

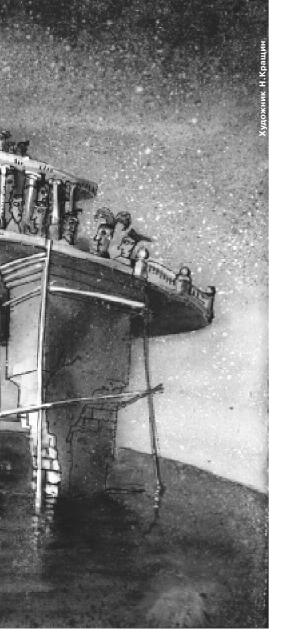
Что такое хемометрика и чем она занимается, можно объяснить по-разному, используя более или менее ученые слова. Общепринятого определения до сих пор нет, и, по-видимому, его никогда и не будет. Наиболее популярное принадлежит Д.Массарту, который писал, что хемометрика — это химическая дисциплина, применяющая математические, статистические и другие методы, основанные на формальной логике, для построения или отбора оптимальных методов измерения и планов эксперимента, а также для извлечения наиболее важной информации при анализе экспериментальных данных. Это определение часто критиковали и предлагали другие дефиниции. Например, С.Волд предложил такое. Хемометрика решает следующие задачи в области химии: как получить химически важную информацию из химических данных, как организовать и представить эту информацию и как получить данные, содержащие такую информацию. Это определение тоже не понравилось, и все сошлись на том, что лучше всего объяснять науку старым

методом, испытанным в других, не менее туманных областях, и заявить, что хемометрика — это то, что делают хемометрики. Вообще-то это явный плагиат — еще С.В.Муррей сказал: «Аналитическая химия — это то, что делают аналитические химики».

А действительно, что же делают хемометрики? Вот маленькая подборка тем публикаций из научных журналов за последние пять лет. Хемометрики контролируют производство полупроводников, аспирина, пива и водки; исследуют причины разрушения документов, написанных старинными галльскими чернилами; проводят допинг-контроль спортсменов; расшифровывают состав косметики Древнего Египта; локализуют месторождение золота в Швеции; идентифицируют подозреваемого в убийстве Анны Линд, министра иностранных дел Швеции; проводят диагностику артрита и рака на ранних стадиях; исследуют органические субстанции в кометном веществе: исследуют кормовой рацион на свинооткормочных комплексах; проверяют, как рацион питания влияет на умственные способности; определяют фальшивые лекарства; определяют происхождение вина, оливкового масла и пигментов, использованных старыми живопис-

Теперь осталось объяснить, как они это делают. Основных принципов их деятельности три. Первый гласит, что лишних данных не бывает — чем больше, тем лучше. На практике это означает, что если вы сумели снять спектр какого-то образца, то нужно не выбрасывать все точки, кроме нескольких характерных длин волн, а использовать все или, по крайней мере, большой кусок. С научной точки зрения это называется многомерным подходом к анализу результатов эксперимента.

Любые данные содержат в себе составляющую, именуемую шумом. Природа шума может быть различной, во многих случаях шум — это просто та часть данных, которая не содержит искомой информации. Что считать шумом, а что — информацией, определяется с учетом поставленных целей. В этом и состоит второй принцип хемометричес-



Доктор физико-математических наук

А.Л.Померанцев



ПРОБЛЕМЫ И МЕТОДЫ НАУКИ

больше из-за роста населения, а на влажных грядках в изобилии водились лягушки.

