Application of Chemometrics and Data Analysis in Pharmaceutical Forensics

Jonas Hoeg Thygesen
Principal Scientist

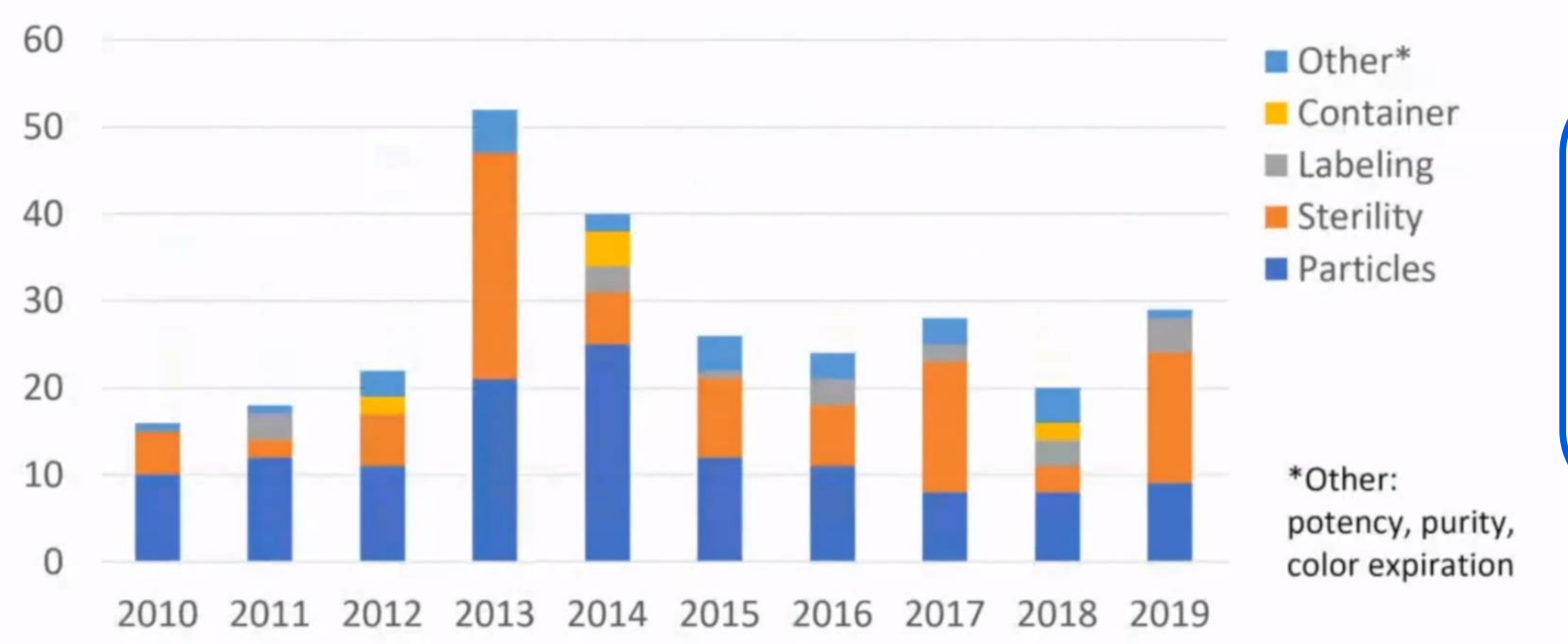
20 November 2020
DSK.2020

Novo Nordisk
Pharmatech A/S

Novo Nordisk Pharmatech Microanalysis Centre

- Centre of excellence within pharmaceutical forensics: Particle analysis, identification of unknown materials and solid state characterization
- Highly specialized range of analytical techniques, specifically tailored to fulfil the Novo Nordisk needs
- Unique insight and understanding of Novo Nordisk products and production processes
- Centralised knowledge hub for Novo Nordisk with regards to visible and subvisible particles

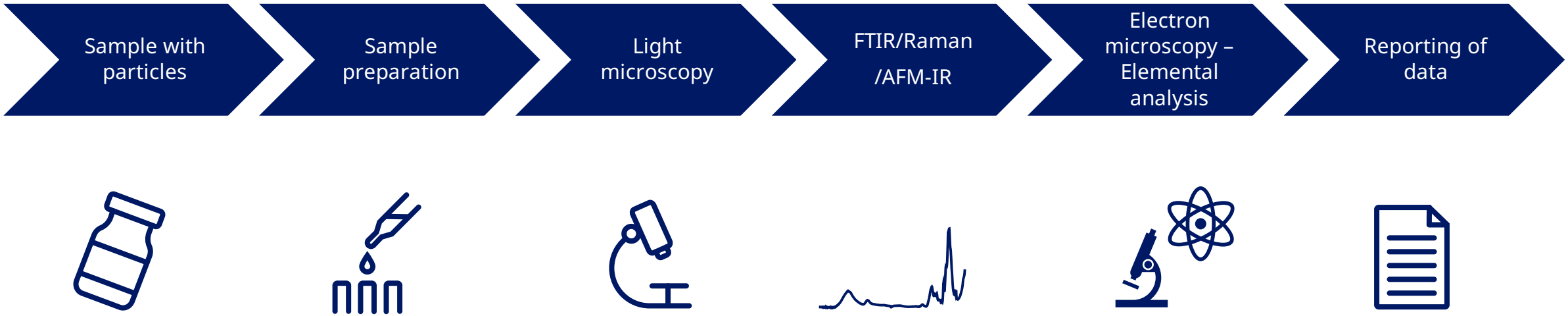
US FDA Recall notices for injectable drug products[1]



