = flat, linear “boomerang”),
      - Mixed multiplicative/additive effects: chemical light absorption and light physical scattering . EMSC etc
      - Sideways motions (in space, wavelength, time): IDLE modelling
    - Global models and local refinements, specular shadow effects, time drift
  - Better sampling for the “machine learning”:
  - Self-aware instruments that always report its predictive uncertainty
  - **Replace the ANN (artificial neural net) by a new,equivalent methodology that is faster, safer and more interpretable, at least for Technical Big Data !**

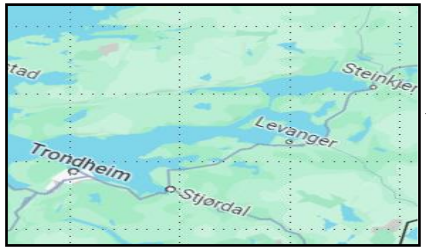
Technical BIG DATA from modern instruments:

- Listen to the music from your measuring instruments:
  - Enjoy the qualities of the live concert
  - Discover the underlying melodies, rhythms and harmonies in the cacophony of raw signals
  - Beware of disharmonies: Ignore the old man's snoring, but react to fire alarms
  - Detect instruments getting out of tune
  - Learn more about what is beind the music: The composition and the composer

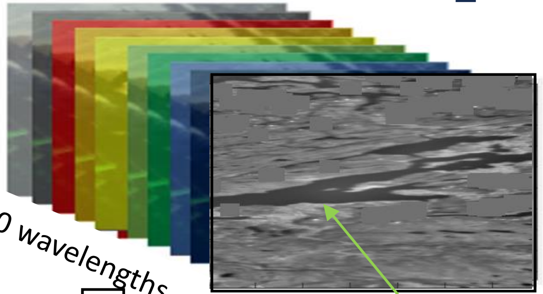
NTNU's HYPSO satellite

Little brother of the
hperspectral camera in the
HYPSO satellite

Flat Earth Society gives good maps

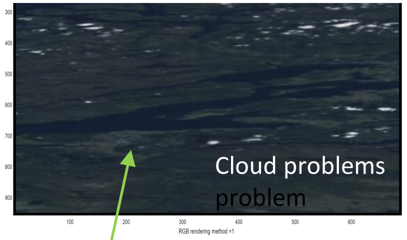


NTNU's HYPSO satellite



≈120 wavelengths

RGB



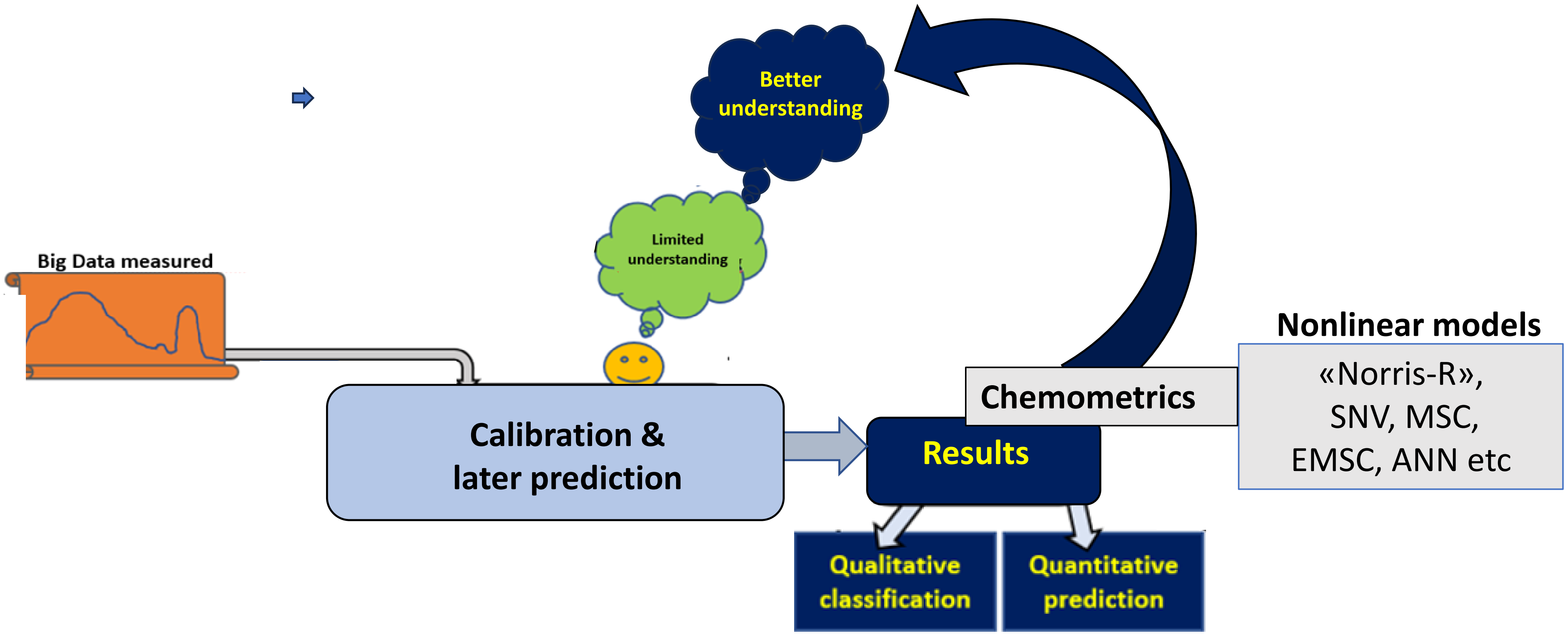
Pragmatic approximasjon

Algeal blooom
In oceatn, etc.

Trondheim!

How to transform the hyperspectral images into quantitative estimates of amoutn of different types of algea ?
Big Data Cybernetics/NTNU

Multichannel NIRS: «Machine learning» since 1983



Big Data: Hyperspectral «video»

A single piece of drying wood:
>350 000 000 VNIR reflectance spectra, 200 channels each:

Idletechs'
physics-informed
machine learning

8 change patterns: $\Rightarrow$ 99.8% NIRS variance explained

Norwegian University of Life Sciences

idletechs

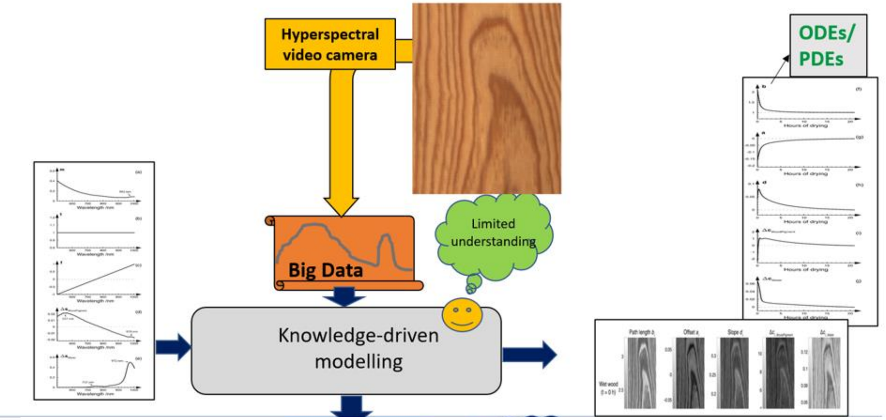
Big Data: Hyperspectral «video»

A single piece of drying wood:

>350 000 000 VNIR reflectance spectra, 200 channels each:

1. EMSC modelling

Modelling
KNOWN
spectral change patterns



3.: Spatial and Temporal Inter-pretation

&

Kinetics

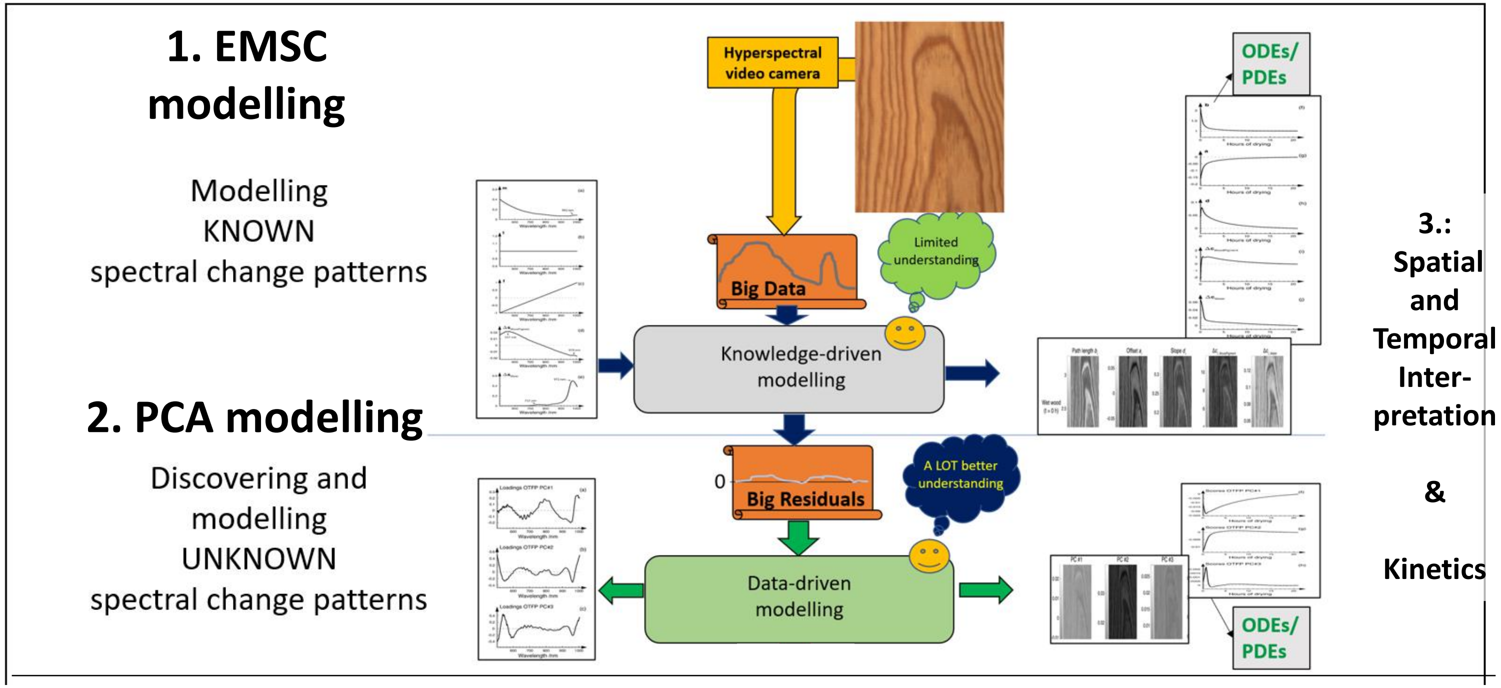
Norwegian University of Life Sciences

idletechs

Big Data: Hyperspectral «video»

A single piece of drying wood:

>350 000 000 VNIR reflectance spectra, 200 channels each:



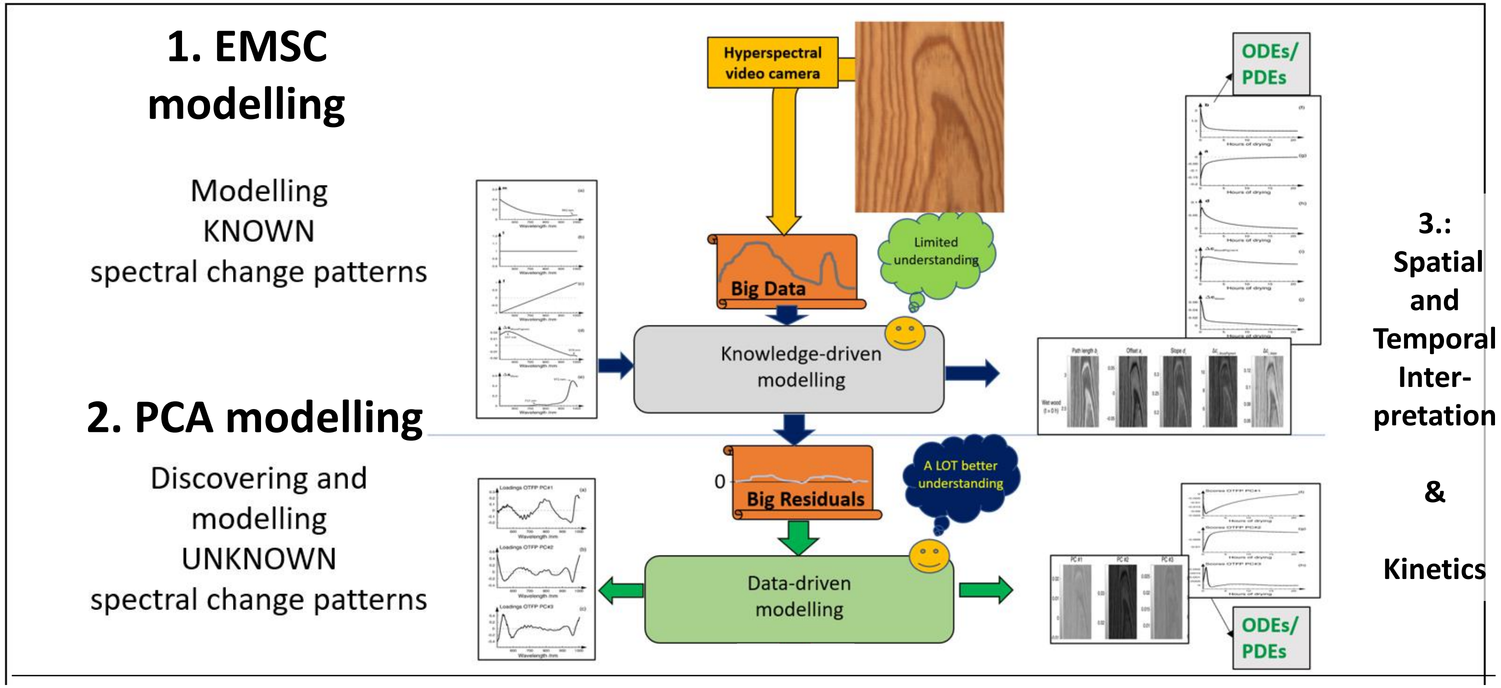
Norwegian University of Life Sciences

idletechs

Big Data: Hyperspectral «video»

A single piece of drying wood:

>350 000 000 VNIR reflectance spectra, 200 channels each:



200 wavelength channnels $\Rightarrow$ 8 change patterns: $\Rightarrow$ 99.8% NIRS variance explained

Norwegian University of Life Sciences

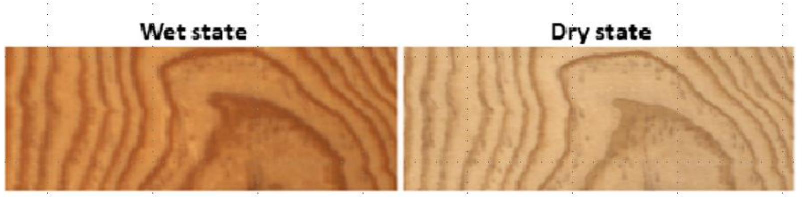
idletechs

Acknowledgements:

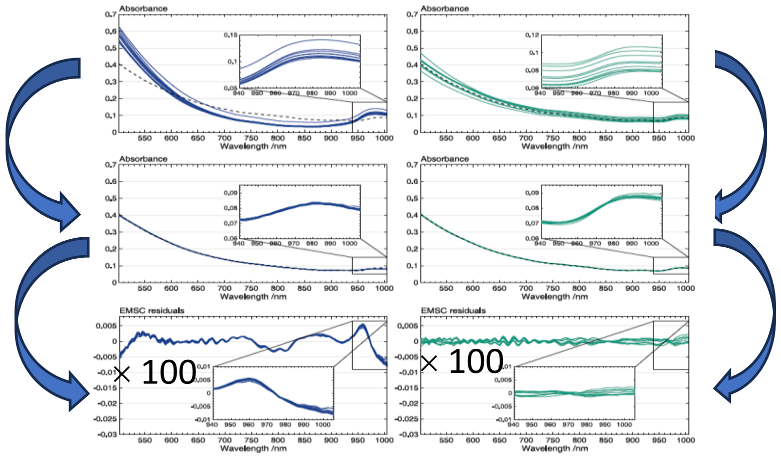
Ingunn Burud & Petter Stefansson, Norwegian University of Life Sciences NMBU, Ås, Norway:
Raffaele Vitale, U. Lille, France

Big Data: Hyperspectral «video»

A single piece of drying wood:
>350 000 000 VNIR reflectance spectra, 200 channels each:
Spectra x space x time



