Все что вы хотели знать о хемометрике – но стеснялись спросить

Содержание

1. Введение
2. Кинетика
3. Промышленность
4. Фармацевтика
5. Многие приложения
6. Информация

Хемометрика: два определения

Дедуктивное

Хемометрика - это научная дисциплина, находящаяся на стыке химии и математики, предметом которой являются математические методы исследования химических данных

сайт Российского хемометрического общества

Индуктивное

Хемометрика – это то, что делают хемометрики.

сайт Международного хемометрического общества

Хемометрики – это такие люди, которые все время пьют пиво и воруют идеи у математиков

Svante Wold

Что делает хемометрика?

- Хемометрика имеет дело с **данными** (зачас-тую с очень большими), поэтому хемометрика - это подраздел информатики (Data mining)
- Данные, которые исследует хемометрика по большей части происходят из **химии**, поэтому хемометрика - это подраздел химии (Analytical chemistry)
- Методы, которые использует хемометрика ориентированы на **формальное** моделирование (Soft modeling)

Почему «хемо-» ?

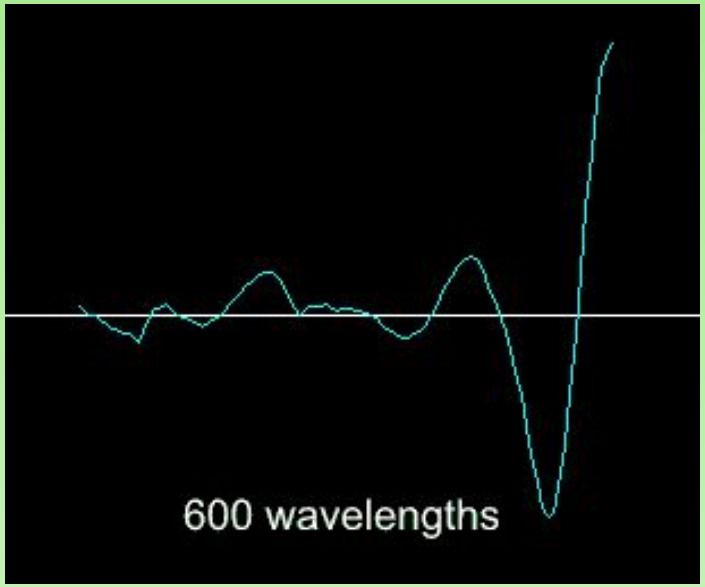
- Хемометрика родилась из задачи анализа химических **спектров**
- Спектроскопия – наилучший метод получения информации по ходу процесса (**on-line**) в режиме реального времени: быстро и без влияния на процесс
- «Хемо» подчеркивает **практическую**, а не статистическую значимость применяемых методов

Почему «-метрика» ?

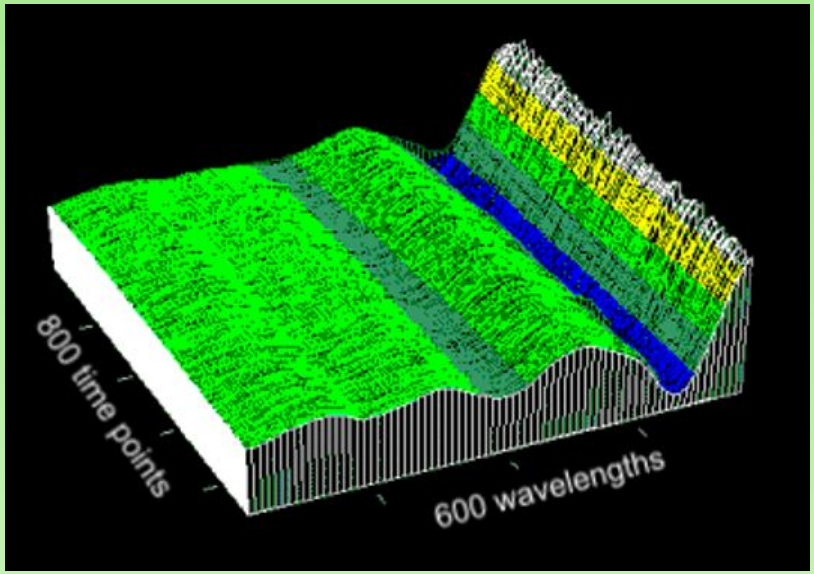
- Хемометрические методы легко и плодотворно **переносятся** в другие области, например, в психологию, биологию, геологию, и т. д.
- Хемометрика активно эксплуатирует **математику** статистику, линейную алгебру
- ‘It is easier to teach a chemist statistics that to teach chemistry to a statistician.’ (Svante Wold)

Много переменных и много измерений

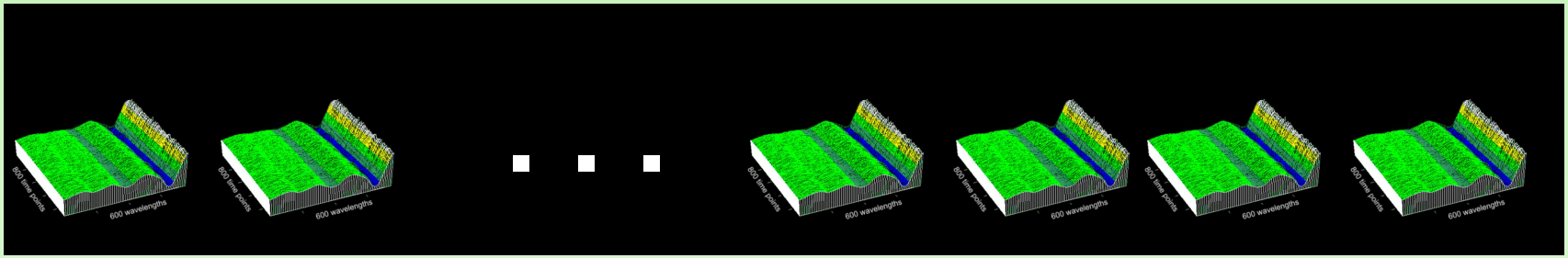
Одно измерение – спектр (600 точек)



Один цикл – 800 спектров (времен)



Один массив данных – 200 образцов (циклов)



Основная задача хемометрики

Заменить **прямые** измерения, которые либо –

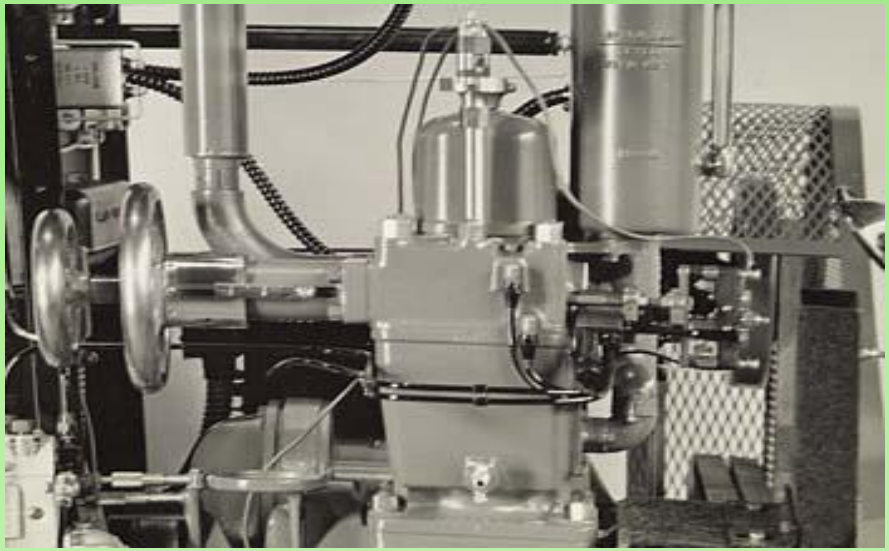
- ❑ невозможны
- ❑ дороги
- ❑ длительны

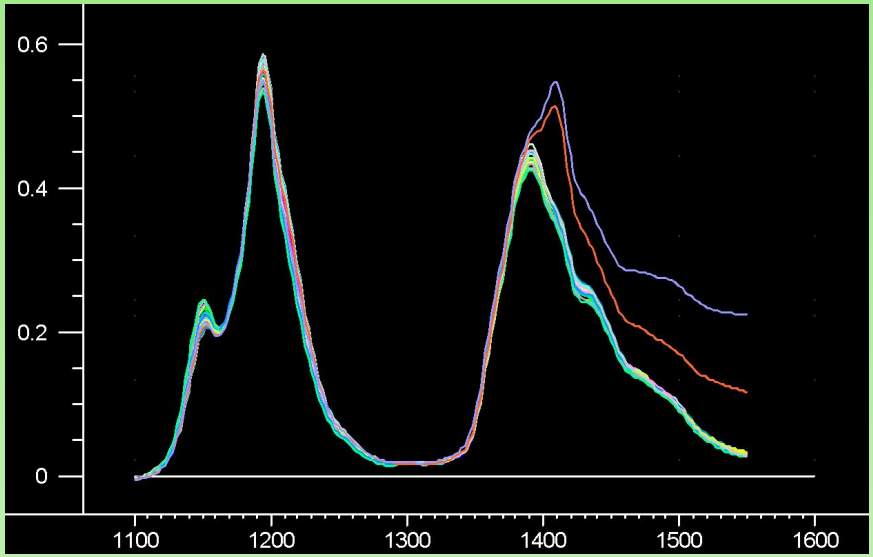
на **косвенные** измерения, которые –

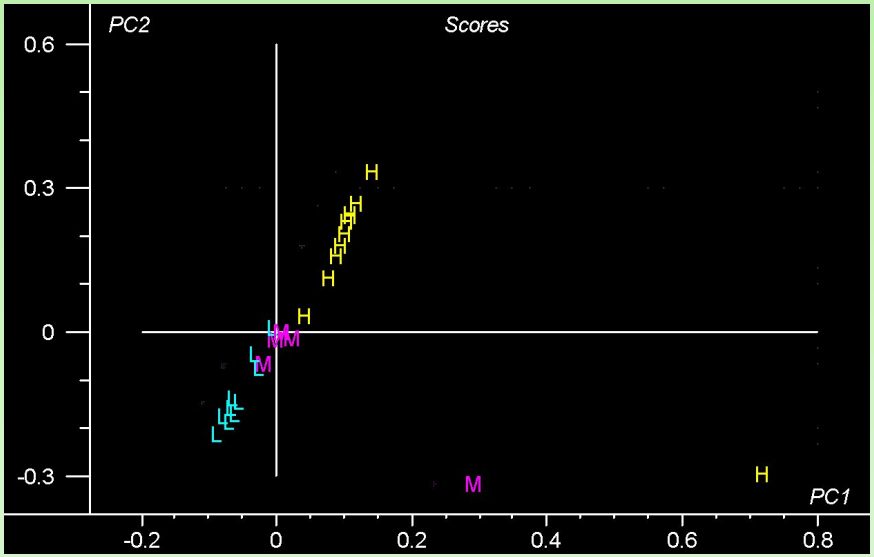
- ❑ доступны
- ❑ дешевы
- ❑ быстры

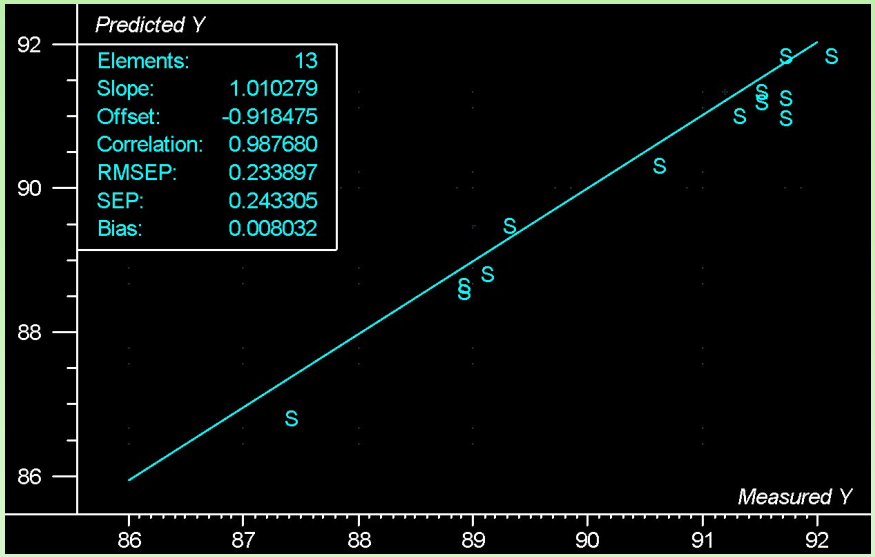
с последующей их **обработкой** (калибровкой).

Определение качества бензина по ИК-спектру









Как определить число ГК?

L6. Defining Multivariate Calibration Model Complexity for Model Selection and Comparison

John Kalivas

Idaho State University, Pocatello, USA

Формальные и содержательные модели

	Содержательные “Hard” models	Формальные “Soft” models
Откуда	Физика, химия, ….	Из данных
Формула	$\mathbf{y}=f(\mathbf{x},\mathbf{a})+\boldsymbol{\varepsilon}$	$\mathbf{y}=\mathbf{X}\mathbf{a}+\boldsymbol{\varepsilon}$
Параметры	Имеют физ. смысл	Физически бессмысленны
Проблемы	Построить модель	Обработать данные
Назначение	Экстраполяция	Интерполяция
Пример	Хим. кинетика	ANOVA

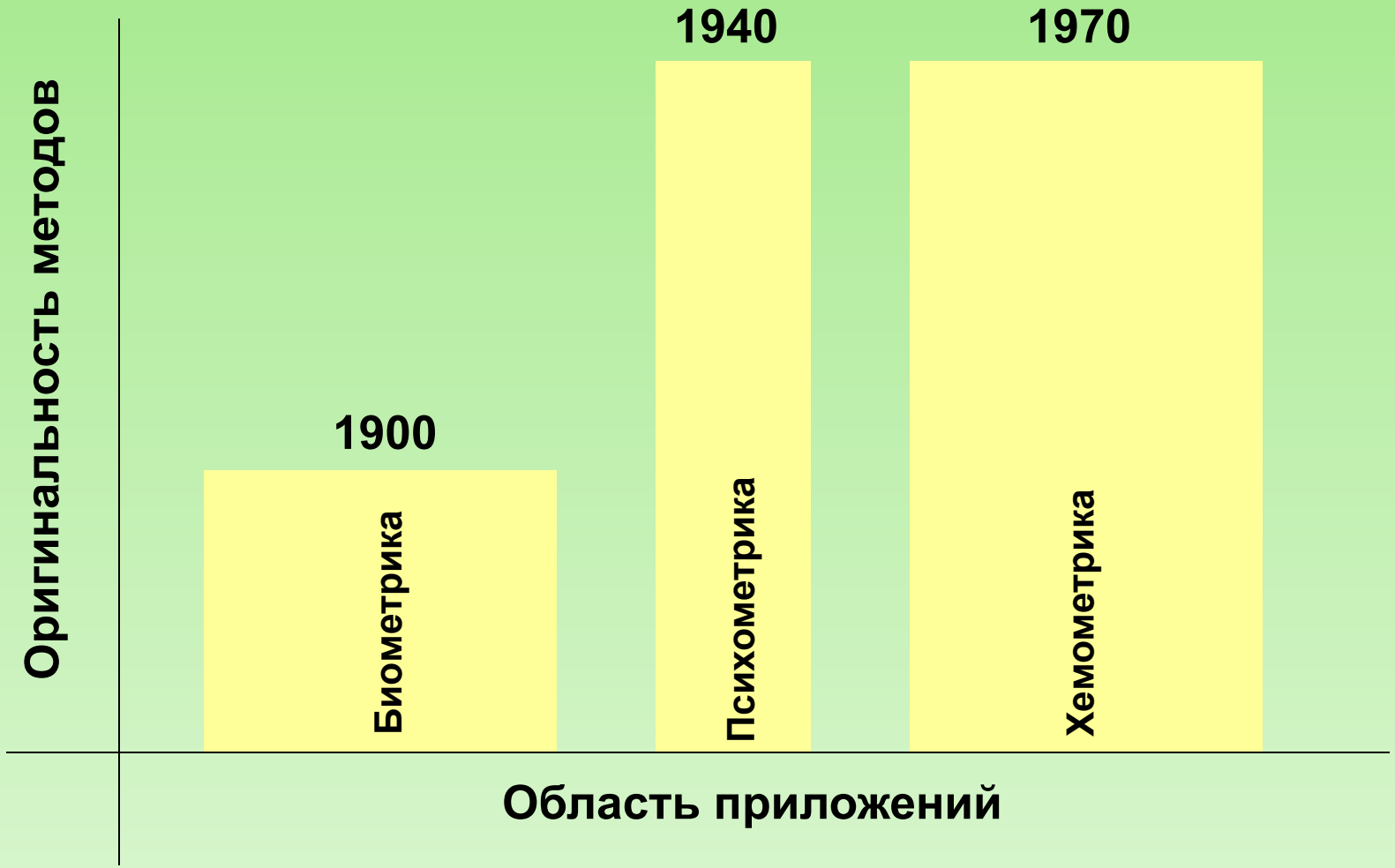
Белые, серые и черные модели

L2. Gray Modelling Approaches to Investigate Chemical Processes

Roma Tauler

Institute of Chemistry and Environmental Research, CSIC, Barcelona, Spain

Хемометрика среди своих сестер



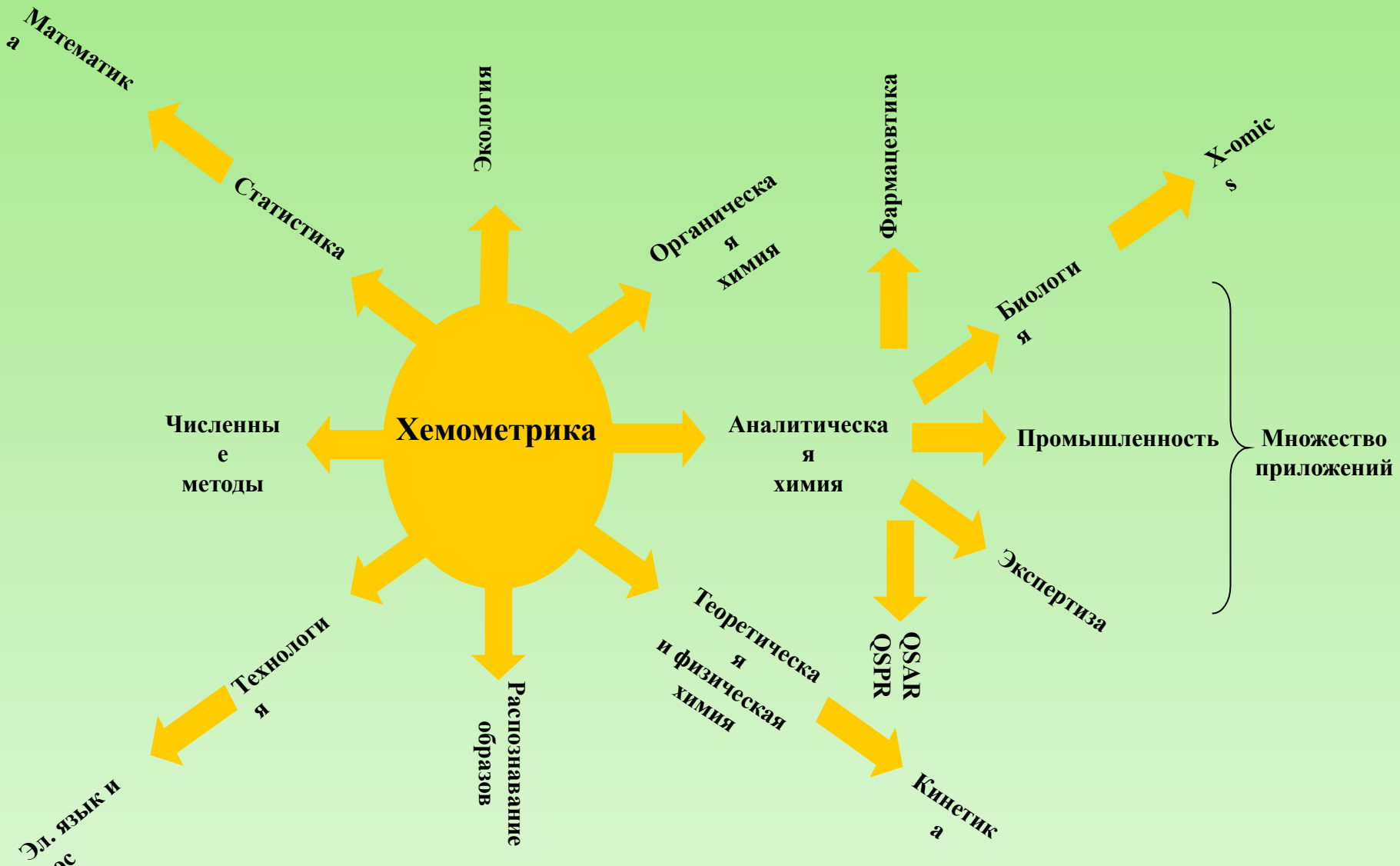
H – принцип. Основы хемометрики

L3.The H-principle of Mathematical Modeling

Agnar Höskuldsson

IPL, DTU, Kgs Lyngby, Denmark

Хемометрика



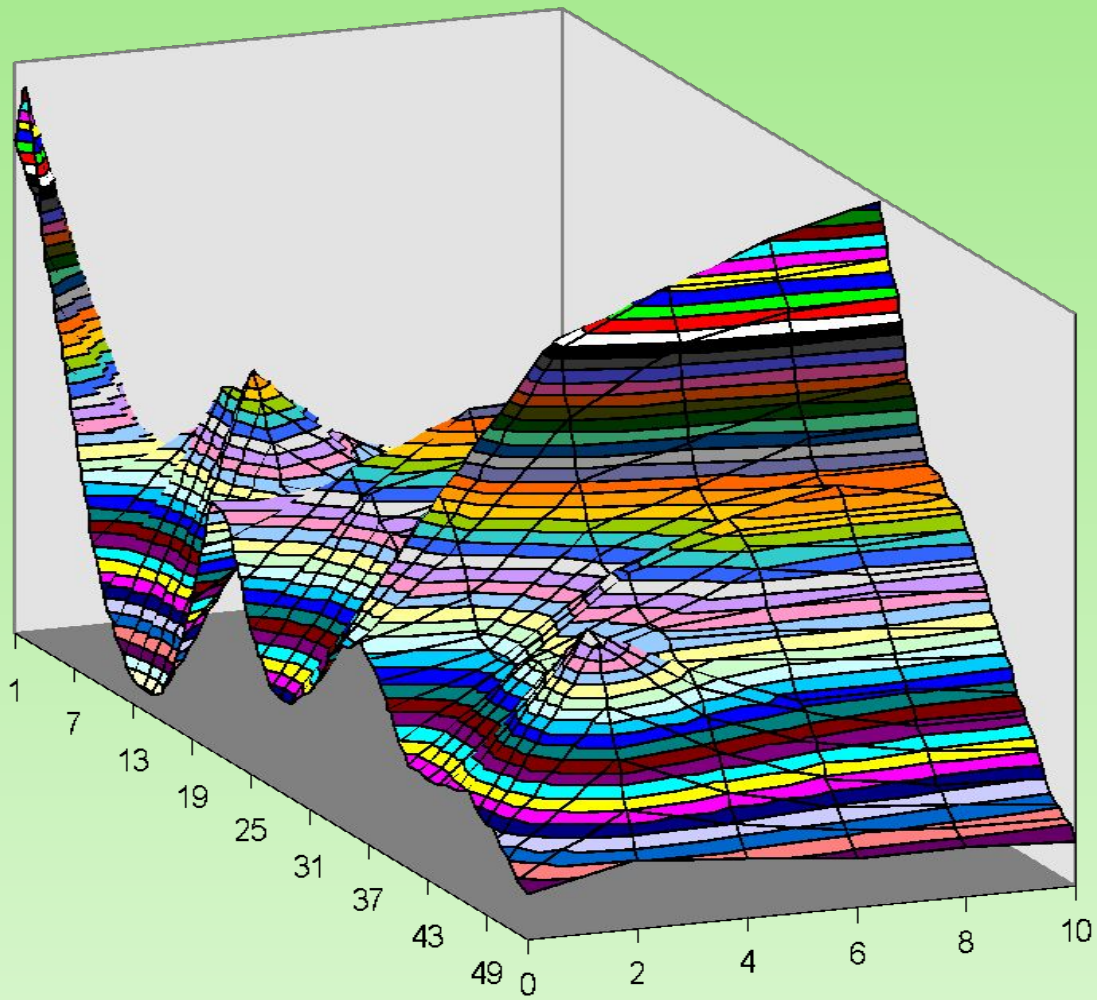
Хемометрика – от археологии до астрономии