- Zero Novo Nordisk particle recalls in the period
- Production processes in control
- Thorough root cause analysis

1: Data by John Shabushnig Insight Pharma Consulting, 2020 PDA Visual Inspection Forum

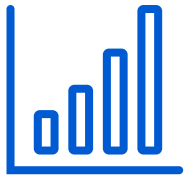
Analytical flow for particles



Data science in pharmaceutical forensics

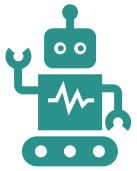
Bussiness Intelligence (BI):

- Data fusion, cleaning and preparation
- Data visualization
- Automation of workflows



Robotic Process Automation (RPA):

- Automation of administrative tasks
- Data scraping/collection



Chemometrics (ML):

- Remove manual decisions
- Obtain objectivity
- Enhanced insight in data

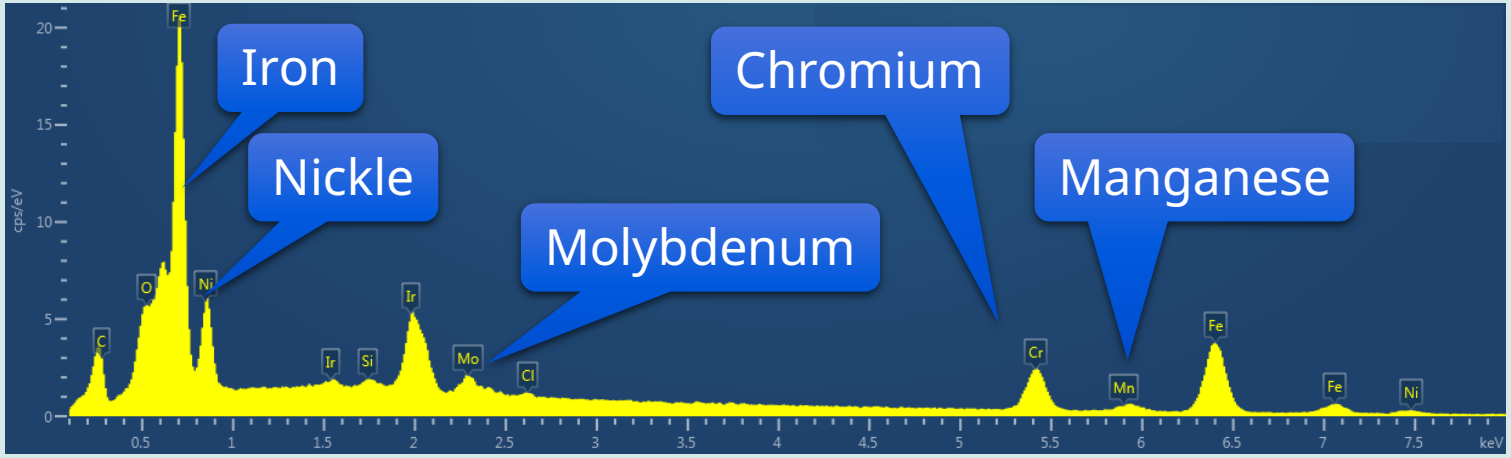
Domain knowledge:

Spectroscopy, electon- and optical microscopy, chemistry, material science and more