Исторический экскурс

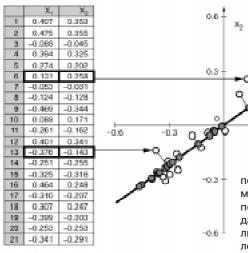
Можно ли точно сказать, когда появилась химия? Ясно, что очень давно, и, весьма вероятно, большую роль в ее развитии, наряду с потребностью в окраске тканей, сыграла необходимость аккуратно отправить в царство теней претендента на трон. Про хемометрику достоверно известно, что она появилась на свет вечером 15 августа 1974 года в небольшом ресторанчике города Сиэтла, что на западном побережье США, в результате шумного разгула, устроенного молодым шведским доцентом Сванте Волдом по случаю его скорого возвращения домой, в город Умеа после длительной стажировки у известных статистиков Бокса и Хантера. Американскую сторону представлял химик Брюс Ковальски со своими учениками, которые тогда интенсивно разрабатывали пакет прикладных программ для анализа химической информации. Собственно, слово «хемометрика» Сванте Волд и его группа из университета Умеа начали использовать еще в начале 70-х, чтобы кратко обозначить то, чем они занимались. А занимались они эксплуатацией идей отца Сванте, Германа Волда, известного математика, разработавшего много методов (в том числе и знаменитый теперь метод проекций на латентные структуры) для анализа данных в психологии. По-видимому, именно по аналогии с психометрикой, биометрикой и другими аналогичными дисциплинами и появился термин «хемометрика». Обстоятельство и место рождения наложили неизгладимый отпечаток как на саму науку, так и на людей, которые ей занимались. Значительно позже Сванте Волд, уже маститый ученый, сказал, что хемометрики — это такие люди, «которые все время пьют пиво и воруют идеи у математиков». А может быть, все дело в том. что первым хемометриком был Уильям Госсет, более известный под псевдонимом Стьюдент («распределение Стьюдента», «критерий Стьюдента»), который работал аналитиком на пивоварне Гин-

Появление хемометрики в России также отмечено веселой историей. После создания в1997 году Российского хемо-

метрического общества мы с коллегами решили, что настало время заявить urbi et orbi о нашем существовании. И тут как раз объявили о приеме докладов на всероссийскую конференцию «Математические методы в химии», которая должна была состояться в летом 1998 года во Владимире. Тогда это был популярный научный форум, организуемый с 1972 года Карповским институтом физической химии. На нем докладывались многие известные ученые, бурно обсуждались новейшие методы применения математики в химии. С этой высокой трибуны мы и надеялись поведать коллегам о том. какие замечательные методы разрабатывают наши друзья по всему миру, как это интересно и перспективно. Доклад был написан и своевременно послан в оргкомитет от имени Российского хемометрического общества. Прошло время, но никакой реакции не последовало. Сроки приближались, уже была опубликована программа, списки докладчиков, но нас там не было. На недоуменные вопросы никто не реагировал, мы решили, что тут какое-то недоразумение и, поскольку Владимир рядом с Москвой, стоит рискнуть и поехать на конференцию без приглашения, чтобы разобраться во всем на месте. Так мы и сделали, явившись в первый день работы в оргкомитет, на регистрацию участников. Вышел грандиозный скандал, сопровождавшийся испугом, криками, заявлениями о недопущении, появлением ответственных товарищей. Дело шло к выдворению с территории университета, жестким административным санкциям, аресту и тюремному заключению. К счастью, через час, стараниями наших старых знакомых по предыдущим конференциям, ситуация разрядилась. Дело оказалось в том, что простодушные владимирские ученые сочли, что хемометрика — это что-то вроде сайентологии, и наше общество пытается смутить нестойкие умы владимирской общественности и отвратить ее от постижения научных ценностей. Спасло нас то, что в библиотеке университета нашлась книга под названием «Хемометрика», изданная еще в 1987 году. Окончательное примирение наступило в ходе заключительного банкета, но доклад нам на всякий случай делать не дали. Вполне возможно, что члены оргкомитета руководствовались суждением Огюста Конта, который еще в 1825 году предупреждал: «Каждая попытка применить мате-

кого подхода к анализу данных: шум — это то, что нам не нужно.

Однако шум и избыточность приводят к тому, что в данных появляются не причинные, а корреляционные связи между переменными. Различие в понятиях причинности и корреляции забавно проиллюстрировали Дж.Е.П.Бокс, Дж.С.Хантер и В.Дж.Хантер. (Об этих отцах хемометрики мы еще будем говорить.) Они привели пример высокой положительной корреляции между числом жителей и количеством аистов в городе Ольденбург (Германия) за период с 1930 по 1936 год. Однако эти две переменные оказались объединены корреляционной связью только потому, что в системе присутствует какая-то третья, скрытая переменная, с которой они обе связаны причинными связями. Поэтому третий принцип хемометрики гласит: ищи скрытые переменные. Что касается аистов - последующие исследования феномена выявили этот скрытый параметр. Им оказалась площадь полей, засаженных капустой в окрестностях города Ольденбурга. Это не шутка: полей стало



матические методы для исследования химических проблем должна рассматриваться как абсолютно абсурдная и противоречащая самому духу химии. Если математический анализ когда-либо займет сколько-нибудь значительное место в химии - извращение, которое, по счастью, почти невероятно, — это повлечет за собой повсеместно быстрое вырождение этой науки».