L5. Exploration and Classification: Applications from Archaeometry to Spectroscopy

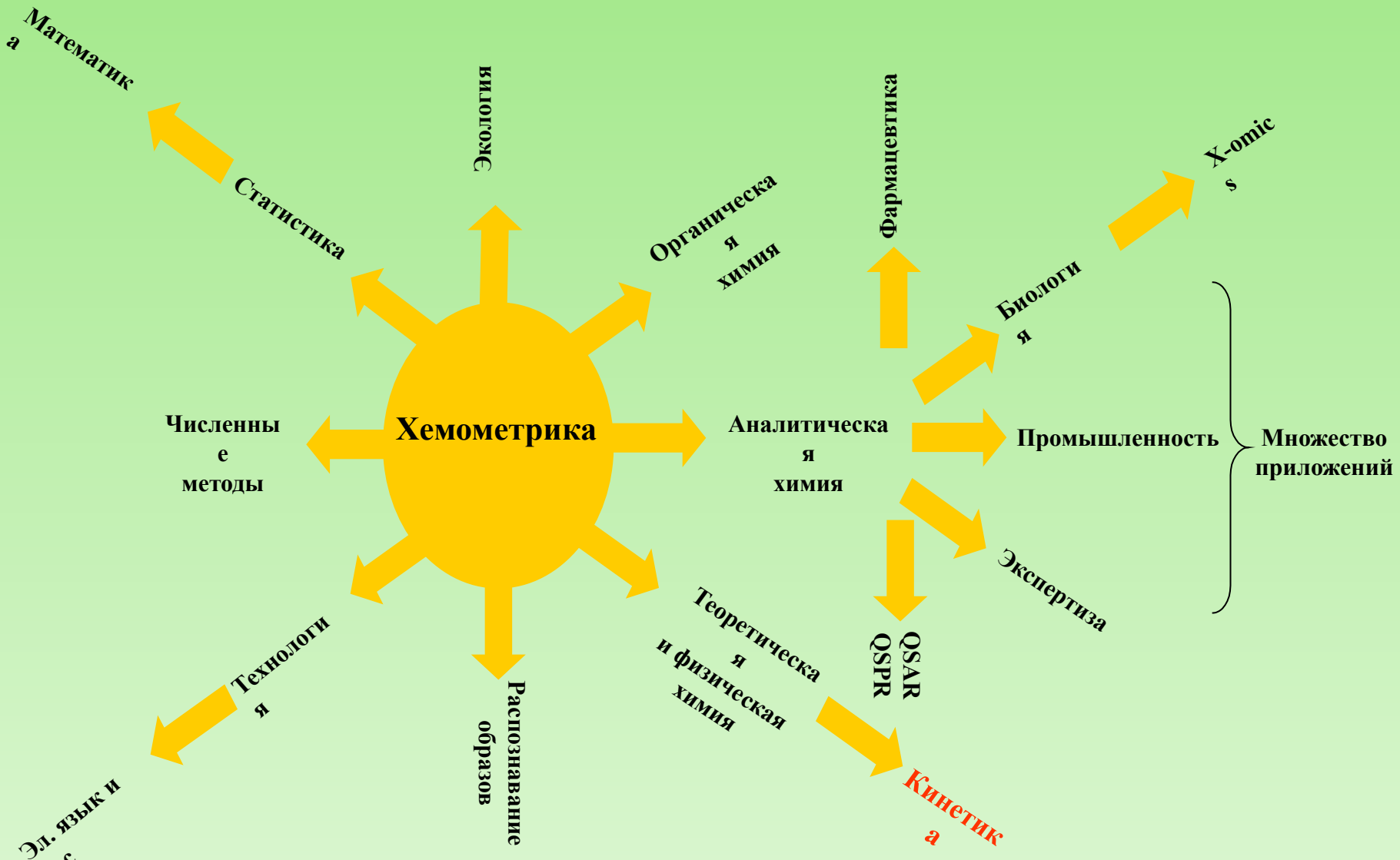
Kurt Varmuza

Vienna Technical University, Vienna, Austria

Кинетика – curve resolution



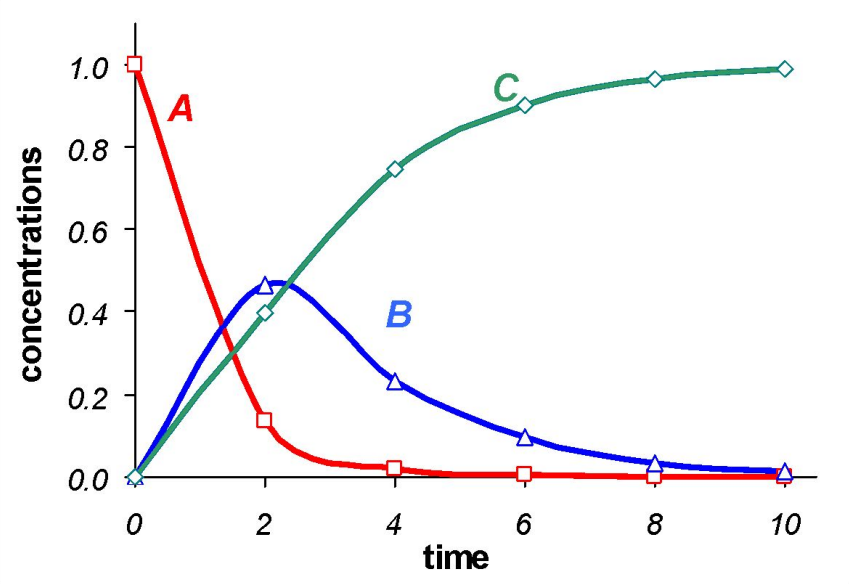
Хемометрика



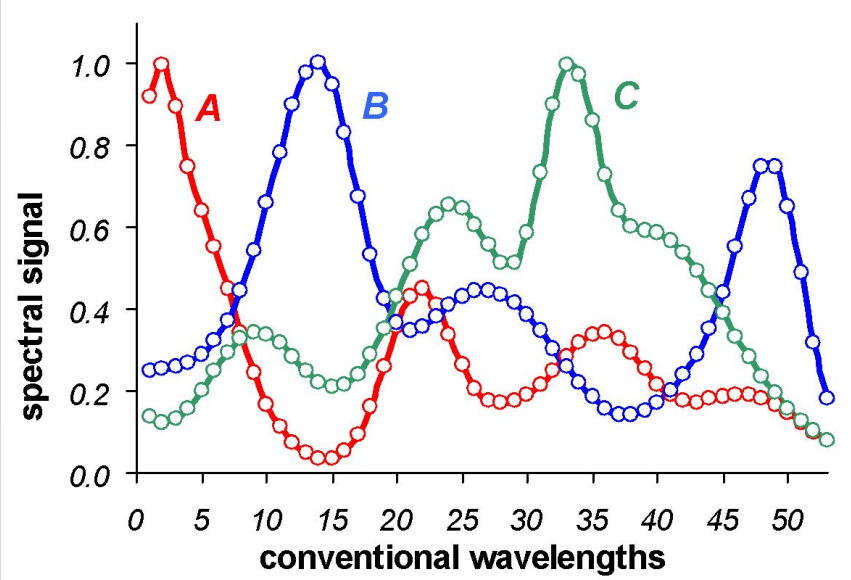
Как найти кинетику по спектрам?

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C$$

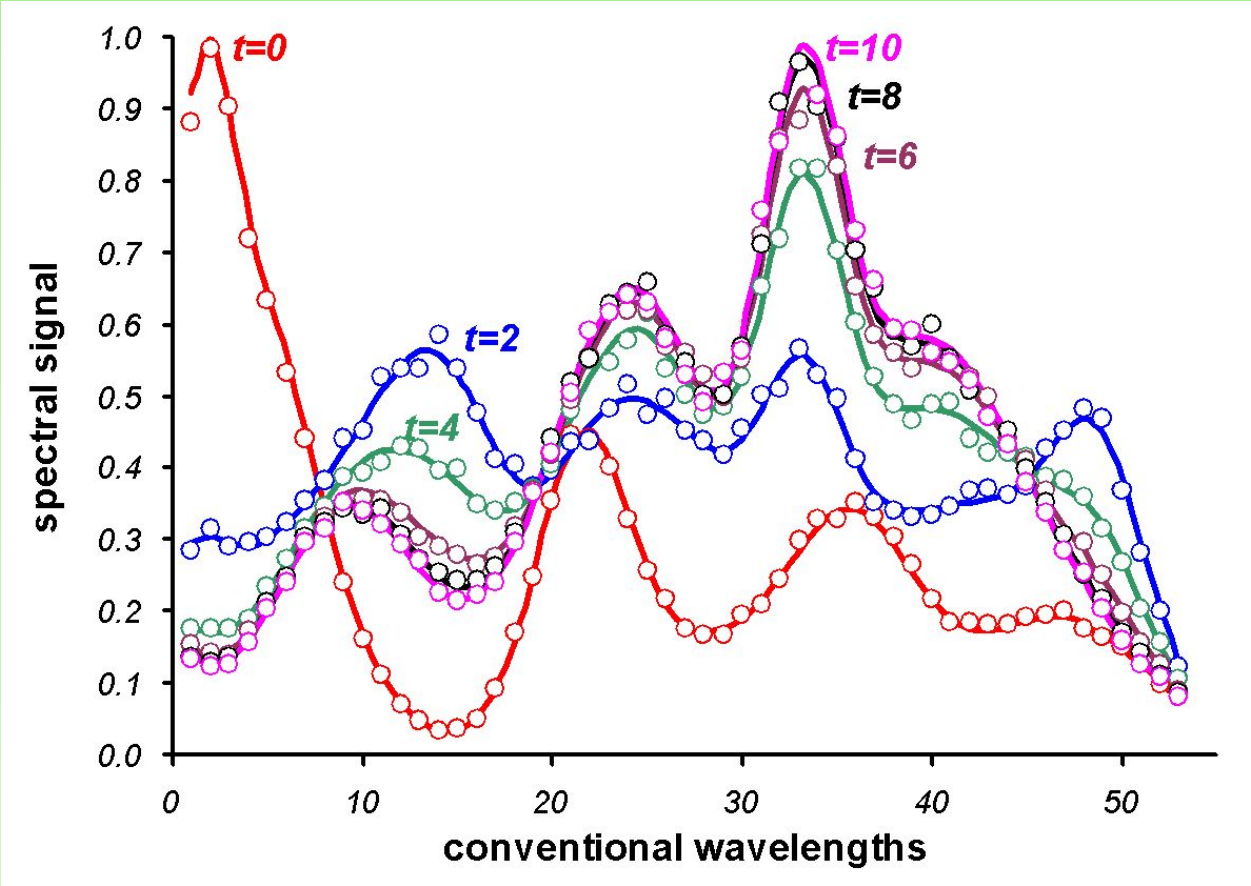
Концентрации



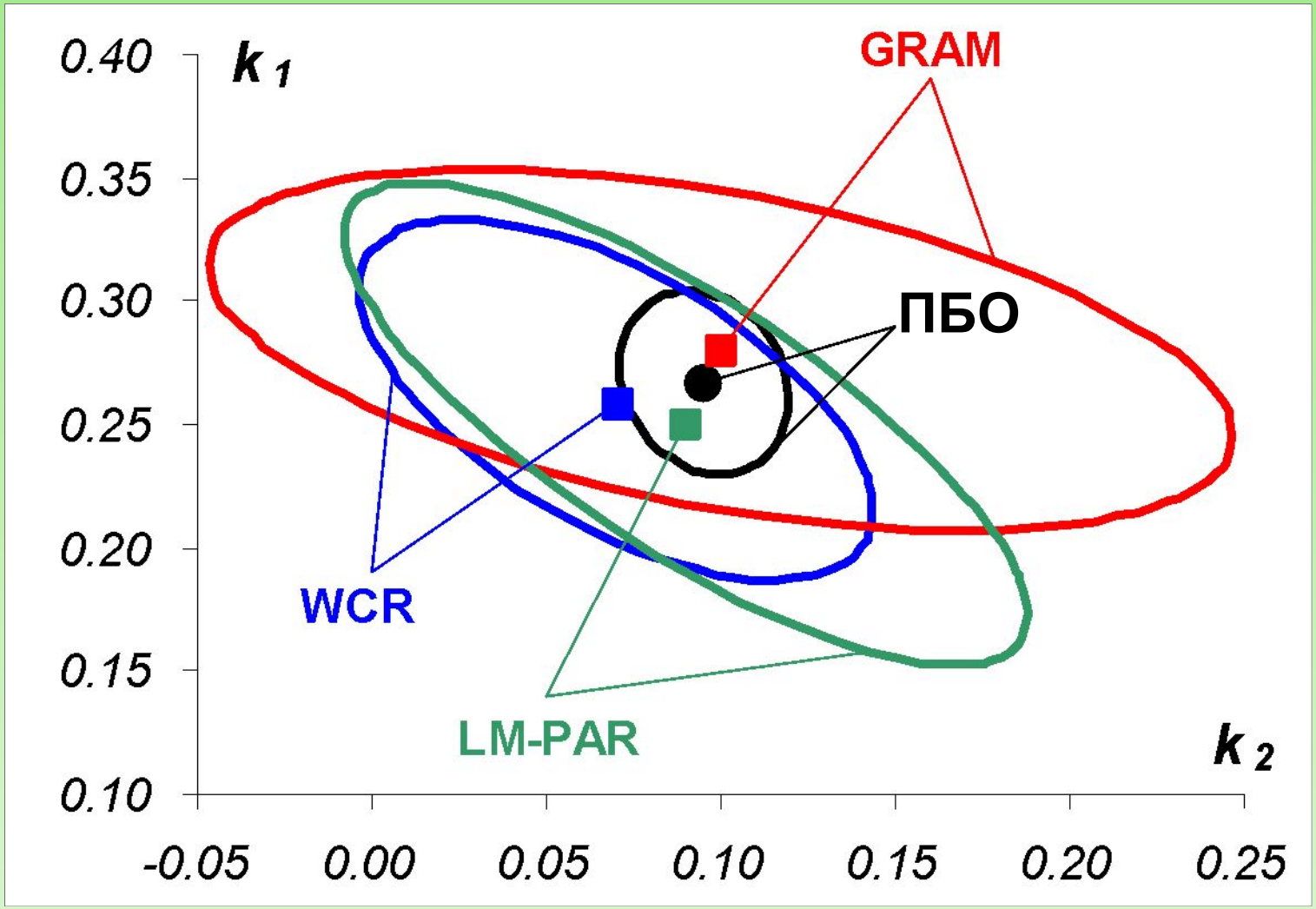
«Чистые» спектры компонент



Эволюция спектра компонент реакции



Оценки кинетических констант



Кинетика и хемометрика

L1.Principal Component Analysis in Photochemistry

Владимир Разумов

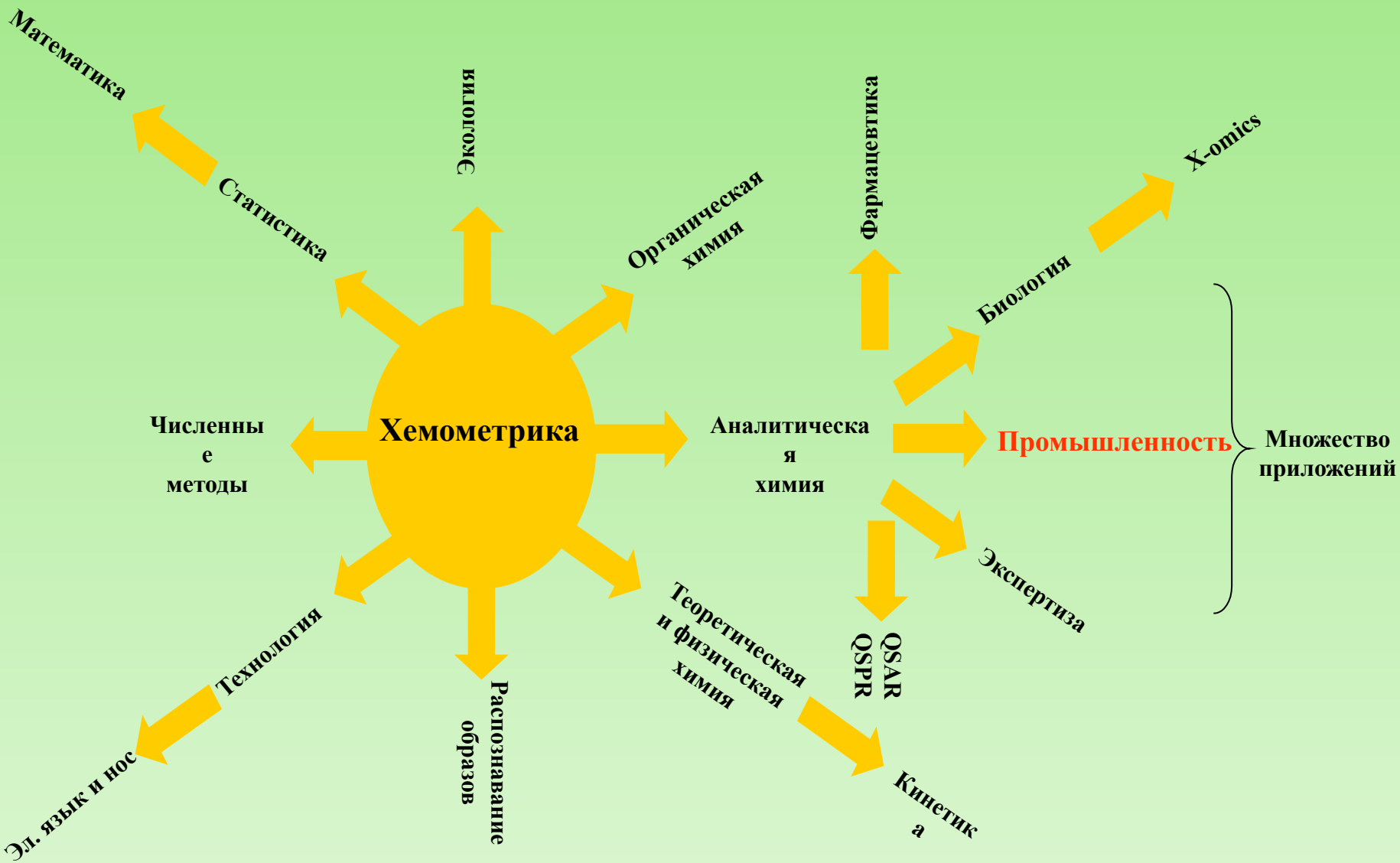
Institute of Problems of Chemical Physics, Chernogolovka, Russia

Хемометрика в промышленности

по материалам Nouna Kettaneh-Wold, Sweeden

и Richard Brereton, UK

Хемометрика



PAT & FDA

Process Analytical Technology (PAT) =
Технология (методы) анализа процессов

PAT =
Статистический контроль процессов (SPS) +
Хемометрика (Chemometrics)

FDA = U.S. Department of Health and Human Services Food and Drug Administration

Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance
Pharmaceutical CGMPs, September 2004

Многомерный статистический контроль процессов (MSPC)

MSPC – это математический анализ реальных исторических данных, характеризующих опыт работы

- **<u>Цель:</u>** Научиться у самих себя принимать оптимальные решения в различных ситуациях.
- **<u>Средства:</u>** Сбор истории работы процесса, а также анализ накопленных данных.
- **<u>Результат:</u>** Снижении затрат при стабилизации качества.

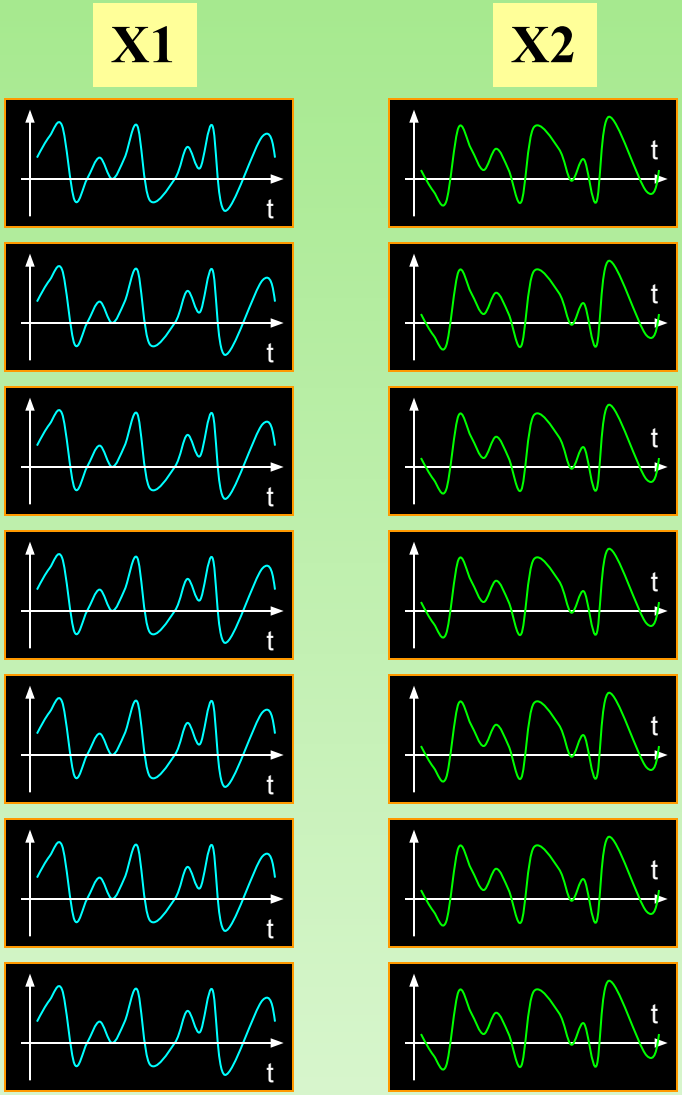
MSPC – это один из методов анализа процессов (PAT)

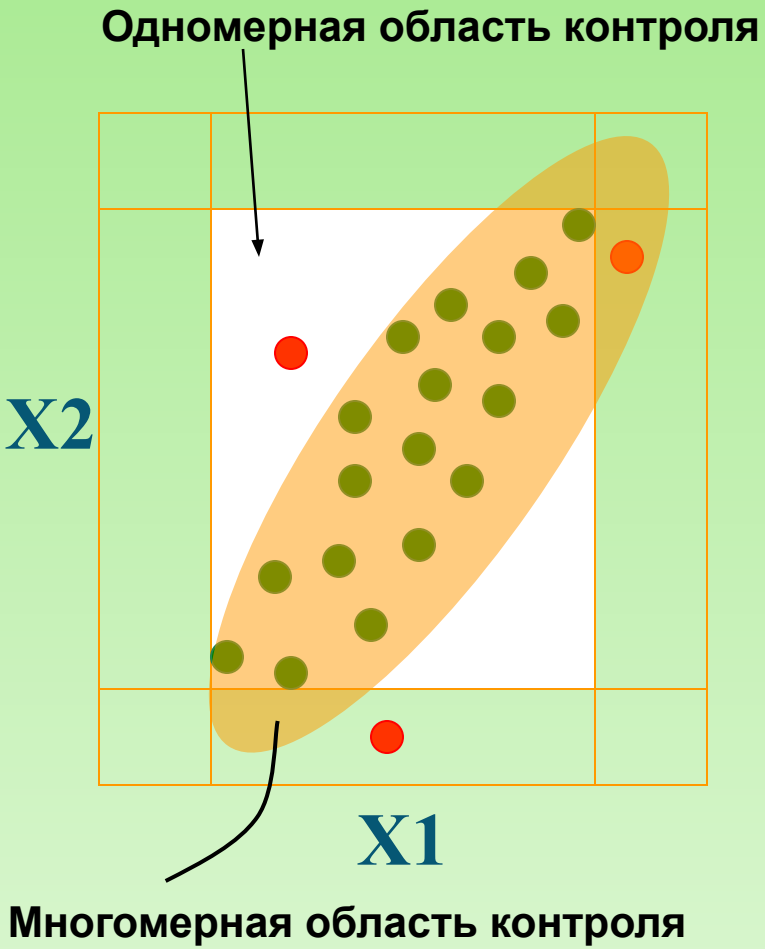
Зачем в MSPC нужна хемометрика?