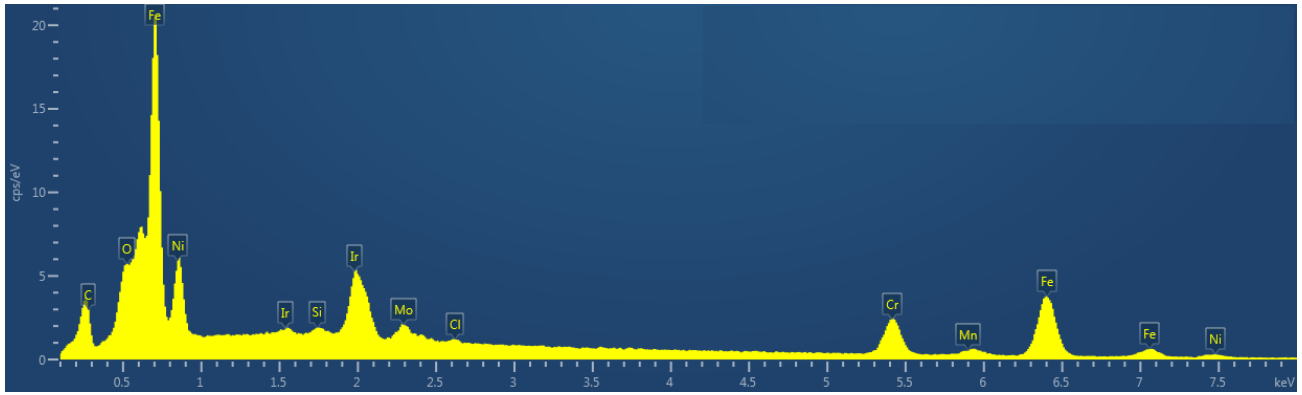
Case study 1: ID of steel

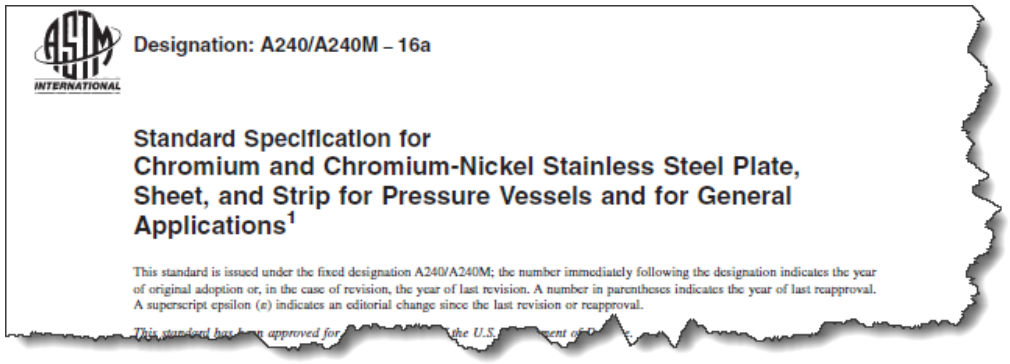
Remove manual decisions
Obtain objectivity

Fe, Cr, Mn, Mo, Ni = Stainless steel



Identification of steel – the manual way





Designation: A240/A240M – 16a

Standard Specification for Chromium and Chromium-Nickel Stainless Steel Plate, Sheet, and Strip for Pressure Vessels and for General Applications[1]

This standard is issued under the fixed designation A240/A240M; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

Elements (% W/W)	Si	V	Cr	Mn	Fe	Co	Ni	Cu	Mo	W
Unknown 1	0.42	0.00	17.7	1.21	71.56	0.00	9.11	0.00	0.30	0.00

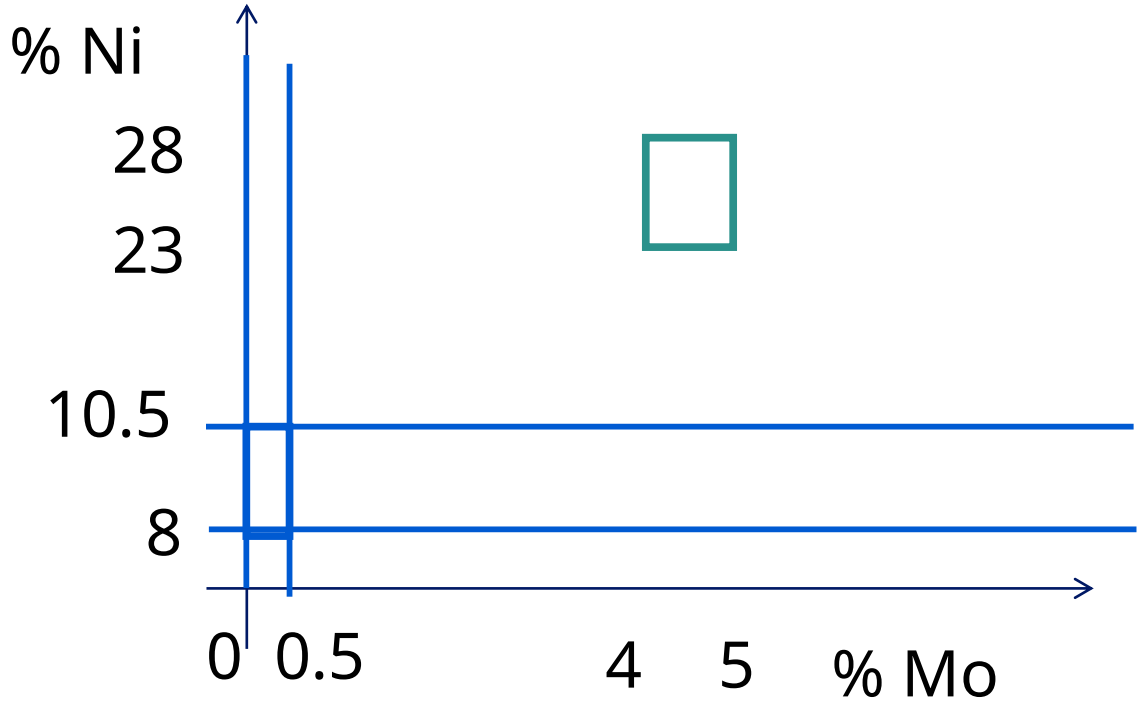
Alloy	Concentration	Si	V	Cr	Mn	Fe	Co	Ni	Cu	Mo	W
AISI 304	Min	0.000	0.000	17.500	0.000	67.005	0.000	8.000	0.000	0.000	0.000
	Max	0.750	0.000	19.500	2.000	74.500	0.000	10.500	0.000	0.000	0.000
AISI 310S	Min	0.000	0.000	24.000	0.000	48.340	0.000	19.000	0.000	0.000	0.000
	Max	1.500	0.000	26.000	2.000	57.000	0.000	22.000	0.000	0.000	0.000
AISI 316	Min	0.000	0.000	16.000	0.000	61.950	0.000	10.000	0.000	2.000	0.000
	Max	0.750	0.000	18.000	2.000	72.000	0.000	14.000	0.000	3.000	0.000
AISI 316L	Min	0.000	0.000	16.000	0.000	61.950	0.000	10.000	0.000	2.000	0.000
	Max	0.750	0.000	18.000	2.000	72.000	0.000	14.000	0.000	3.000	0.000
AISI 904L	Min	0.000	0.000	19.000	0.000	38.800	0.000	23.000	1.000	4.000	0.000
	Max	1.000	0.000	23.000	2.000	53.000	0.000	28.000	2.000	5.000	0.000
SAF2507	Min	0.000	0.000	24.000	0.000	58.095	0.000	6.000	0.000	3.000	0.000
	Max	0.800	0.000	26.000	1.200	66.760	0.000	8.000	0.500	5.000	0.000
Hastelloy C-276	Min	0.000	0.000	14.500	0.000	4.000	0.000	50.990	0.000	15.000	3.000
	Max	0.080	0.350	16.500	1.000	7.000	2.500	63.500	0.000	17.000	4.500