Метод главных компонент

Для того чтобы объяснить, как хемометрические принципы применяются на практике, рассмотрим один из старейших методов анализа данных, называемый методом главных компонент. Мы не будем отвлекаться на математические подробности, а постараемся передать суть этого подхода, используя интуитивно-понятную геометрическую интерпретацию. Начнем с простейшего случая, когда исследовано несколько образцов, для каждого из которых измерялись всего две переменные х, и х,. Такие данные легко изобразить на плоскости (рис. 1). Каждой строке исходной таблицы (то есть образцу) соответствует точка на плоскости с соответствующими координатами. Они обозначены пустыми кружками. Проведем через них прямую, так, чтобы вдоль нее происходило максимальное изменение данных, она называется первой главной компонентой — ГК1. Затем спроецируем все исходные точки на эту ось. Получившиеся точки закрашены. Теперь мы можем предположить, что на самом деле все наши экспериментальные точки и должны были лежать на этой новой оси. Просто какие-то неведомые силы отклонили их от правильного, идеального положения, а мы вернули их на место. Тогда все отклонения от новой оси можно считать шумом, то есть ненужной нам информацией.

В общем, многомерном случае процесс выделения главных компонент происходит так. Ишется центр облака данных и туда переносится новое начало координат — это нулевая главная ком-

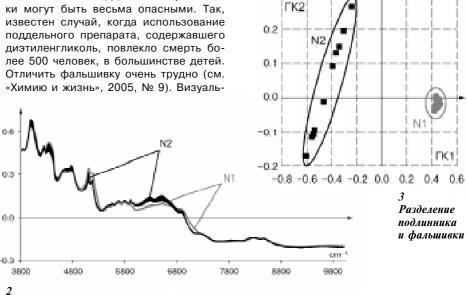
Метод главных компонент для двумерных данных

понента (ГКО). Выбирается направление максимального изменения данных — это первая главная компонента (ГК1). Если данные описаны не полностью (шум велик), то выбирается еще одно направление (ГК2) — перпендикулярное к первому, так, чтобы описать оставшееся изменение в данных, и так далее. В результате мы переходим от большого количества переменных к новому представлению, размерность которого значительно меньше. Часто удается упростить данные на порядки: от 1000 переменных перейти всего к двум. При этом ничего не выбрасывается, все идет в дело. Так работает первый принцип все переменные учитываются. В то же время несущественная для сути дела часть данных отделяется, превращается в шум. Так применяется второй принцип — долой шум. Найденные при этом главные компоненты и дают нам искомые скрытые переменные, тайные пружины, управляющие устройством данных. А это и есть третий принцип — найди скрытую причину. Теперь посмотрим. как применяется этот метод.

Фальшивые лекарства

По данным Всемирной организации здравоохранения, более 70% всех подделок продаются в развивающихся странах. По некоторым данным, более 12% лекарств на российских прилавках представляют собой подделки. Эти фальшивки могут быть весьма опасными. Так, известен случай, когда использование поддельного препарата, содержавшего диэтиленгликоль, повлекло смерть более 500 человек, в большинстве детей. Отличить фальшивку очень трудно (см. «Химию и жизнь», 2005, № 9). Визуальный контроль, тесты на растворимость могут определить только очень грубые подделки. Сложные химические методы, такие, как тонкослойная хроматография, более эффективны, но они требуют большого времени и дорогих реактивов. Кроме того, даже они пасуют в тех случаях, когда подделка содержит нужное количество активного ингредиента, но отличается присутствием других, часто очень опасных добавок, как в случае с диэтиленгликолем. Очевидна и неэффективность всевозможных кампаний по оклейке упаковок разнообразными марками, подделать которые еще проще, чем сами лекарства.

