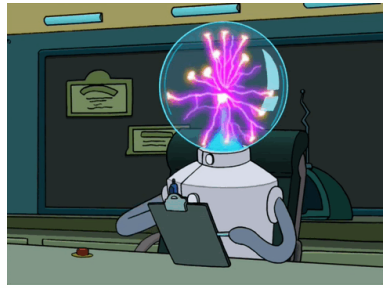
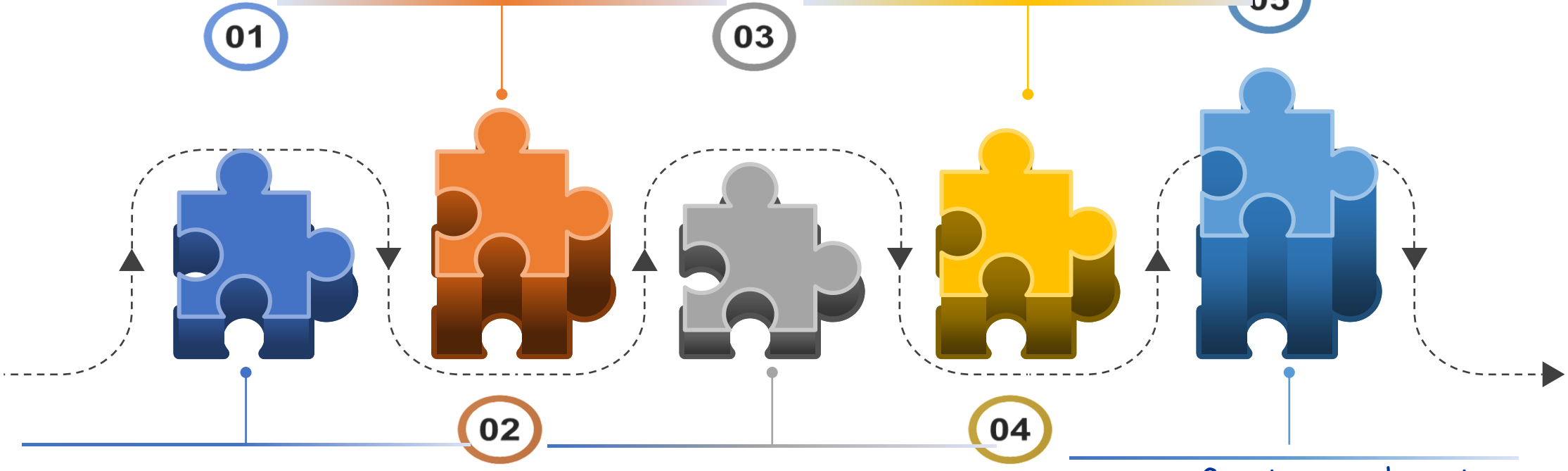
# Data mining, machine learning and data science: what is the interface with metrology?



Prof. Dr. Werickson F.C. Rocha
wfrocha@inmetro.gov.br

# Who Am I ?



01

Chemistry

UNICAMP 2001-2005

02

Master's degree in science

UNICAMP 2005-2007

03

Master's degree in engineering

UNICAMP 2007-2008

04

Phd in Science

UNICAMP 2007-2010

05

-Guest-reseacher at Nist (2014) in statistics

- Guest-reseacher at Nist (2015-2017) in Machine learning applied chemical and biochemical data


http://inmetro.gov.br

Researcher at Inmetro in 2010 and the main interest is in the use of machine learning methods in metrology, with special emphasis on pattern recognition analysis, classification and quantitative determinations using multivariate calibration.

# Topics

1) Some definitions

2) Data: new gold

3) Generating trusted data for machine learning

4) What does the data say?

5) Examples

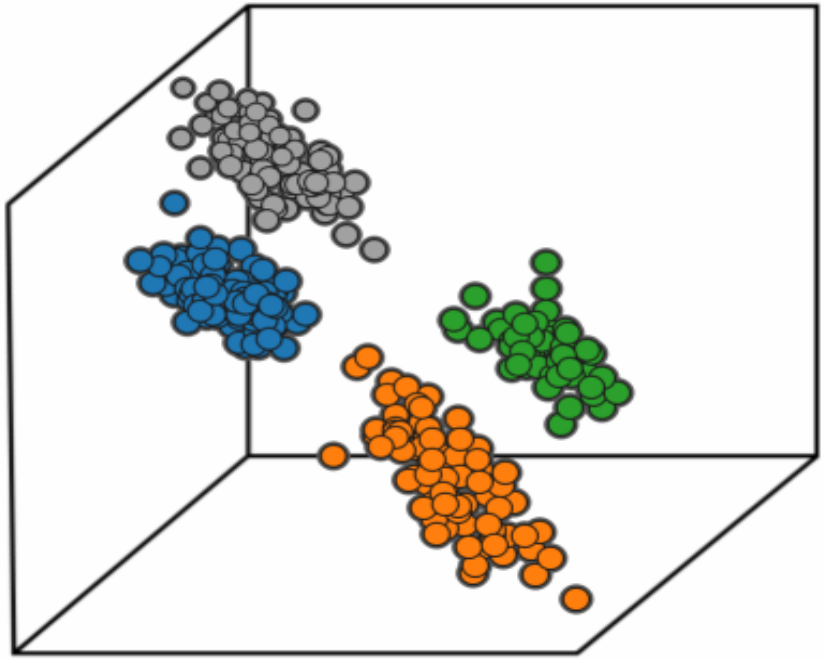
6) Conclusions

# 1) Definition: data mining

Data mining refers to the application of algorithms for extracting patterns from data*.

Cluster analysis

- PCA
- Kohonen



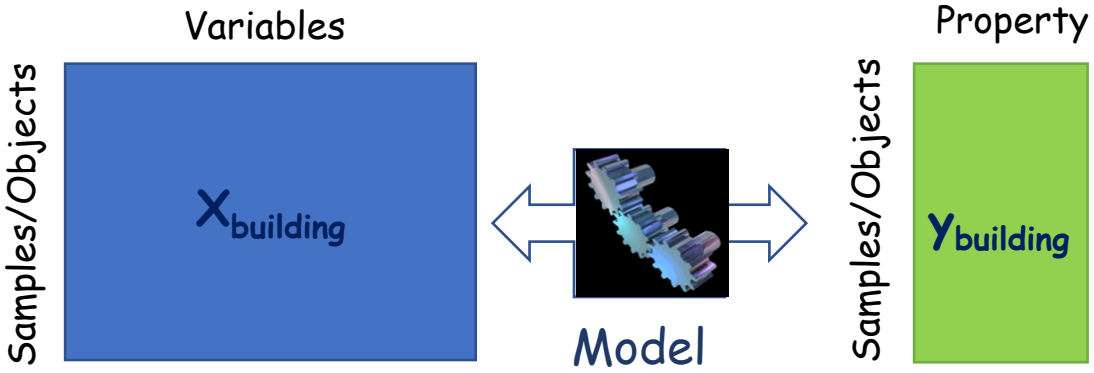
*Kulin, M.; Kazaz, T.;De Poorter, E.; Moerman, I. A Survey on Machine Learning-Based Performance Improvement of Wireless Networks: PHY, MAC and Network Layer. Electronics **2021**, 10, 318

# 1) Definition: machine learning

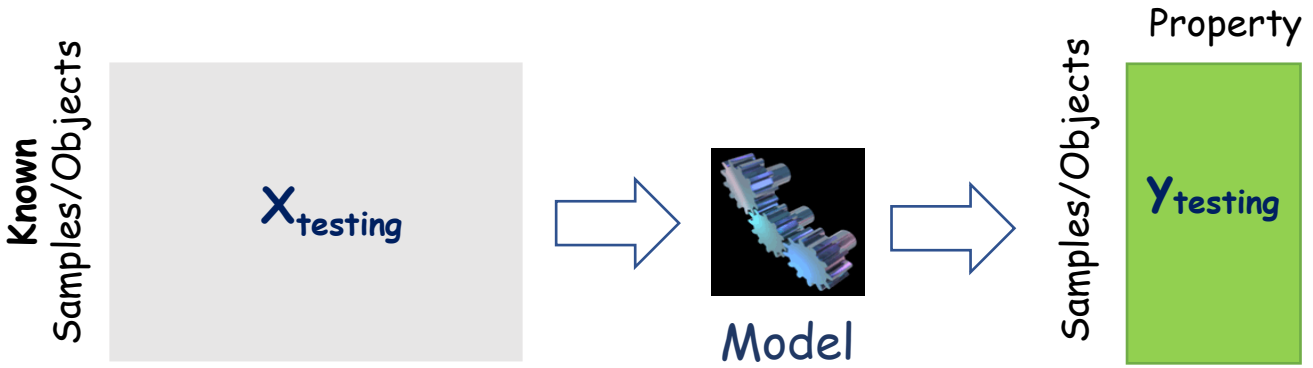
Machine Learning is normally defined as a series of methods that learn from the data to make or construct a model that can make informed decisions based on what is learned*.

learning process

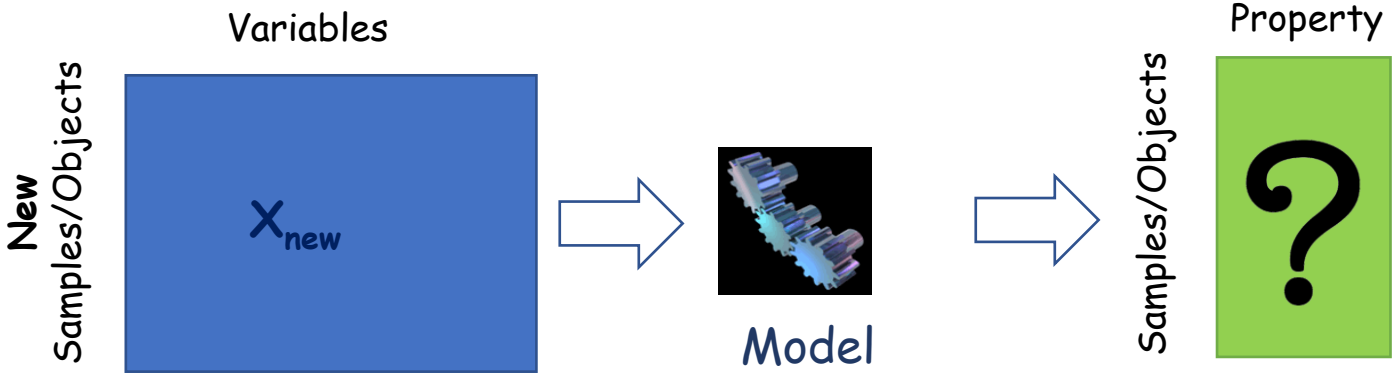
## a) Building the model




## b) Testing the model



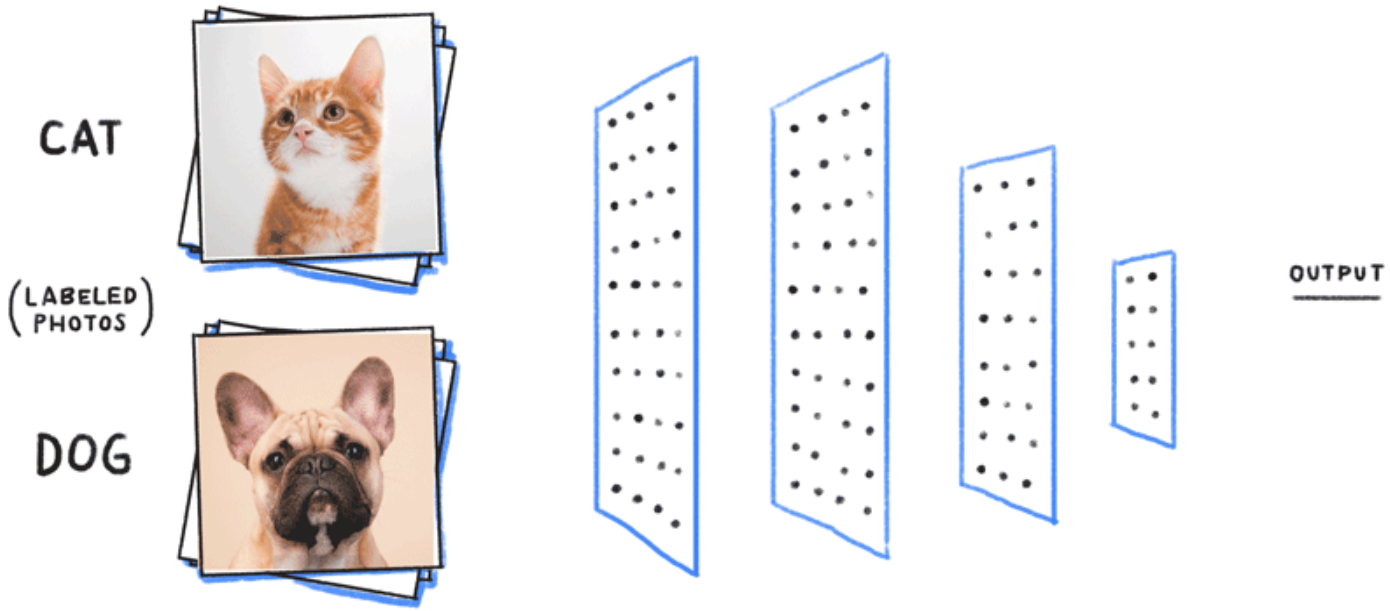

## c) Applying the model




*Amigo, J. M. **Data Mining, Machine Learning, Deep Learning, Chemometrics. Definitions, Common Points and Trends (Spoiler Alert: VALIDATE your models!)**. *Braz. J. Anal. Chem.*, 2021, *8* (32), pp 45-61. doi: http://dx.doi.org/10.30744/ brjac.2179-3425.AR-38-2021