Objectivity?
User-friendliness?

A graphical approach to ID of steel

Alloy	Concentration	Si	V	Cr	Mn	Fe	Co	Ni	Cu	Mo	W
AISI 304	Min	0.000	0.000	17.500	0.000	67.006	0.000	8.000	0.000	0.000	0.000
	Max	0.750	0.000	19.500	2.000	74.500	0.000	10.500	0.000	0.000	0.000
AISI 310S	Min	0.000	0.000	24.000	0.000	48.340	0.000	19.000	0.000	0.000	0.000
	Max	1.500	0.000	26.000	2.000	57.000	0.000	22.000	0.000	0.000	0.000
AISI 316	Min	0.000	0.000	16.000	0.000	61.950	0.000	10.000	0.000	2.000	0.000
	Max	0.750	0.000	18.000	2.000	72.000	0.000	14.000	0.000	3.000	0.000
AISI 316L	Min	0.000	0.000	16.000	0.000	61.950	0.000	10.000	0.000	2.000	0.000
	Max	0.750	0.000	18.000	2.000	72.000	0.000	14.000	0.000	3.000	0.000
AISI 904L	Min	0.000	0.000	19.000	0.000	38.800	0.000	23.000	1.000	4.000	0.000
	Max	1.000	0.000	23.000	2.000	53.000	0.000	28.000	2.000	5.000	0.000
SAF2507	Min	0.000	0.000	24.000	0.000	58.096	0.000	6.000	0.000	3.000	0.000
	Max	0.800	0.000	26.000	1.200	66.760	0.000	8.000	0.500	5.000	0.000
Hastelloy C-276	Min	0.000	0.000	14.500	0.000	4.000	0.000	50.990	0.000	15.000	3.000
	Max	0.080	0.350	16.500	1.000	7.000	2.500	63.500	0.000	17.000	4.500

	Si	V	Cr	Mn	Fe	Co	Ni	Cu	Mo	W
Unknown 1	0.42	0.00	17.7	1.21	71.56	0.00	9.11	0.00	0.00	0.00

