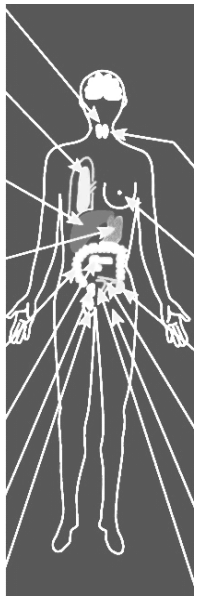
Handling large amounts of data

Applications in applied fluorescence spectroscopy

Rasmus Bro

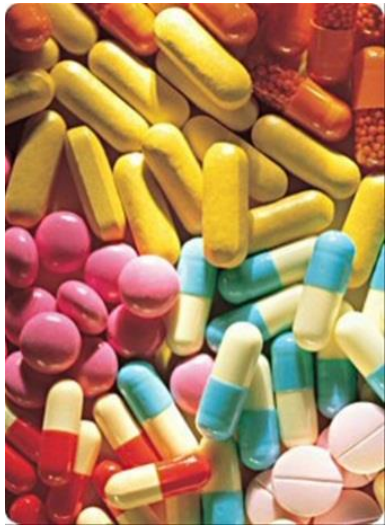
Dept. Food Science
University of Copenhagen

Dioxin,
Environment,
Dose-response



Metabolomics,
Proteomics
Systems biology,
Cancer,
Diabetes,
Pharma
…

Food quality, gastronomy
Raw material influence,
Production optimization,
end point detection



What we work with

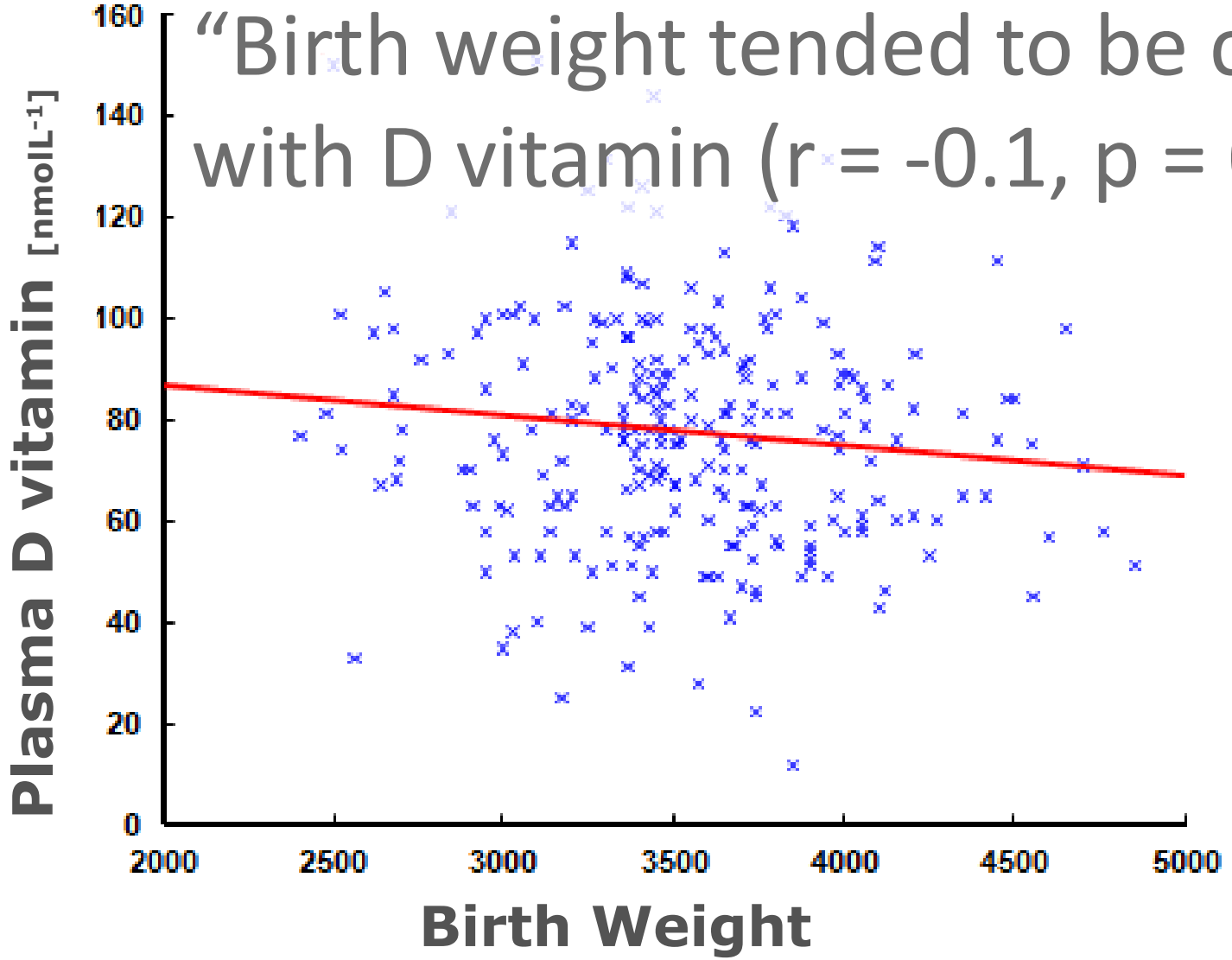
Fluorescence
High resolution NMR
Mass spectrometry
Near-infrared
Raman spectroscopy
Ultrasound
Hyperspectral imaging
Chromatography
…

Data

"Birth weight tended to be correlated with D vitamin ($r = -0.1$, $p = 0.06$)"

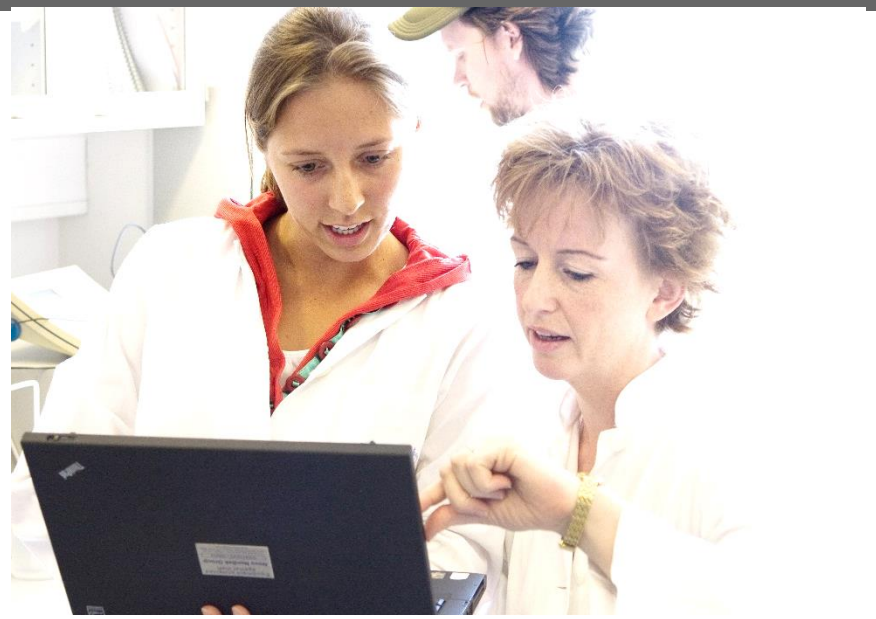
We often hide behind statistics

"Birth weight tended to be correlated with D vitamin (r = -0.1, p = 0.06)"