# 1) Definition: deep learning

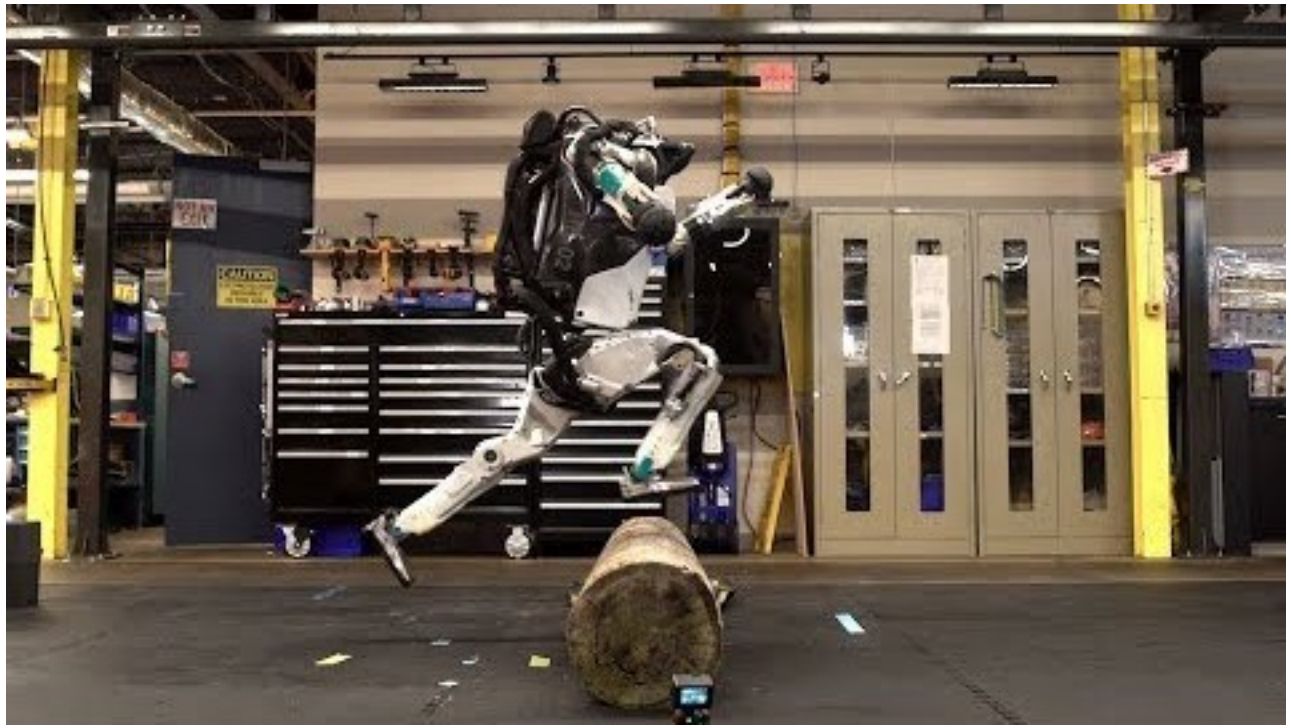
Deep learning is a subset of ML, in which data is passed via multiple number of non-linear transformations to calculate an output*.




*Kulin, M.; Kazaz, T.;De Poorter, E.; Moerman, I. A Survey on Machine Learning-Based Performance Improvement of Wireless Networks: PHY, MAC and Network Layer. Electronics **2021**, 10, 318

# 1) Definition: artificial intelligence

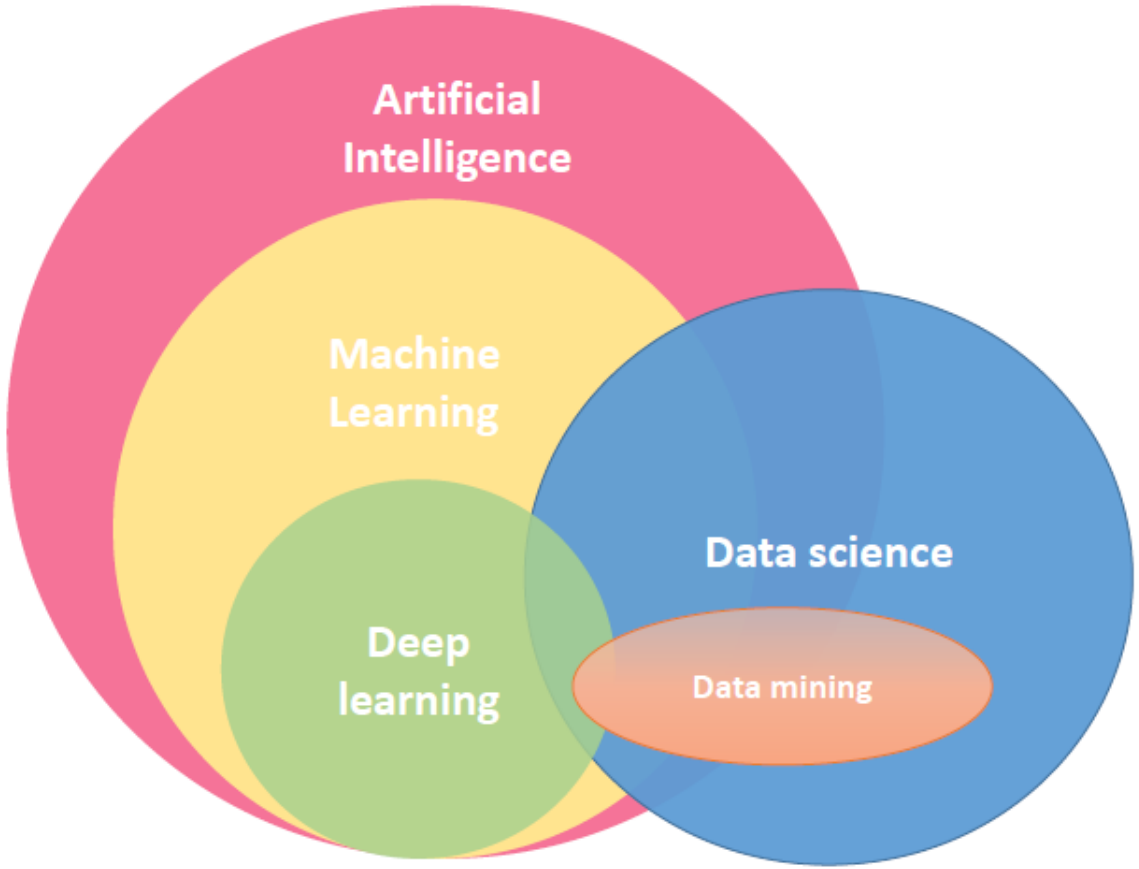
The science and engineering of making intelligent machines, especially computer systems by reproducing human intelligence through learning, reasoning and self-correction/adaption*.



*Kulin, M.; Kazaz, T.;De Poorter, E.; Moerman, I. A Survey on Machine Learning-Based Performance Improvement of Wireless Networks: PHY, MAC and Network Layer. Electronics **2021**, 10, 318

# 1) Definition: data science

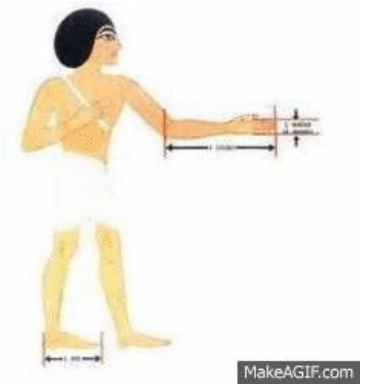
Data science is the study of the generalizable extraction of knowledge from data*.




*Kulin, M.; Kazaz, T.;De Poorter, E.; Moerman, I. A Survey on Machine Learning-Based Performance Improvement of Wireless Networks: PHY, MAC and Network Layer. Electronics **2021**, 10, 318

# 1) Definition: metrology

Metrology is “the science of measurement, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science and technology,” as defined by the International Bureau of Weights and Measures (BIPM, 2004).



# Topics

1) Some definitions

2) Data: new gold

3) Generating trusted data for machine learning

4) What the data says?

5) Examples

6) Conclusions

# 2) Data: new gold

Data is considered nowadays as the new gold.

-The volume of business data around the world duplicate every 1.2 years*.


*Fonte: https://www.bigdatanews.datasciencecentral.com/profiles/blogs/a-comprehensive-list-of-big-data-statistics

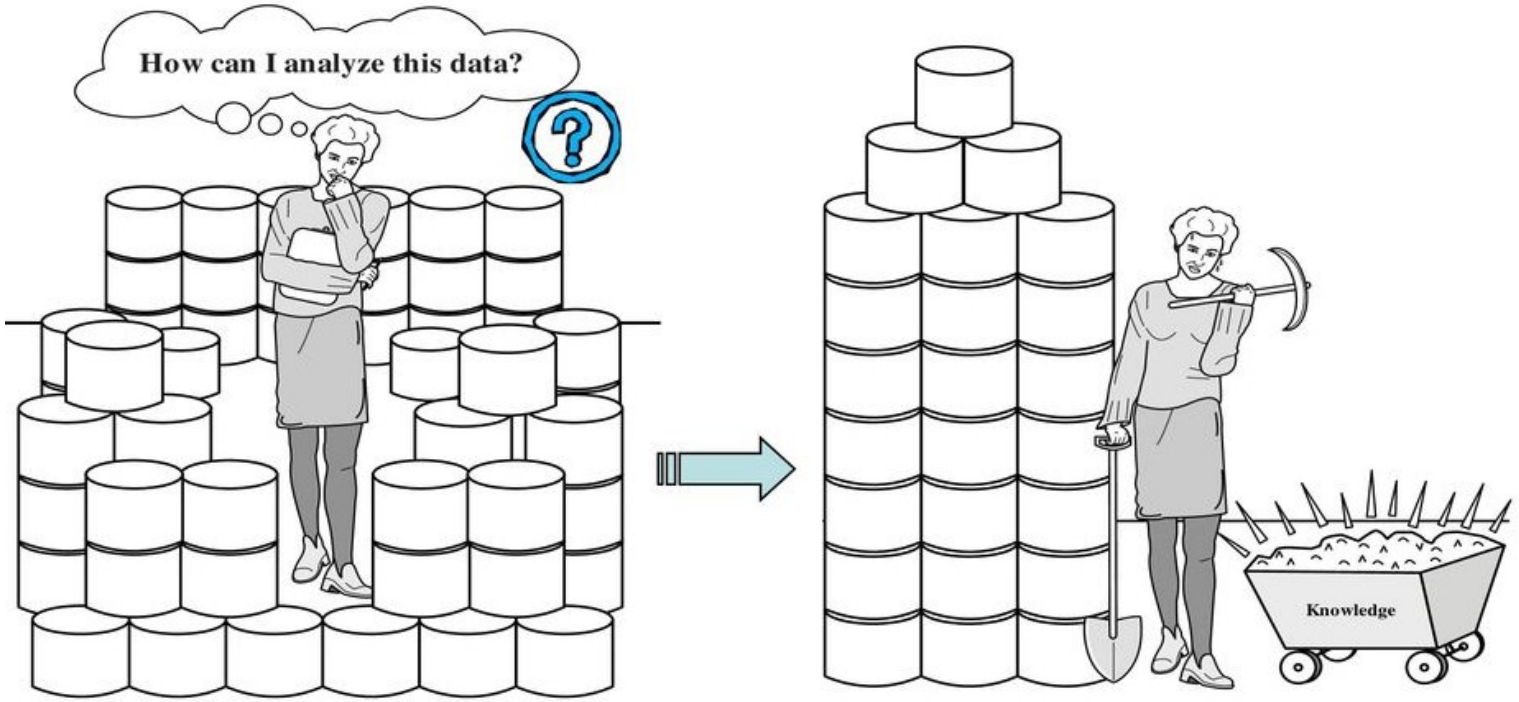
# 2) Data: new gold

- Farmers from **Iowa** to **India** are using data from seeds, satellites, sensors, and tractors to make better decisions about what to grow, when to plant, how to track food freshness from farm to fork, and how to adapt to changing climates**.