Потому, что все больше данных о процессах получают с помощью современных, эффективных приборов

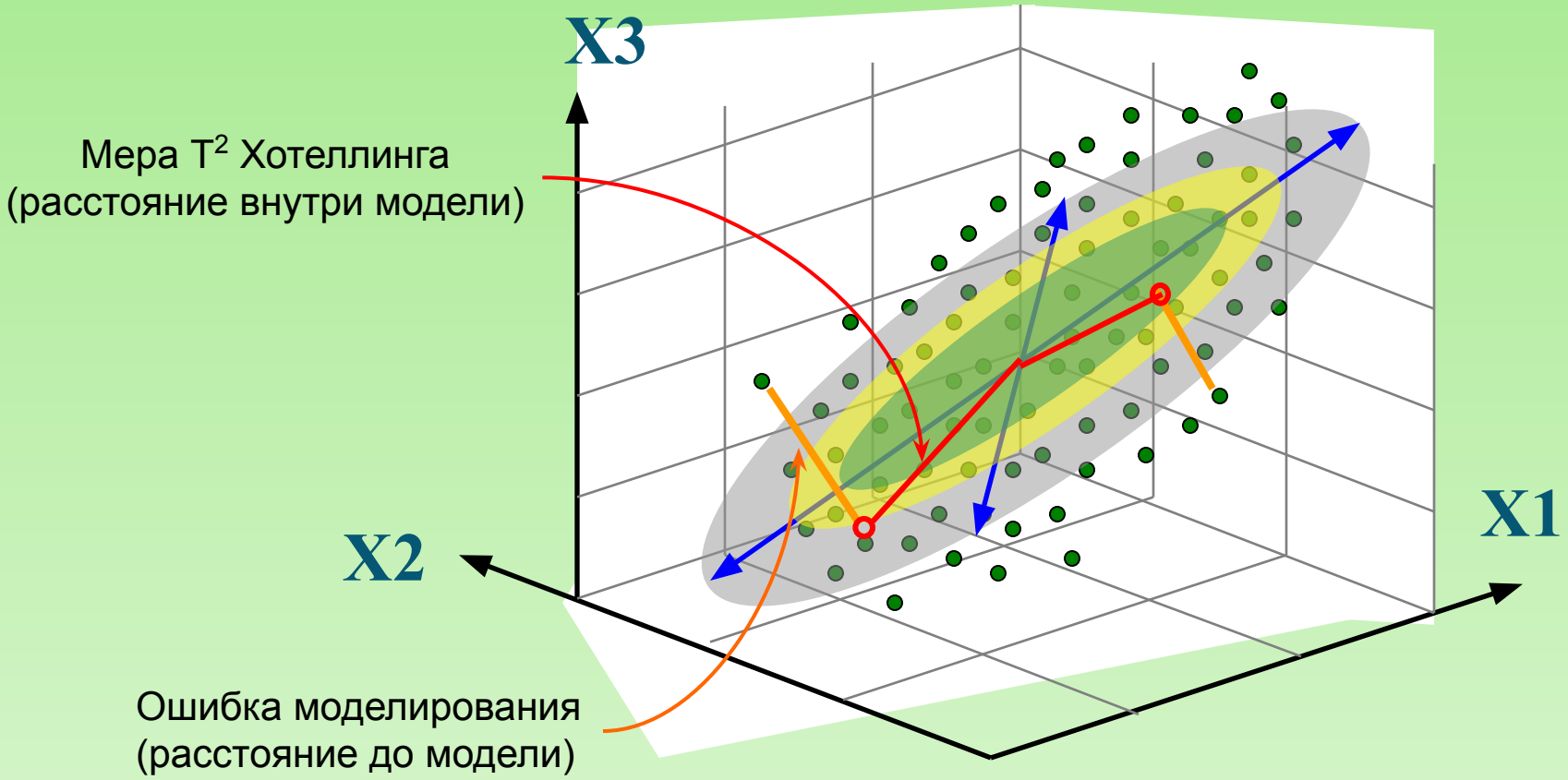
- Макропеременные (температура, pH, давление, ...)
- Спектроскопия (УФ, ИК, БИК, ...)
- Хроматография (ЖХ, ГХ, ...)
- Гибридные методы (ЖХ/МС, ЖХ/ЯМР, ...)
- Видео образы и гиперспектры (2D, 3D)

Контроль с помощью SPC





Контроль с помощью MSPC



Периодические (batch) процессы

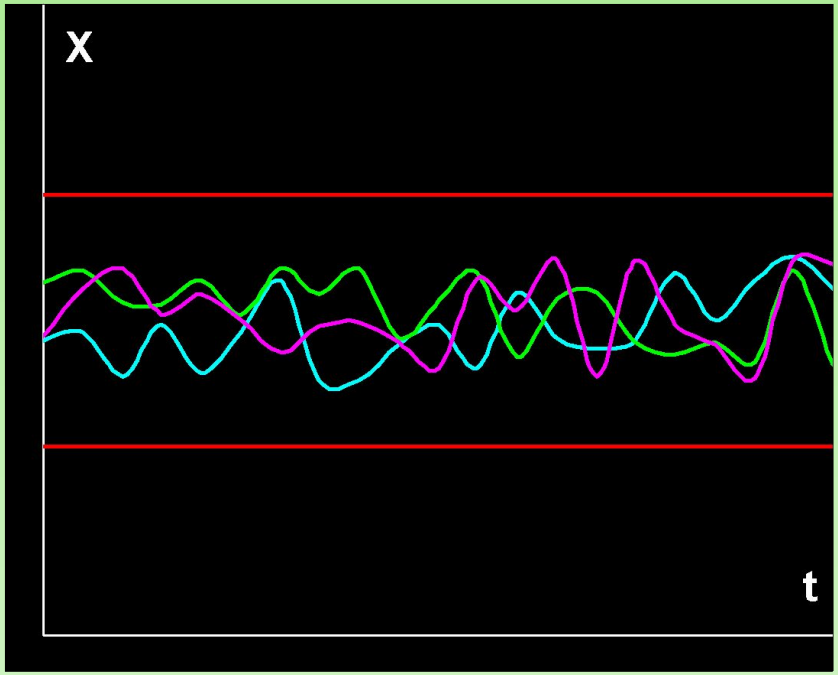
Периодические процессы применяются при производстве лекарств, полимеров, пищевых продуктов, и т. п. Они характеризуются –

- Конечной продолжительностью
- Кинетикой протекания
- Изменчивостью, как внутри, так и между циклами
- Частой сменой производимых продуктов

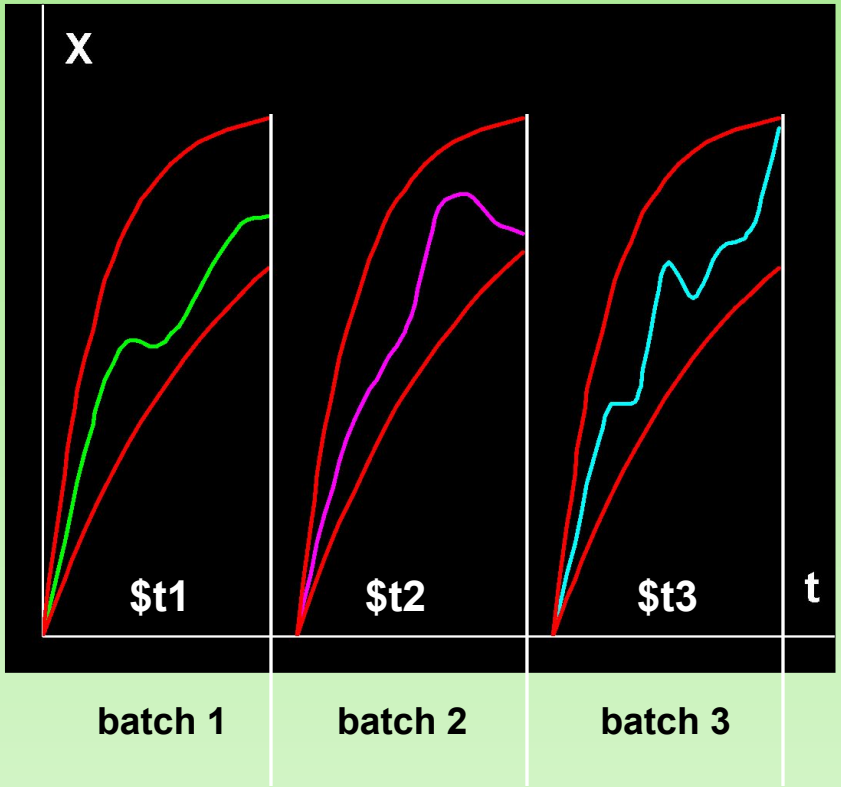
В непрерывных процессах важны только взаимоотношения между переменными, тогда как в периодических важно знать как переменные **меняются во времени**. Поэтому здесь особенно важен контроль в реальном времени.

Сравнение двух типов процессов

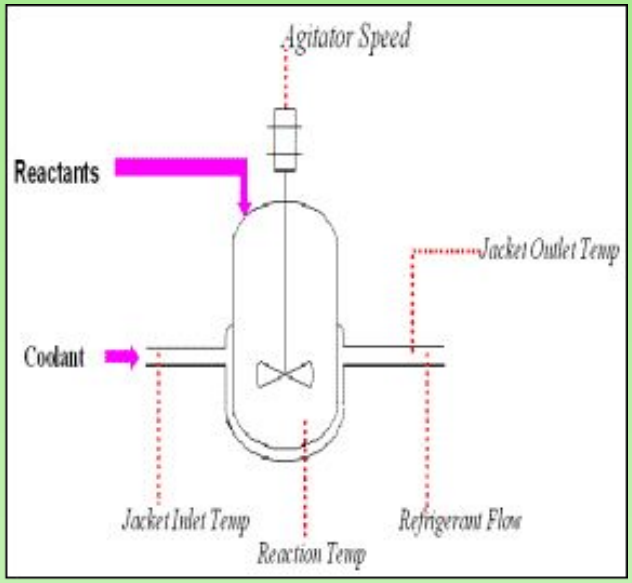
Непрерывный процесс

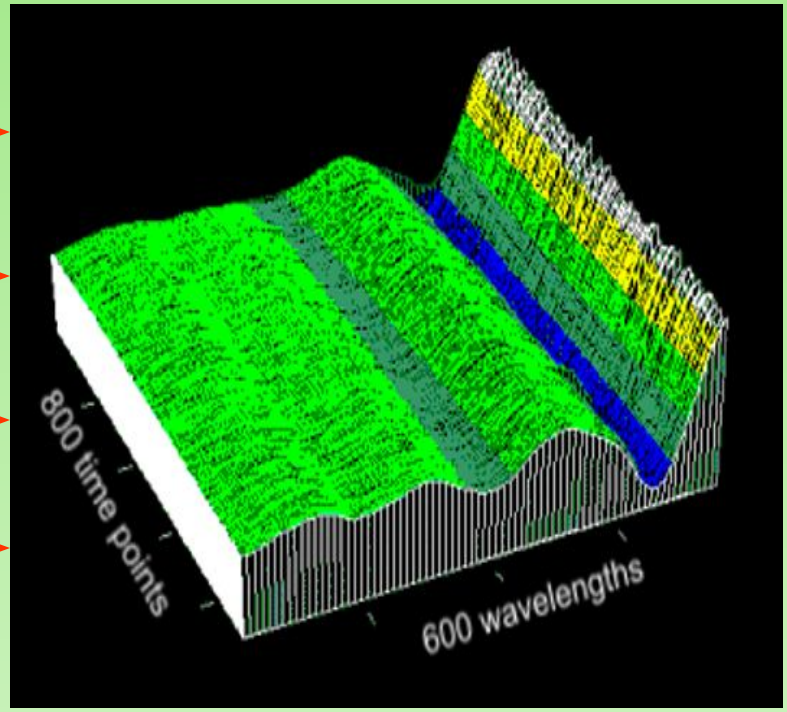


Периодический процесс

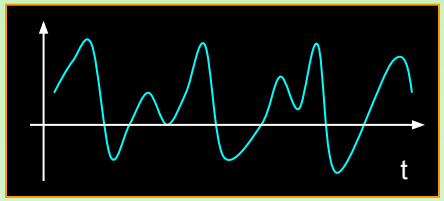


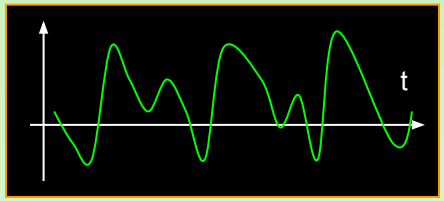
Контроль в фармацевтике





Макропараметры: температура, давление





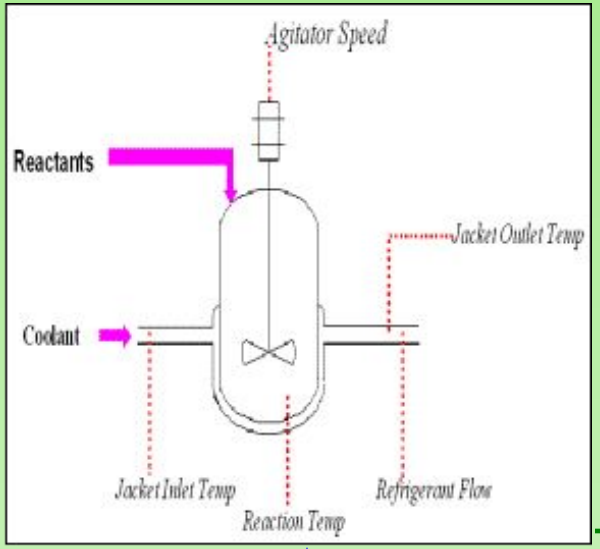
$$\frac{dA}{dt} = -k_1 A; \qquad A(0) = A_0$$

$$\frac{dB}{dt} = k_1 A - k_2 B; \qquad B(0) = B_0$$

$$\frac{dC}{dt} = k_2 B; \qquad C(0) = C_0$$

MSPC в фармацевтике

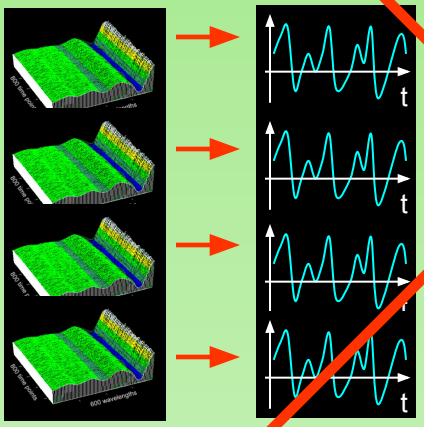
Реактор



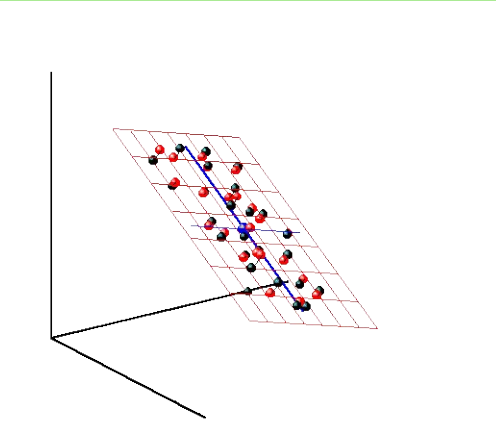
История процесса

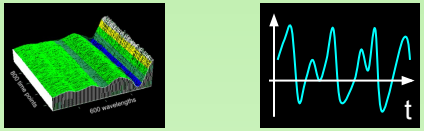
Спектры

Макро



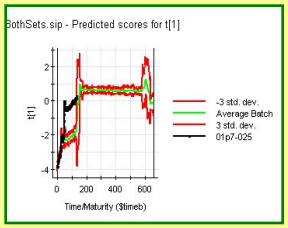
Проекционная модель

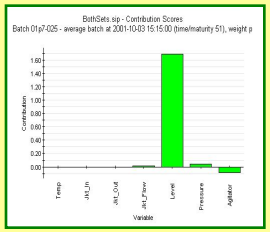




Текущий цикл

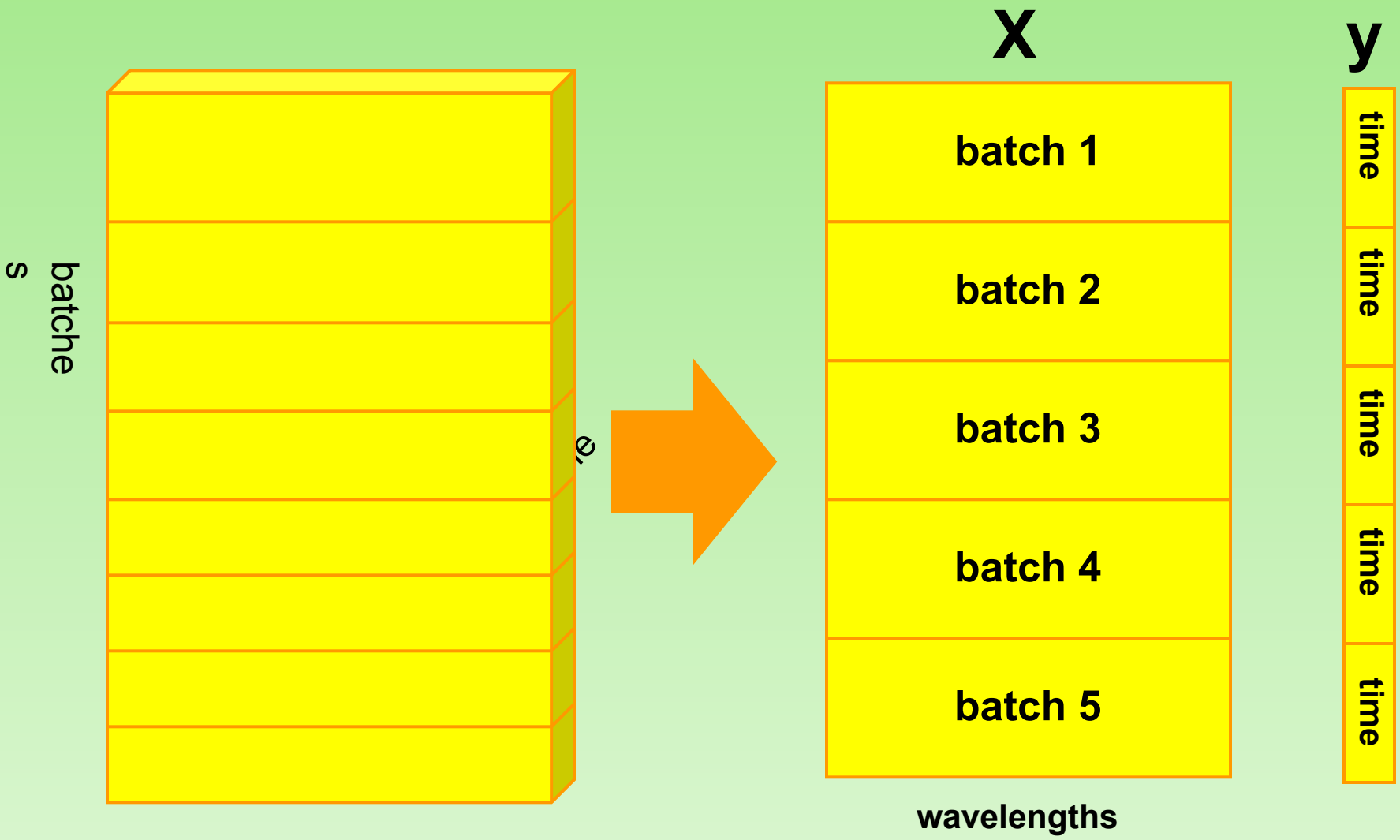
MSPC Контроль



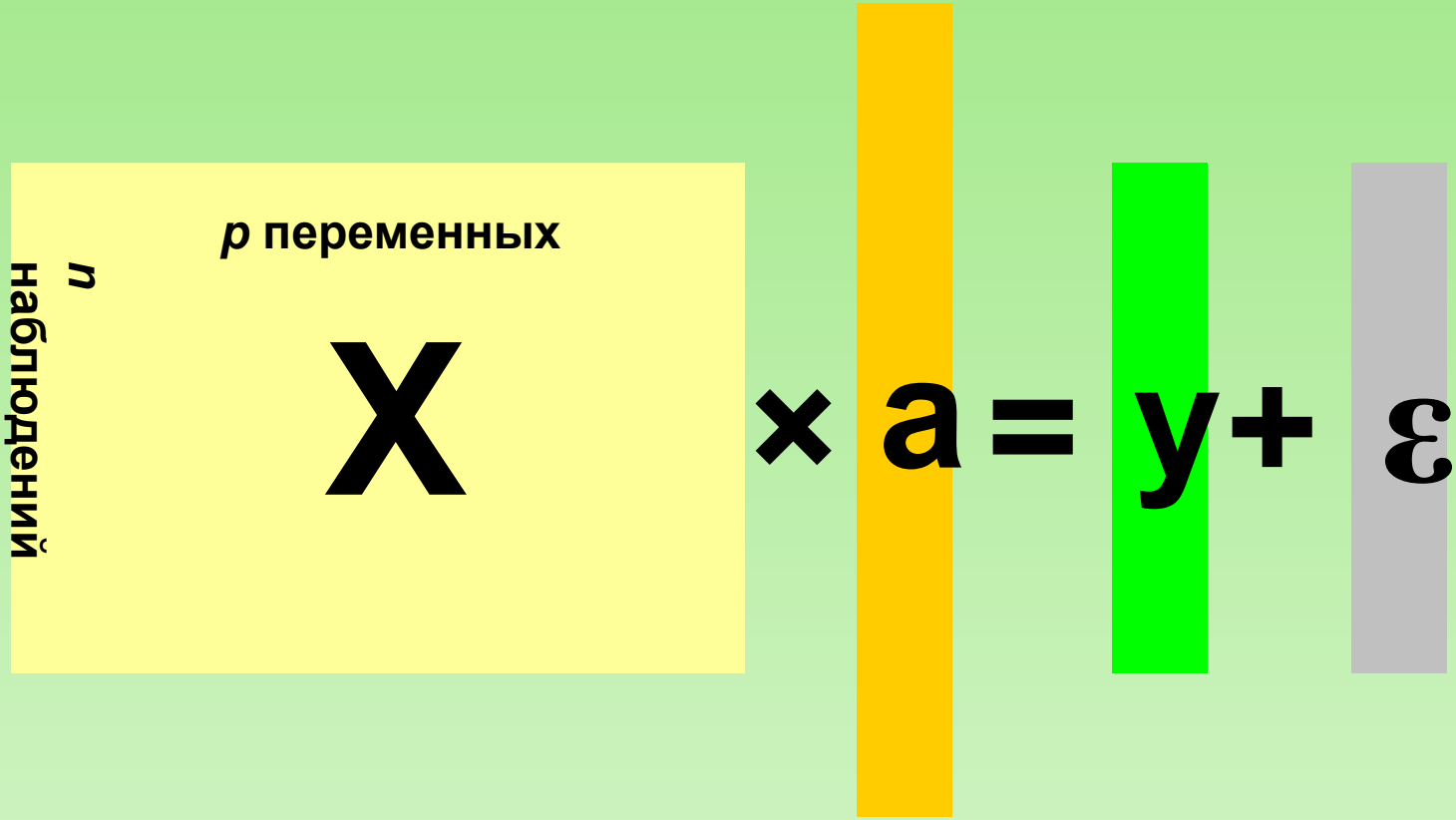


Обратная связь

Трехмодальные (3-way) данные



Регрессия



$p >> n$ rank (X) < p

Вполне реальный пример

Real Time Quality Control (Batch monitoring) of Pharmaceutical Production Processes.