Задачу можно сформулировать так: разработать быстрый (не более пяти минут), мобильный (все оборудование должно умещаться в портфеле), недорогой (не более 20 000 долларов начальных затрат и не более 30 рублей за один тест) метод, позволяющий определить подделку даже в тех случаях, когда в ней присутствует нужное количество активного компонента. Именно так и была поставлена проблема перед небольшой группой хемометриков из России, Швеции, Дании и Финляндии, собравшихся в 2003 году в Пушкинских Горах попеть веселые песни и поговорить о науке. Они предложили применить для решения проблемы «БИК-анализ» — метод ближней инфракрасной спектроскопии, причем данные обрабатывать методом главных компонент. Идея БИК-спектроскопии очень проста. Образец (например, таблетка) помещается в прибор, где на него светят ИК-излучением, которое, попадая на таблетку, частично рассеивается, частично поглощается, частично проходит насквозь. Свет, побывавший в образце, несет информацию о том, где он был и что он там встретил. Его надо только аккуратно собрать и разложить по длинам волн. Для



Спектры настоящих (N1) и фальшивых (N2) таблеток

этого потребуются всякие хитрые устройства, вроде призм, диодных детекторов и прочего, в которых разбираются специалисты. Для нас важен только финальный результат — спектр, в котором интенсивность собранного света изображена в зависимости от длины волны или от его частоты. Вот усредненные спектры для 10 образцов истинных (№ 1) и 10 поддельных (№ 2) таблеток некоего лекарства — рис. 2 (образцы были получены от фирмы-производителя).

Спектры очень похожи, так как исследовался фальсификат высокого качества, неразличимый традиционными химическими методами. Можно ли, используя метод главных компонент, различить их?

На рис. З показаны данные, обработанные методом главных компонент. Четко видны две группы точек, соответствующих истинным и фальсифицированным таблеткам. Заметим, что разброс точек в группе № 2 (контрафакт) существенно больше, чем в группе № 1 (подлинник). Это объясняется лучшим контролем при легальном производстве. В этом примере достаточно использовать только две главные компоненты, которые объясняют 99% данных. Обмануть спектрофотометр невозможно. Гдето всегда найдется диапазон длин волн, в котором проявится отличие.

Итак, что мы имеем в итоге? Отличить подделку можно очень быстро, без специального приготовления образца. Причем еще пять лет назад мы смогли бы сделать БИК-анализ только в лабораторных условиях — обычный спектрофотометр с трудом умещался на столе. Но электроника творит чудеса сейчас в коробку из-под обуви можно уложить два таких прибора и еще останется место. Стоимость одного устройства 15 000 долларов, процедура тестирования проста: положи таблетку в ячейку, нажми на кнопку и получи результат. Так и видятся летучие отряды борцов с подделками, вооруженные БИК-анализаторами, которые рыщут по стране в поисках злоумышленных фальшивотаблетчиков... На самом деле до реализации этой мечты еще очень далеко — ведь с приборами «договориться» труднее, чем с теми, кто выдает разрешения, делает и клеит соответствующие марки.

Познать — это значит анализировать

Что такое наука? Как появляется научное знание? Какими средствами можно решить насущную проблему? Какие проблемы являются насущными? Как сформулировать, как поставить задачу? Чего может и чего не может достичь научное исследование? Все это — вечные вопросы, ответы на которые образуют систему современных научных методов, концепций, идей, кратко называемую парадигмой. Парадигма — это концеп



туальная схема, используемая для постановки проблем и их решения, господствующая в научном сообществе в течение длительного времени.

Оглядываясь во тьму веков, мы увидим, что первой парадигмой, зародившейся в V-III веках до н. э., была геометрическая парадигма. Познать — значит нарисовать. Такова была концепция научного знания, господствовавшая во времена Евклида, Пифагора и Архимеда. Трисекция угла, квадратура круга, удвоение куба — именно эти великие задачи будоражили умы ученых начиная с античного времени и вплоть до конца XIX века. И методы были под стать проблемам — начертить с помощью циркуля и линейки. Такое научное знание ценилось превыше всего. Недаром на могиле Архимеда был изображен цилиндр со вписанным в него шаром - определение соотношения их объемов он почитал своим величайшим достижением. Прошли столетия. На смену первой, геометрической, парадигме пришли другие, но та. первая, жива и поныне. Все ученые уснащают свои работы схемами, графиками, диаграммами, и эта статья — не исключение. Считается, что только так можно правильно объяснить, проиллюстрировать свои идеи. Да что рядовые ученые, даже великий Гаусс, так много сделавший для науки в рамках совсем другой парадигмы, тоже завещал нарисовать на своем надгробии круг со вписанным в него 17-угольником.