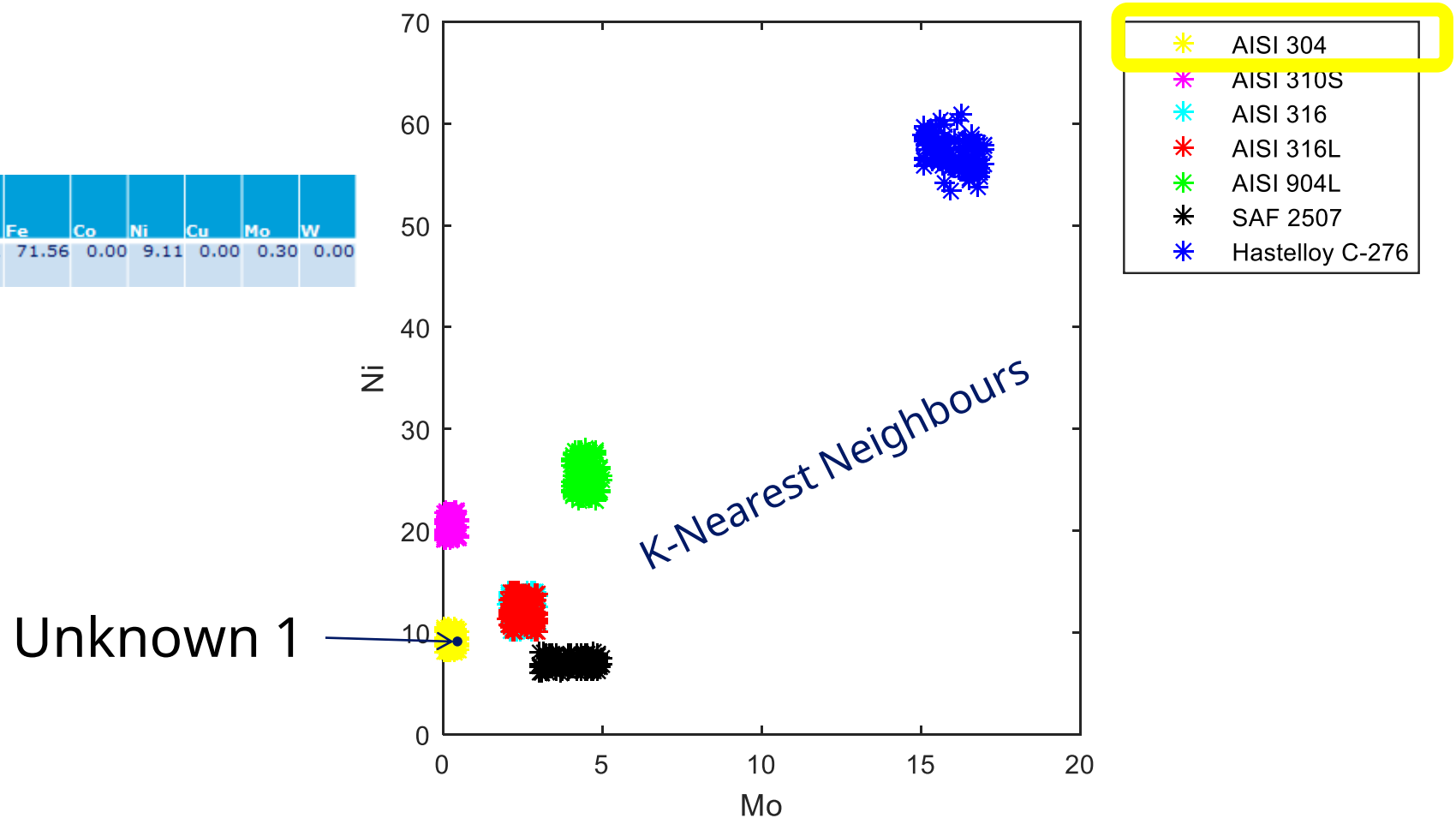


Window estimation:

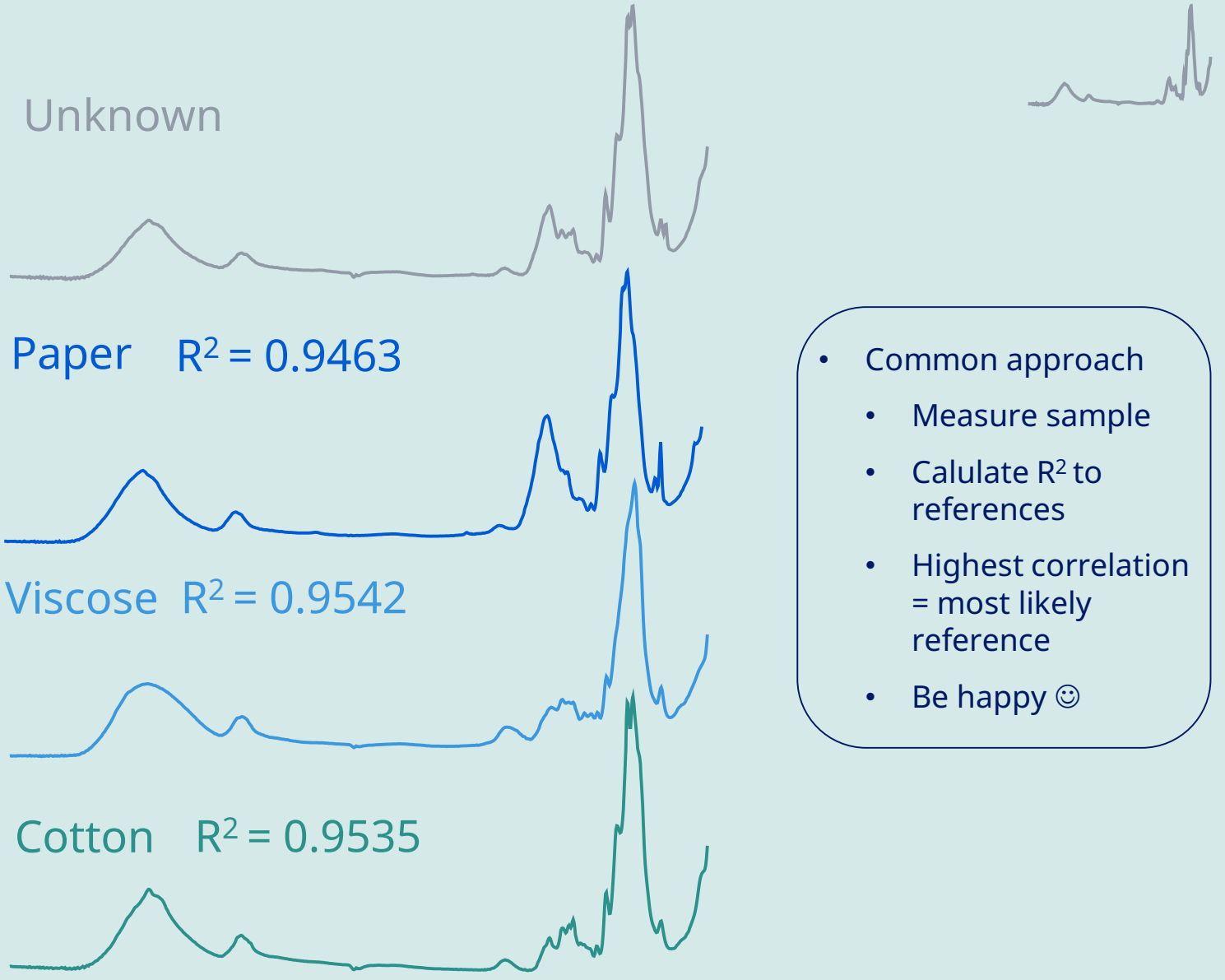
- Monte Carlo Simulation
- 100 dummy samples in each window