** **Source:** "What's the big deal with data", BSA,https://data.bsa.org/

# 2) Data: new gold

Reference data are assessed by experts and are trustworthy such that people can use the data with confidence and base significant decisions on the data.




We have rich data,
but poor information

Data mining-searching for knowledge
(interesting patterns) in your data.

HAN, J; KAMBER, M. Data Mining: Concepts and Techniques. Elsevier, 2006.

# Topics

1) Some definitions

2) Data: new gold

3) Generating trusted data for machine learning

4) What does the data say?

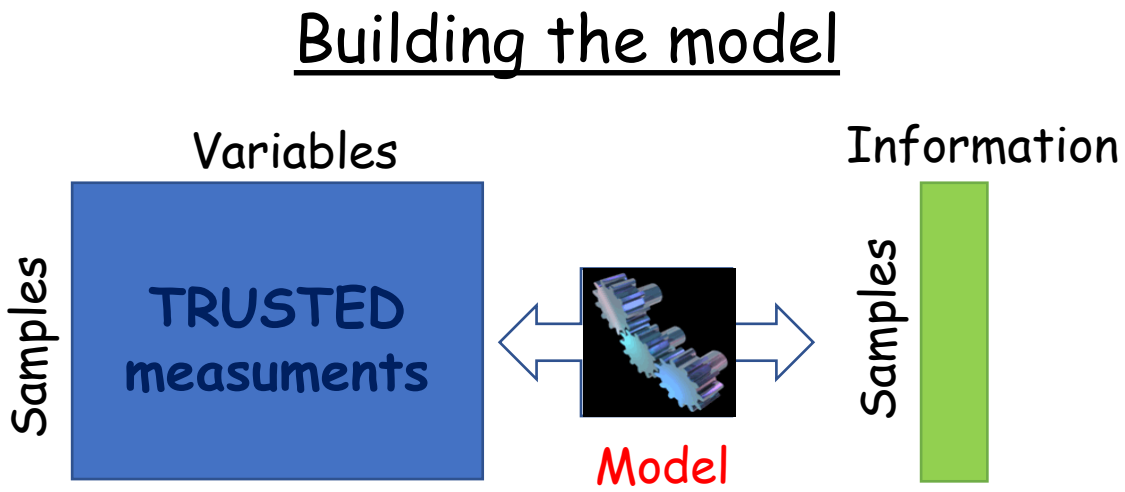
5) Examples

6) Conclusions

# 3) Generating trusted data for machine learning

How can I obtain these data ?

-Trusted measurements
-Appropriate instruments for the purpose
-Calibrated instruments
-Trained operador
- Reference material
- Standard reference material




Ensure the metrological reliability of the developed models.

It allows you to make a reliable and accurate decision.

# 3) Generating trusted data for machine learning

INMETRO

https://www.buzzfeednews.com/article/stephaniemlee/dan-ariely-honesty-study-retraction

BuzzFeed News

REPORTING TO YOU

SIGN IN ABOUT US GOT A TIP? SUPPORT US BUZZFEED.COM SECTIONS

TRENDING Texas Abortion Law Hurricane Ida R. Kelly Trial Leaving Afghanistan California Wildfires Miss Mercedes Morr

SCIENCE

## A Famous Honesty Researcher Is Retracting A Study Over Fake Data

Renowned psychologist Dan Ariely literally wrote the book on dishonesty. Now some are questioning whether the scientist himself is being dishonest.

**Stephanie M. Lee**
BuzzFeed News Reporter

# 3) Generating trusted data for machine learning

Instrumental variables
(Wavenumber,
retention time, etc)

Variables

Samples

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.66 | 0.90 | 0.78 | 0.28 | 0.72 | 0.83 | 0.43 | ... | 0.96 | 0.53 | 0.11 | 0.58 | 0.1 | | | | | | | |
| 0.58 | 0.43 | 0.05 | 0.39 | 0.67 | 0.79 | 0.03 | ... | 0.69 | 0.96 | 0.39 | 0.58 | 0.8 | | | | | | | |
| 0.06 | 0.69 | 0.96 | 0.39 | 0.58 | 0.85 | 0.93 | ... | 0.24 | 0.82 | 0.65 | 0.22 | 0.0 | | | | | | | |
| 0.74 | 0.20 | 0.64 | 0.83 | 0.35 | 0.28 | 0.11 | ... | 0.95 | 0.13 | 0.73 | 0.17 | 0.5 | | | | | | | |
| 0.28 | 0.48 | 0.57 | 0.24 | 0.16 | 0.26 | 0.27 | ... | 0.48 | 0.57 | 0.24 | 0.16 | 0.2 | | | | | | | |
| 0.44 | 0.27 | 0.14 | 0.12 | 0.29 | 0.17 | 0.48 | ... | 0.27 | 0.14 | 0.12 | 0.29 | 0.17 | 0.48 | 0.12 | ... | 0.17 | 0.48 | 0.44 | 0.27 | 0.14 |
| 0.18 | 0.95 | 0.13 | 0.73 | 0.17 | 0.53 | 0.79 | ... | 0.22 | 0.90 | 0.36 | 0.44 | 0.77 | 0.76 | 0.73 | ... | 0.53 | 0.79 | 0.49 | 0.22 | 0.90 |
| 0.09 | 0.24 | 0.82 | 0.65 | 0.22 | 0.07 | 0.26 | ... | 0.01 | 0.67 | 0.78 | 0.70 | 0.21 | 0.62 | 0.65 | ... | 0.07 | 0.26 | 0.57 | 0.01 | 0.67 |
| 0.90 | 0.95 | 0.50 | 0.20 | 0.67 | 1.00 | 0.21 | ... | 0.43 | 0.05 | 0.39 | 0.67 | 0.79 | 0.03 | 0.20 | ... | 1.00 | 0.21 | 0.58 | 0.43 | 0.05 |
| 0.99 | 0.48 | 0.66 | 0.85 | 0.40 | 0.75 | 0.28 | ... | 0.13 | 0.51 | 0.02 | 0.89 | 0.12 | 0.38 | 0.85 | ... | 0.75 | 0.28 | 0.59 | 0.13 | 0.51 |
| 0.59 | 0.13 | 0.51 | 0.02 | 0.89 | 0.12 | 0.38 | ... | 0.90 | 0.78 | 0.28 | 0.72 | 0.83 | 0.43 | 0.02 | ... | 0.12 | 0.38 | 0.66 | 0.90 | 0.78 |
| | | | | | | | | 0.86 | 0.12 | 0.19 | 0.34 | 0.43 | 0.93 | 0.11 | ... | 0.16 | 0.21 | 0.69 | 0.86 | 0.12 |
| | | | | | | | | 0.20 | 0.64 | 0.83 | 0.35 | 0.28 | 0.11 | 0.19 | ... | 0.43 | 0.93 | 0.74 | 0.20 | 0.64 |
| | | | | | | | | 0.95 | 0.50 | 0.20 | 0.67 | 1.00 | 0.21 | 0.36 | ... | 0.77 | 0.76 | 0.90 | 0.95 | 0.50 |
| | | | | | | | | 0.48 | 0.66 | 0.85 | 0.40 | 0.75 | 0.28 | 0.78 | ... | 0.21 | 0.62 | 0.99 | 0.48 | 0.66 |

Food
Chemicals
Environmental samples
Pharmaceutical samples

...

Metrology for Digital Transformation
SIM-MWG-14

# Topics

1) Some definitions

2) Data: new gold

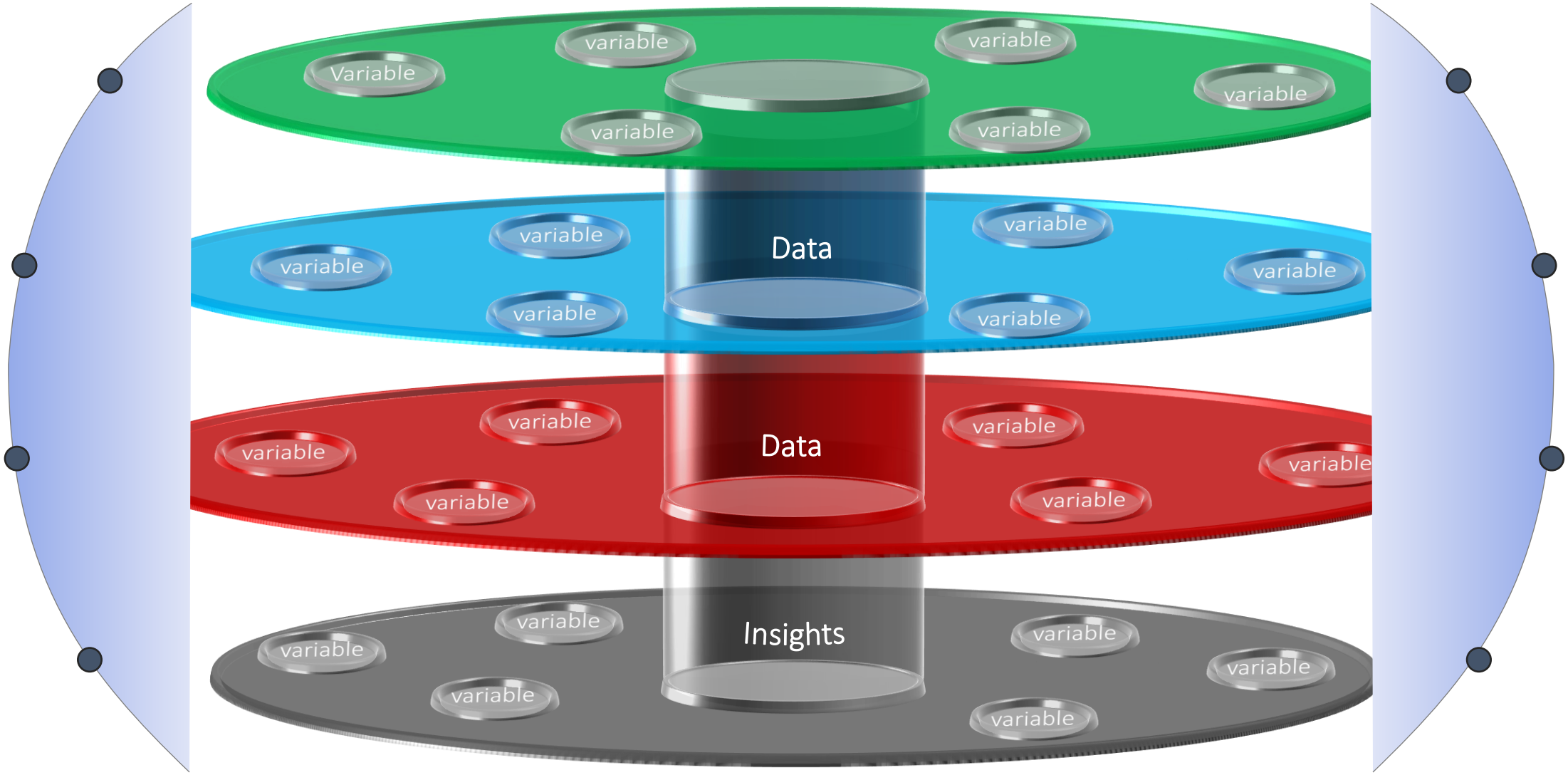
3) Generating trusted data for machine learning

4) What does the data say?