Ход исторического прогресса привел к смене парадигмы. Перед наукой стали совсем другие задачи, для решения которых была выстроена новая, алгебраическая парадигма. Познать — значит посчитать. Так думали ученые, заложившие в X-XV веках концепцию новых научных методов: папа Сильвестр II, Лука Пачоли, Франсуа Виет. На смену циркулю и линейке пришли таблица умножения, затем таблица логарифмов и методы решения алгебраических уравнений. Число заменило точку, уравнение кривую. Завершил стройное здание алгебраической парадигмы Декарт, погрузивший геометрию в систему координат и показавший, что старая геометрическая парадигма — это частный случай новой, алгебраической. Но мы все еще любим числа.

Третья, современная, парадигма, которую можно назвать дифференциальной, появилась в XVII веке стараниями Ньютона и Лейбница. Они создали но-

вый метод — дифференциальное исчисление, и с тех пор познать — означает составить и решить дифференциальное уравнение. Эта парадигма принесла фантастические, невероятные результаты. Впервые ученые смогли не только фиксировать и объяснять происходящее, но и предсказывать будущее. Расчет движения небесных тел, кинетика химических и биологических процессов, моделирование экономических и социальных явлений — везде применялись дифференциальные уравнения. Возникла система численных методов для их решения, были созданы мощнейшие компьютеры для воплошения задуманного. Все кажется прекрасно, но...

Сделаем небольшое отступление и поговорим о телевидении. За последние 10 лет мы стали свидетелями того, как на экране последовательно сменяли друг друга парадигмы в решении насущной проблемы — что будущее нам сулит. В начале 90-х господа астрологи с помощью необычайно убедительных чертежей объясняли, что позиция Сатурна в четвертом квадранте неотвратимо сулит нам скорое процветание. Вскоре на смену чертежникам пришел господин нумеролог из Афганистана, который с легкостью превращал слова в числа и обратно. Числа эти сулили нам то же самое, но слегка попозже. Потом в передаче Гордона появился доктор технических наук, академик Академии какихто очень стратегических знаний, который показал всем дифференциальное уравнение первого порядка (Риккати), из которого вытекало все то же самое, но еще более убедительно. А некоторые говорят, что наше телевидение деградирует!

Горение водорода, то есть известная со школьной скамьи реакция 2H₂+O₂ = 2Н₂О, наверное, менее сложная задача, чем прогноз будущего страны, но профессор В.И.Димитров в своей книге «Простая кинетика» насчитал 30 прямых и столько же обратных стадий этого процесса. Каждая из них описывается соответствующим дифференциальным уравнением, которые вместе дают систему из 60 уравнений. Заметим, что рассматривается самый простой изотермический случай, без учета граничных эффектов, переноса тепла и масс. Решение этой системы затрудняется тем, что большинство кинетических параметров нам неизвестны и их нужно определять эксперименталь-



ПРОБЛЕМЫ И МЕТОДЫ НАУКИ

но, но многие стадии очень быстротечны, идут с участием короткоживущих радикалов. И вдобавок это только начало проблем — дальше нам предстоит интегрировать эту систему, которая является очень жесткой, пронизанной внутренними корреляционными связями. И все это еще нужно как-то связать с практикой. А что говорить о других задачах? Ну как, например, применить дифференциальный подход к рассмотренной выше задаче дискриминации фальшивых лекарств?

Все это свидетельствует о слабости дифференциальной парадигмы, в которой каждое явление должно быть описано содержательной моделью, имеющей самостоятельный смысл, опирающейся на некоторые фундаментальные, априорные знания. Им на смену приходит новая, четвертая, парадигма, в основе которой лежит формальное моделирование, основанное только на эксперименте. Именно из эксперимента, из данных, и только из них, извлекается новое знание. Эта информация спрятана там, как золото в руде, и цель ученого — собрать данные, подвергнуть их анализу и найти эту информацию. Познать - это значит собрать и проанализировать данные. Такова, по-видимому, будет формула новой, четвертой, парадигмы, при зарождении которой мы присутствуем. В отличие от третьей, дифференциальной, эта парадигма не стремится к предсказанию поведения систем за пределами экспериментального опыта. Она сосредоточена на осмыслении и интерполяции уже имеющихся данных, на постижении скрытых причин изменения окружающей нас природы.