We often hide behind statistics

Need new tools that

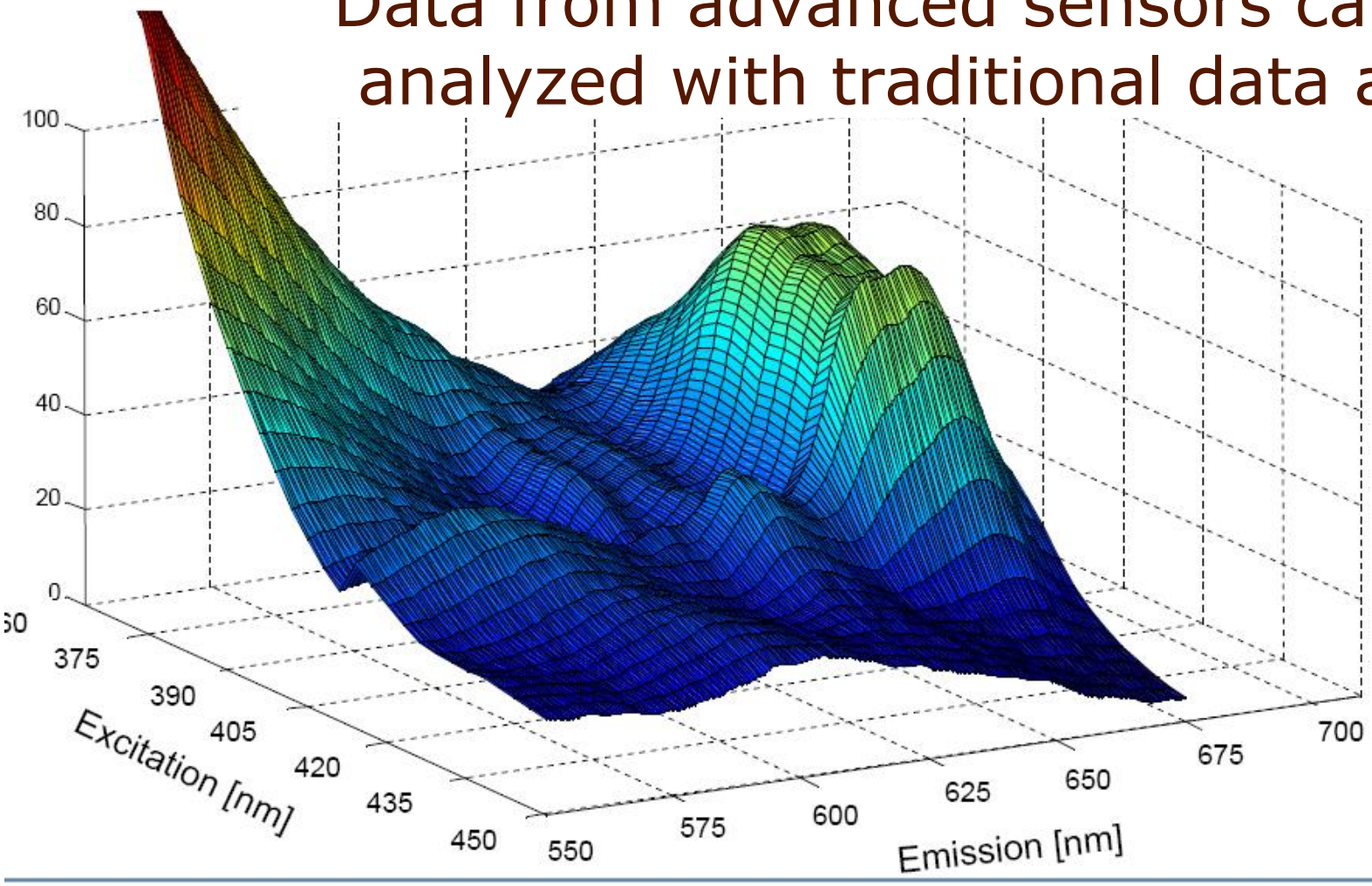


Enable the scientist to be critical

Connects the competences
(technician, operator, biologist, engineer, doctor, …)

Sensors – multivariate

Data from advanced sensors can not be analyzed with traditional data analysis



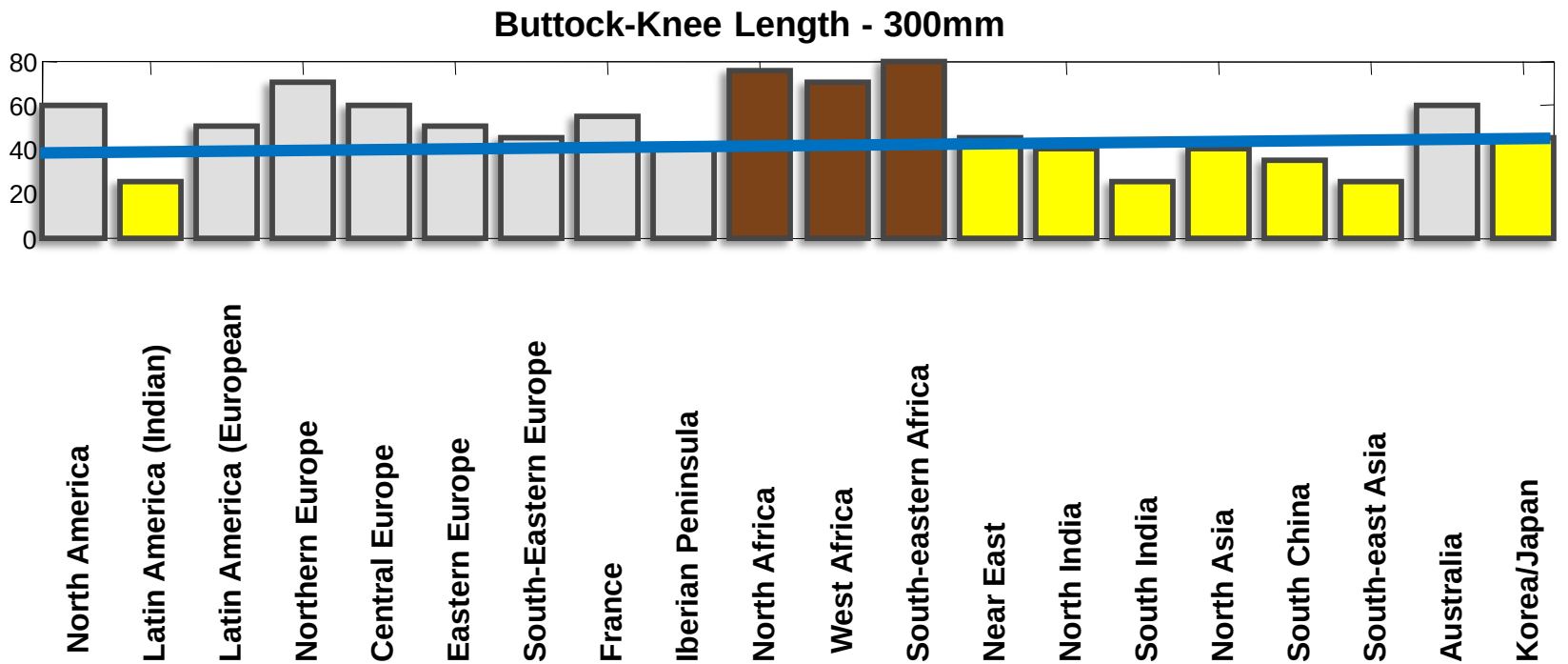
Human pattern recognition uses all available data

Using a single variable is:

Wrong, incorrect, suboptimal, oldfashioned, ..!

Korea overlaps with Caucasian

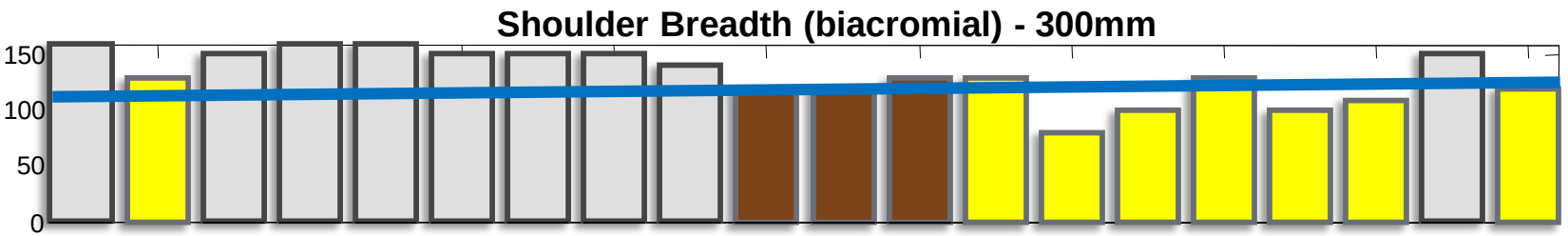
Caucasian
African
Asian

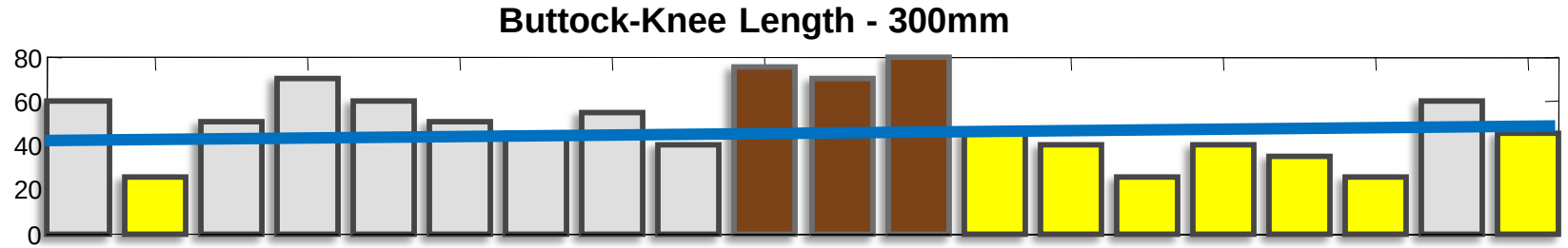


Using a single variable is:

Wrong, incorrect, suboptimal, oldfashioned, ..!