VNIR:400-1000 nm

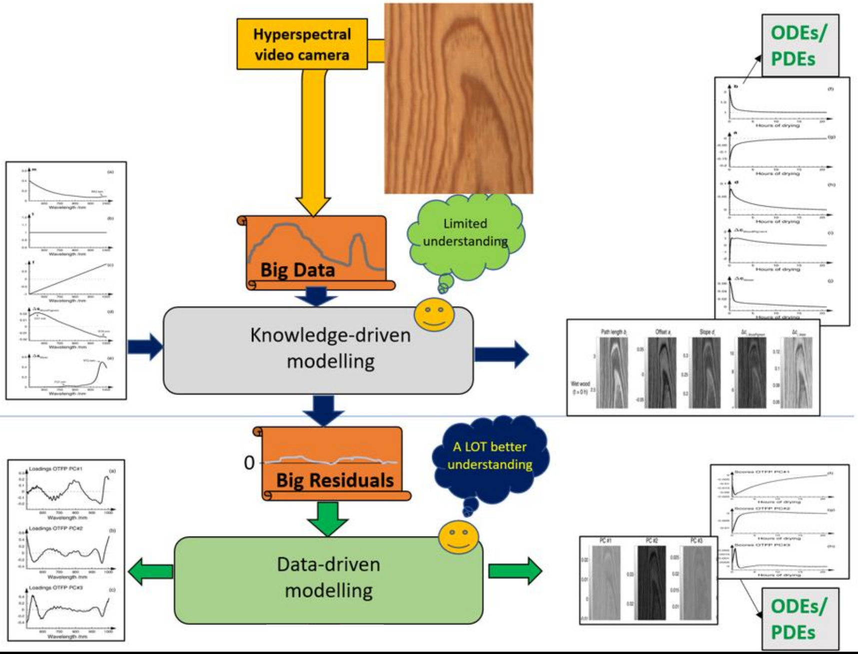


EMSC modelling

Modelling
KNOWN
spectral change patterns

2. PCA modelling

Discovering and
modelling
UNKNOWN
spectral change patterns



3.: Spatial and Temporal Inter-pretation

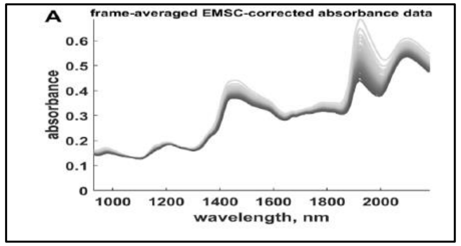
&

Kinetics

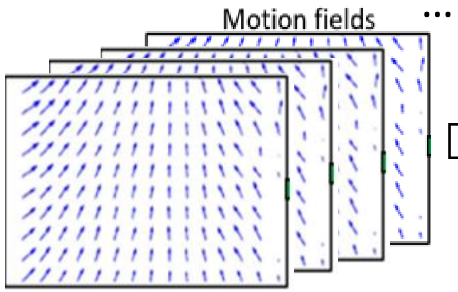
Acknowledgements:

Ingunn Burud & Petter Stefansson, Norwegian University of Life Sciences NMBU, Ås, Norway:
Raffaele Vitale, U. Lille, France

SWIR: 900-2500 nm



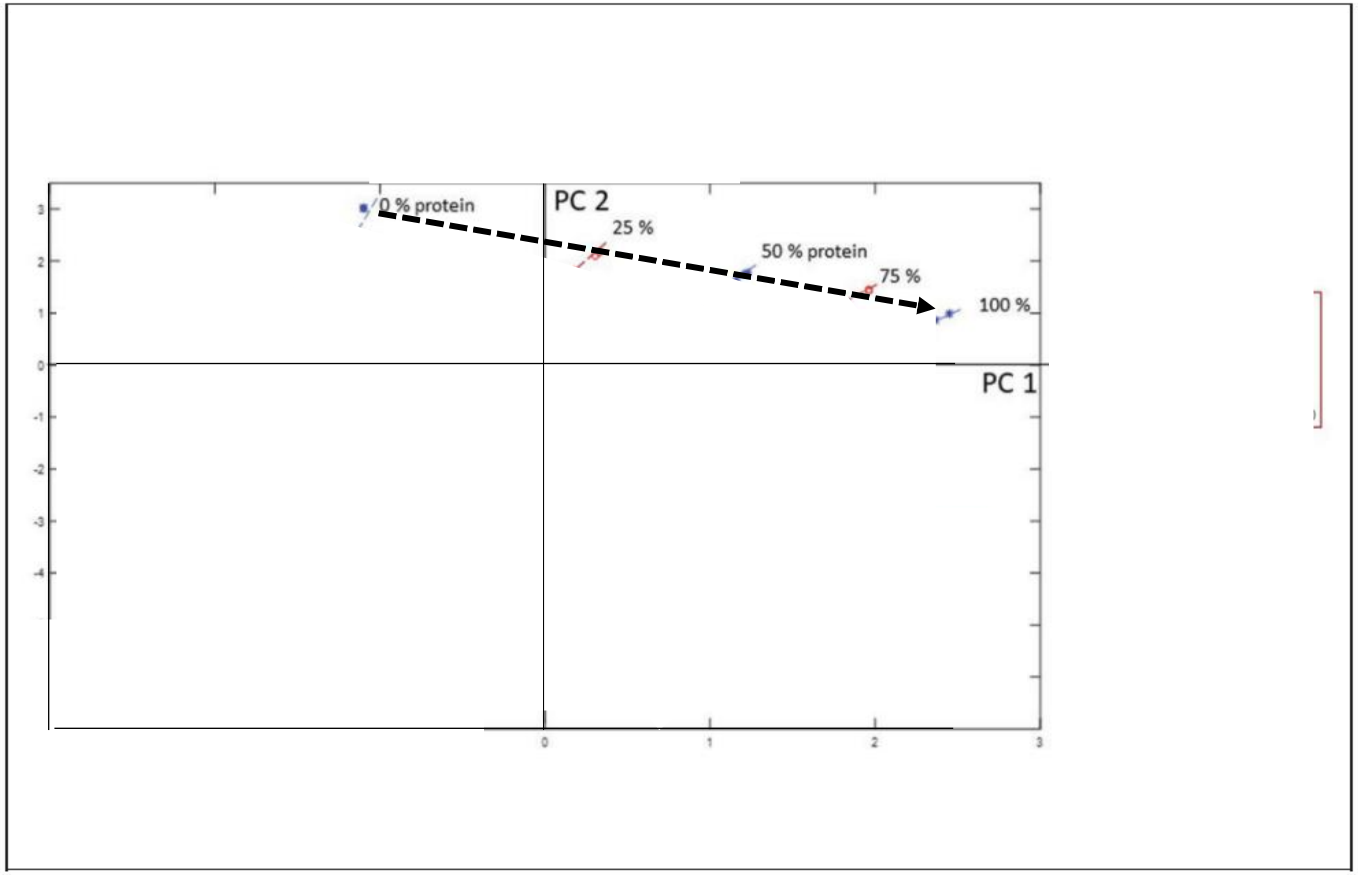
Discovering and
modelling
UNKNOWN
sensor motion and
sample shrinkage