5) Examples

6) Conclusions

# 4) What does the data say ?




Metrology for Digital Transformation
SIM-MWG-14

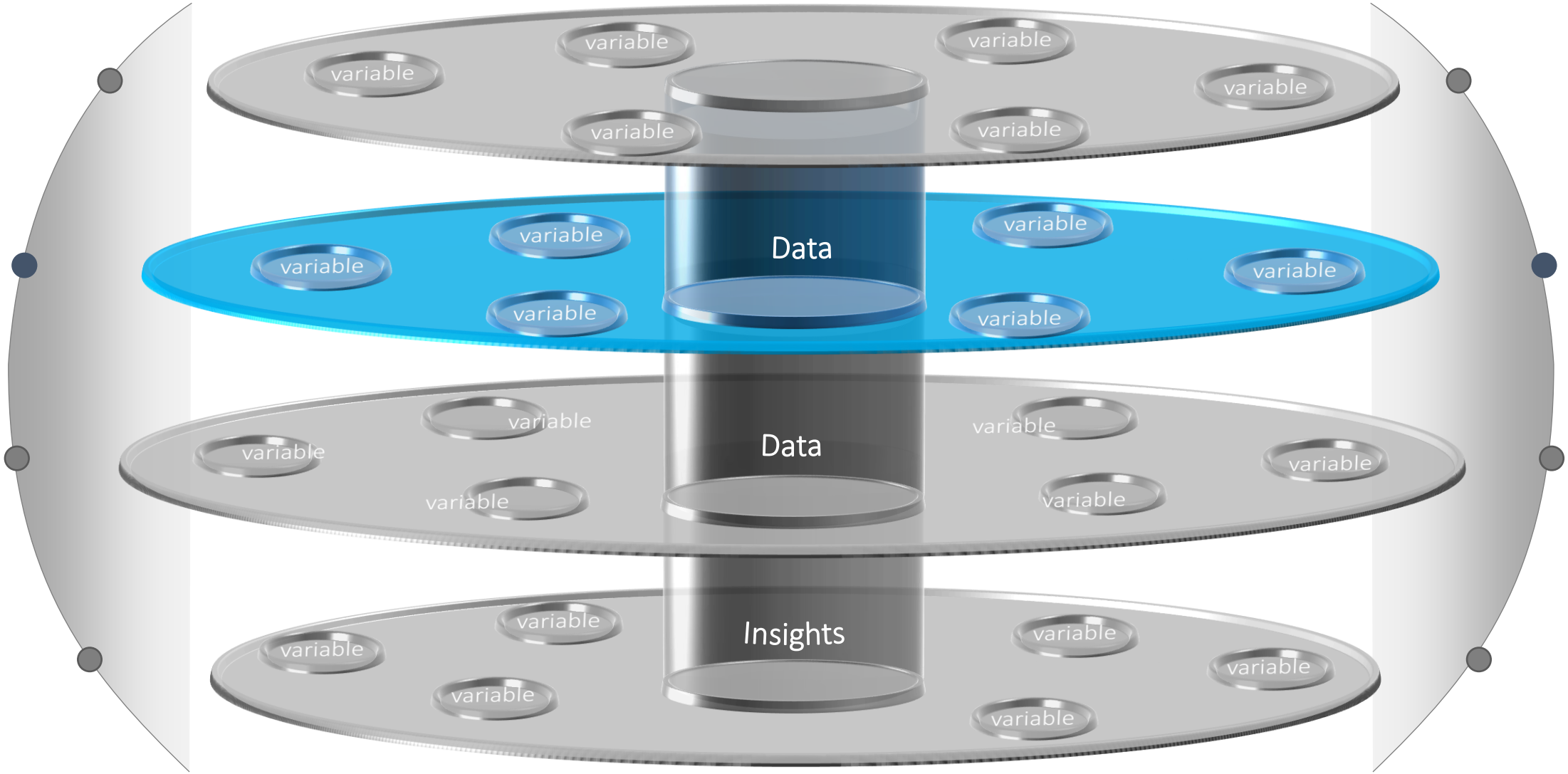
# 4) What does the data say ?

## a) Is the person diabetic ?


Metrology for Digital Transformation
SIM-MWG-14

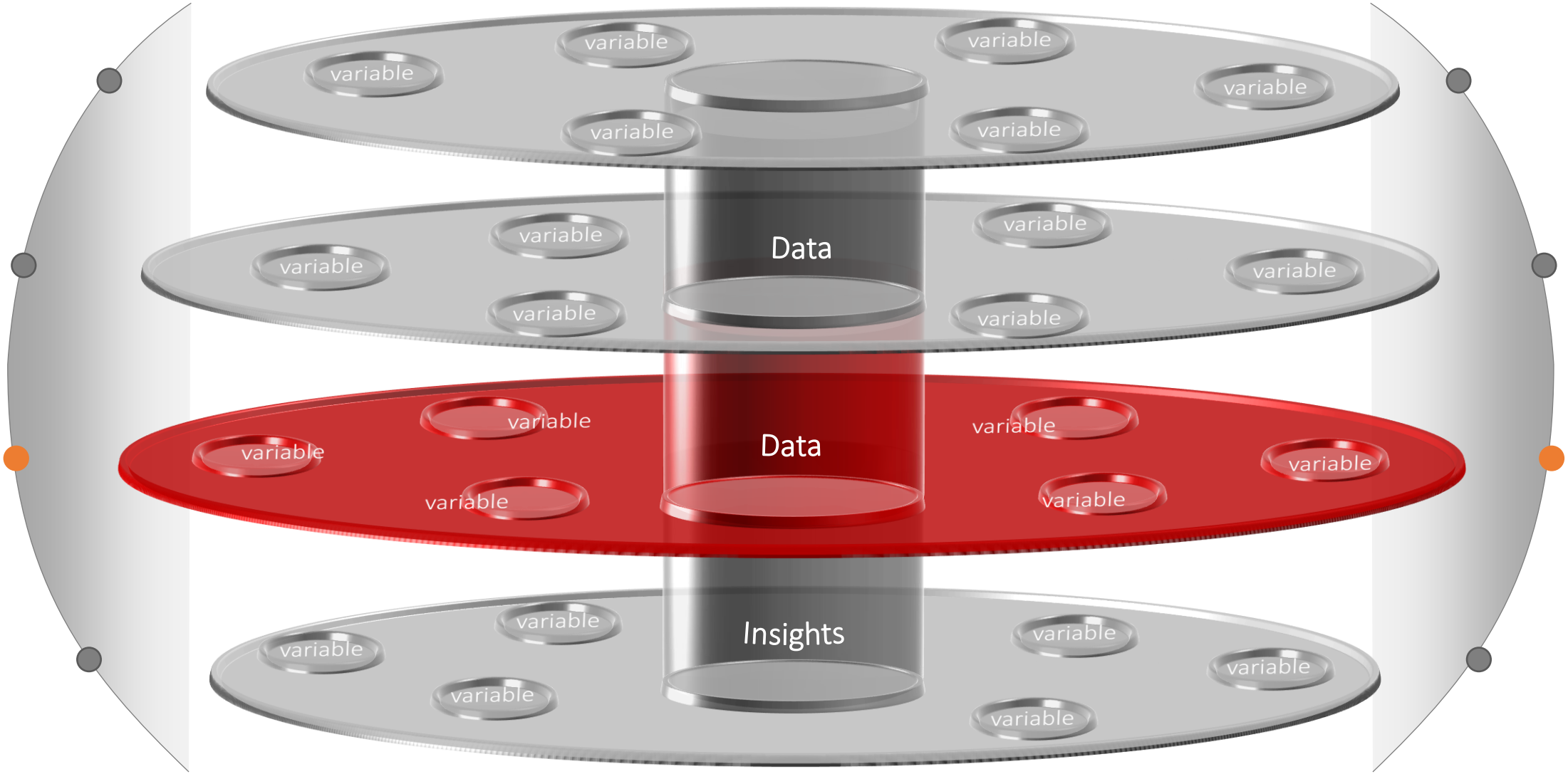
# 4) What does the data say ?

## b) How many soils are there?



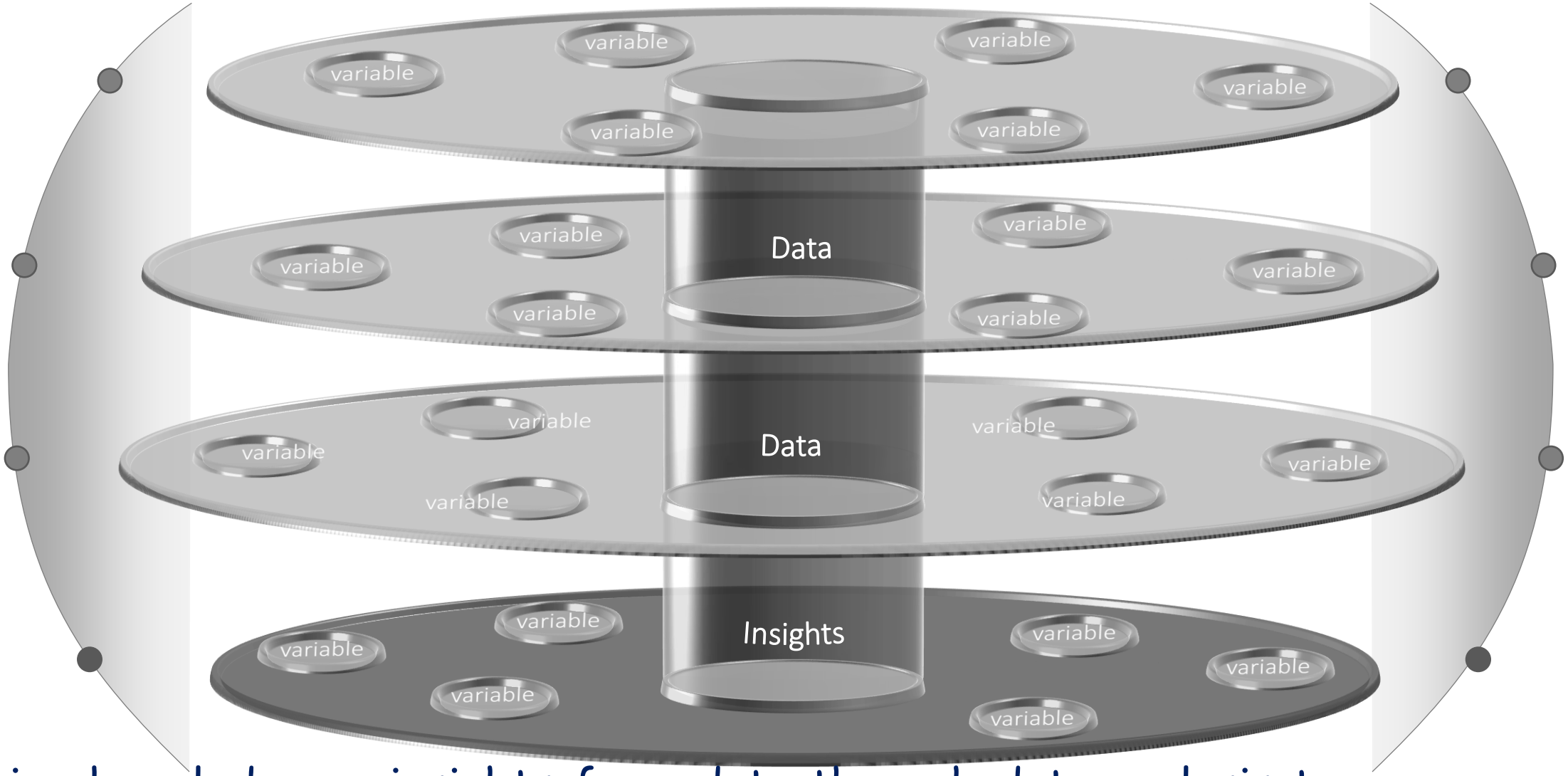
# 4) What does the data say ?