Chris Ambrozic and John Parsons
Umetrics Inc., Kinnelon, NJ 07405

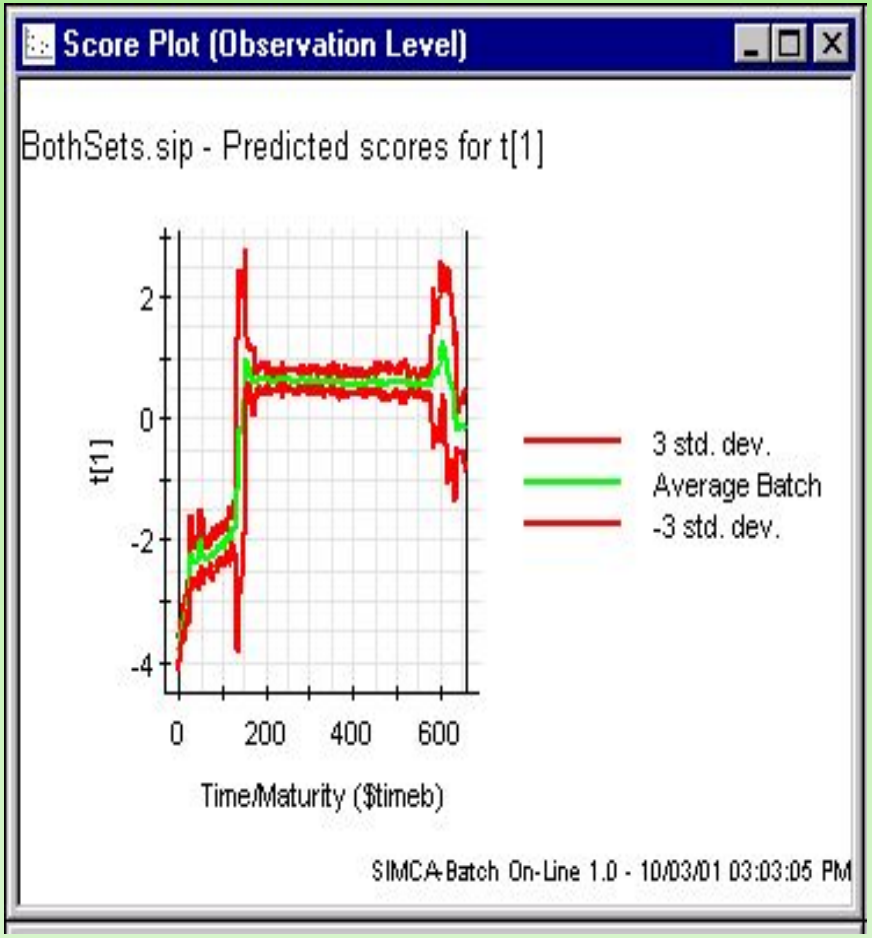
www.umetrics.com

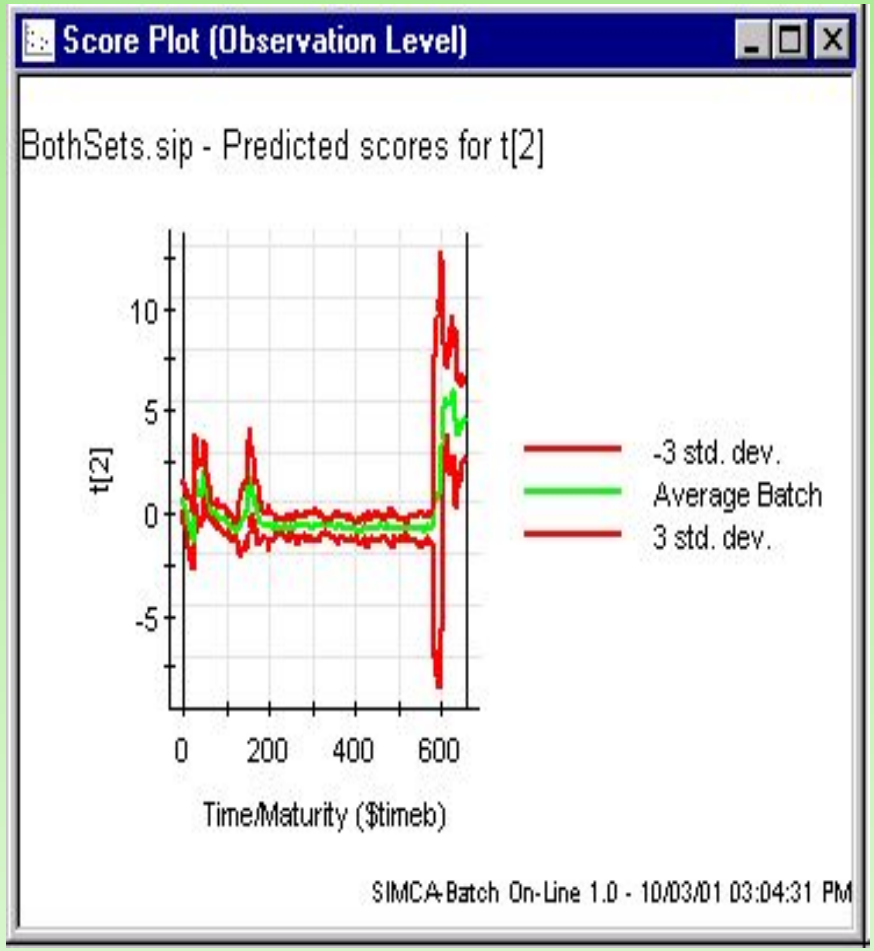
Presented at FDA, 28 Nov., 2001

UMETRICS

MSPC Контроль по первым двум ГК

Средняя траектория хороших циклов (зеленая) ± 3σ (красные).
Хорошие циклы должны лежать в этих пределах.

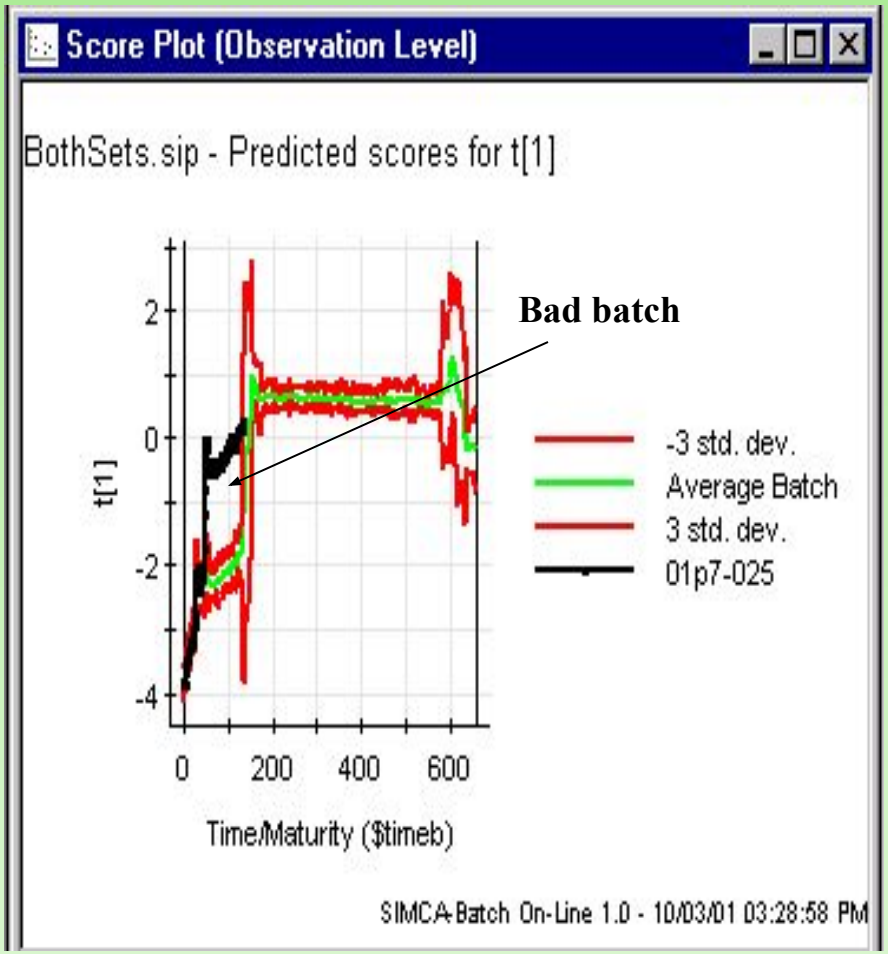


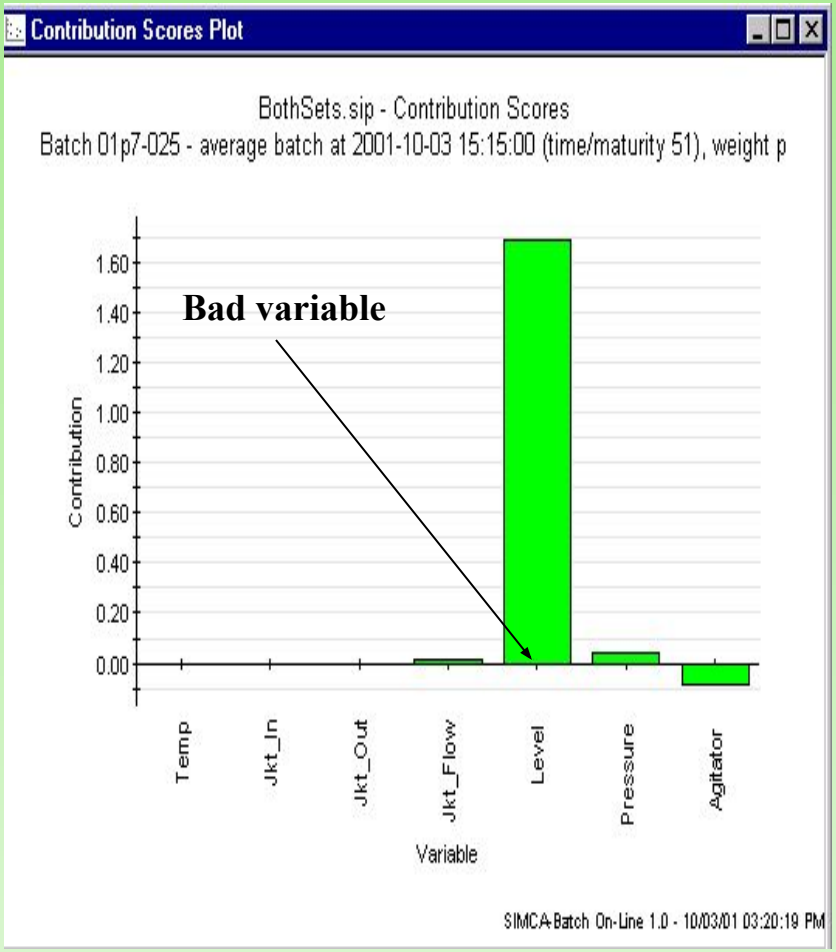


Контроль процесса в реальном времени

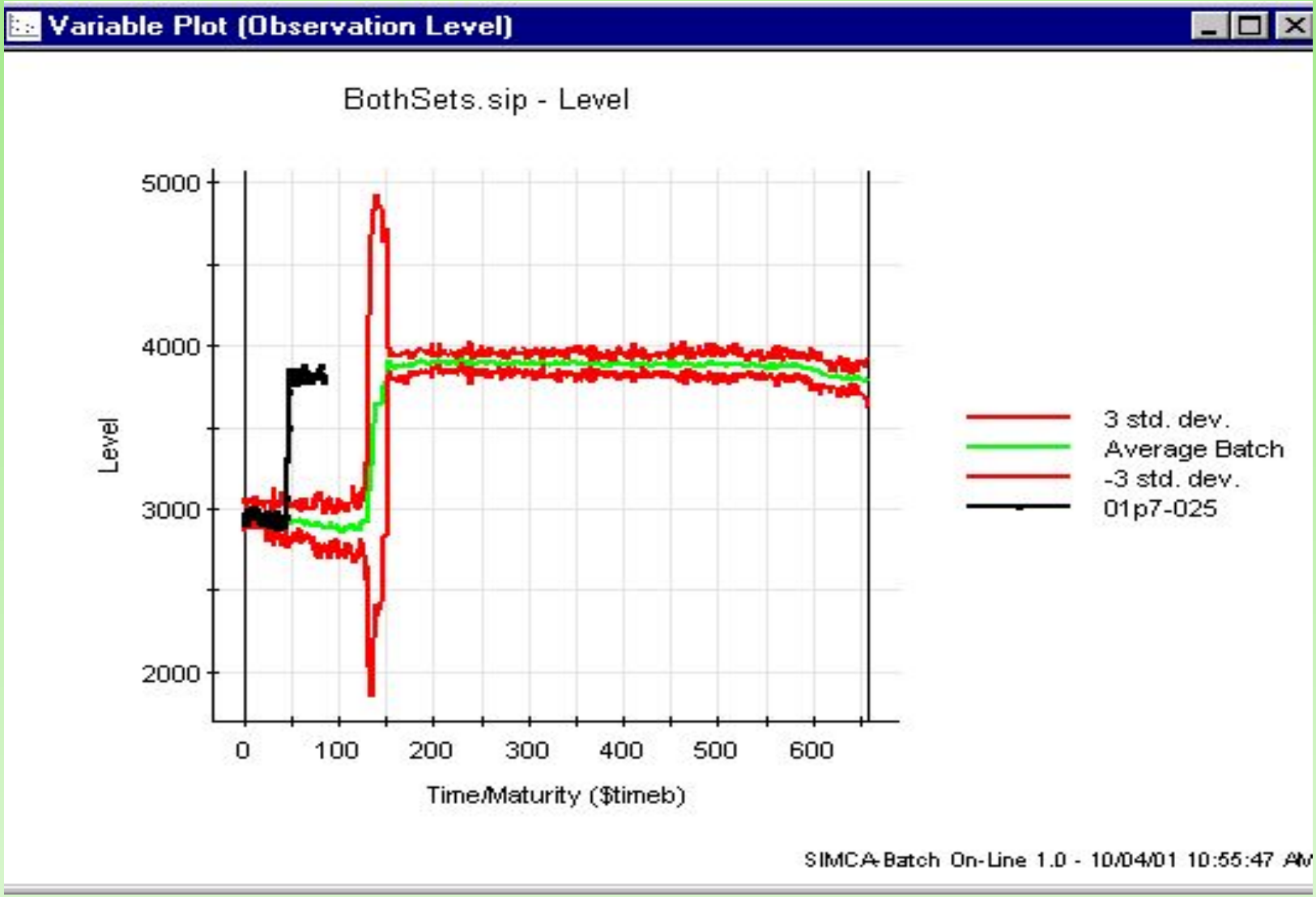
Плохой процесс (**черный**) выходит за пределы допустимых отклонений по ГК1.

Причина выясняется с помощью графика вкладов переменных в первую ГК.





Контрольная карта для этой переменной

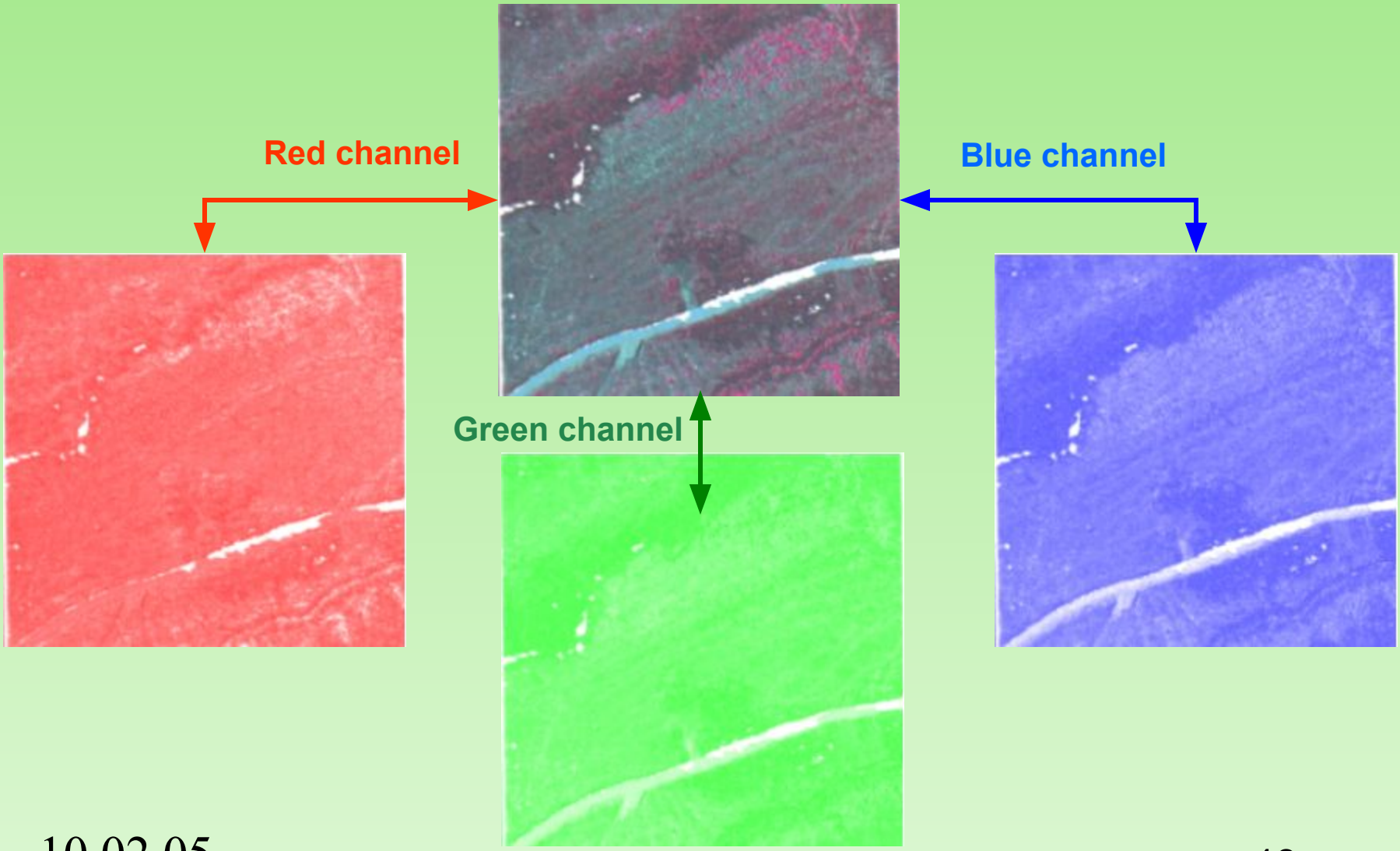


Многомерный анализ изображений (MIA)