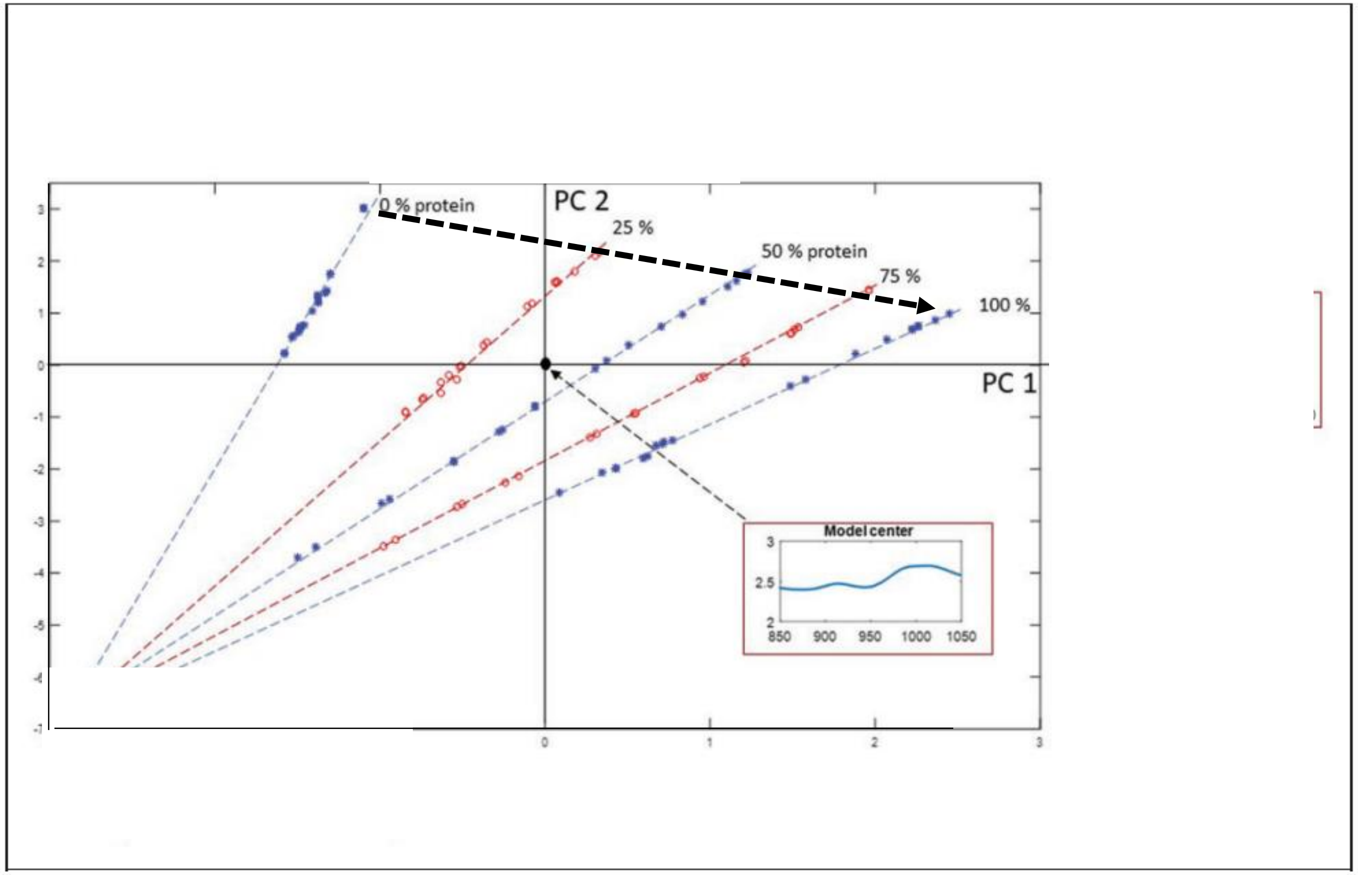
PCA

Example of sample perturbation experiment to learn how light interacts