## c) How much sugar do I have in my soda?




Metrology for Digital Transformation
SIM-MWG-14

# 4) What does the data say ?




Extracting knowledge or insights from data through data analysis to answer questions related to the metrological area under study.

# Topics

1) Some definitions

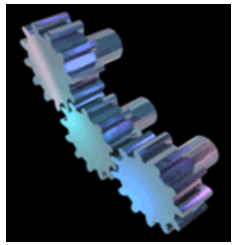
2) Data: new gold

3) Generating trusted data for machine learning

4) What does the data say?

5) Examples

6) Conclusions

1 

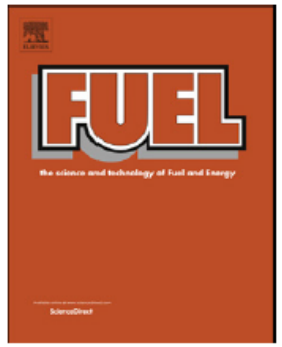
# Example

Fuel 243 (2019) 413–422

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel



Full Length Article

# Determination of physicochemical properties of petroleum derivatives and biodiesel using GC/MS and chemometric methods with uncertainty estimation

Werickson Fortunado de Carvalho Rocha[a], David A. Sheen[b,*]

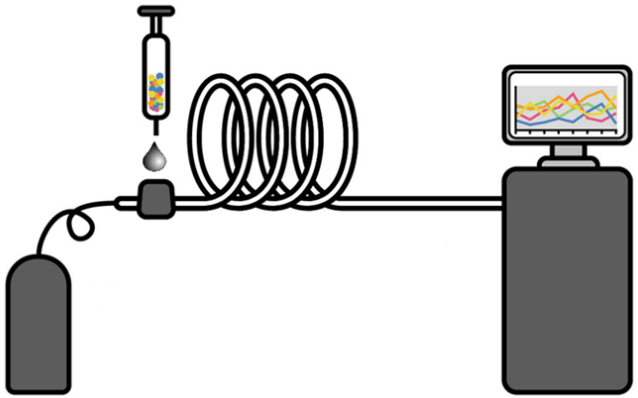
[a] National Institute of Metrology, Quality, and Technology-INMETRO, Division of Chemical Metrology, 25250-020 Duque de Caxias, RJ, Brazil
[b] Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

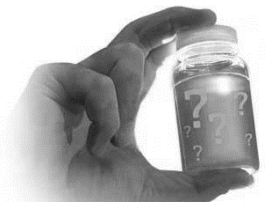
# Goal

The objective of this work is to build models with uncertainty estimation to predict the physicochemical properties of the fuels.

# Experimental data

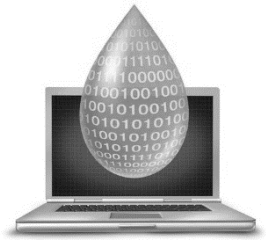
**Data acquisition by GC-MS**





**sample**

**standard ASTM method**



**Chemometric analysis with uncertainty estimation**

**Making decision**

**Physical-chemical properties of fuels with reliability**

Metrology for Digital Transformation
SIM-MWG-14

# Experimental data

**Table 1. Physicochemical properties and methods for sample characterization.**

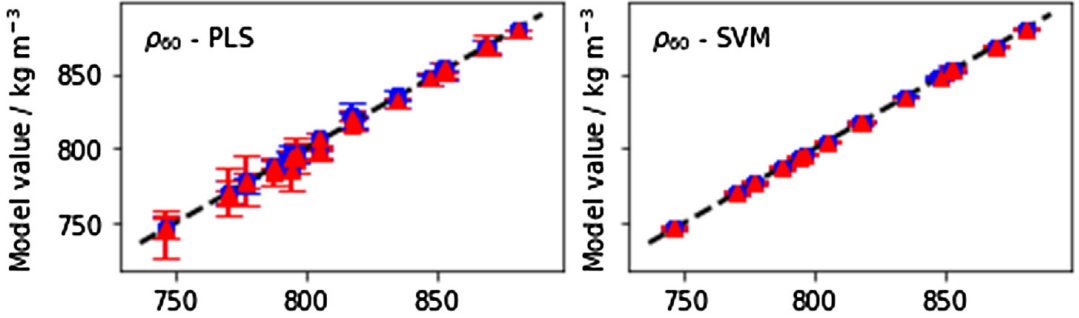
| Property | Symbol | ASTM Method |
|---|---|---|
| Density at 15°C (kg/m³) | $\rho_{15}$ | ASTM D4052 |
| Density at 60°C (kg/m³) | $\rho_{60}$ | ASTM D4052 |
| Kinematic viscosity at 40 °C (mm²/s) | $\nu_{40}$ | ASTM D445 |
| Kinematic viscosity at 60 °C (mm²/s) | $\nu_{60}$ | ASTM D445 |
| Pour point (°C) | $T_p$ | ASTM D5949 |
| Acid number (mg KOH/g) | $N_A$ | ASTM D664 |
| Base number (mg KOH/g) | $N_B$ | ASTM D4739 |

**Table 2. Main characteristics and properties of samples used in this study. The values are an average of duplicate measurements.**

| Name | Type | $\rho_{15}$ | $\rho_{60}$ | $\nu_{40}$ | $\nu_{60}$ | $T_p$ | $N_A$ | $N_B$ |
|---|---|---|---|---|---|---|---|---|
| 87 Octane Gasoline | Gasoline | 743.55 | ----- | 0.2704 | ----- | -69 | 0.15 | 1.4 |
| Jet Fuel A | Kerosene | 803 | 769.7 | 1.265 | 0.7304 | -58.5 | 0.175 | 1.45 |
| JP5 | Kerosene | 826.3 | 793.8 | 1.555 | 0.6244 | -66 | 0.18 | 1.1 |
| JP8 | Kerosene | 779.7 | 746 | 1.079 | 0.652 | -63 | 0.18 | **1.95** |
| SRM 1616b | Kerosene | 827.6 | 794.2 | 1.578 | 2.557 | -55.5 | 0.16 | 0.9 |
| SRM 1617b | Gasoline | 809.1 | 777 | 1.401 | 1.487 | -58.5 | 0.225 | 1.55 |
| SRM 1620c | Residual Oil | ----- | ----- | ----- | ----- | ----- | 0.52 | 2.6 |
| SRM 1623c | Residual Oil | 902.2 | 868.9 | 5.239 | 1.968 | -3 | 0.31 | 1.1 |
| SRM 1624d | Diesel Oil | 849.7 | 818.1 | 3.179 | 2.308 | -9 | 0.29 | 1.05 |
| SRM 2721 | Crude Oil | 881.25 | 834.8 | 9.158 | 2.793 | -70 | 0.645 | 1.35 |
| SRM 2722 | Crude Oil | 911 | 880.8 | 8.584 | 2.324 | -9 | 0.44 | 1.05 |
| SRM 2723b | Diesel Oil | 848.8 | 817 | 2.763 | 1.14 | -36 | 0.23 | **1.95** |
| SRM 2770 | Diesel Oil | 818.6 | 787.3 | 3.128 | 2.3095 | -16.5 | 0.17 | 0.9 |
| SRM 2771 | Diesel Oil | 835.1 | 804.95 | 3.054 | 2.72 | -54 | 0.09 | 1.1 |
| SRM 2772 | Biodiesel | 885 | 852.5 | 2.55 | ----- | 0 | 0.18 | 0.8 |
| SRM 2773 | Biodiesel | 879.9 | 847.3 | 4.414 | 1.73 | 9 | 0.43 | 1.15 |
| SRM 2779 | Crude Oil | 844.35 | 796.3 | 2.461 | 2.712 | -60 | 0.16 | 1.25 |

# Results and discussion

# Performance of the PLS and SVM regression models for selected properties.



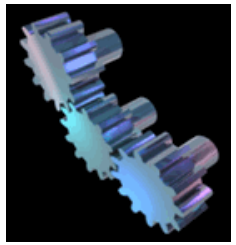

| Property | Symbol | ASTM Method |
|---|---|---|
| Density at 60°C (kg/m³) | $\rho_{60}$ | ASTM D4052 |
| Kinematic viscosity at 60 °C (mm²/s) | $v_{60}$ | ASTM D445 |
| Pour point (°C) | $T_p$ | ASTM D5949 |
| Acid number (mg KOH/g) | $N_A$ | ASTM D664 |

# Conclusions

In this work, linear and non-linear methods with uncertainty were used to estimate physicochemical properties of fuels using GC-MS data.

The properties were measured using ASTM standard methods, and partial least squares and support vector machines were used to determine a relation between the GC-MS data and the measured properties.

2



# Example

ELSEVIER

Fuel

Volume 197, 1 June 2017, Pages 248-258

Full Length Article

# Unsupervised classification of petroleum Certified Reference Materials and other fuels by chemometric analysis of gas chromatography-mass spectrometry data

Werickson Fortunato de Carvalho Rocha [a], Michele M. Schantz [b], David A. Sheen [b], Pamela M. Chu [b], Katrice A. Lippa [b]

Metrology for Digital Transformation
SIM-MWG-14

# Goal

- The objective of this work is to build chemometric models for unsupervised classification of transportation fuels using GC-MS.