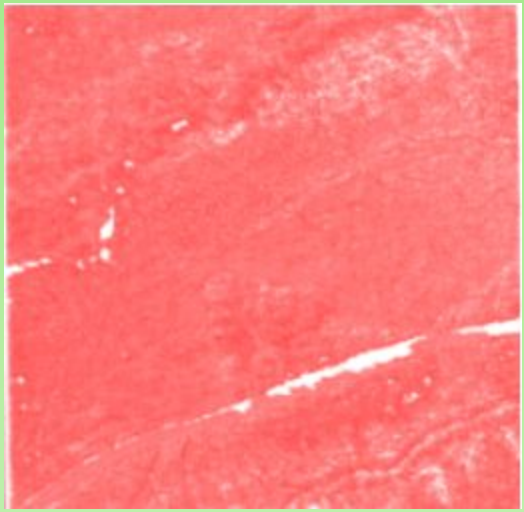
по материалам Paul Geladi, Sweeden

Хемометрика

Разложение изображения по каналам

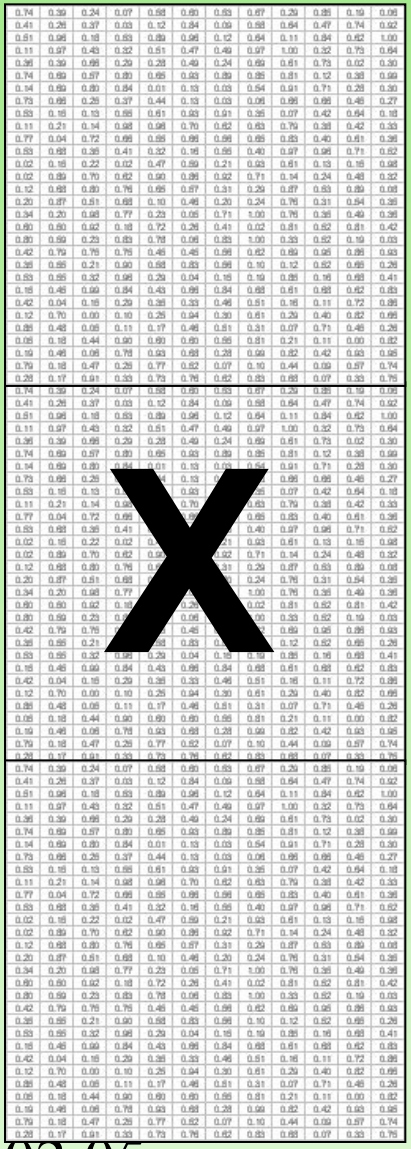


Оцифровка изображения



0.74	0.39	0.24	0.07	0.68	0.60	0.63	0.67	0.29	0.85	0.19	0.06
0.41	0.26	0.37	0.03	0.12	0.84	0.09	0.88	0.64	0.47	0.74	0.92
0.51	0.96	0.18	0.53	0.89	0.96	0.12	0.64	0.11	0.84	0.62	1.00
0.11	0.97	0.43	0.32	0.51	0.47	0.49	0.97	1.00	0.32	0.73	0.64
0.36	0.39	0.66	0.29	0.28	0.49	0.24	0.69	0.61	0.73	0.02	0.30
0.74	0.89	0.57	0.80	0.65	0.93	0.89	0.85	0.81	0.12	0.38	0.99
0.14	0.89	0.80	0.84	0.01	0.13	0.03	0.54	0.91	0.71	0.28	0.30
0.73	0.66	0.26	0.37	0.44	0.13	0.03	0.06	0.66	0.66	0.46	0.27
0.53	0.16	0.13	0.66	0.61	0.93	0.91	0.35	0.07	0.42	0.64	0.18
0.11	0.21	0.14	0.98	0.96	0.70	0.62	0.63	0.79	0.38	0.42	0.33
0.77	0.04	0.72	0.66	0.65	0.66	0.66	0.65	0.83	0.40	0.61	0.36
0.53	0.68	0.36	0.41	0.32	0.16	0.66	0.40	0.97	0.96	0.71	0.82
0.02	0.16	0.22	0.02	0.47	0.59	0.21	0.93	0.61	0.13	0.16	0.98
0.02	0.89	0.70	0.62	0.90	0.86	0.92	0.71	0.14	0.24	0.48	0.32
0.12	0.68	0.80	0.76	0.65	0.57	0.31	0.29	0.87	0.53	0.89	0.08
0.20	0.87	0.51	0.68	0.10	0.46	0.20	0.24	0.76	0.31	0.54	0.36
0.34	0.20	0.98	0.77	0.23	0.05	0.71	1.00	0.76	0.36	0.49	0.36
0.60	0.60	0.92	0.18	0.72	0.26	0.41	0.02	0.81	0.52	0.81	0.42
0.80	0.59	0.23	0.83	0.78	0.06	0.83	1.00	0.33	0.52	0.19	0.03
0.42	0.79	0.76	0.76	0.46	0.46	0.96	0.62	0.89	0.96	0.86	0.93
0.36	0.66	0.21	0.90	0.68	0.83	0.96	0.10	0.12	0.52	0.66	0.26
0.53	0.66	0.32	0.98	0.29	0.04	0.16	0.19	0.86	0.16	0.68	0.41
0.16	0.46	0.99	0.84	0.43	0.66	0.84	0.68	0.61	0.68	0.62	0.83
0.42	0.04	0.16	0.29	0.36	0.33	0.46	0.51	0.16	0.11	0.72	0.86
0.12	0.70	0.00	0.10	0.26	0.94	0.30	0.61	0.29	0.40	0.82	0.66
0.86	0.48	0.06	0.11	0.17	0.46	0.51	0.31	0.07	0.71	0.46	0.26
0.06	0.18	0.44	0.90	0.60	0.60	0.66	0.81	0.21	0.11	0.00	0.82
0.19	0.46	0.06	0.78	0.93	0.68	0.28	0.99	0.82	0.42	0.93	0.96
0.79	0.18	0.47	0.26	0.77	0.62	0.07	0.10	0.44	0.09	0.57	0.74
0.28	0.17	0.91	0.33	0.73	0.76	0.62	0.83	0.68	0.07	0.33	0.76

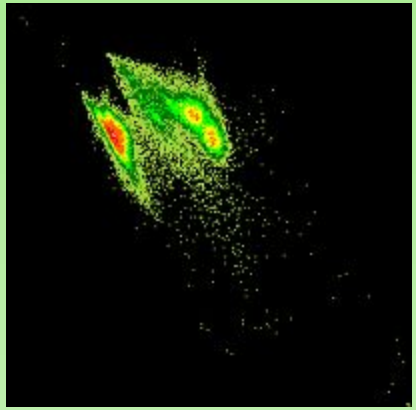
Применение МГК для анализа



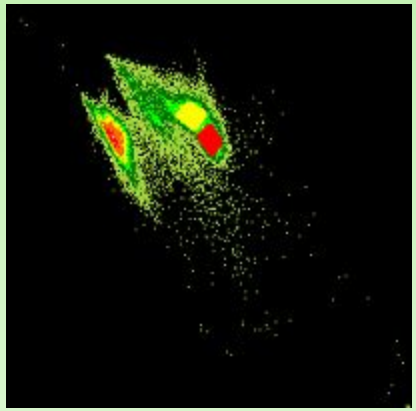
X

PCA

Образ ГК1



Образ ГК2



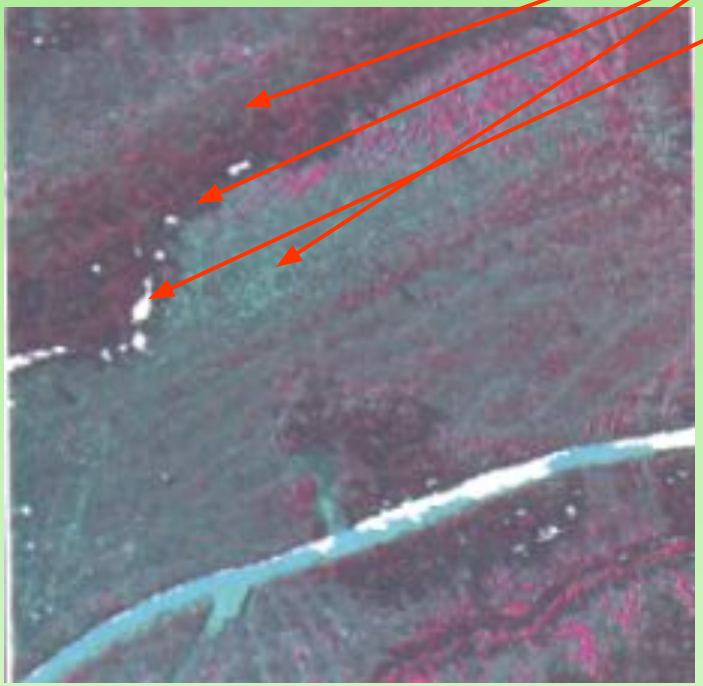
Исследование состояния лесов (Канада)

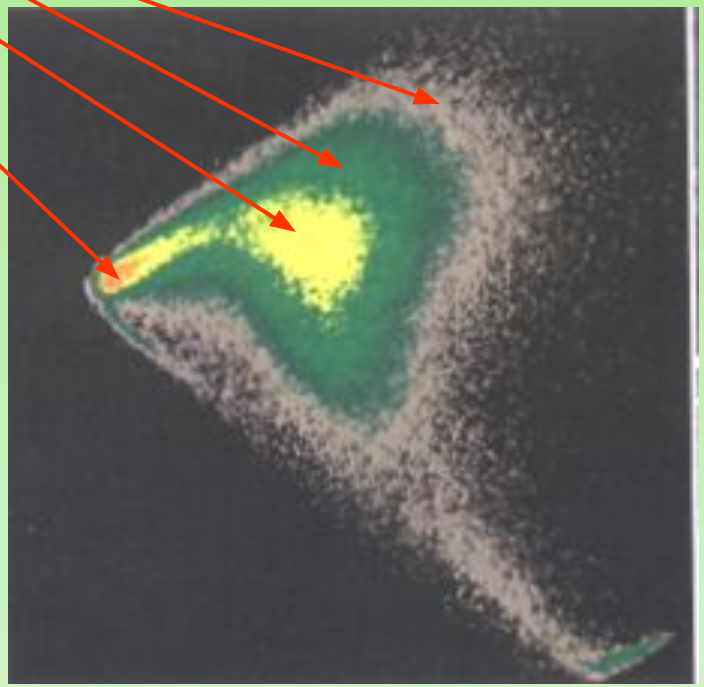
Старые деревья

отражения

Исходный аэроснимок

Он же в пространстве ГК1

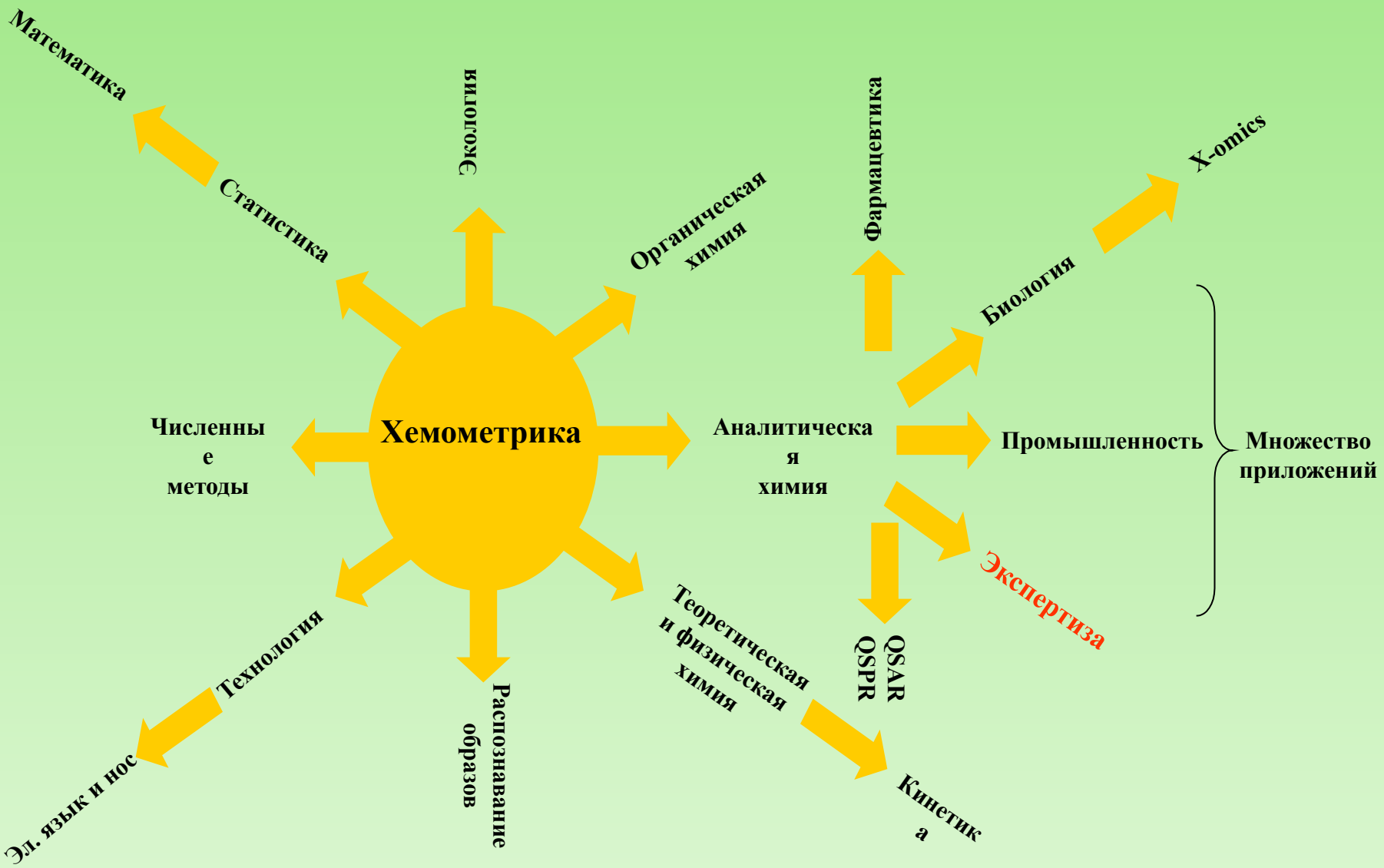




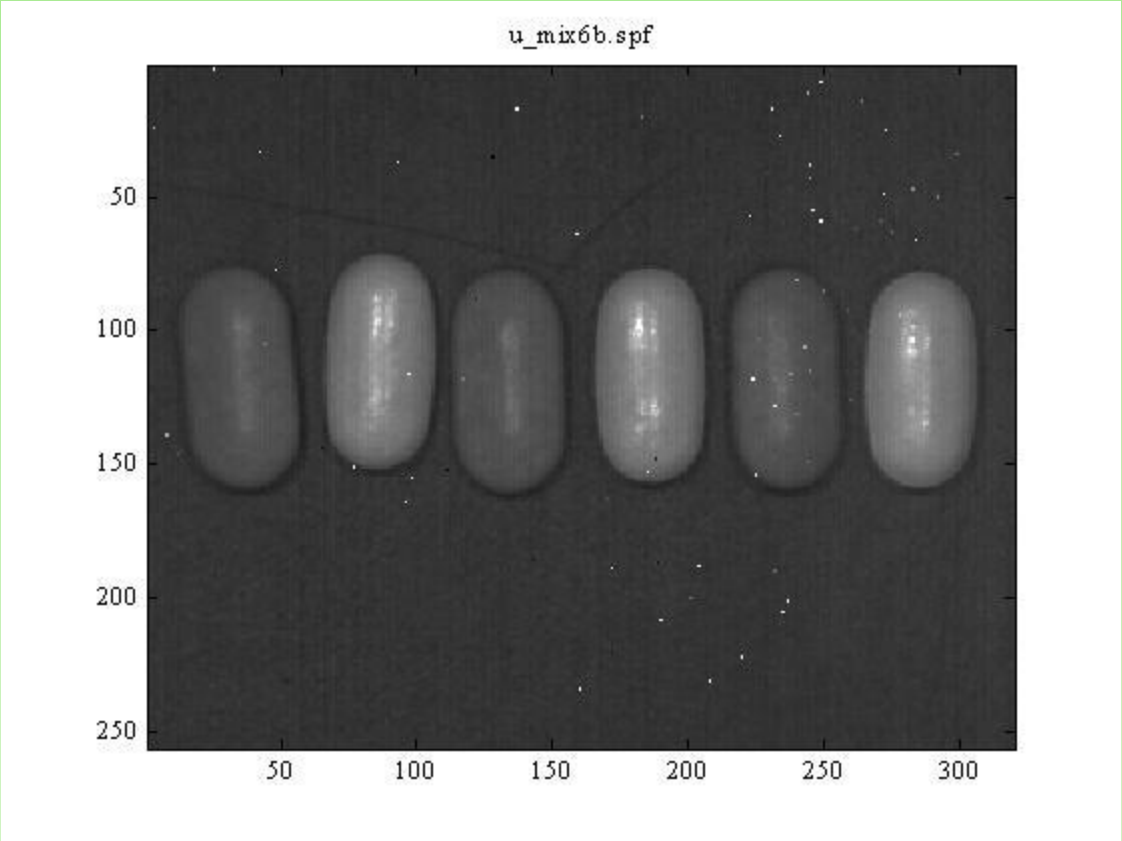
Хемометрика и экспертиза

по материалам Оксаны Родионовой, Россия
и Lars P. Houmøller, Denmark

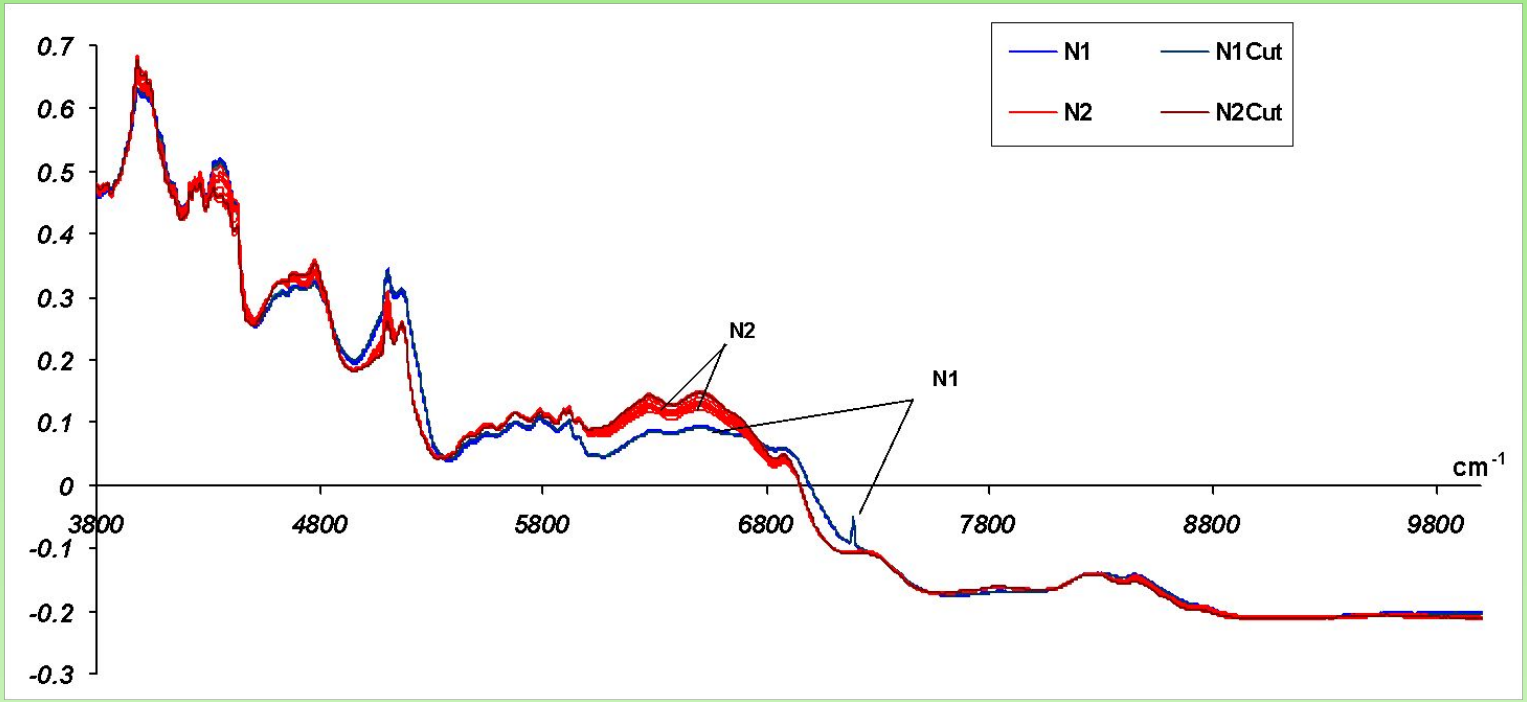
Хемометрика



Фальшивые таблетки



ИК Спектры таблеток

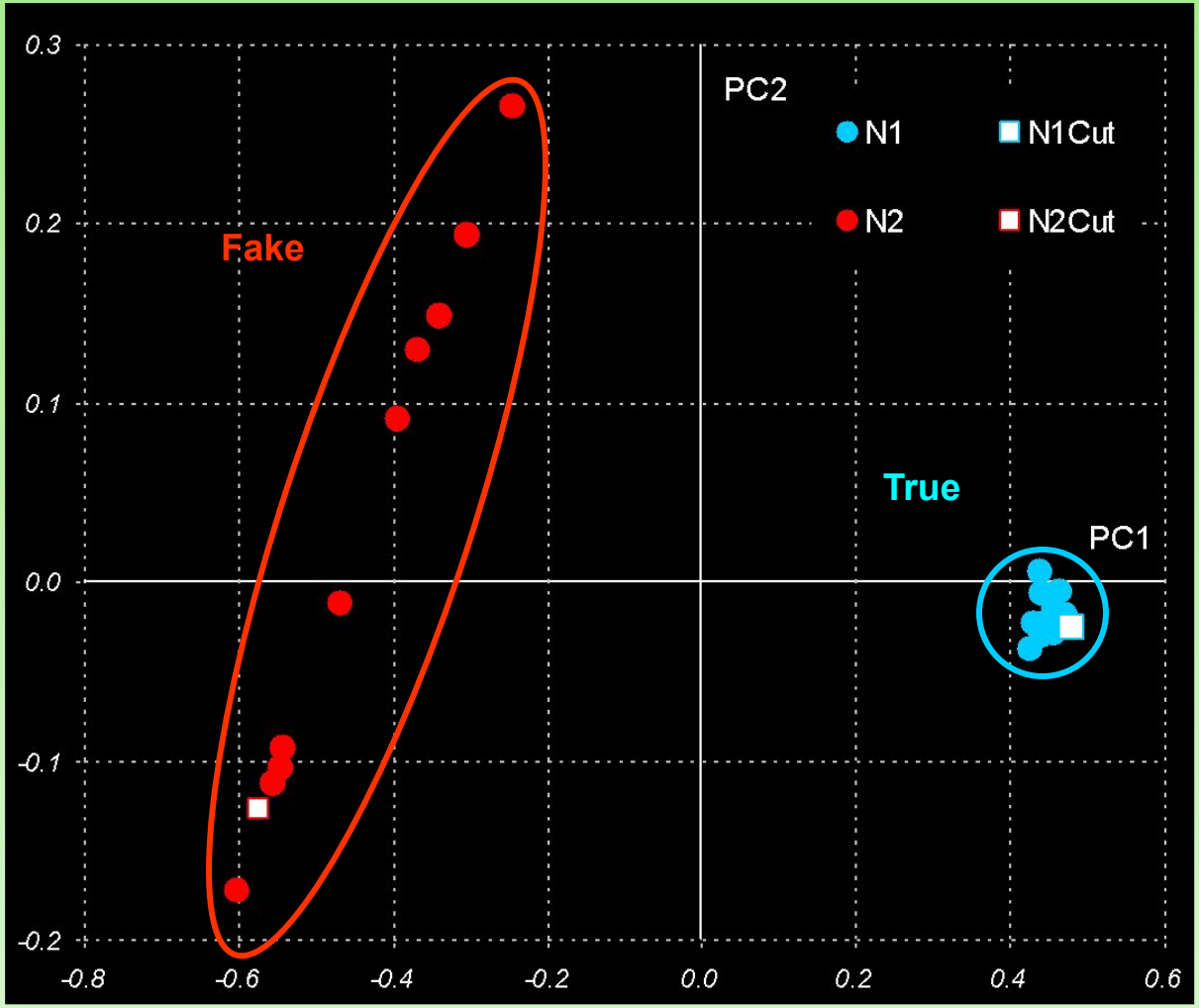


Прибор Bomem MB160 в диапазоне 3800 – 10000 cm^{-1} (ближний ИК спектр); 1609 значений.