Classification of ”Unknown 1”

Elements (% W/W)	Si	V	Cr	Mn	Fe	Co	Ni	Cu	Mo	W
Unknown 1	0.42	0.00	17.7	1.21	71.56	0.00	9.11	0.00	0.30	0.00



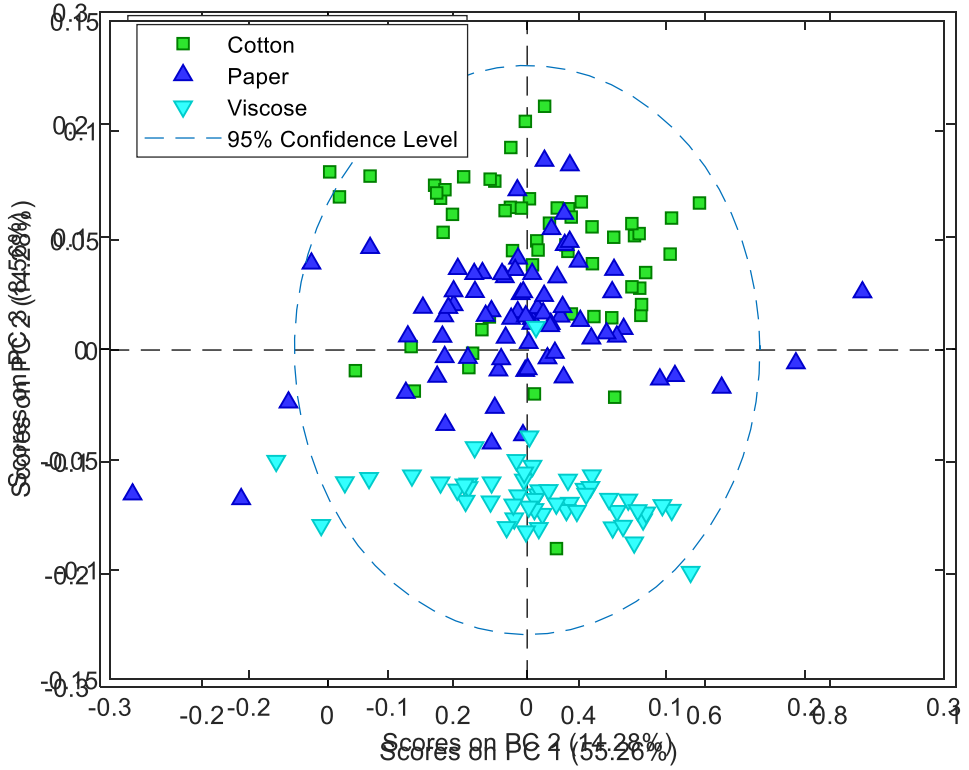
Case study 2: FTIR of cellulose fibres



- Common approach
  - Measure sample
  - Calulate R^2 to references
  - Highest correlation = most likely reference
  - Be happy ☺

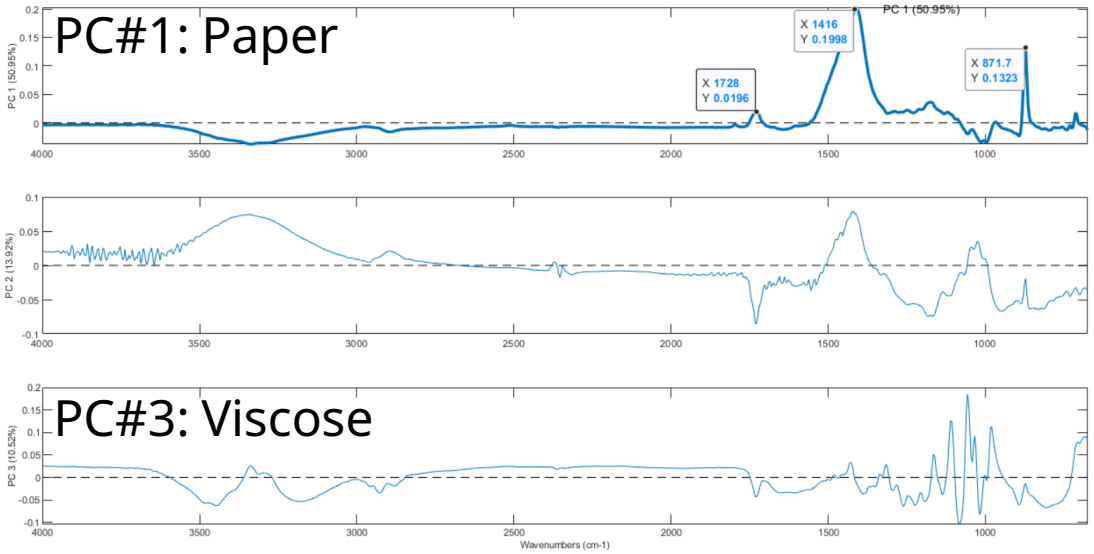
New (multivariate?) analytics required

- Scope of project:
  - Cotton
  - Paper
  - Viscose/Rayon
- Measurement of reference catalogue
- Use of multivariate statistics for enhanced knowledge from data