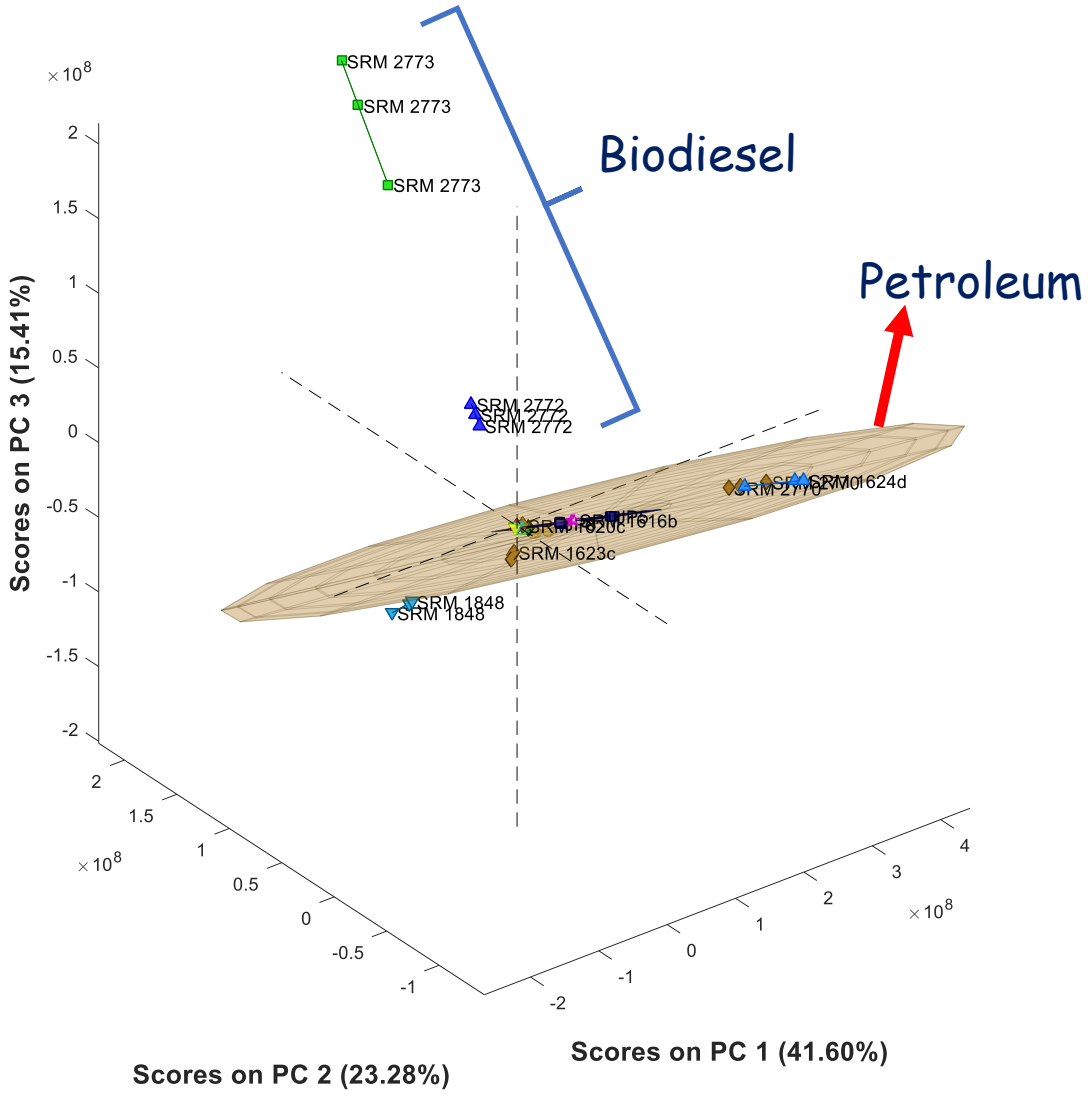
# Experimental part: GC-MS analysis

**Table 1.** List of Materials Analyzed in This Study

| SRM # (if applicable) | Identification | Description |
|---|---|---|
| SRM 1615 | Gas Oil | |
| SRM 1616b | Sulfur in Kerosene (Low-Level) | Special low sulfur kerosene (No.1-K) for nonflue-connected applications |
| SRM 1617b | Sulfur in Kerosene (High-Level) | High sulfur kerosene |
| SRM 1620c | Sulfur in Residual Fuel Oil (4 %) | Commercial “No. 6” residual fuel oil |
| SRM 1623c | Sulfur in Residual Fuel Oil (0.3 %) | Commercial “No. 4 (light)” residual fuel oil |
| SRM 1624d | Sulfur in Diesel | Commercial “No. 2 D” distillate fuel oil |
| SRM 1848 | Motor Oil Additive | Additive used in manufacture of lubricating oil for gasoline engines |
| SRM 2299 | Sulfur in Gasoline | Commercial reformulated unleaded gasoline |
| SRM 2721 | Crude Oil (Light-Sour) | Light-sour Texas crude oil |
| SRM 2722 | Crude Oil (Heavy-Sweet) | Heavy-sweet Texas crude oil |
| SRM 2723b | Low Sulfur Diesel | Commercial “No. 2 D” distillate fuel oil |
| SRM 2770 | Sulfur in Diesel Fuel Oil | Commercial “No. 2 D” distillate fuel oil |
| SRM 2771 | Zero Sulfur Diesel | Commercial diesel fuel blend stock |
| SRM 2772 | Biodiesel (Soy-Based) | Commercial 100 % biodiesel produced from soy |
| SRM 2773 | Biodiesel (Animal-Based) | Commercial 100 % biodiesel produced from animal products |
| SRM 2779 | Gulf of Mexico Crude Oil | Collected from 2010 Deepwater Horizon oil site |
| Gasoline | Commercial 87-octane gasoline sold in 2015 | |
| Jet A | Jet fuel from Air Force Research Laboratory (AFRL) | |
| JP5 | Jet fuel from AFRL | |
| JP8 | Jet fuel from AFRL | |

# MPCA

# MPCA Scores




These are **Kerosene-based fuels** and consist mostly of **aliphatic** and **aromatic hydrocarbons** of length C8–C17+.

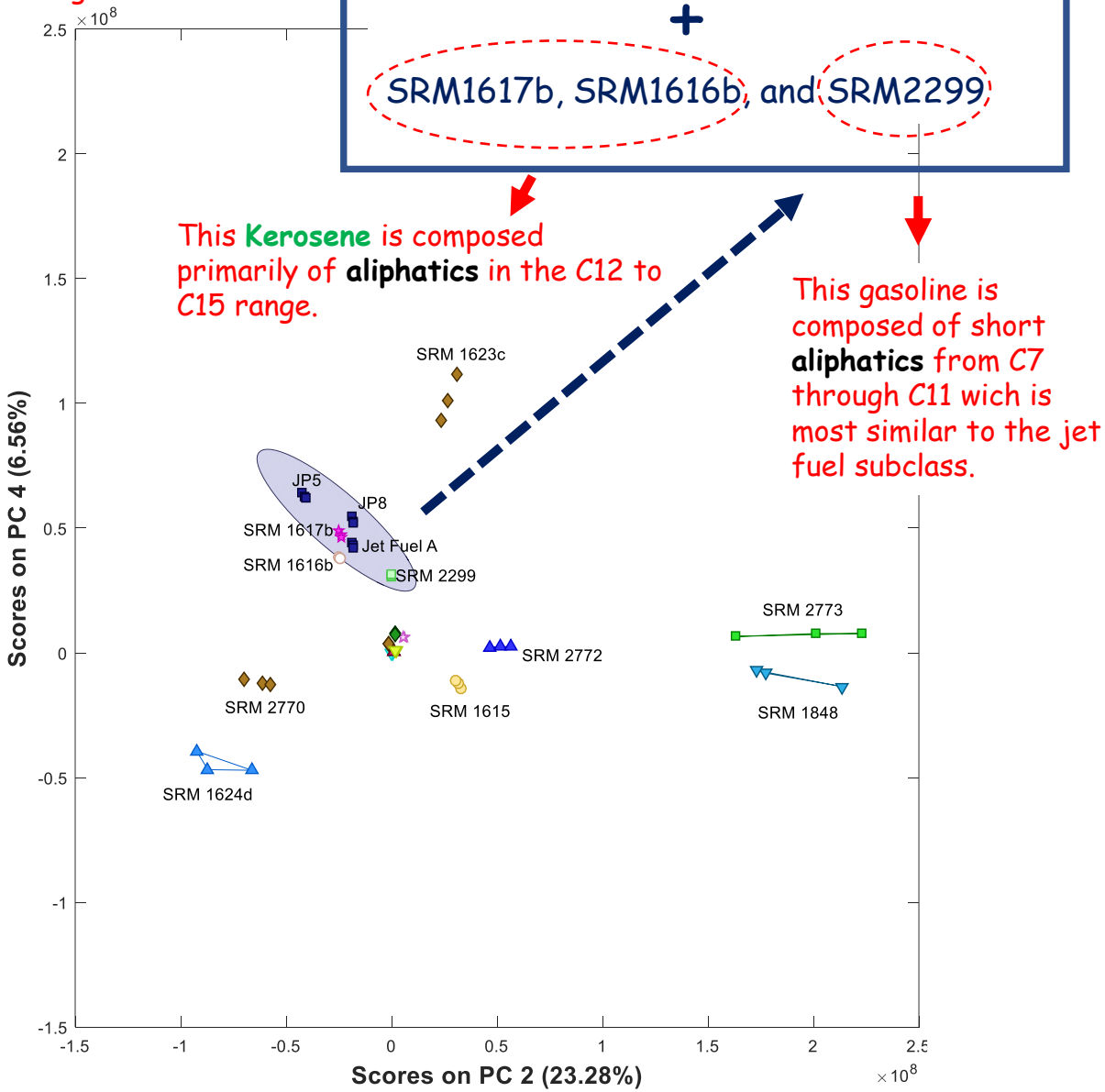
JET FUEL class: JP5, JP8, and jet fuel A

+

SRM1617b, SRM1616b, and SRM2299

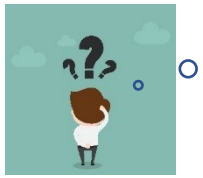
This **Kerosene** is composed primarily of **aliphatics** in the C12 to C15 range.

This gasoline is composed of short **aliphatics** from C7 through C11 wich is most similar to the jet fuel subclass.



# Kohonen

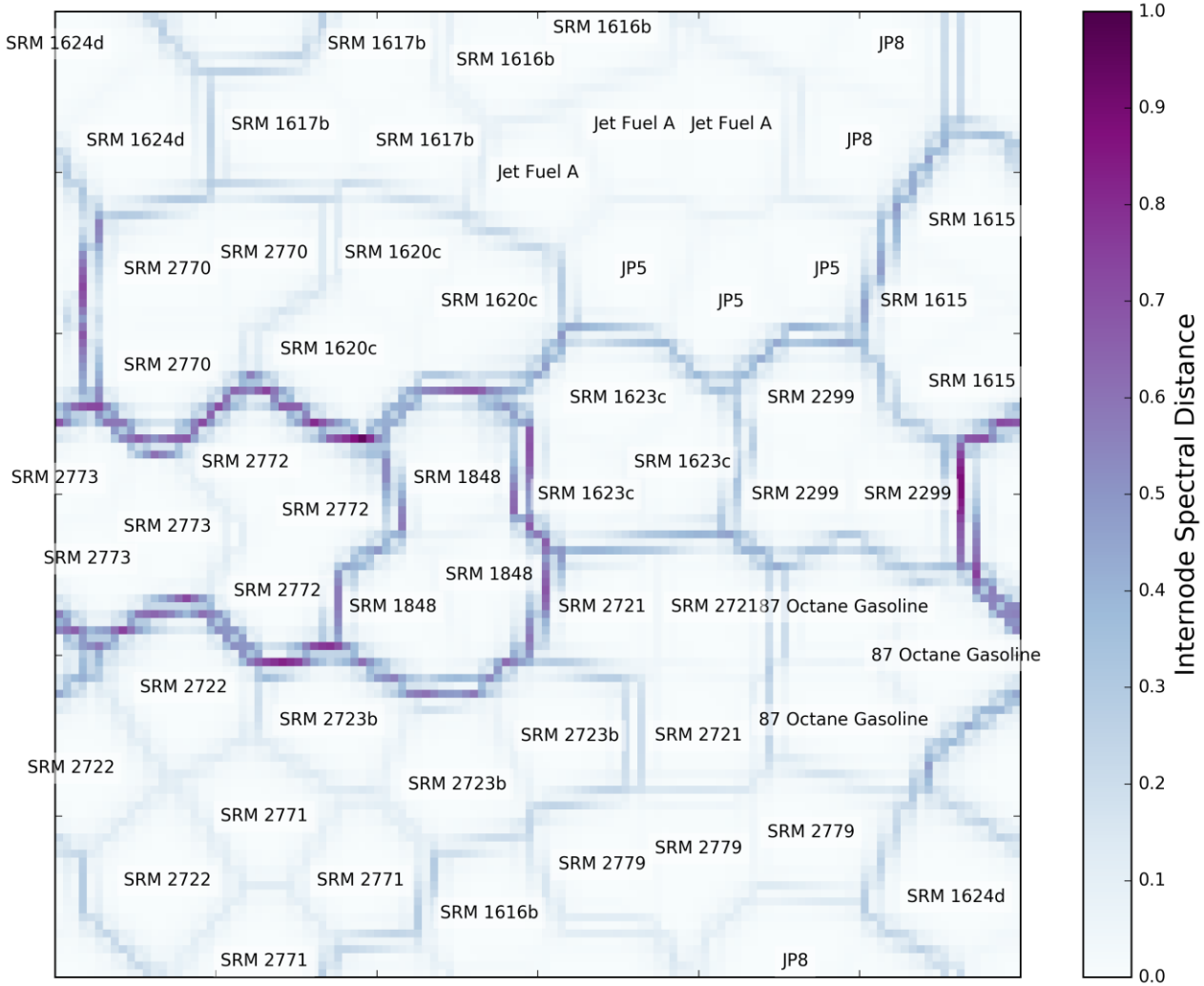
How many classes do we have ?