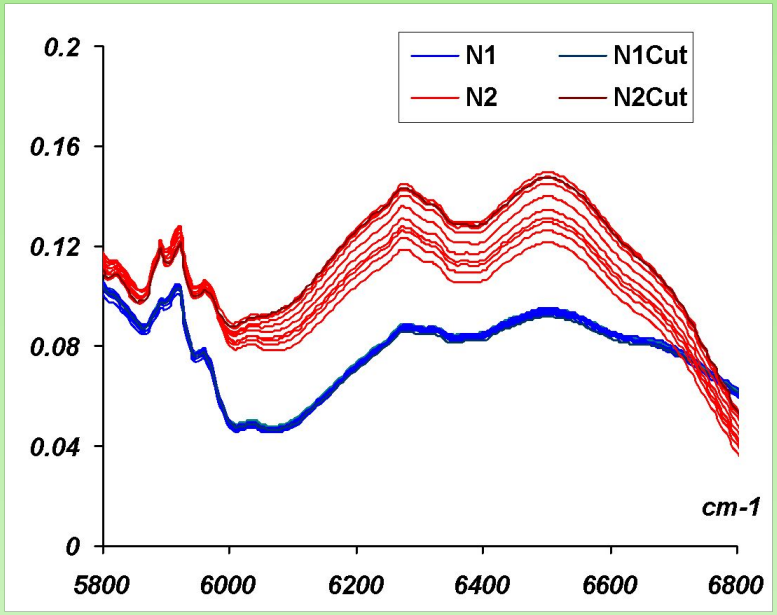
N1 - таблетки от производителя –10 штук;
N1Cut – образец N1 разрезанный

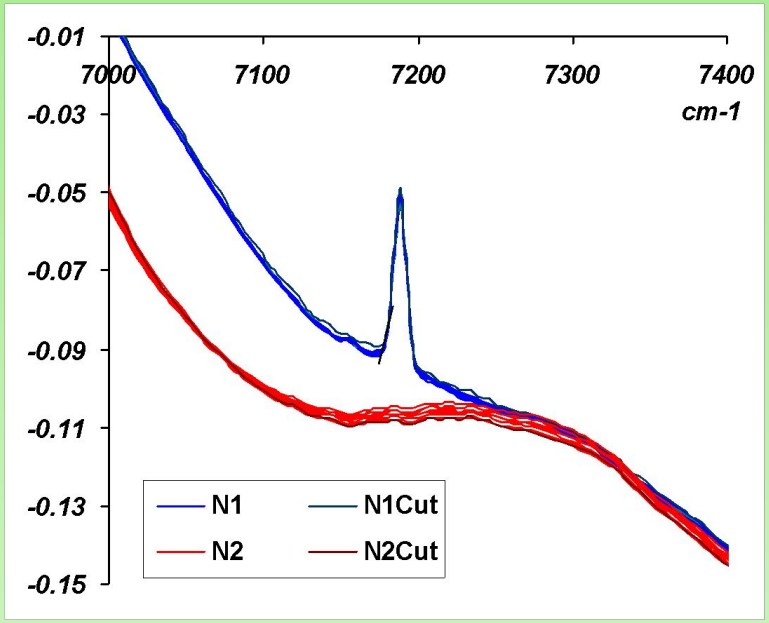
N2 - таблетки фальсифицированные – 10 штук;
N2Cut – образец N2 разрезанный

PCA полного спектра

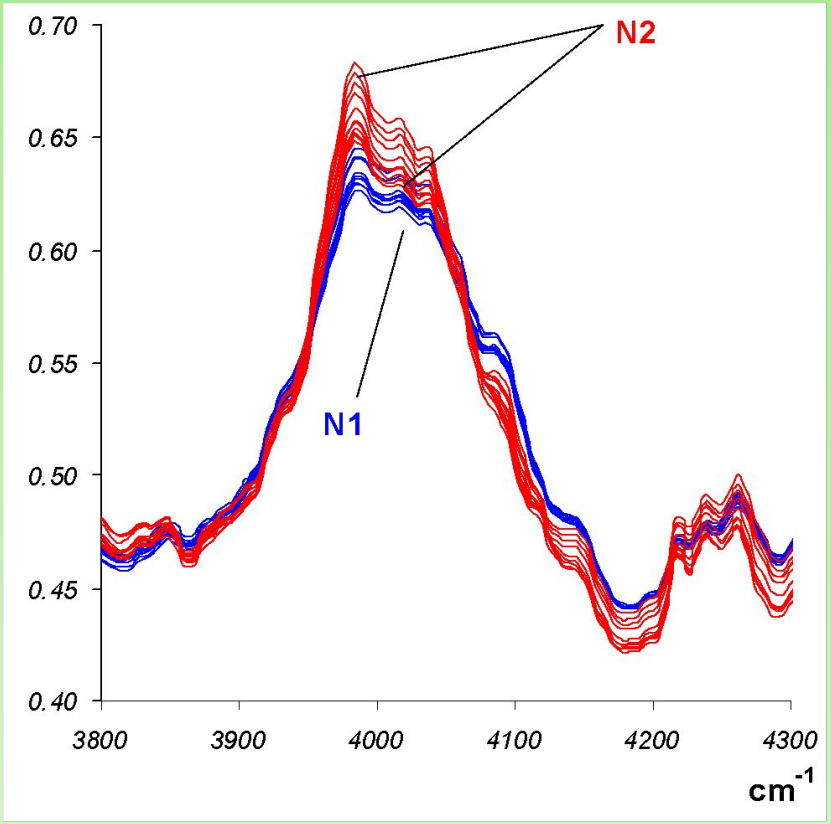


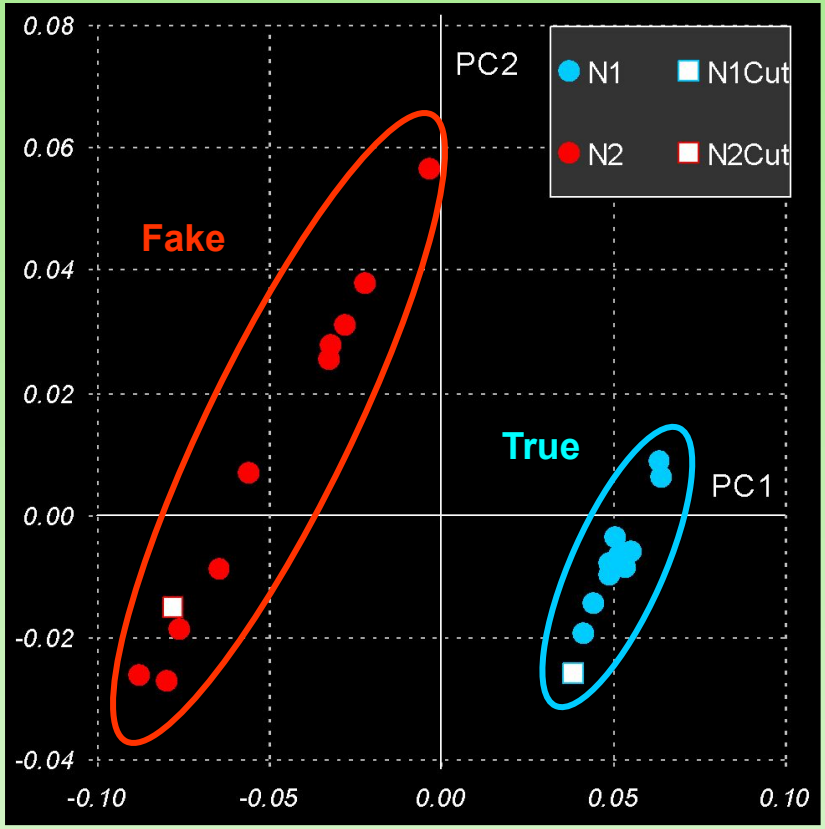
Отдельные участки спектра





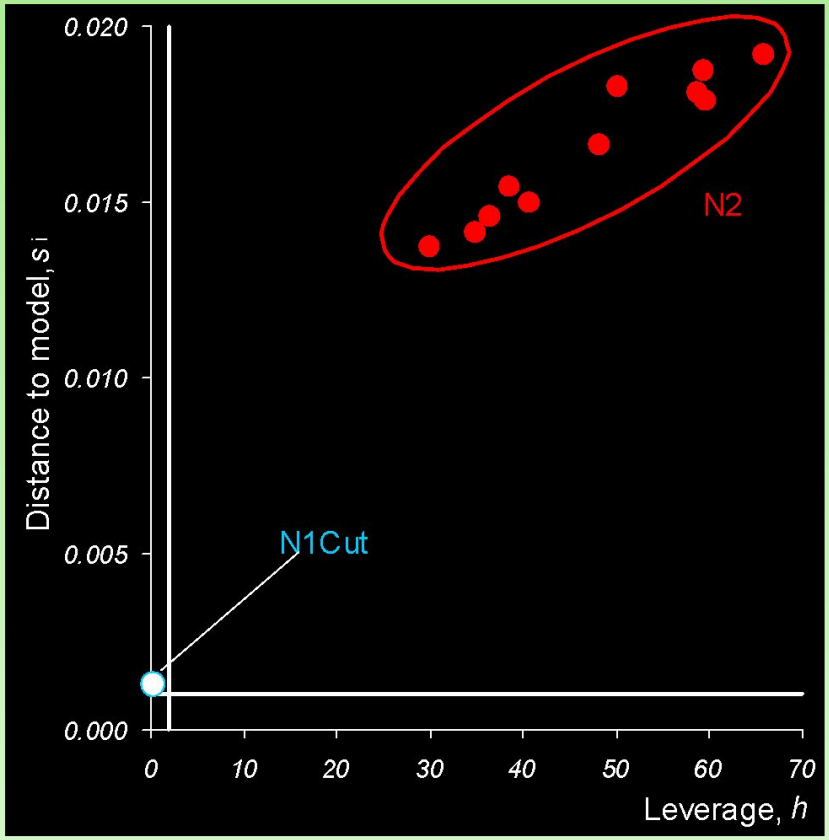
Анализ «наихудшего» участка спектра



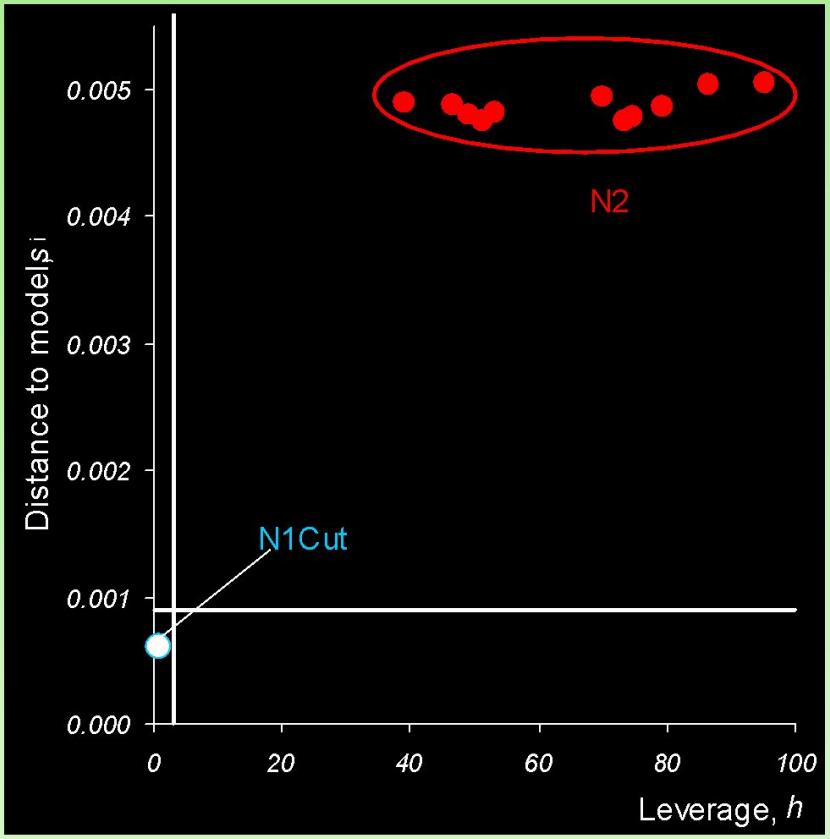


SIMCA

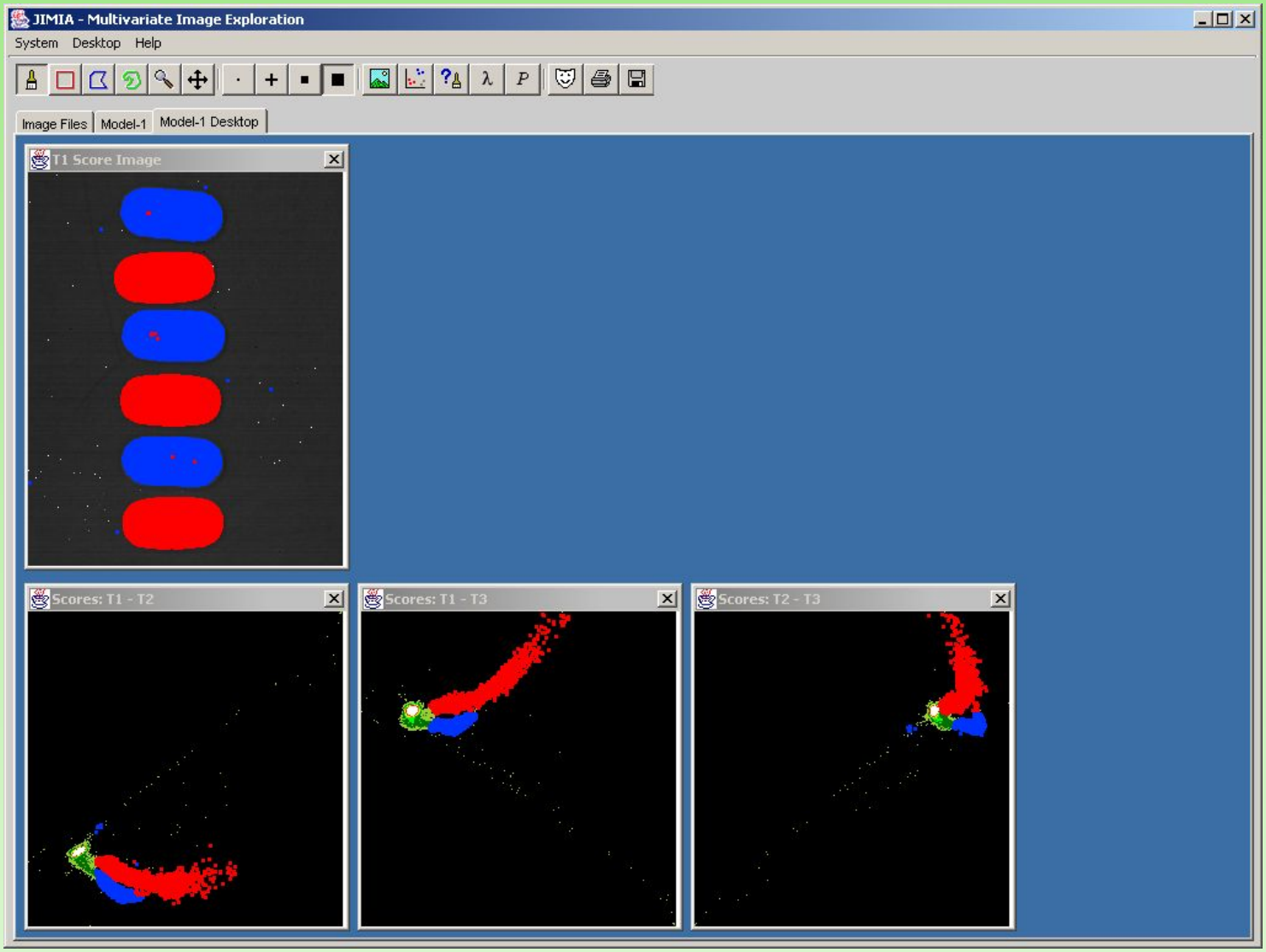
Полный спектр (1609 точек)



Участок спектра (130 точек)



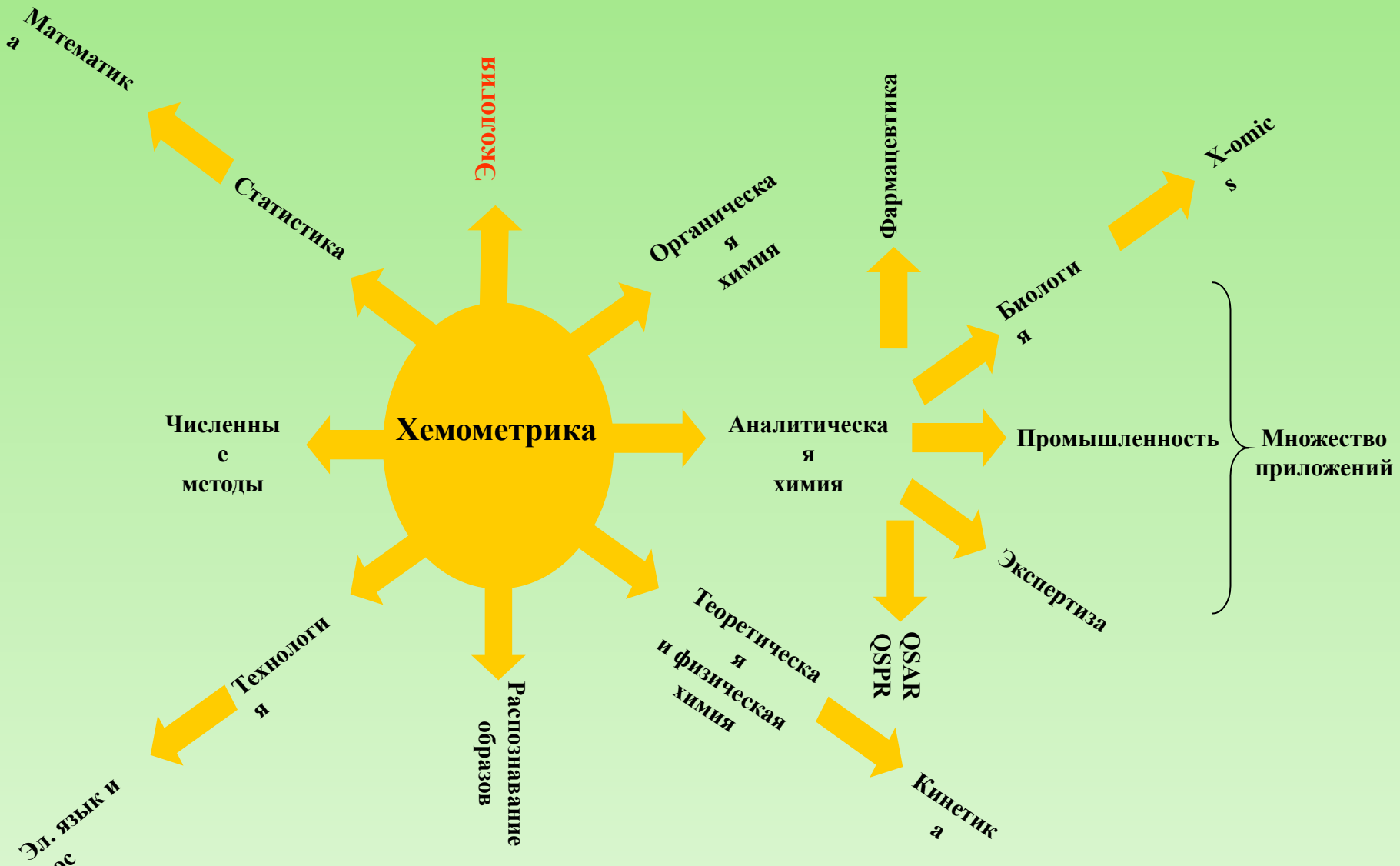
Гипреспектры таблеток



Хемометрика и экология

по материалам Pentti Minkkinen, Finland

Хемометрика



Влияние ванадиевой пыли на людей

Испытательная группа 17 человек подверженных пыли V_2O_5 на фабрике

Контрольная группа 18 человек

Измерялись 26 клинических показателей плазмы крови

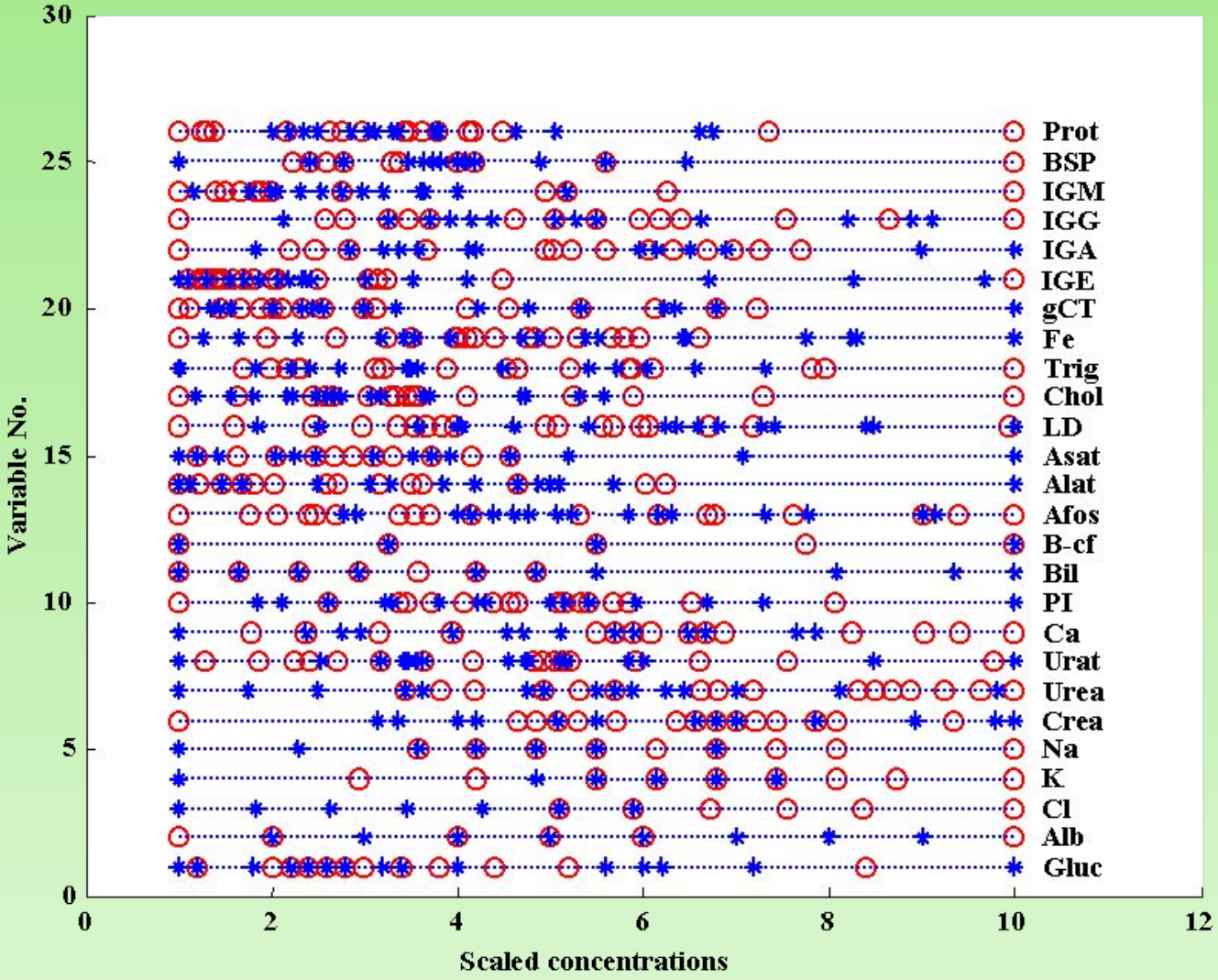
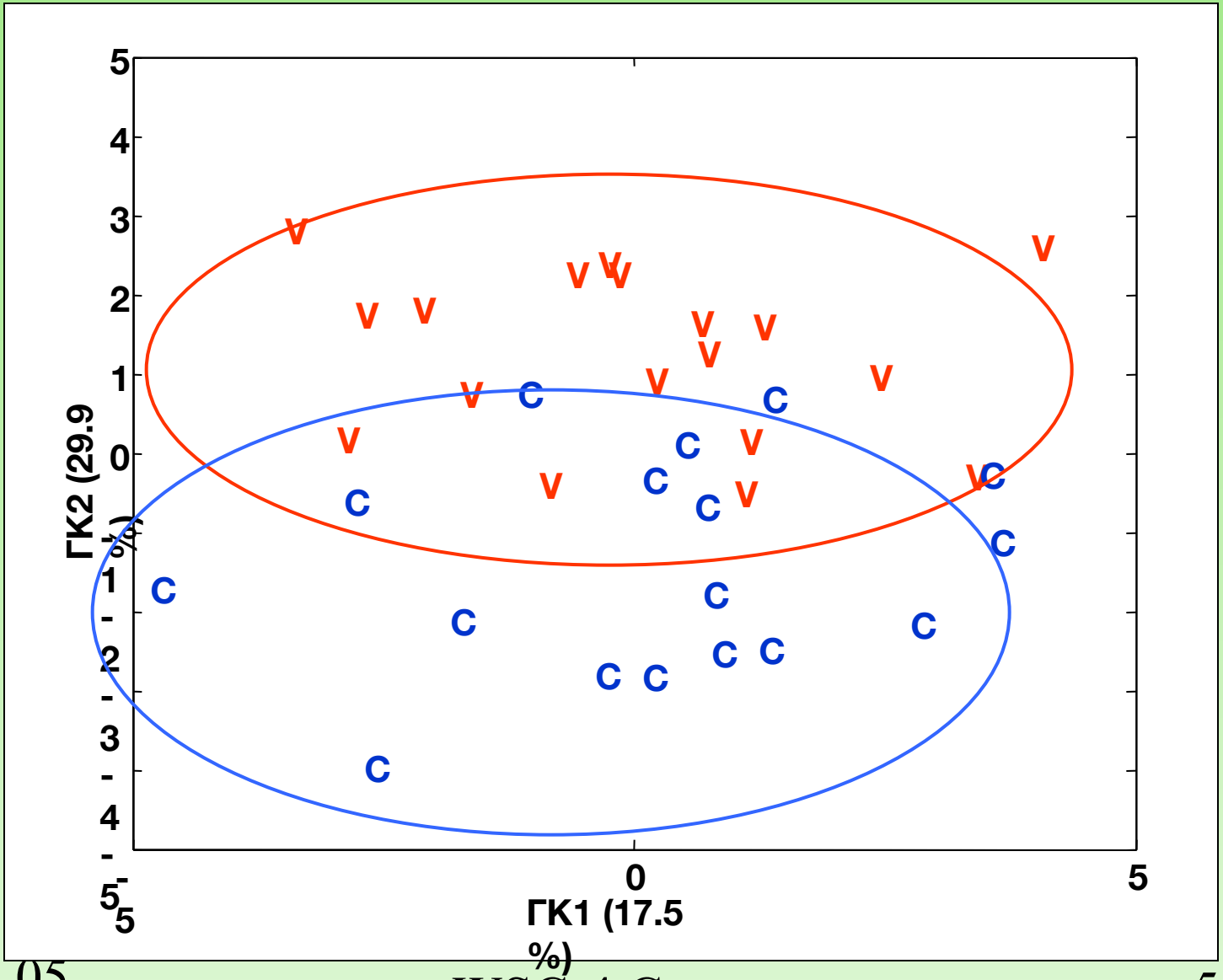