with one's sample type:

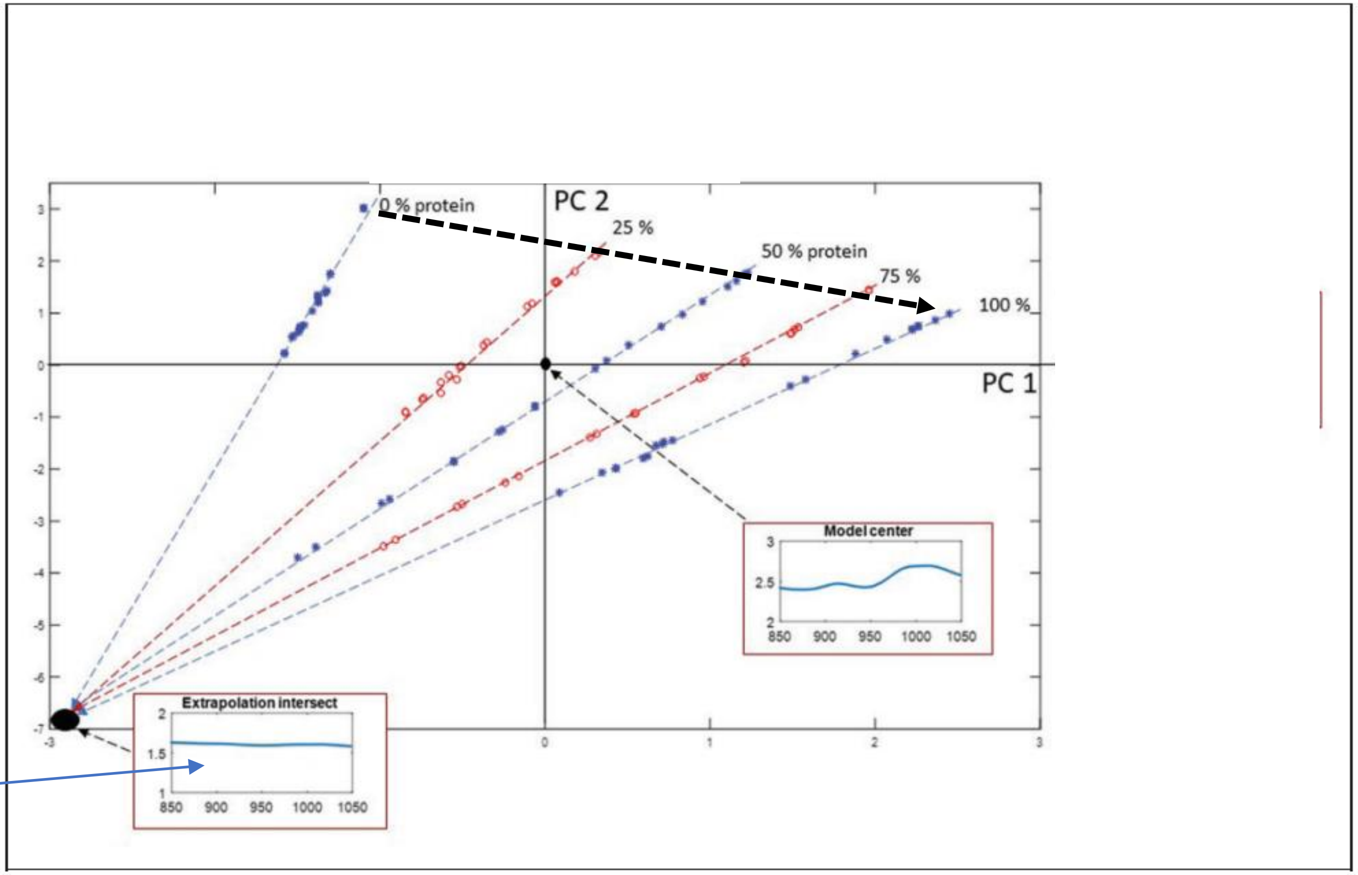
Different sample compositions change the absorbance