Caucas.
African
Asian

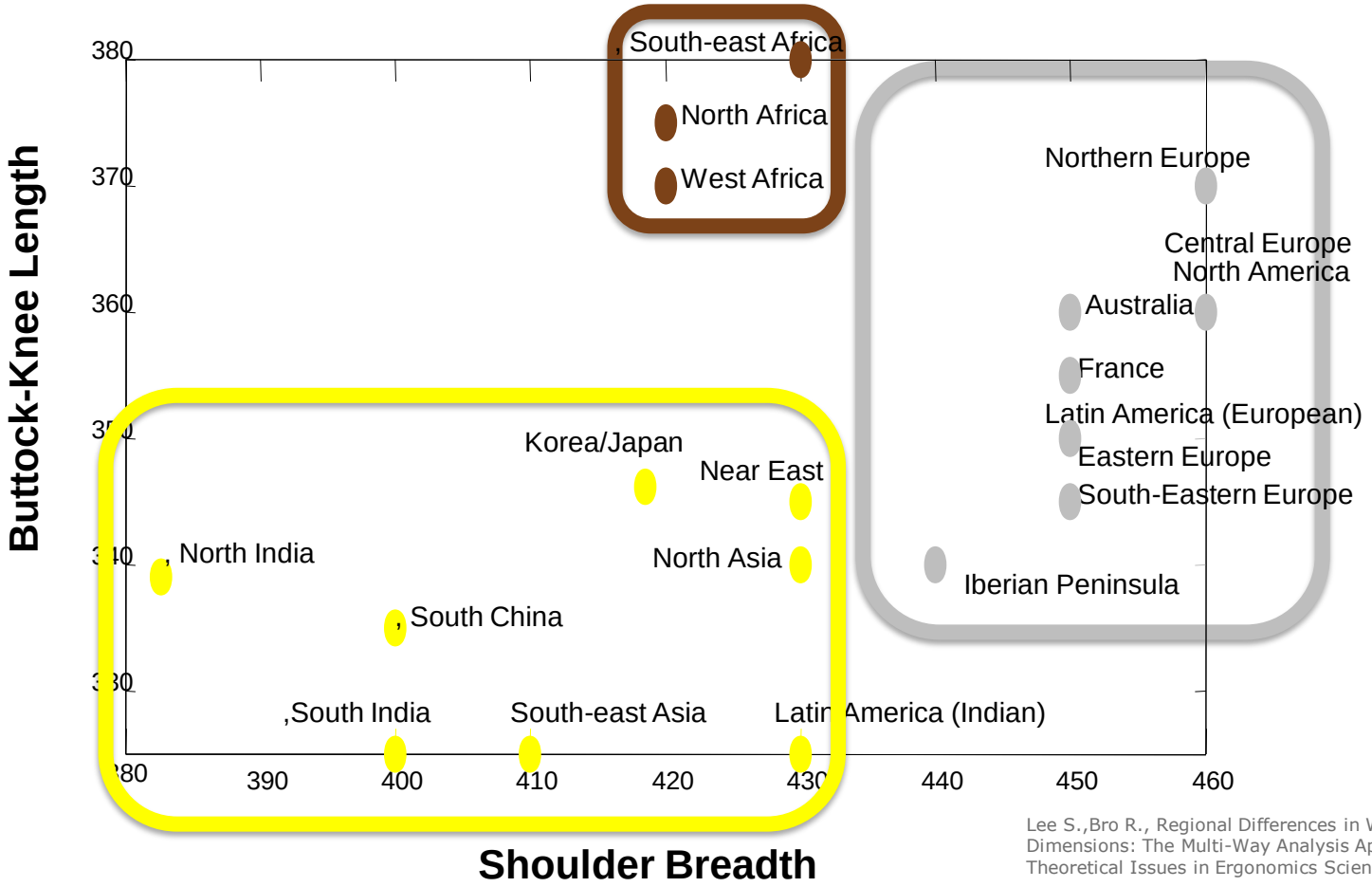




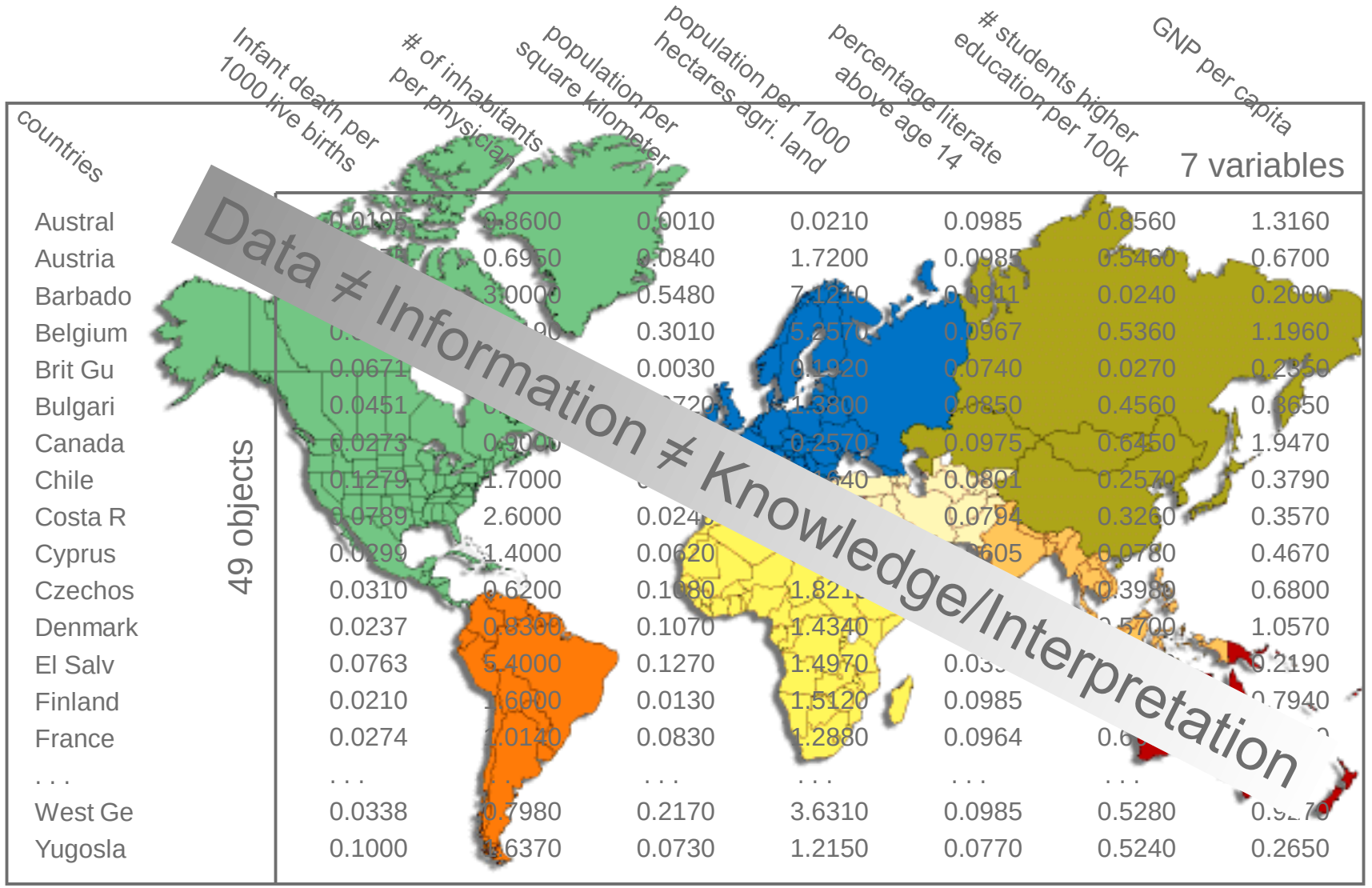
North America
Latin America (Indian)
Latin America (European
Northern Europe
Central Europe
Eastern Europe
South-Eastern Europe
France
Iberian Peninsula
North Africa
West Africa
South-eastern Africa
Near East
North India
South India
North Asia
South China
South-east Asia
Australia
Korea/Japan

Simply plot the two versus each other

Co-variation = new information that is *not* available in the individual variables



Lee S.,Bro R., Regional Differences in World Human Body Dimensions: The Multi-Way Analysis Approach, Theoretical Issues in Ergonomics Science, 2007



countries	Infant death per 1000 live births	# of inhabitants per physician	population per square kilometer	population per 1000 hectares agri. land	percentage literate above age 14	# students higher education per 100k	GNP per capita
	49 objects						7 variables
Austral	0.0195	0.8600	0.0010	0.0210	0.0985	0.8560	1.3160
Austria	[illegible]	0.6950	0.0840	1.7200	0.0985	0.5460	0.6700
Barbado	[illegible]	3.0000	0.5480	7.1210	0.0911	0.0240	0.2000
Belgium	[illegible]	[illegible]	0.3010	5.2570	0.0967	0.5360	1.1960
Brit Gu	0.0671	[illegible]	0.0030	0.1920	0.0740	0.0270	0.2850
Bulgari	0.0451	[illegible]	[illegible]	1.3800	0.0850	0.4560	0.3650
Canada	0.0273	0.9000	[illegible]	0.2570	0.0975	0.6450	1.9470
Chile	0.1279	1.7000	[illegible]	[illegible]	0.0801	0.2570	0.3790
Costa R	0.0789	2.6000	0.024[illegible]	[illegible]	0.0794	0.3260	0.3570
Cyprus	0.0299	1.4000	0.0620	[illegible]	[illegible]605	0.0780	0.4670
Czechos	0.0310	0.6200	0.1080	1.821[illegible]	[illegible]	0.3980	0.6800
Denmark	0.0237	0.8300	0.1070	1.4340	[illegible]	[illegible]5700	1.0570
El Salv	0.0763	5.4000	0.1270	1.4970	0.03[illegible]	[illegible]	0.2190
Finland	0.0210	1.6000	0.0130	1.5120	0.0985	[illegible]	0.7940
France	0.0274	1.0140	0.0830	1.2880	0.0964	0.6[illegible]	[illegible]
. . .	. . .	. . .	. . .	. . .	. . .	. . .	
West Ge	0.0338	0.7980	0.2170	3.6310	0.0985	0.5280	0.9[illegible]
Yugosla	0.1000	1.6370	0.0730	1.2150	0.0770	0.5240	0.2650

Data ≠ Information ≠ Knowledge/Interpretation

Gunst & Mason (1980) Regression analysis and its applications: A data-oriented approach, NY, Marcel Dekker, p. 358