График ГК



PLS-D Дискриминация

Матрица дескрипторов **X**

x_{11}	x_{12}	…					…	x_{1k}
x_{21}	x_{22}	…					…	x_{2k}
…	…						…	…
…	…						…	…
x_{i1}	x_{i2}	…					…	x_{ik}
$x_{i+1,1}$	$x_{i+1,2}$	…					…	$x_{i+1,k}$
…	…							…
…	…							…
x_{n1}	x_{n2}							x_{nk}

CLASS C

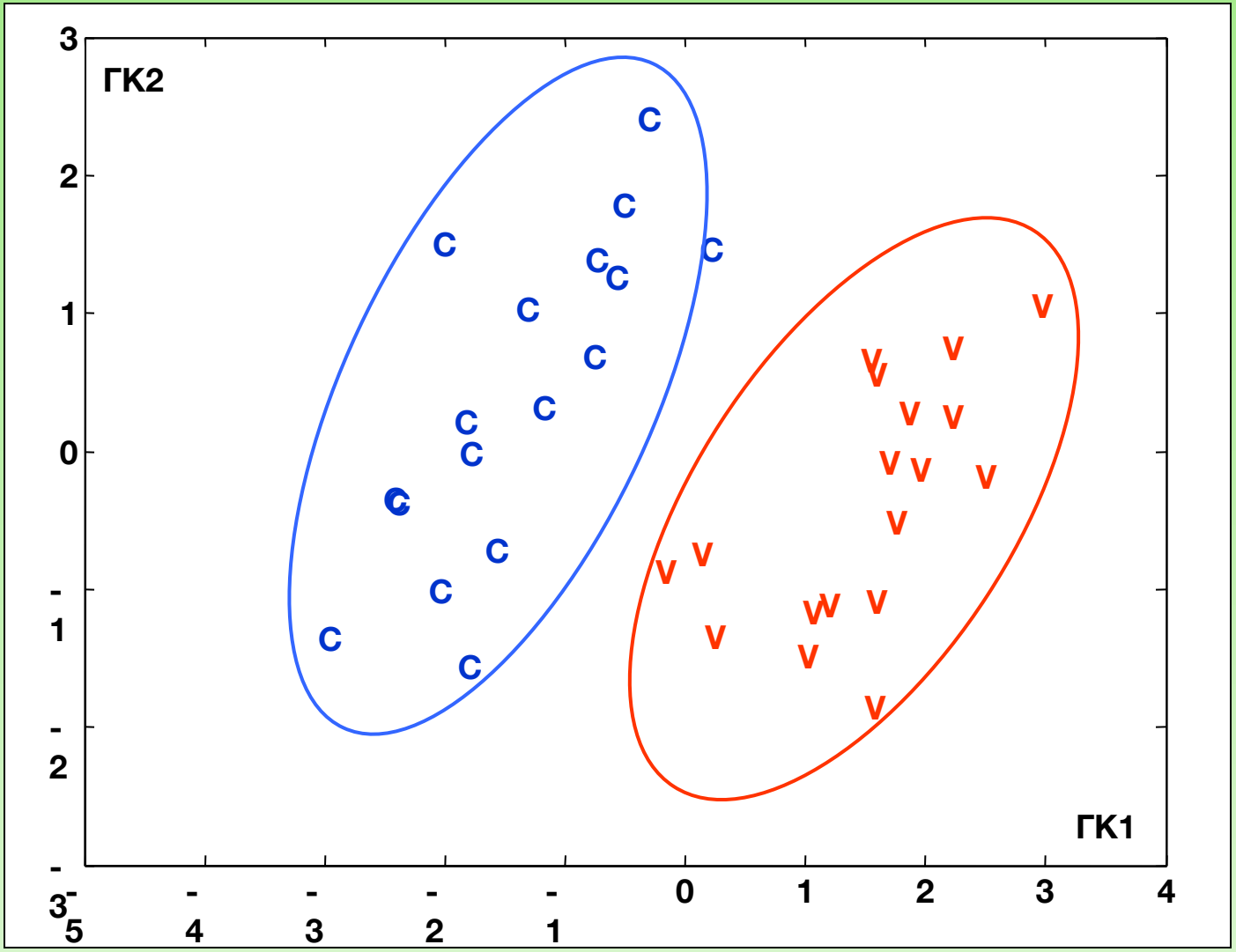
CLASS V

PLS2

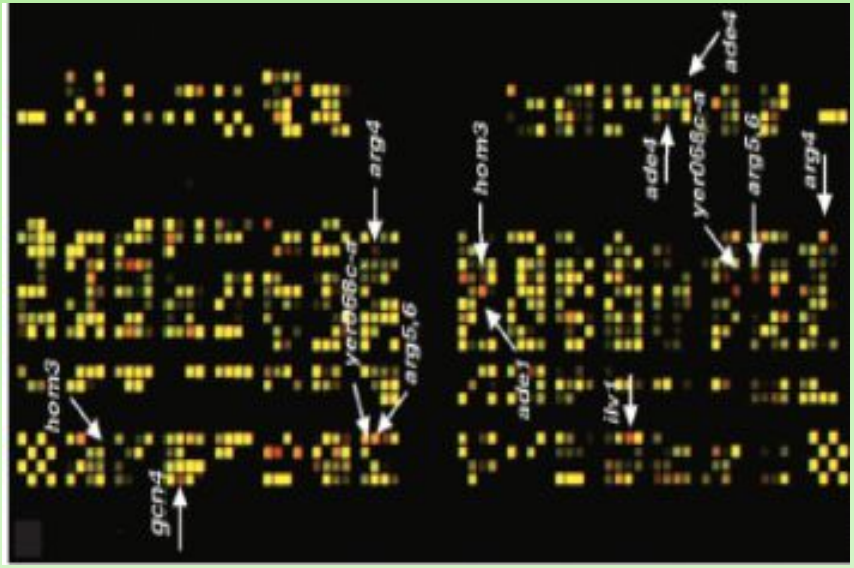
Матрица индикаторов **Y**

1	0
1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1
0	1

График ГК в PLS-D



Хемометрика в биологии



по материалам David O. Nelson and Matt Coleman, USA

Хемометрика

Новая биология: X- omics

Традиционный подход:

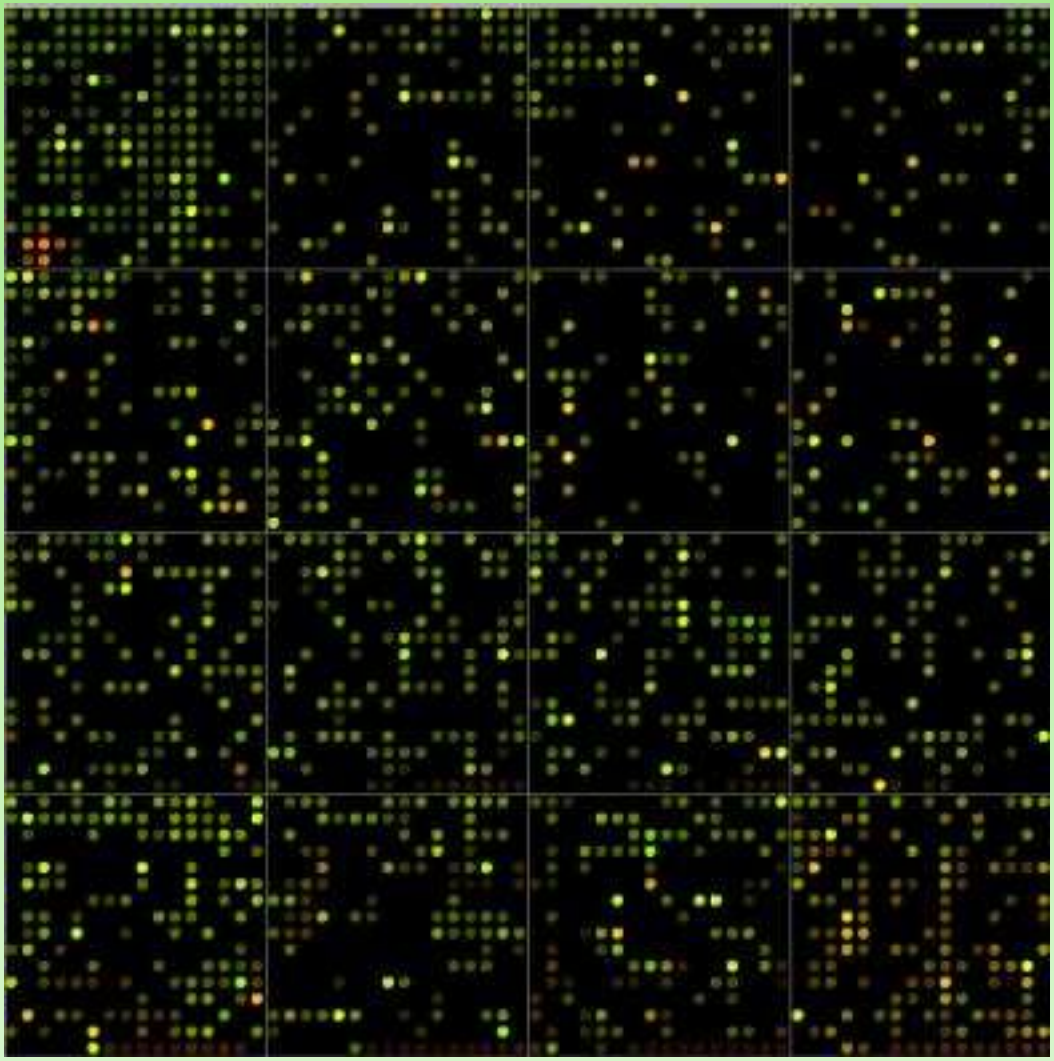
- Один ген/белок/реакция за раз
- Исследование отдельной модели

Новый «системный» подход:

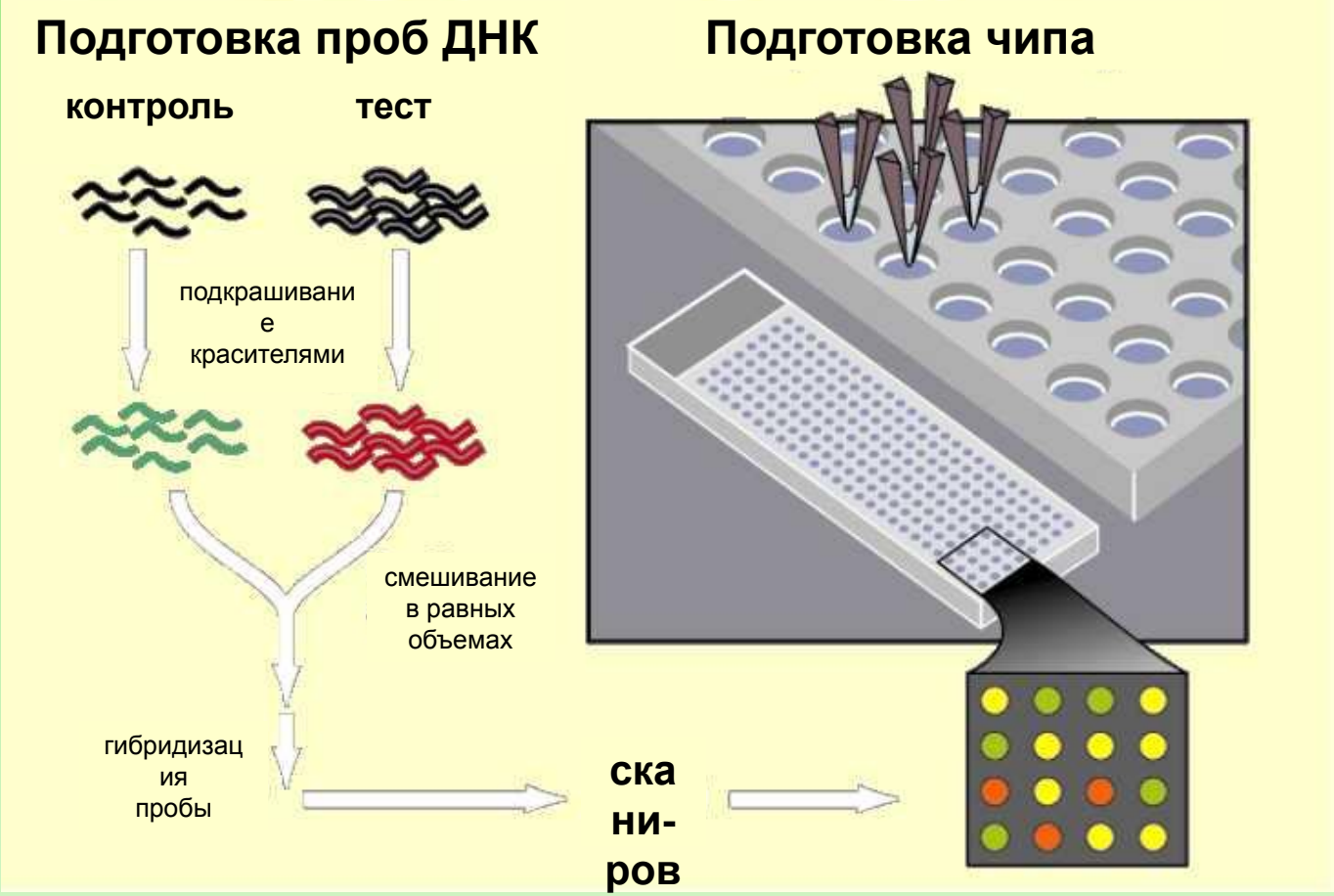
- Все ДНК/РНК/белки исследуются вместе
- Глобальный анализ больших данных

Что такое microarray - чип?

- Каждый микрореактор содержит ДНК (или белок)
- Обычно в каждом реакторе разные ДНК, с повторами для контроля точности
- За один раз несколько микрореакторов гибридизируются с экстрактами тканей
- Результат виден как образ, состоящий из множества цветных пятен



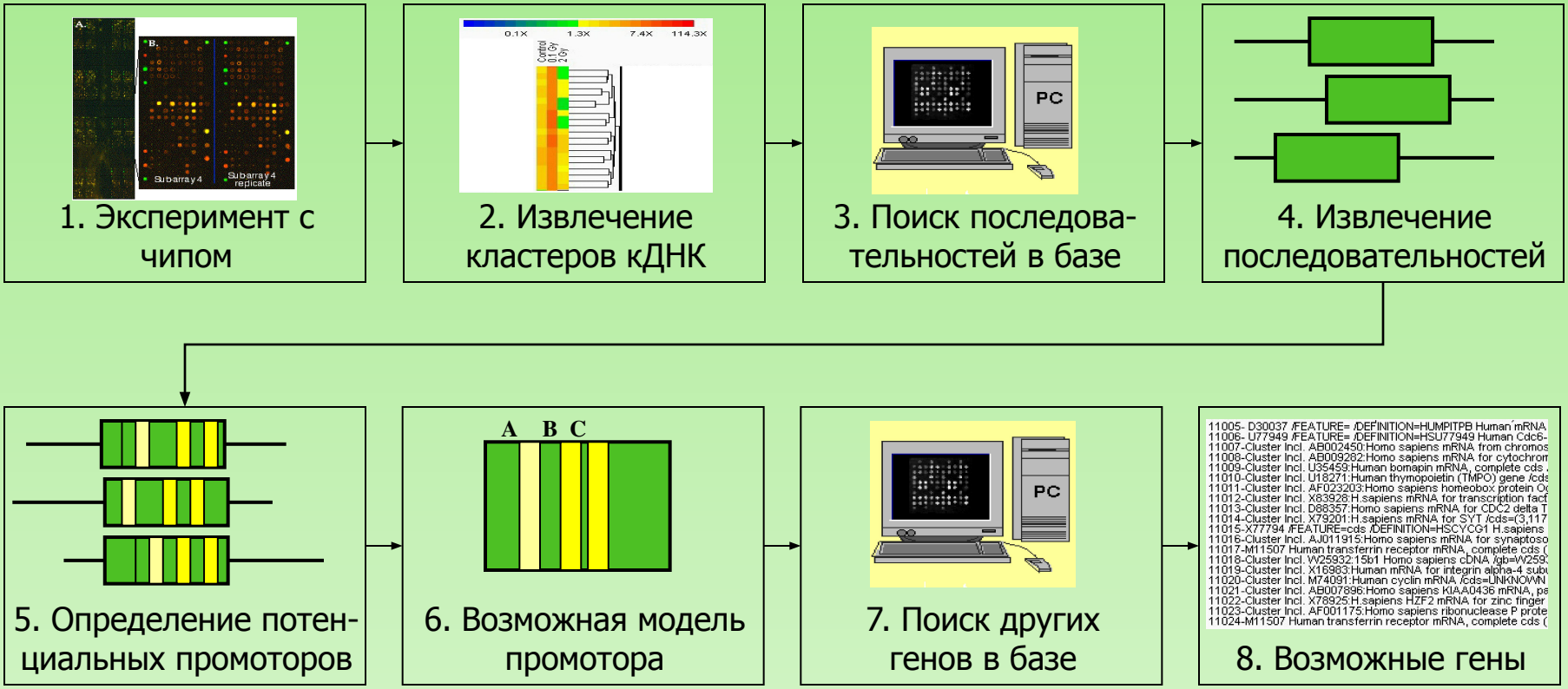
Эксперименты с ДНК



Сравниваются две пробы ДНК. Они подкрашиваются разными красителями

Результат эксперимента это интенсивности через два фильтра - красный и зеленый

Схема обнаружения промоторов



Из Thomas Werner Biomolecular Engineering, **17**: 87-94 (2001)

QSAR и QSPR

по материалам Альфреда Сулейманова, Россия

Хемометрика

Что такое QSAR/QSPR?