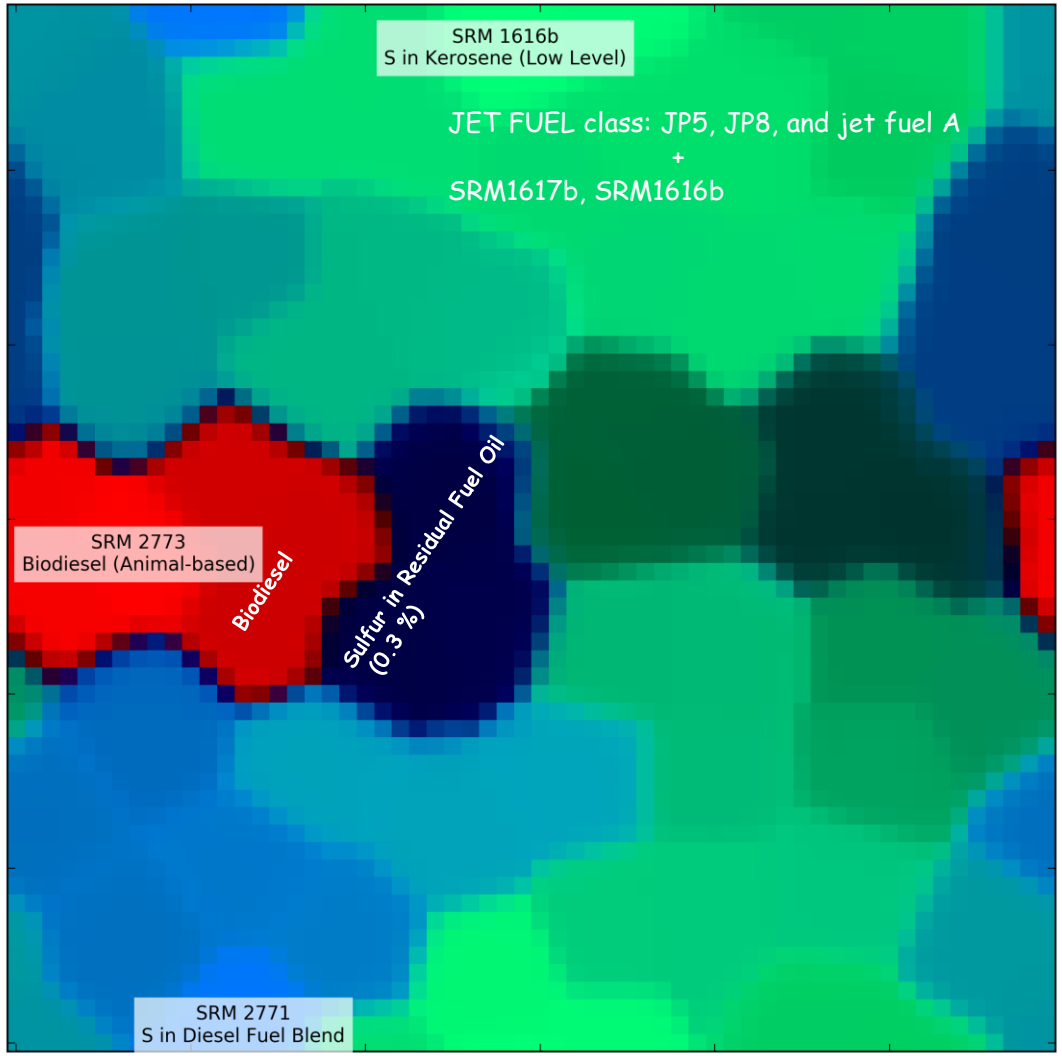
Minimum 3

RGB

# U-matrix



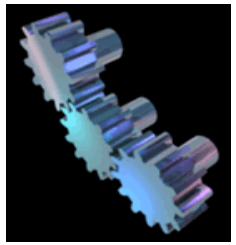

# Distance-matrix



# Conclusion

The results show that the combination of GC×MS and chemometric analysis can be employed as a general tool for the differentiation of petroleum Standard Reference Materials and other fuels. This procedure was tested with many fuels class. Also, many classes were properly differentiated through pattern recognition by MPCA and Kohonen.

3



# Example

Accred Qual Assur (2011) 16:523–528
DOI 10.1007/s00769-011-0807-9

PRACTITIONER'S REPORT

## Use of multivariate statistical analysis to evaluate experimental results for certification of two pharmaceutical reference materials

**Werickson Fortunato de Carvalho Rocha · Raquel Nogueira**

Received: 12 May 2011 / Accepted: 5 July 2011 / Published online: 22 July 2011

In this paper, we purpose the use of the multivariate statistical analysis for an easy and quick evaluation of the homogeneity test data, followed by selection of the replicates to be considered for $u_{bb}$ calculation. It should be noted that the paper has also the intention to encourage metrologists to apply chemometric techniques on their activities, for instance at different stages of the production and certification of reference materials, and also in proficiency testing. Few publications exist in this field, from

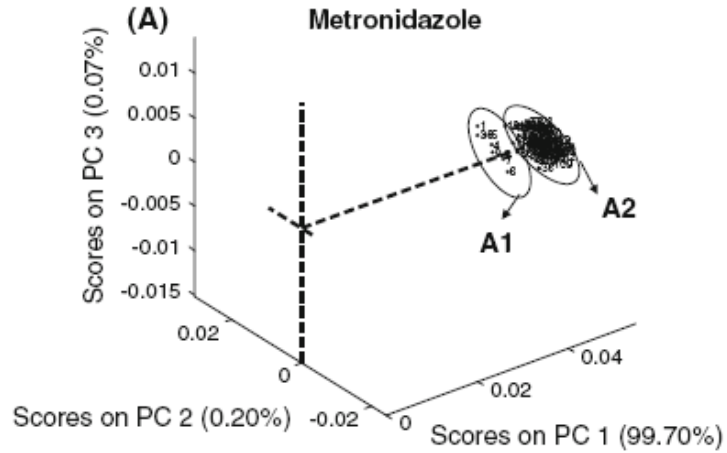



**Fig. 1** PCA score plots for the organic impurities content of the pharmaceutical candidate CRMs. **a** Metronidazole and **b** Captopril. *Groups A1* and *B1* are formed by outlier results and diverge from the main groups of results A2 and B2

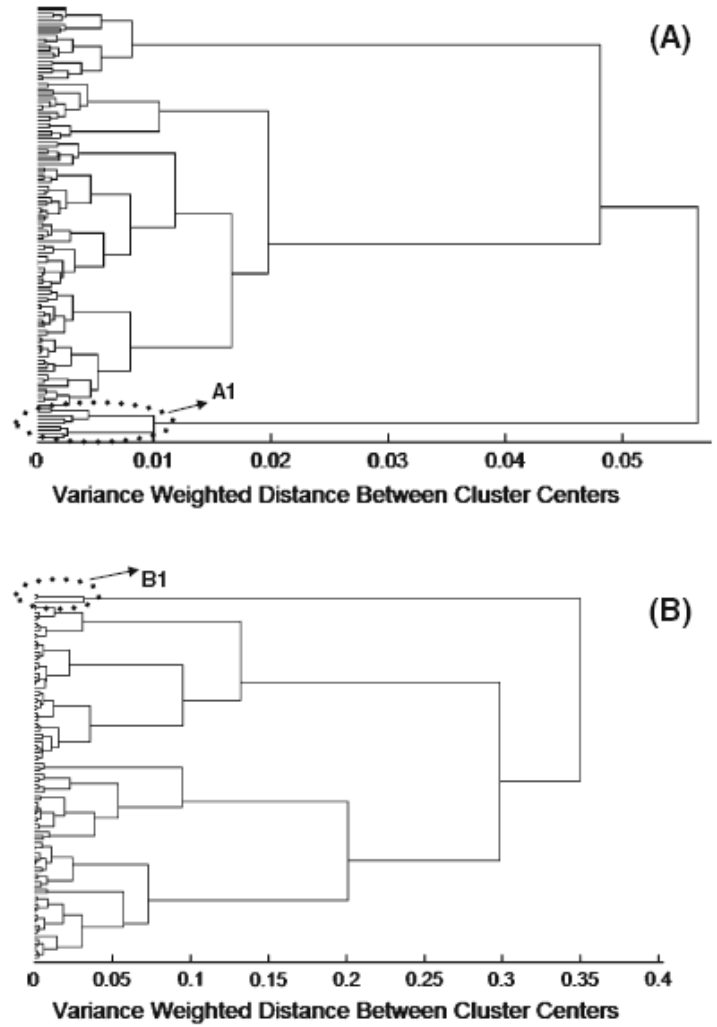


**Table 1** Standard uncertainty due to between-bottle(in) homogeneity ($u_{bb}$) for the metronidazole and captopril candidate CRMs with our without deletion of outliers

| | $u_{bb}$ (g/100 g) | |
|---|---|---|
| | Metronidazole | Captopril |
| All data points | 0.002917 | 0.015218 |
| Data points excluding outliers | 0.001504[a] | 0.014877[b] |

[a] 9 of 117 results were excluded (replicates 1, 2, and 3 of one sample; 3 HPLC injections each)

[b] 3 of 102 results were excluded (replicate 3 of one sample; 3 HPLC injections)

48,44%

2,24%

## Conclusions

For the metronidazole and captopril CRMs, the results from the multivariate analysis methods of PCA and HCA revealed that some replicate results diverged from those of the main groups of homogeneity test results. Considering that these observed differences are small, they cannot be easily detected. The multivariate analysis was able to indicate, with a reliability of 95% confidence, which samples should not be taken into consideration for homogeneity evaluation, leading to an $u_{bb}$ reduction by 48.44 and 2.24% for metronidazole and captopril CRMs, respectively.

4

Microchemical Journal xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

# A comparison of three procedures for robust PCA of experimental results of the homogeneity test of a new sodium diclofenac candidate certified reference material

Werickson Fortunato de Carvalho Rocha *, Raquel Nogueira, Gisele Estevão Baptista da Silva, Suzane Maio Queiroz, Gabriel Fonseca Sarmanho

*National Institute of Metrology, Quality and Technology (Inmetro), Directorate of Industrial and Scientific Metrology, Chemical Metrology Division, 25250-020, Xerém, Duque de Caxias, RJ, Brazil*

PCA PCA 1 PCA 2 PCA 3

**Table 2**
Standard uncertainty due to between-bottle (in)homogeneity ($u_{bb}$) and its contribution C for the sodium diclofenac candidate CRM, with our without deletion of outliers.

| | All data | Classic PCA | Projection pursuit-PP | ROBPCA | Spherical PCA |
|---|---|---|---|---|---|
| $MS_{between}$ (g 100 g$^{-1}$)[2a] | $1.779673\times10^{-4}$ | $1.511814\times10^{-4}$ | $1.328008\times10^{-4}$ | $1.448167\times10^{-4}$ | $1.499739\times10^{-4}$ |
| $MS_{within}$ (g 100 g$^{-1}$)[2b] | $2.185600\times10^{-5}$ | $1.146937\times10^{-5}$ | $1.136185\times10^{-5}$ | $1.140589\times10^{-5}$ | $1.151570\times10^{-5}$ |
| $n_0$[c] | 9 | 9 | 9 | 9 | 9 |
| $u_{bb}$ (g 100 g$^{-1}$) | 0.0041666 | 0.0039419 | 0.0036753 | 0.0038521 | 0.0039242 |
| $C$ (%) | 0.0041666 | 0.0039416 | 0.0036748 | 0.0038517 | 0.0039239 |

[a] Mean square between CRM batch units from single factor ANOVA.
[b] Mean square within CRM batch units from single factor ANOVA.
[c] number of replicates.

0.0039419

0.0038521

0.0036753

0.0039242

0.004166

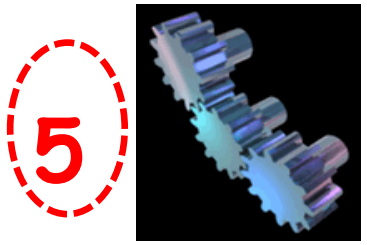
Reduction in the uncertainty of homogeneity 11.8 %.

## 4. Conclusions

In chemical metrology, robust statistical methods find several applications, including dimension reduction, modeling and model evaluation, and outlier detection. Robust methods focus on modeling the data majority, and they downweight deviating or outlying observations.