Data analysis

7 variables; 49 objects

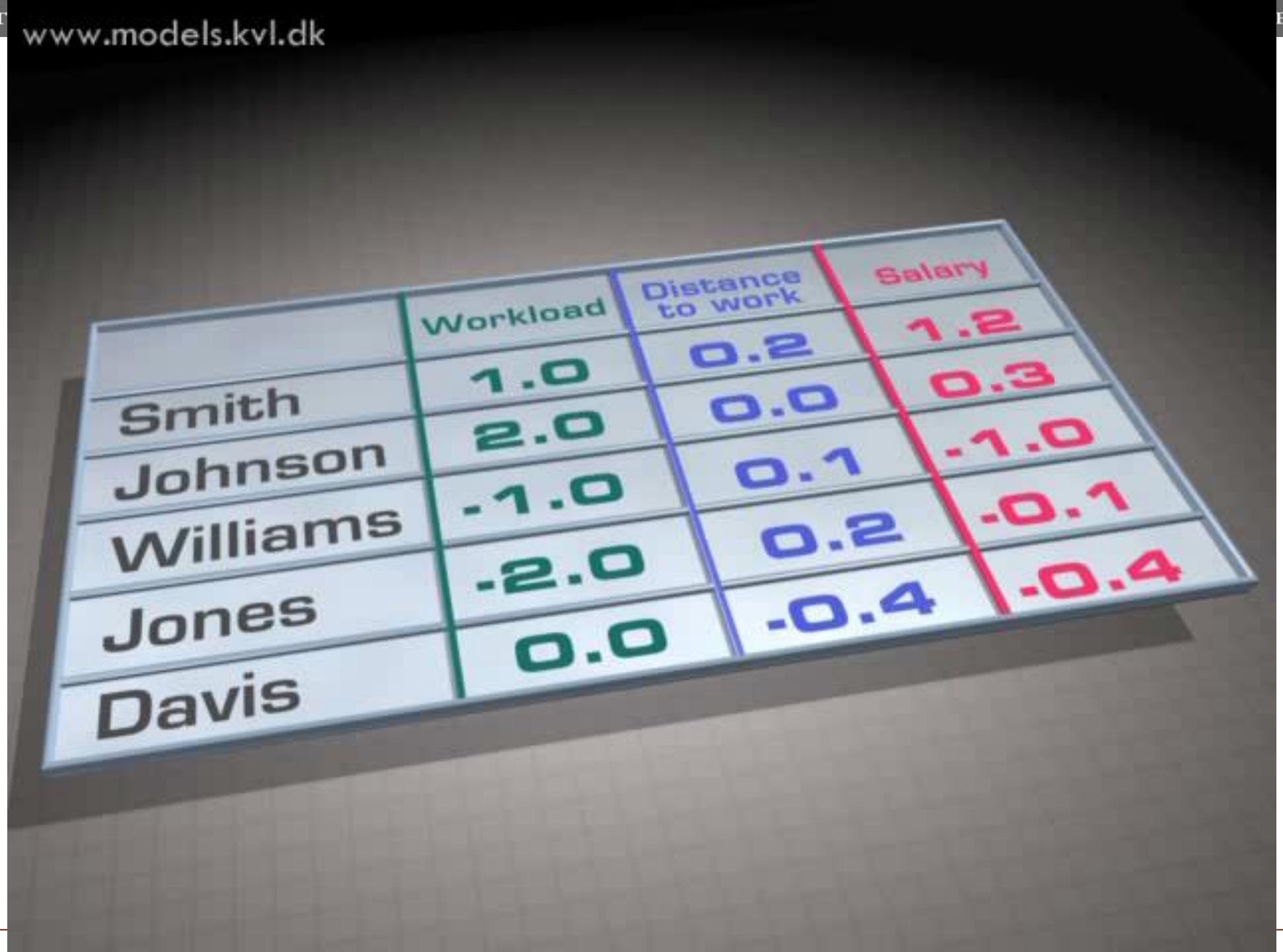
countries	Infant death per 1000 live births	# of inhabitants per physician	population per square kilometer	population per 1000 hectares agri. land	percentage literate above age 14	# students higher education per 100k	GNP per capita
Austral	0.0195	0.8600	0.0010	0.0210	0.0985	0.8560	1.3160
Austria	0.0375	0.6950	0.0840	1.7200	0.0985	0.5460	0.6700
Barbado	0.0604	3.0000	0.5480	7.1210	0.0911	0.0240	0.2000
Belgium	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.1960
Brit Gu	0.0671	3.9000	0.0030	0.1920	0.0740	0.0270	0.2350
Bulgari	0.0451	0.7400	0.0720	1.3800	0.0850	0.4560	0.3650
Canada	0.0273	0.9000	0.0020	0.2570	0.0975	0.6450	1.9470
Chile	0.1279	1.7000	0.0110	1.1640	0.0801	0.2570	0.3790
Costa R	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.3260	0.3570
Cyprus	0.0299	1.4000	0.0620	1.0420	0.0605	0.0780	0.4670
Czechos	[illegible]	[illegible]	0.1080	1.8210	0.0975	0.3980	0.6800
Denmark	[illegible]	0.8300	0.1070	1.4340	0.0985	0.5700	1.0570
El Salv	0.0763	5.4000	0.1270	1.4970	0.0394	0.0890	0.2190
Finland	0.0210	1.6000	0.0130	1.5120	0.0985	0.5290	0.7940
France	0.0274	1.0140	0.0830	1.2880	0.0964	0.6670	0.9430
...	...	...	...	...	...	...	...
West Ge	[illegible]	[illegible]	[illegible]	[illegible]	0.0985	0.5280	0.9270
Yugosla	0.1000	1.6370	0.0730	1.2150	0.0770	0.5240	0.2650

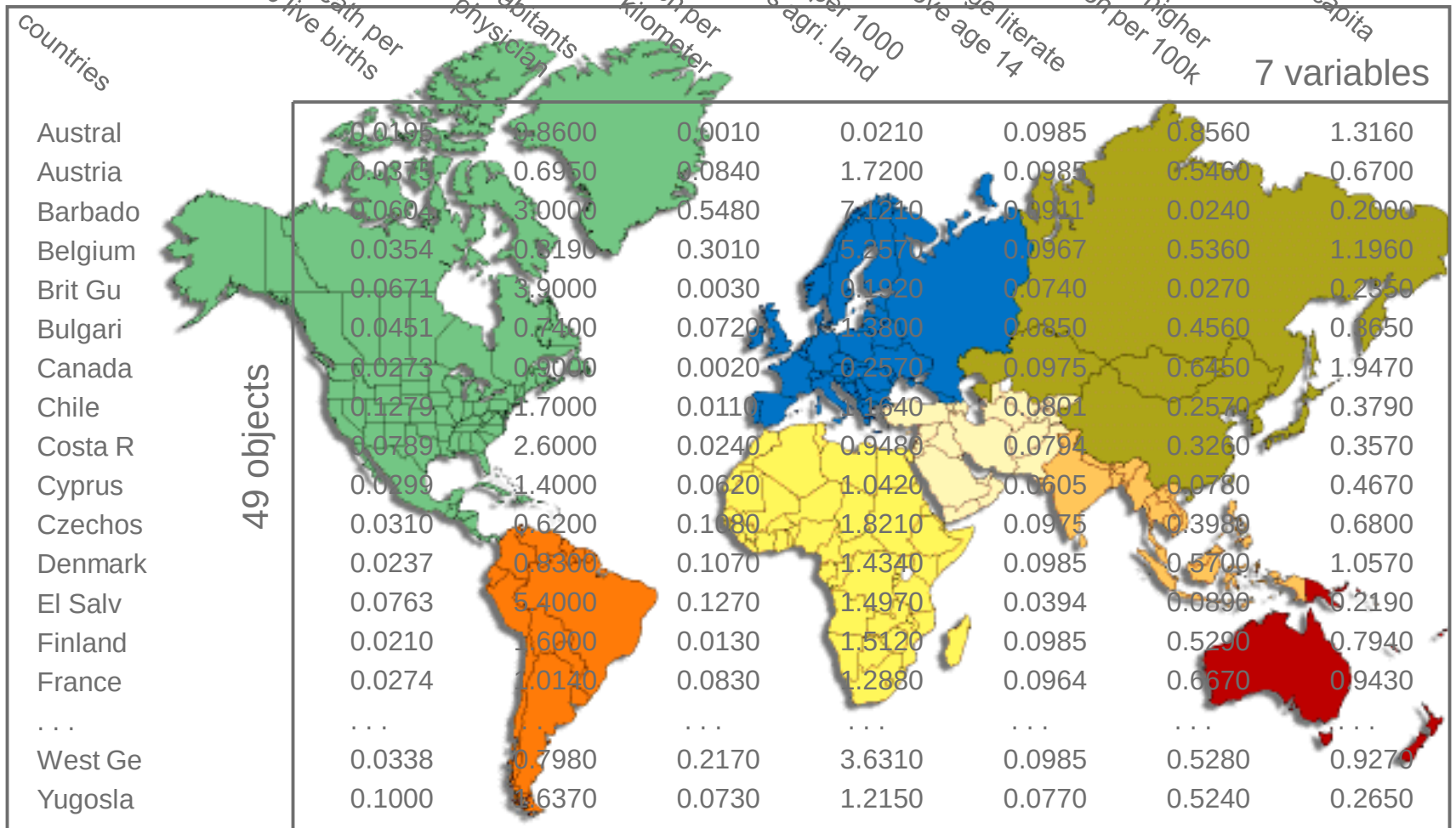
Incredibly simple questions

What European country is most similar to Japan?

What country is most bizarre?

Data analysis


www.models.kvl.dk
Workload
Distance to work
Salary
Smith
1.0
0.2
1.2
Johnson
2.0
0.0
0.3
Williams
-1.0
0.1
-1.0
Jones
-2.0
0.2
-0.1
Davis
0.0
-0.4
-0.4