QSAR/QSPR

Quantitative structure-activity/property relationships

Количественная связь «структура-активность/свойство»

Цель – создание новых продуктов с заданными свойствами

Разработка изделий из ПВХ (QSPR)

Матрица дескрипторов **X**

Модификаторы		Стабилизаторы		Прочие добавки
FM22	Дакрилан	Интерстаб	Нафтомикс	
	$X_{training}$			

Матрица свойств **Y**

Белизна	Разрывное растяжение	Ударная прочность	Текучесть расплава
	$Y_{training}$		

PLS

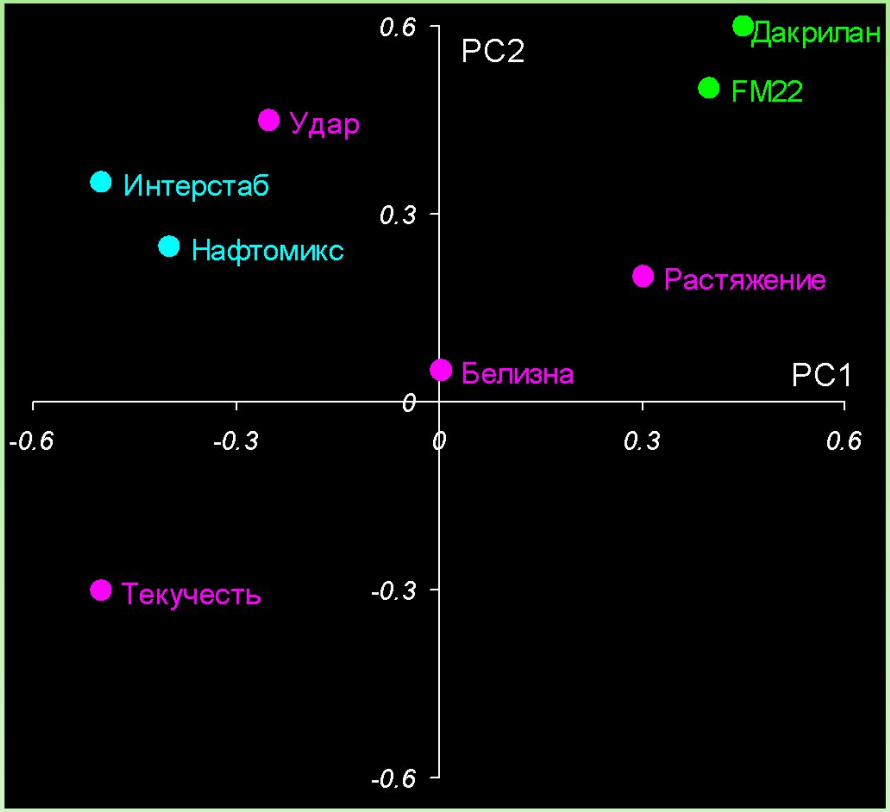
	X_{new}			

Co

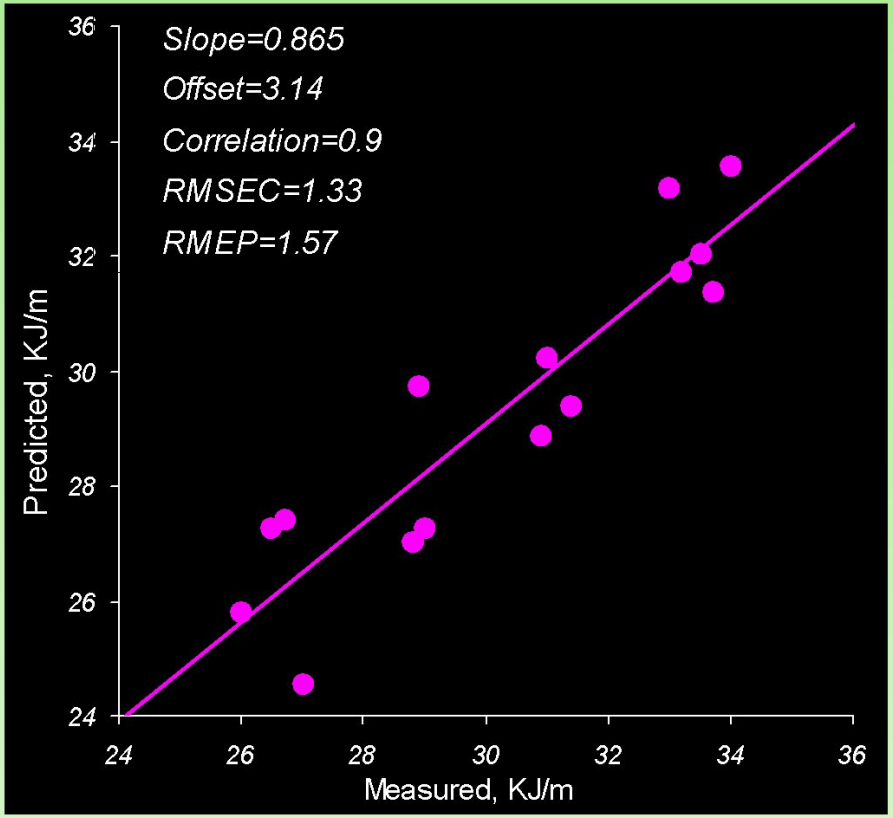
	Y_{new}		

Связь «состав-свойства»

График X и Y нагрузок



Прогноз ударной прочности



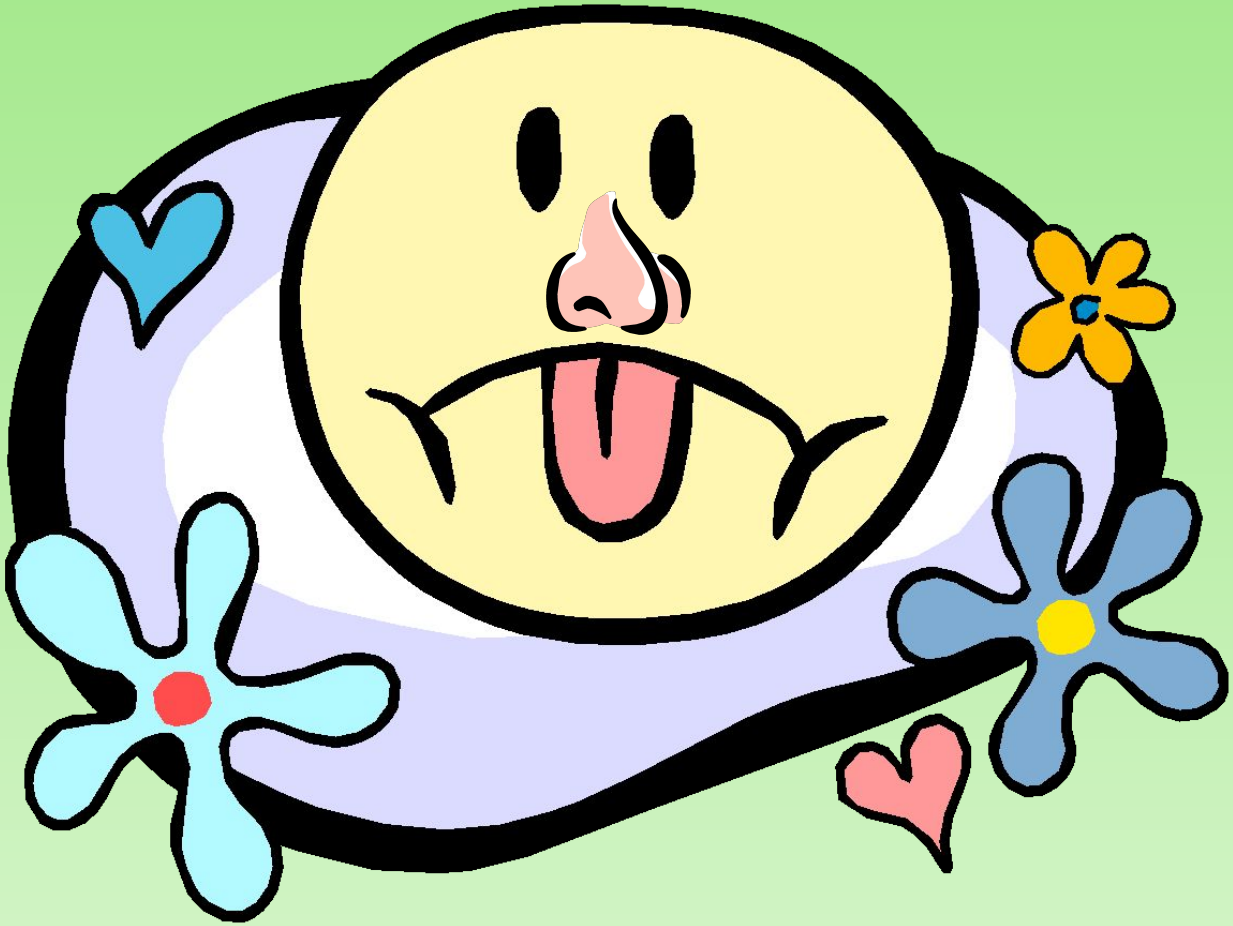
Подробнее о QSAR/QSPR

T14. QSAR/QSPR: Universal Approach to the Prediction of Properties of Chemical Compounds and Materials

Владимир Палюлин

Moscow State University, Russia

Электронные язык и нос



по материалам Алисы Рудницкой, Россия

Хемометрика

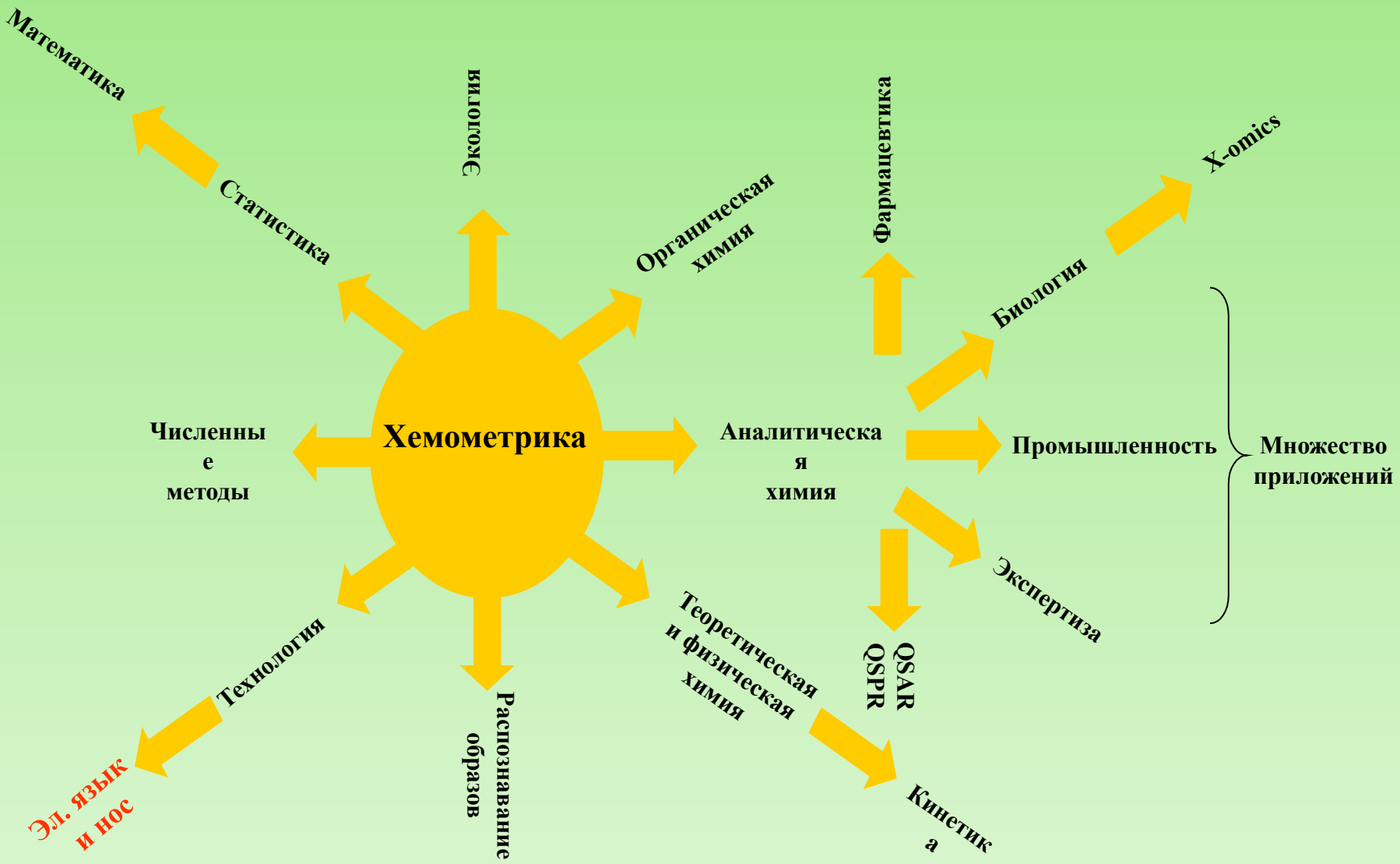
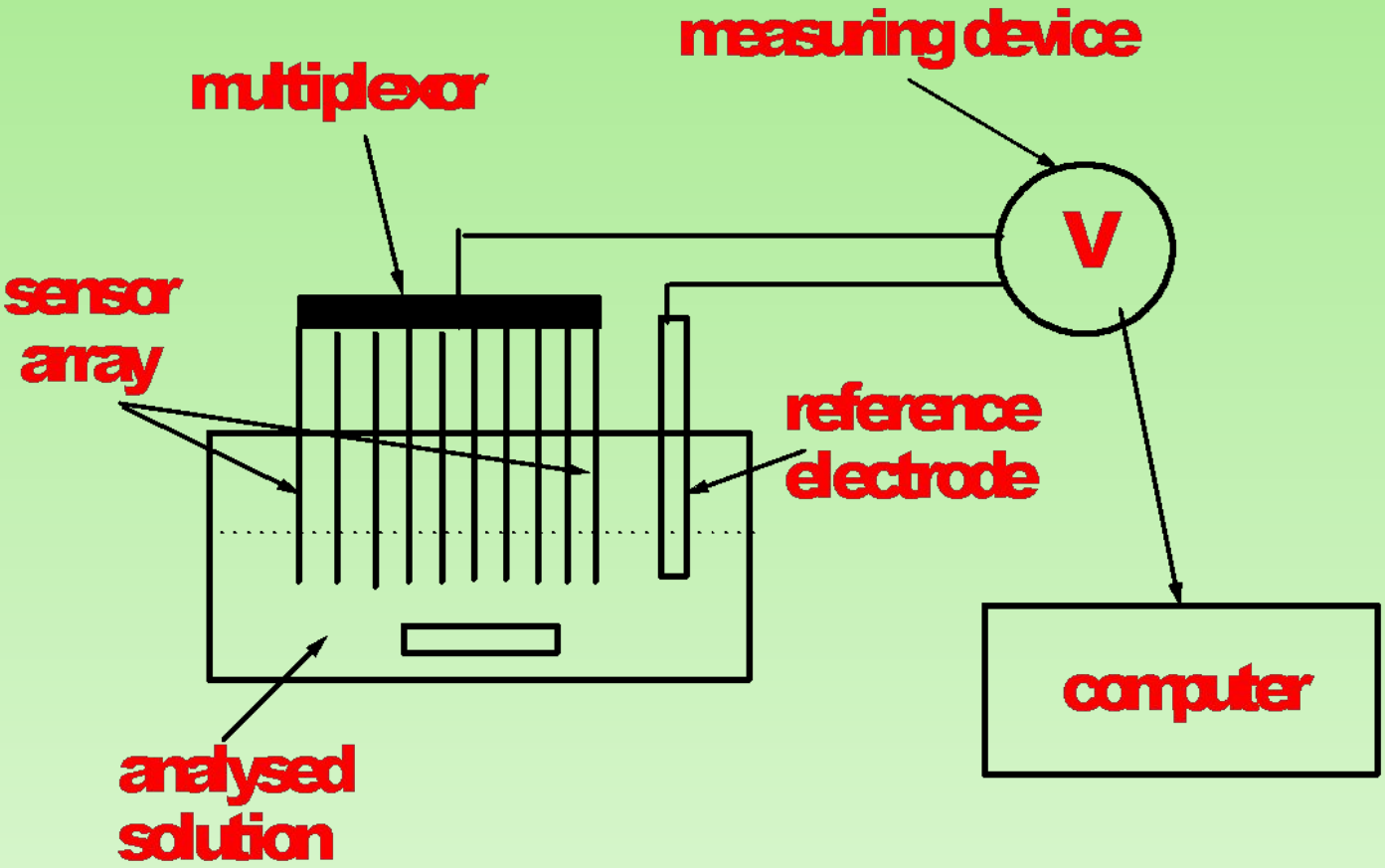
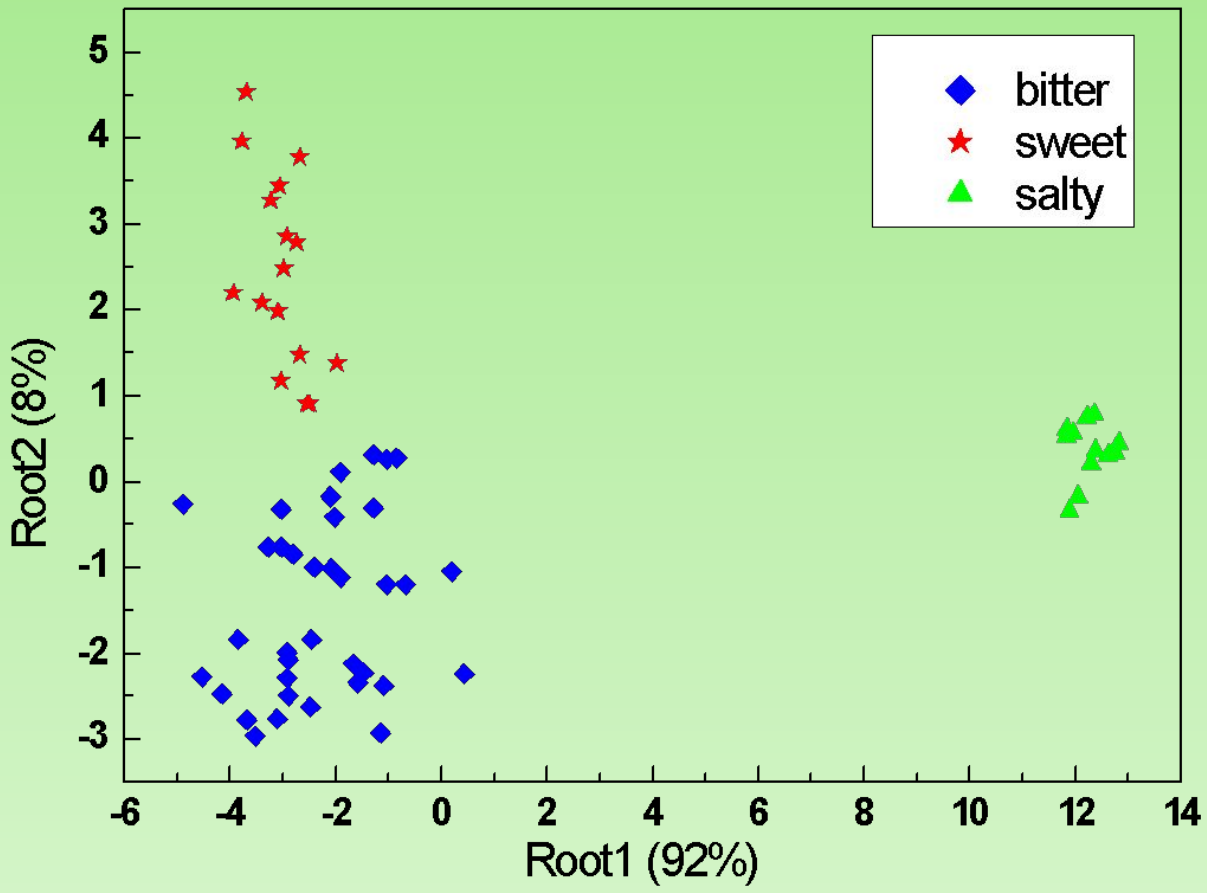


Схема электронного языка



Реальный вид

Корреляция «человек-сенсор»



Подробнее о языке и носе

T5. Comparison of PLS Regression and Artificial Neural Network for the Processing of the Electronic Tongue Data from Fermentation Growth Media Monitoring

Алиса Рудницкая

Sr. Petersburg University, Russia

Пробоотбор

по материалам Pentti Minkkinen, Finland

Теория пробоотбора

Пробоотбор

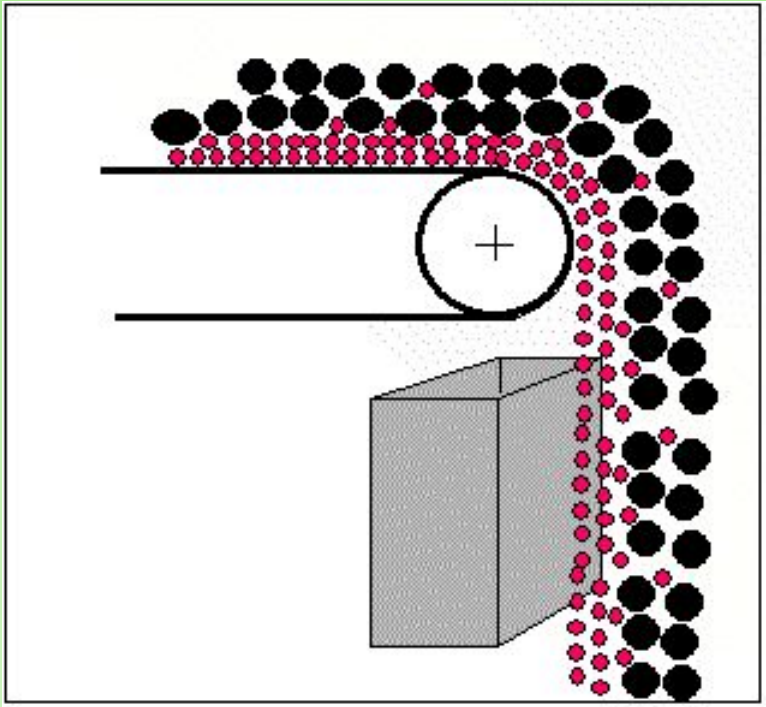
- Искусство отделения малой части из большой массы и подготовки этой пробы к анализу

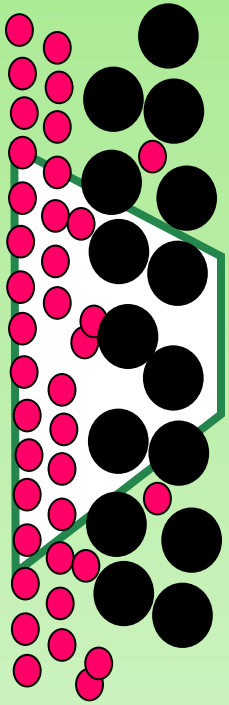
Лозунг дня

- Результат анализа всегда хуже, чем проба, по которой он сделан
- Проба должна быть представительной

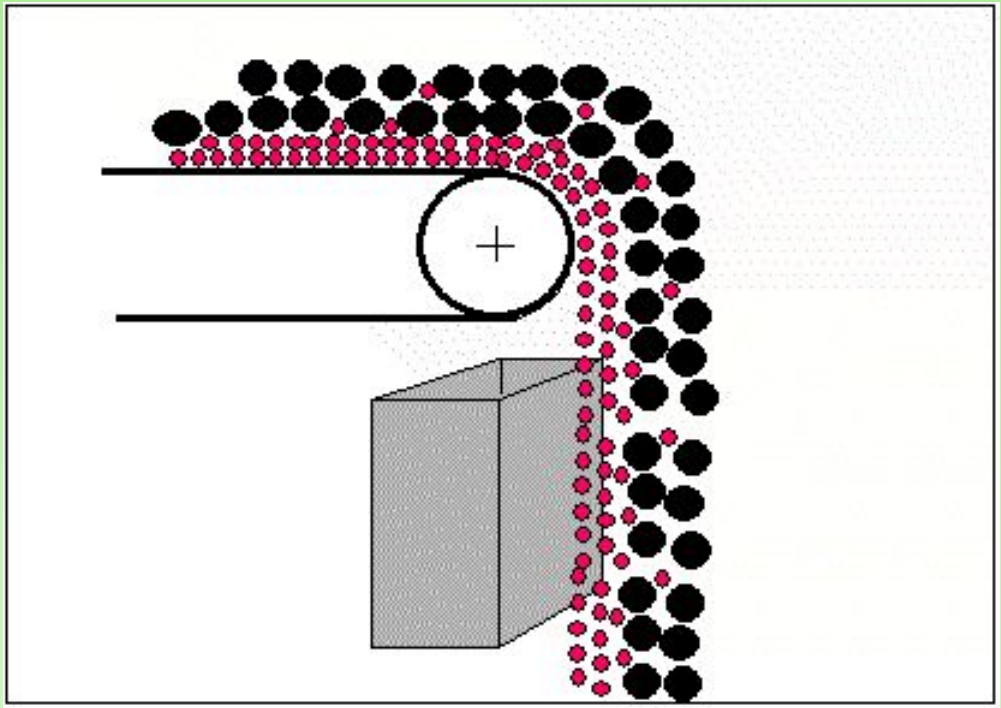
Теория, соединяющая техническую и статистическую стороны была разработана
***Pierre Gy: Sampling for Analytical Purposes*, Wiley, 1998**
и теперь развивается в работах
Kim Esbensen & Pentti Minkkinen

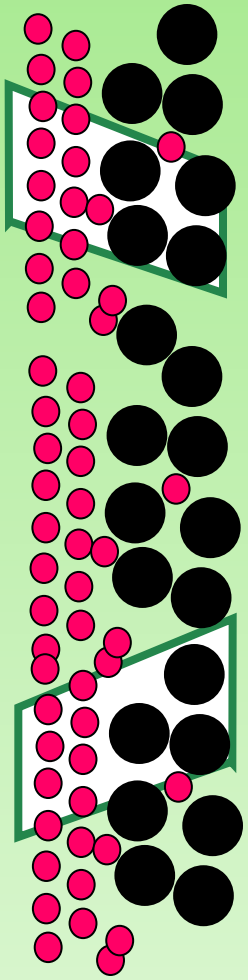
Неправильный отбор образца





Правильный отбор образца





Выводы

Хемометрика находит приложение во многих, очень многих областях

Конференция Analytica Expo 2005

Москва, Выставочный Центр «Сокольники», 19 апреля 2005 г.

APPLICATIONS OF CHEMOMETRICS TO PAT

Richard G. Brereton

Centre for Chemometrics University of Bristol United Kingdom