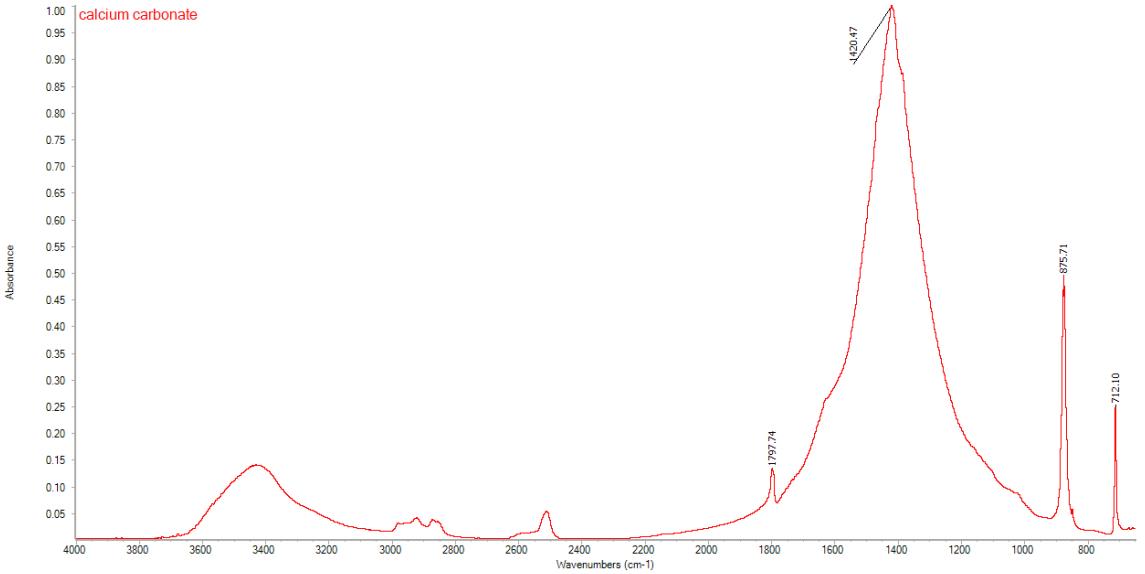


Power of chemometrics – explains the why

PCA loadings



Calcium carbonate reference spectrum



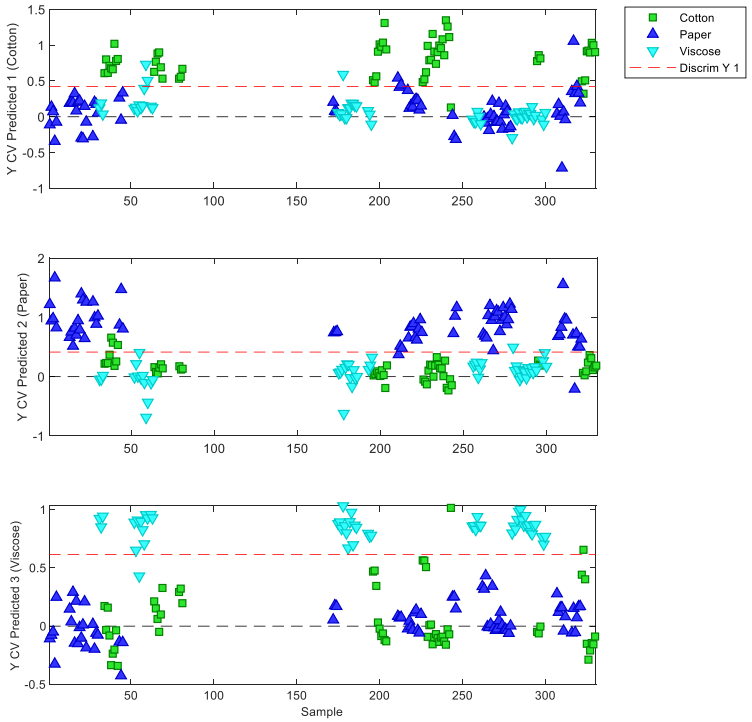
Chemically meaningfull clustering in PCA -> PLS-DA candidate method?