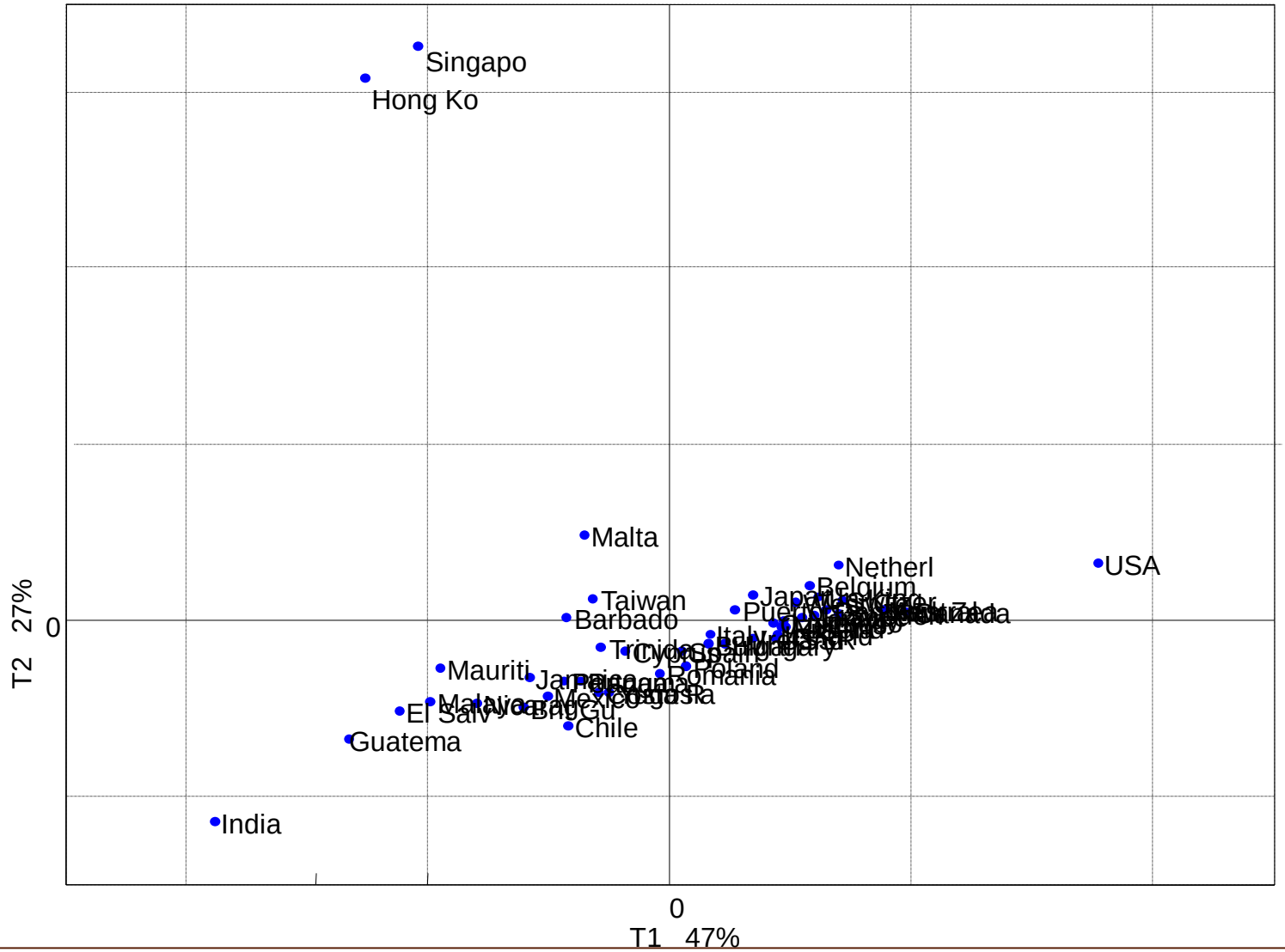


countries	Infant death per 1000 live births	# of inhabitants per physician	population per square kilometer	population per 1000 hectares agri. land	percentage literate above age 14	# students higher education per 100k	GNP per capita
	49 objects						7 variables
Austral	0.0195	0.8600	0.0010	0.0210	0.0985	0.8560	1.3160
Austria	0.0375	0.6950	0.0840	1.7200	0.0985	0.5460	0.6700
Barbado	0.0604	3.0000	0.5480	7.1210	0.0911	0.0240	0.2000
Belgium	0.0354	0.8190	0.3010	5.2570	0.0967	0.5360	1.1960
Brit Gu	0.0671	3.9000	0.0030	0.1920	0.0740	0.0270	0.2850
Bulgari	0.0451	0.7400	0.0720	1.3800	0.0850	0.4560	0.3650
Canada	0.0273	0.9000	0.0020	0.2570	0.0975	0.6450	1.9470
Chile	0.1279	1.7000	0.0110	1.1640	0.0801	0.2570	0.3790
Costa R	0.0789	2.6000	0.0240	0.9480	0.0794	0.3260	0.3570
Cyprus	0.0299	1.4000	0.0620	1.0420	0.0605	0.0780	0.4670
Czechos	0.0310	0.6200	0.1080	1.8210	0.0975	0.3980	0.6800
Denmark	0.0237	0.8300	0.1070	1.4340	0.0985	0.5700	1.0570
El Salv	0.0763	5.4000	0.1270	1.4970	0.0394	0.0890	0.2190
Finland	0.0210	1.6000	0.0130	1.5120	0.0985	0.5290	0.7940
France	0.0274	1.0140	0.0830	1.2880	0.0964	0.6670	0.9430
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
West Ge	0.0338	0.7980	0.2170	3.6310	0.0985	0.5280	0.9270
Yugosla	0.1000	1.6370	0.0730	1.2150	0.0770	0.5240	0.2650

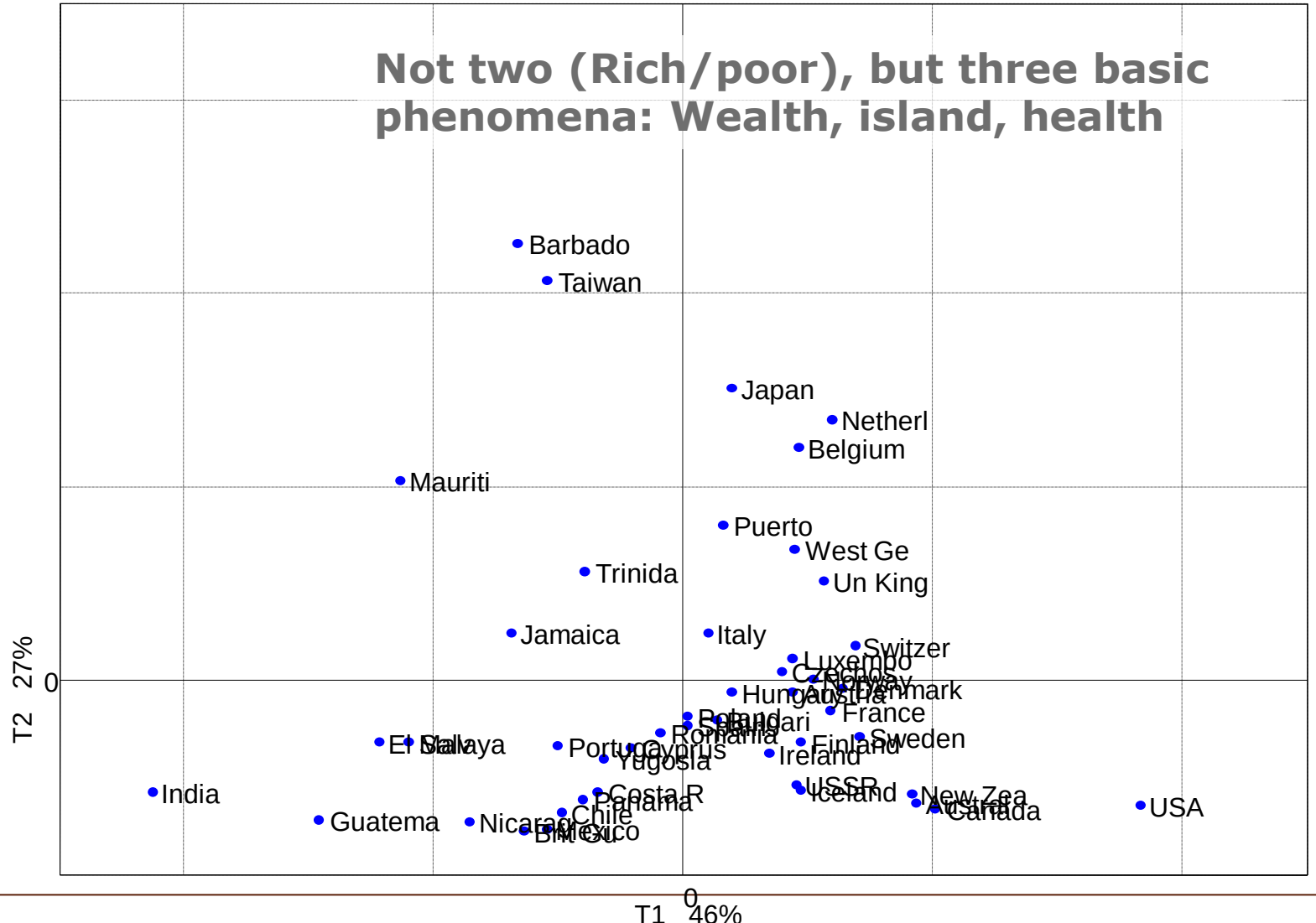
Gunst & Mason (1980) Regression analysis and its applications: A data-oriented approach, NY, Marcel Dekker, p. 358