Different sample treatments change the effective optical path length in sample

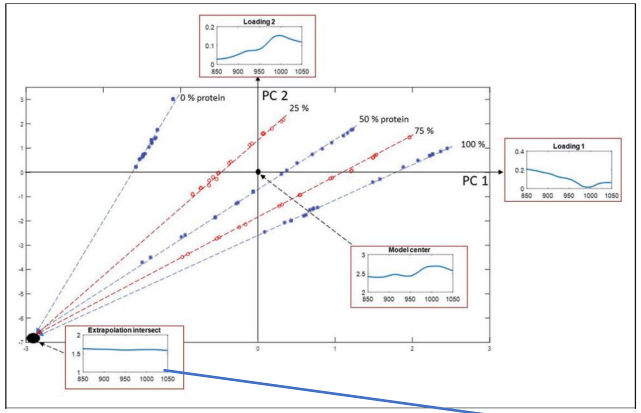


The intersect represents zero effective optical path length in sample

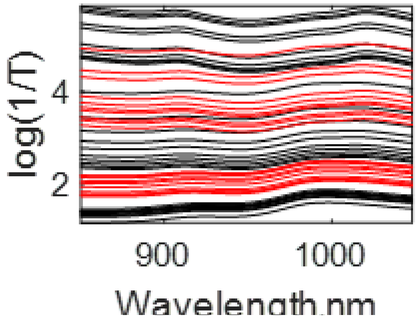


Estimating and subtracting this «instrument constant» makes interpretation much easier!

The intersect represents zero effective optical path length in sample

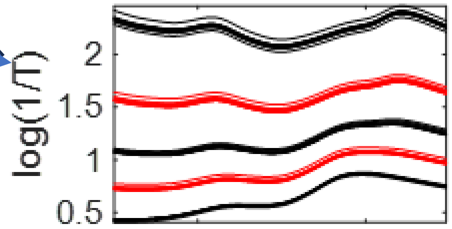


Scaled input data



Scaled input data

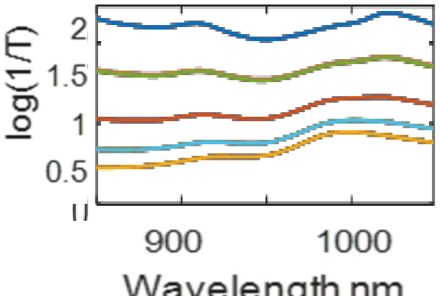
Conventional EMSC

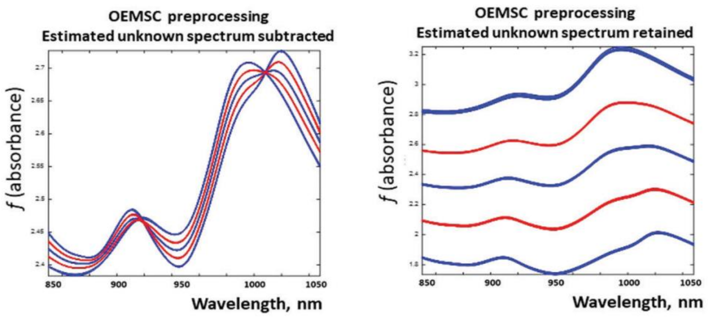


Subtraction of instrument absorbance offset spectrum

Further pre-processing

Scaled input data





$\Rightarrow$ If possible,
measure **each sample under different conditions** (path length, temperature, moisture content, TiO2 added, etc.). More info, more robust predictions.

What we do at NTNU

NTNU: Teach how to generate and use real-world Technical BIG DATA

NTNU:

- Education BDC
- Philosophyy of science, technical and societal cybernetic
- HYPSO flat earth society

Idletechs:

- Methods for modelling and visualization and warnings, linking instr. To scada
- Software for modelling, display, prediction classification outliers control, compression
- White label, proprietary. All standard protocols
  - Basic:

What we do at NTNU

- NTNU: Research and education
    - Teach students how to generate and use real-world Technical BIG DATA
        - Experimental design
        - Multivariate metamodeling to speed up complex mechanistic models
        - Various types of machine learning, including chemometric /NIR work-horses like EMSC, PCA and PLSR