Performance of PLS-DA

Cotton
Not cotton

Paper
Not paper

Viscose
Not Viscose



Sensitivity 95-98%

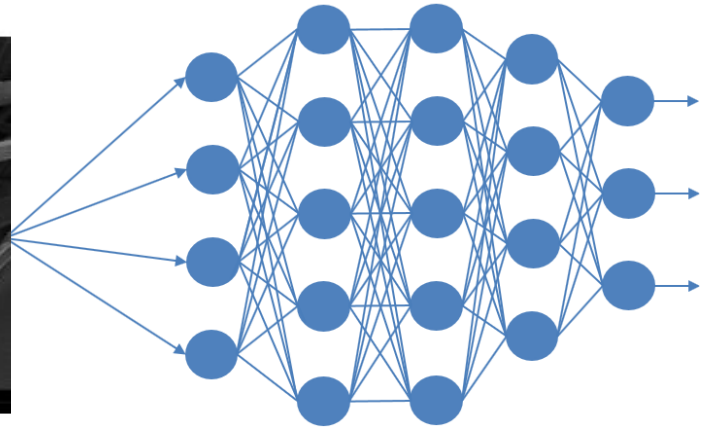
Specificity 98%

Precision 96-97%

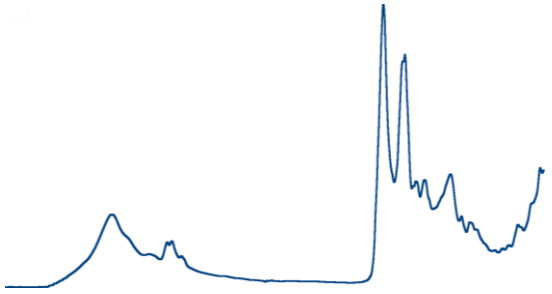
```
CV RESULTS
Confusion Matrix (CV):
    Class:      TPR         FPR         TNR         FNR         N      Err         P          F1
   Cotton     0.94737     0.01613     0.98387     0.05263     57     0.02762     0.96429     0.95575
   Paper      0.97183     0.01818     0.98182     0.02817     71     0.02210     0.97183     0.97183
   Viscose    0.98113     0.01563     0.98438     0.01887     53     0.01657     0.96296     0.97196

Confusion Table (CV):
                                Actual Class
                               Cotton   Paper    Viscose
Predicted as Cotton              54       2        0
Predicted as Paper                1      69        1
Predicted as Viscose              2       0       52
Predicted as Unassigned           0       0        0
```

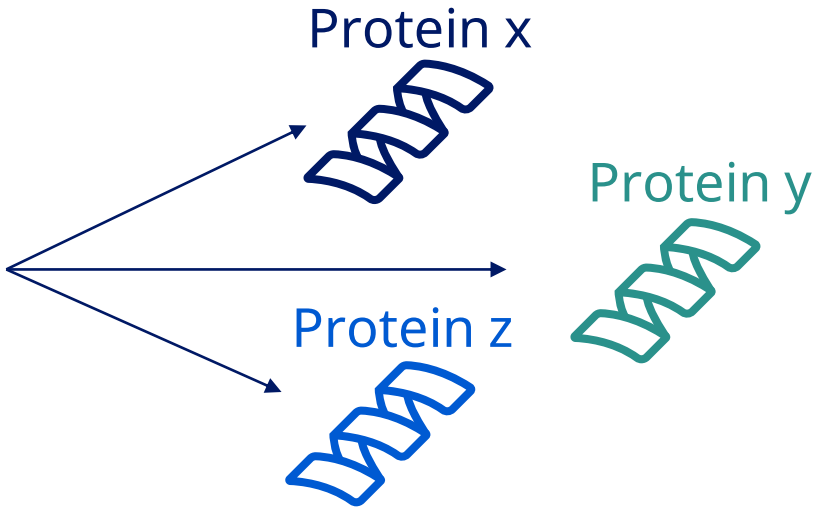
Future projects in scope

Image classification



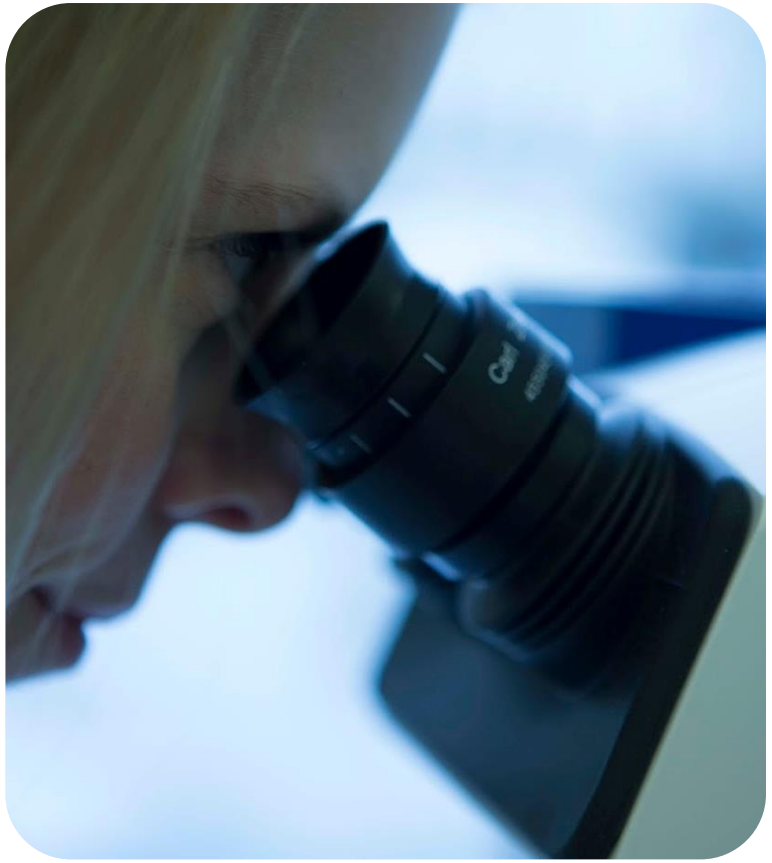
Spectral classification





Take home messages

- Chemometrics and modern data analytics are powerful tools
- The use of such tools enables further insight in data
- Large potential for use of these tools in pharmaceutical forensics



Come join us!

- Internship in Novo Nordisk Pharmatech from Feb 2021
  - Digitalization
  - Customer interaction

- MSc Project at Novo Nordisk Pharmatech?
  - ID of proteins
  - Image classification
  - PAT

Jonas Hoeg Thygesen
Principal Scientist
R&D - Microanalysis

+45 3037 0021 (mobile)
jhet@novonordiskpharmatech.com

Novo Nordisk Pharmatech A/S

novo nordisk®