Principal Component Analysis

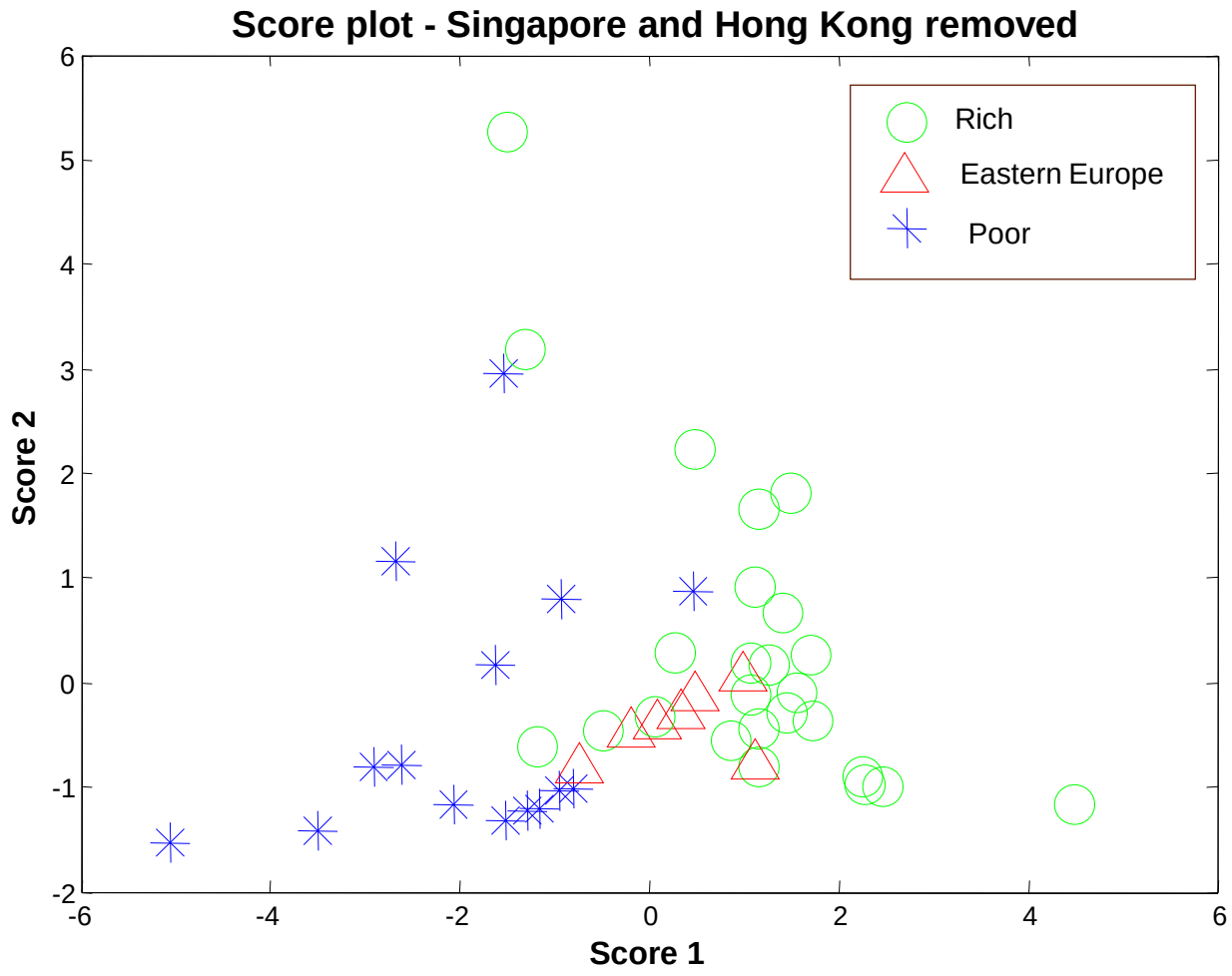
Outliers are easily spotted



Principal Component Analysis
The exploratory aspect



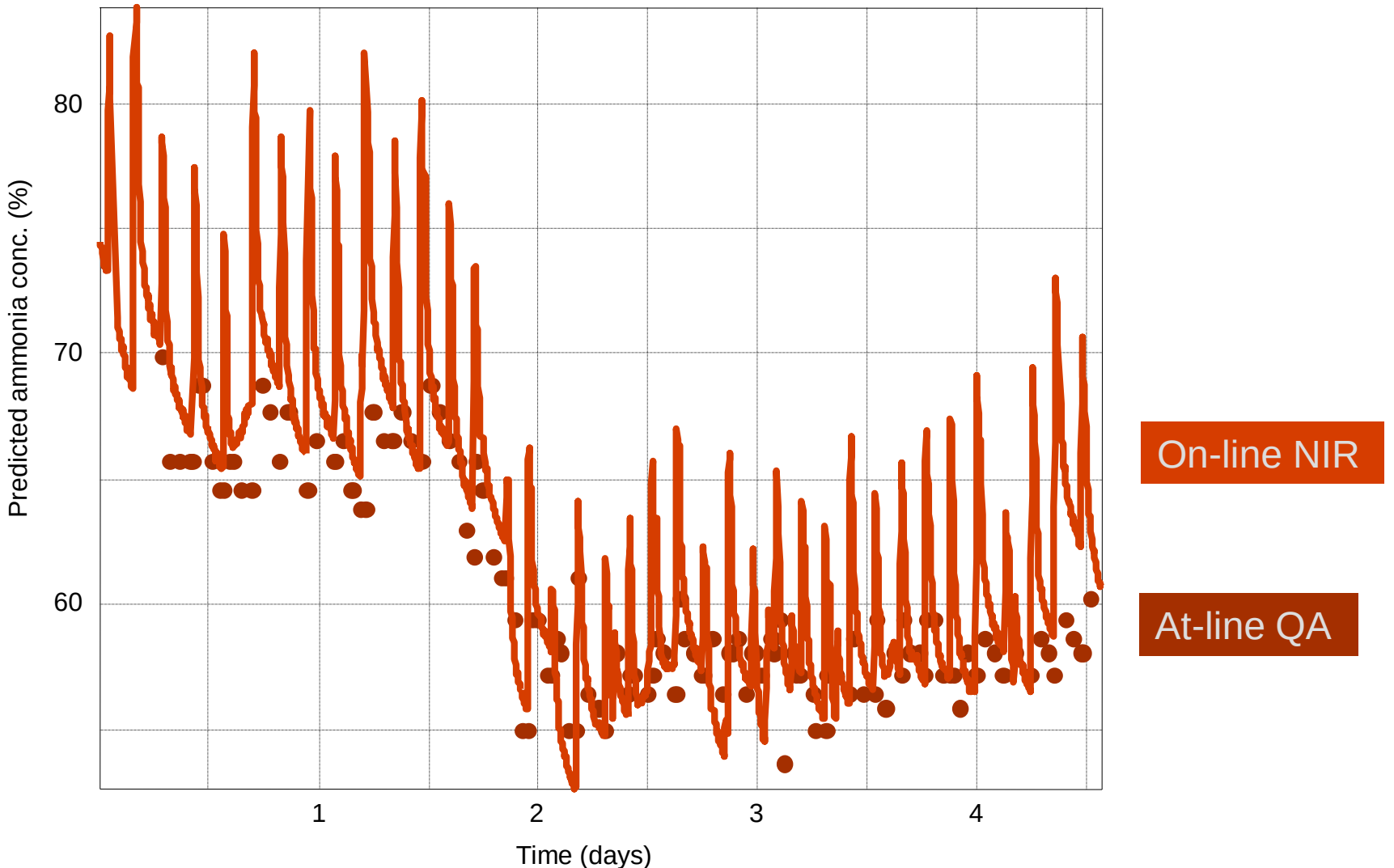
Principal Component Analysis
Using additional information



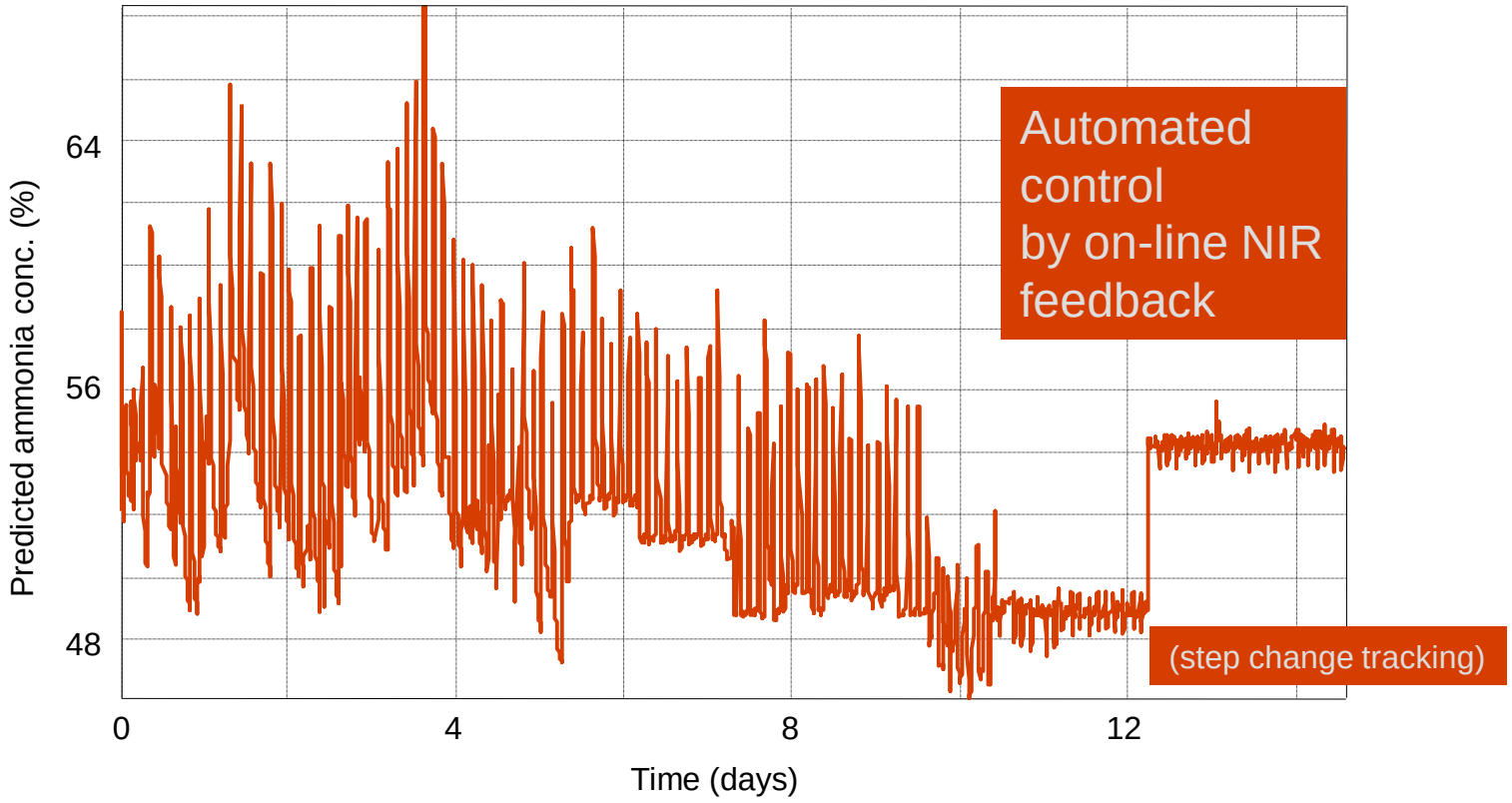
- handle many variables
- detect outliers
- find new patterns
- do true fingerprinting
- generate new hypotheses