In this paper, it was demonstrated that PCA modeling was an efficient tool to identify outlying results in the data series from the homogeneity test for certification of a new pharmaceutical candidate CRM. The identification and further elimination of outliers guaranteed the accuracy of the estimated uncertainty due to due to between-bottle (in) homogeneity ($u_{bb}$) and of the CRM standard uncertainty ($u_{CRM}$).

The robust PCA based on PP showed the best performance for identification of outliers, leading to reductions in the contribution of (in)homogeneity ($C$) and $u_{bb}$ values by 11.8%. This multivariate statistical method can be considered an important tool to evaluate experimen-




# Example

Sensors and Actuators B 158 (2011) 327–332

Contents lists available at ScienceDirect

Sensors and Actuators B: Chemical

journal homepage: www.elsevier.com/locate/snb

## Evaluation study of different glass electrodes by an interlaboratory comparison for determining the pH of fuel ethanol

Mary Ane Gonçalves, Fabiano Barbieri Gonzaga*, Isabel Cristina Serta Fraga, Carla de Matos Ribeiro, Sidney Pereira Sobral, Paulo Paschoal Borges, Werickson Fortunato de Carvalho Rocha

*Electrochemistry Laboratory, Chemical Metrology Division, National Institute of Metrology, Standardization and Industrial Quality, Inmetro, Av. Nossa Senhora das Graças, 50, Xerém, 22250-020 Duque de Caxias, RJ, Brazil*

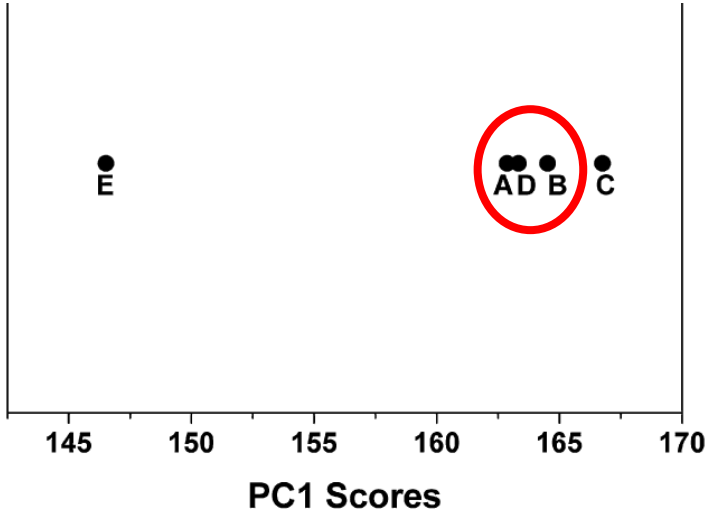



**Fig. 2.** Scores plot in the first principal component (PC1), resulting from PCA, comparing the different electrodes. The PCA was carried out using all the data without any pre-treatment (electrodes as rows or samples and raw pH data as columns or variables).

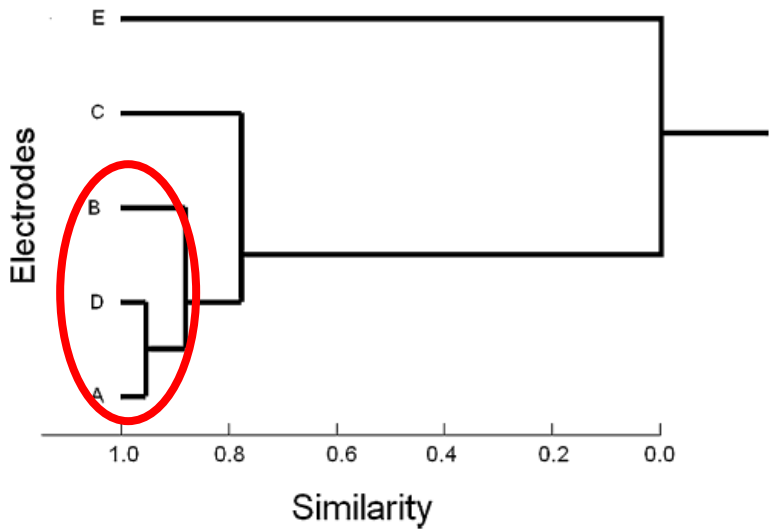


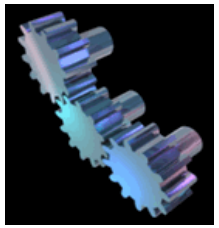

**Fig. 3.** Dendrogram plot, resulting from HCA, comparing the different electrodes. The similarity data ($S$) are related to the distance between the electrodes in the multidimensional space (electrodes as rows or samples and raw pH data as columns or variables) according to the equation $S = [1 - (d/d_{max})]$, where $d$ is the Euclidean distance between two specific electrodes and $d_{max}$ is the maximum Euclidean distance considering all the electrodes [23].

## Conclusion:

It can be concluded that the electrodes A, B and D have similar results.

The work suggests a revision of ASTM D6423 which indicates that the use of electrode (Orion) is required.

6



Example

*metabolites*

MDPI

*Article*

# Metabolomics Test Materials for Quality Control: A Study of a Urine Materials Suite

Daniel W. Bearden [1,†], David A. Sheen [1,*], Yamil Simón-Manso [2], Bruce A. Benner, Jr. [1], Werickson F. C. Rocha [1,3], Niksa Blonder [1], Katrice A. Lippa [1], Richard D. Beger [4], Laura K. Schnackenberg [4], Jinchun Sun [4], Khyati Y. Mehta [5], Amrita K. Cheema [5,6], Haiwei Gu [7], Ramesh Marupaka [8], G. A. Nagana Gowda [9] and Daniel Raftery [9]

1 Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA; danbearden@metabolomicspartners.com (D.W.B.); bruce.benner@nist.gov (B.A.B.J.); wfrocha@inmetro.gov.br (W.F.C.R.); niksa.blonder@nist.gov (N.B.); katrice.lippa@nist.gov (K.A.L.)

2 Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA; yamil.simon@nist.gov

3 National Institute of Metrology, Quality, and Technology—INMETRO, 25250-020 Duque de Caxias, RJ, Brazil

4 Division of Systems Biology, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079, USA; richard.beger@fda.hhs.gov (R.D.B.); laura.schnackenberg@fda.hhs.gov (L.K.S.); jinchun.sun@fda.hhs.gov (J.S.)

5 Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC 20057, USA; kym8@georgetown.edu (K.Y.M.); akc27@georgetown.edu (A.K.C.)

6 Departments of Oncology and Biochemistry, Molecular and Cellular Biology, Georgetown University Medical Center, Washington, DC 20057, USA

7 College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA; haiweigu@asu.edu

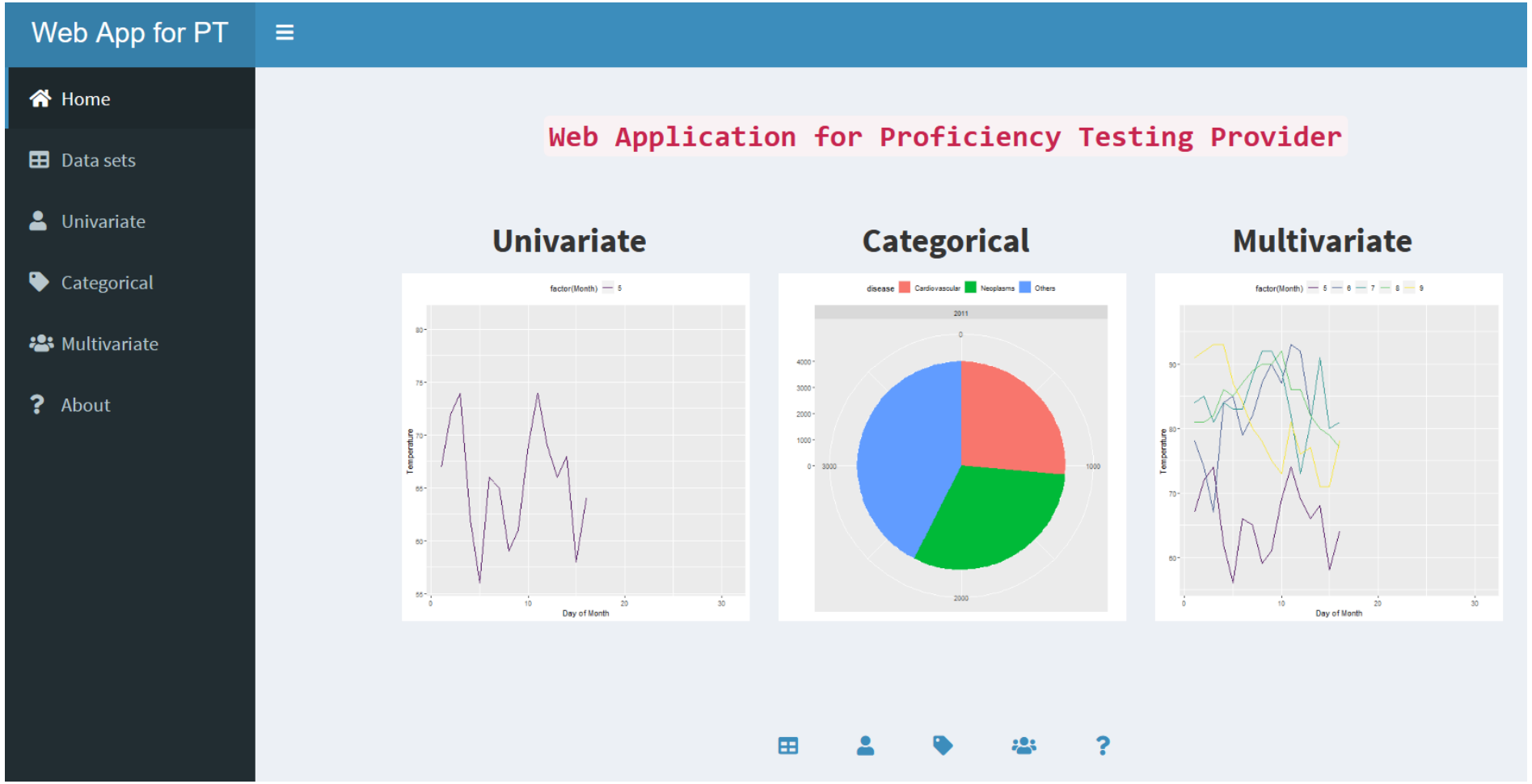
8 Clinical Toxicology at CIAN Diagnostics, Frederick, MD 21703, USA; rameshmrpk@gmail.com

9 Department of Anesthesiology and Pain Medicine, Mitochondria and Metabolism Center, University of Washington, Seattle, WA 98109, USA; ngowda@uw.edu (G.A.N.G.); draftery@uw.edu (D.R.)

* Correspondence: david.sheen@nist.gov; Tel.: +1-301-975-2603

† Retired.

# 7 Building softwares for metrology



# Topics

1) Some definitions

2) Data: new gold

3) Generating trusted data for machine learning

4) What does the data say?

5) Examples

6) Conclusions

# 6) Conclusions

- Data science methods have been used in metrological activities in order to transform complex data into relevant information.
- There are several applications that use data science methods to perform metrological activities that provide better understanding of the results.

# Acknowledgement

-The Inter-American Metrology System (SIM)

-Rodolfo Saboia Souza-Head of Division at Inmetro


SIM
SISTEMA INTERAMERICANO DE METROLOGÍA

Metrology for Digital Transformation
SIM-MWG-14

**IADB SIM RESEARCH ENGAGEMENT OPPORTUNITY 2021**

**Metrological evaluation of lung ultrasound using virtual vector machine for diagnosis of acute respiratory distress syndrome (ME-LUS-VMM-DARDS)**

- We are looking for NMIs that have people that work with machine learning to collaborate in the project.
- If you are interested, contact me at wfrocha@inmetro.gov.br or Rodrigo Costa-Felix rpfelix@inmetro.gov.br – Coordinator

Team:
Inmetro, CENAM, INTI and Nist

If you want to join us, let me know.

# Thank you!

Werickson Fortunato de Carvalho Rocha
Researcher
*wfrocha@inmetro.gov.br*