With PCA we can …

On-line NIR versus at-line QA measurements



On-line NIR versus at-line QA measurements

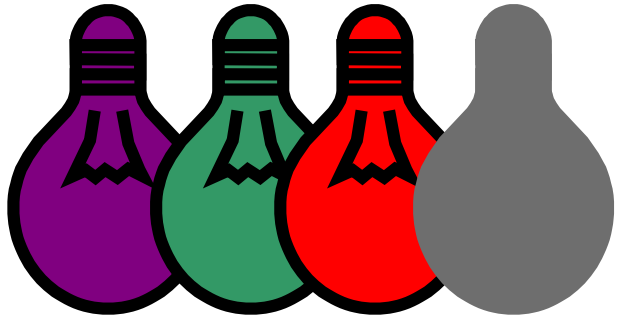


Zachariassen et al, *Use of NIR spectroscopy and chemometrics for on-line process monitoring of ammonia in Low Methoxylated Amidated pectin production* Chemometrics and Intelligent Laboratory Systems 76(2005)149-161

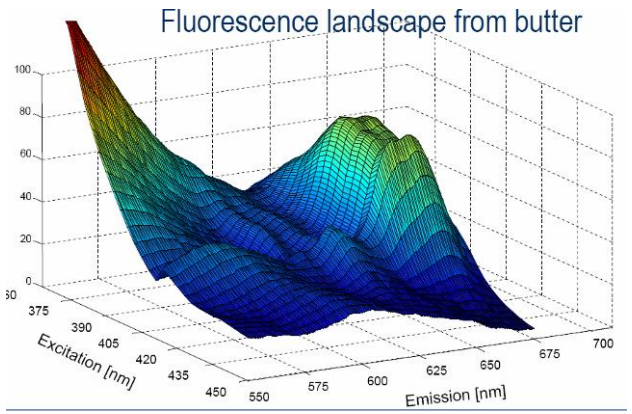
Understanding oxidation of butter and cheese

- Oxidation from light causes rancid taste of butter
- Important for packaging and shelf-storage
- Riboflavin thought to be important

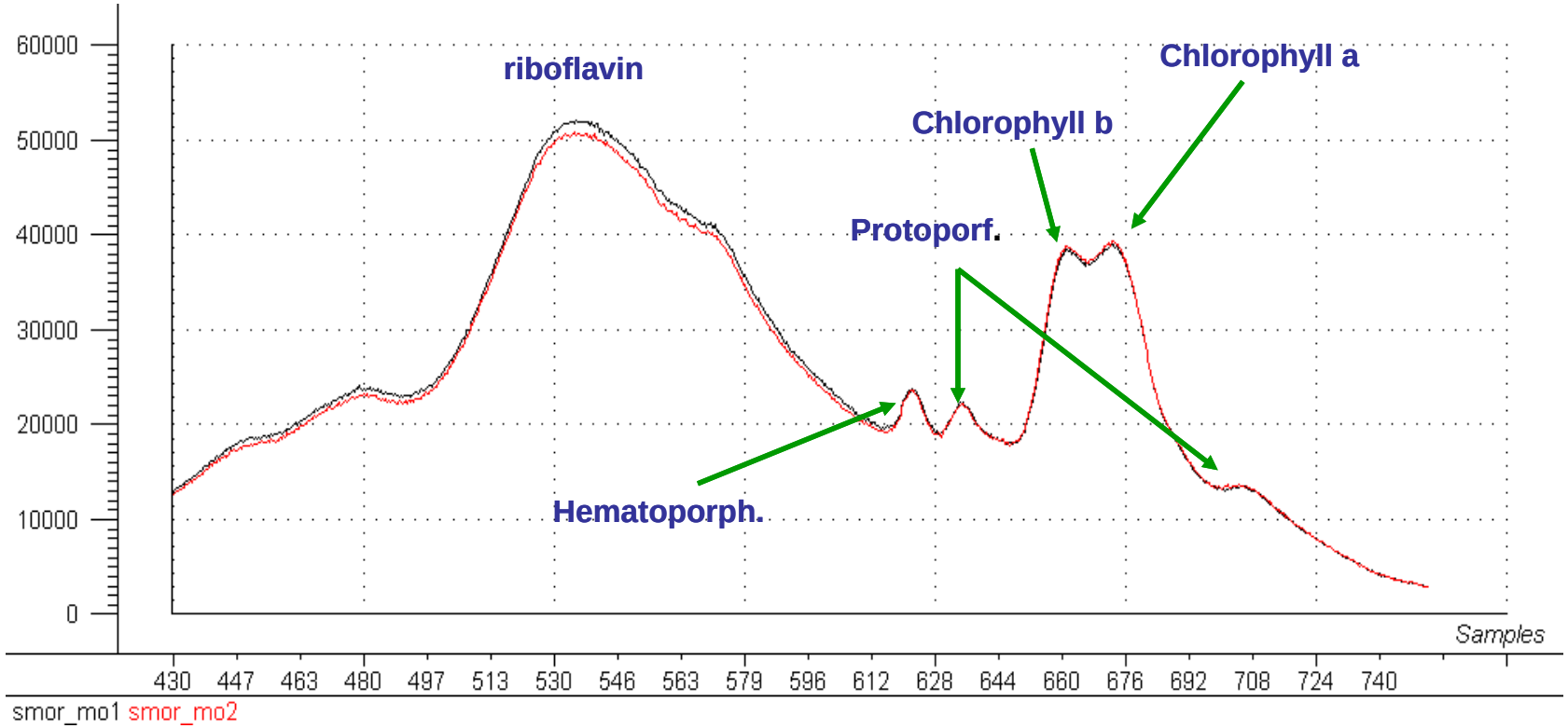
Butter at:
Different light
With/without O_2
Different storage time



- Sensory analysis (quality)
- Fluorescence EEM

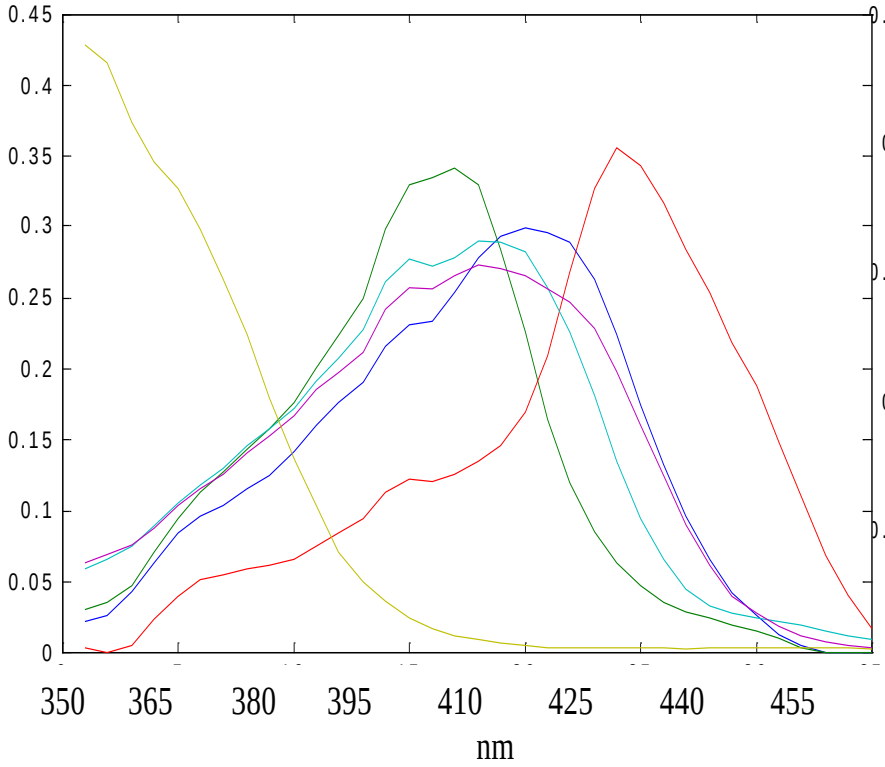


Typical data

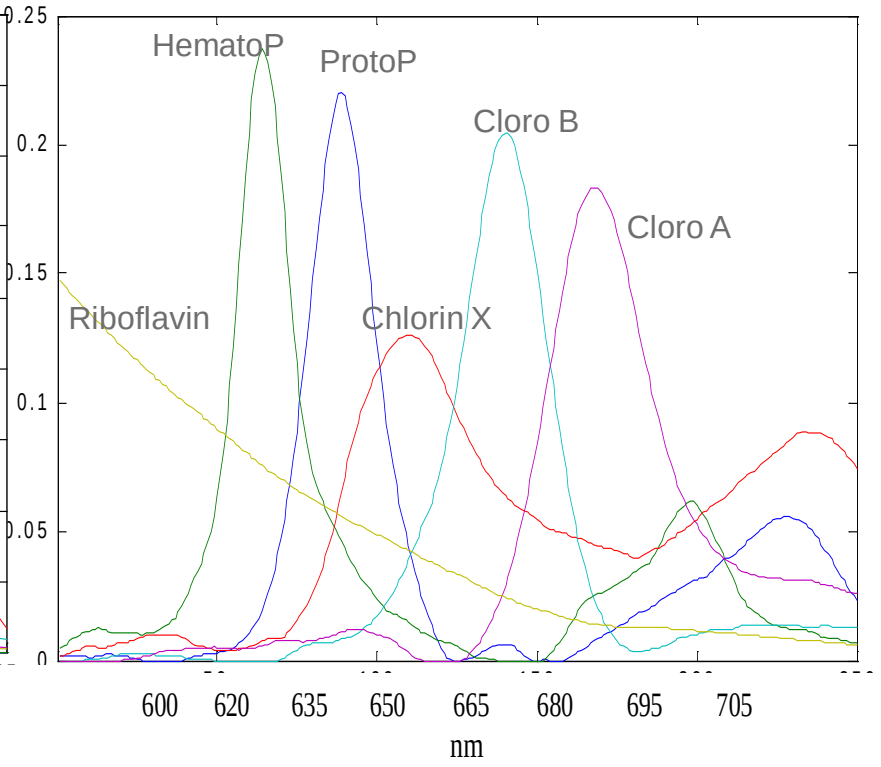


Understanding the chemistry

Excitation

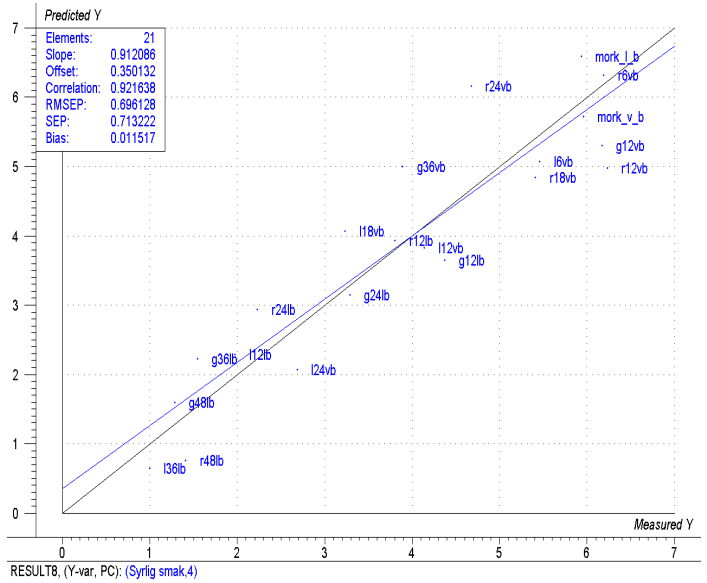


Emission



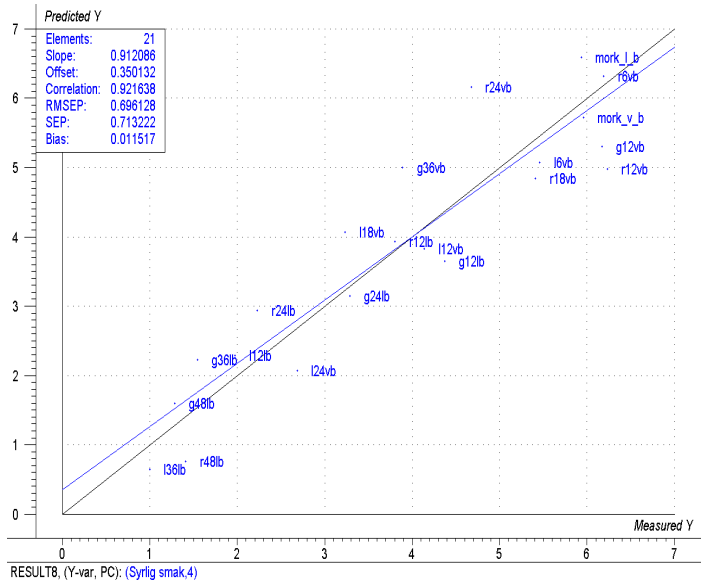
Relation between rancid taste and concentrations

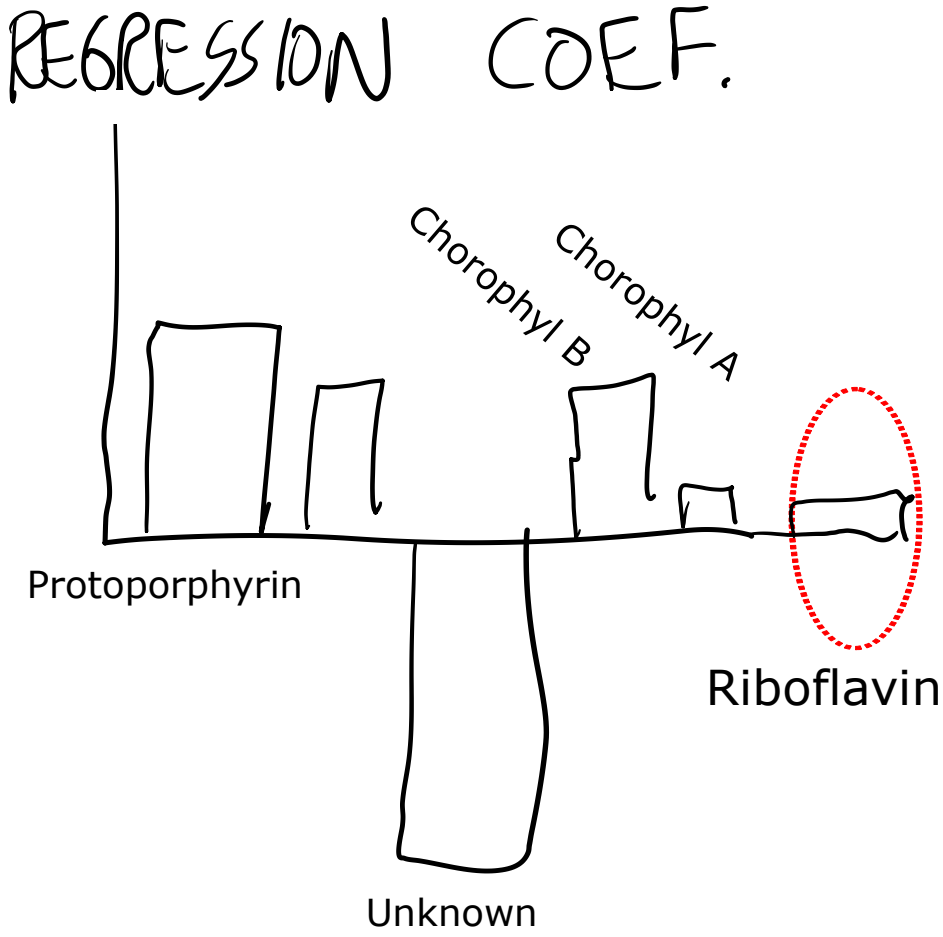
Rancid taste

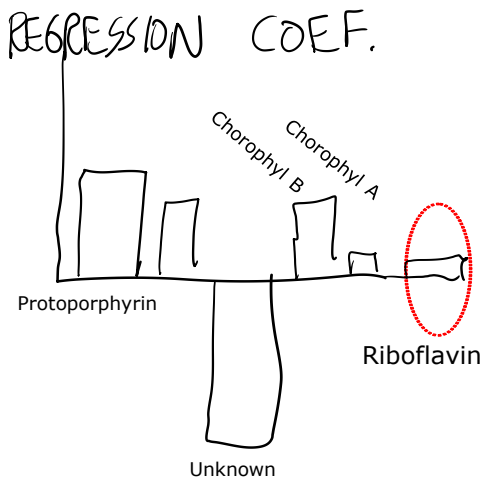


Relation between rancid taste and concentrations

Rancid taste





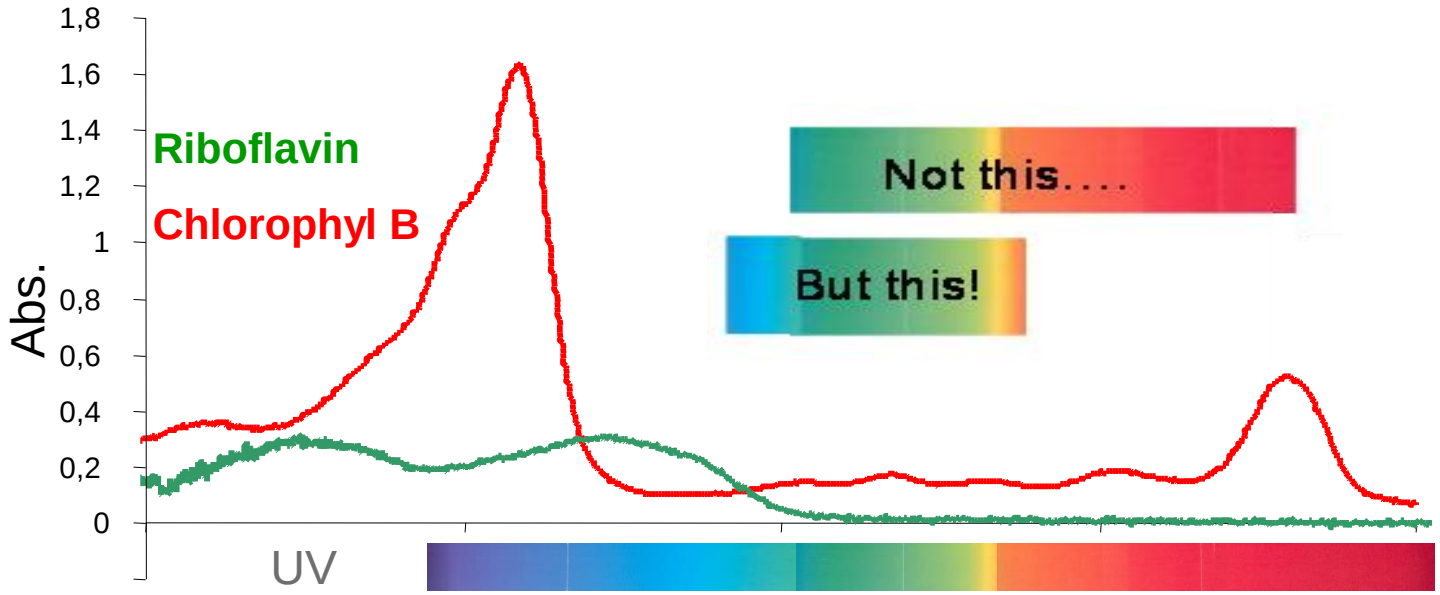


Oxidation:

Not just riboflavin

'New' ones seem more important

Affects optimal packaging



Examples

- Predicting consumer liking
- Detecting adulteration
- Understanding protein structure
- Optimizing flavor of cheese
- Quality control of raw material

www.models.life.ku.dk

M-files Papers Courses

Data sets And many other things