Ссылаться на Маркса теперь не принято. Дескать, зря абсолютизировал способ производства. Но кроме парадигмы науки действительно есть еще и парадигма производства, технологии. И так уж выходит, что накладываются они друг на друга замечательно, можно сказать, коррелируют: аграрно-рабовладельческая с геометрической, мануфактурно-феодальная с алгебраической, конвейерно-капиталистическая с дифференциальной. Наука никогда не существовала отдельно от производства, так что их прогресс был общим - научнотехническим. Что же происходит сейчас, какая технологическая революция стоит за новой, четвертой парадигмой науки? Да, есть такая революция, она уже

началась! Это произошло 10 сентября 2004 года в США.

Чтобы понять суть происходящего, рассмотрим, как устроена нынешняя технология на примере производства автомобилей. В Воронеже делают болты, в Самаре — гайки, а в Тольятти их соединяют друг с другом. Благодаря стандартизации, унификации и ОТК воронежские гайки накручиваются на самарские болты, и получается много-много машин. Однако такая концепция совершенно не подходит для биологических, химических, фармацевтических и пищевых производств. Стандартная столовская еда всегда будет проигрывать домашним котлеткам. Дело, в частности, в контроле над процессом приготовления, который осуществляется бабушкой не только в начале и конце процесса, как это предусмотрено технологическим регламентом, а непрерывно, на каждой стадии, в режиме on-line.

Способов контроля процессов всего четыре: off-, at-, on- и in-line. Это вот что такое. Допустим, вы варите суп и хотите узнать, достаточно ли в нем соли. Если отлить суп в баночку и послать образец в химическую лабораторию, то это — off-line, и результат будет известен через две недели. Если притащить лабораторию на кухню и проверять образец рядом с плитой — это at-line, и результат будет получен через час. А если у вас есть на кухне специализированный быстродействующий датчик (бабушка), в который помещается образец, то это уже on-line. Вариант in-line предусматривает вмонтирование этого датчика прямо в кастрюльку для непрерывного мониторинга ситуации.

Как получить лекарство высокого качества? Совершенно ясно, что сложный биохимический процесс его производства надо контролировать непрерывно. по ходу, то есть в режиме on- или inline. Попытки применить здесь дифференциальную парадигму, то есть описать все стадии системой дифференциальных уравнений, а потом следить только за легкодоступными показателями вроде температуры и давления, предпринимались неоднократно. Нет числа статьям и диссертациям на эту тему. Но пример с окислением водорода ясно демонстрирует всю тщетность этих попыток.

А что можно и нужно делать в такой сложной ситуации? Да просто оснастить реактор разнообразными датчиками, вроде того спектрофотометра, о котором шла речь выше, в примере с фальшивыми лекарствами. Пусть они непрерывно, по ходу процесса фиксируют все возможные показатели, характеризующие его ход. Эти данные будут нести в себе информацию о всех сторонах процесса. Надо только научиться извлекать ее оттуда и использовать для управления процессом. Для этого сначала необходимо накопить достаточный исторический опыт - собрать и систематизировать информацию об удачных, посредственных и совсем провальных попытках производства. А уже потом, по ходу процесса, нужно все время сопоставлять данные о нем, с тем, что было раньше, проверяя на своем прошлом опыте, все ли идет как надо, нет ли отклонений. Тут самое время опять вспомнить бабушку, которая именно так и варит суп, пробуя и улучшая его в соответствии с накопленным опытом.

Именно эту, «бабушкину», технологию, основанную на непрерывном контроле качества по ходу процесса, сопоставлении текущих данных с историческим опытом, применении датчиков, дающих многомерную, косвенную информацию о состоянии системы, и предписано внедрять на фармацевтических и аналогичных им предприятиях. В сентябре 2004 года Федеральное агентство США по контролю фармпрепаратов и продуктов питания (FDA) выпустило документ «Guidance for Industry. PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance, FDA, 2004», или «Директива для промышленности. Аналитический контроль процессов — схема инноваций для разработки, производства и контроля качества в фармацевтике». В нем изложены основные принципы этого подхода, основной лозунг которого - «сделать качество неотъемлемым свойством продукта». Нельзя сказать, что до этого ничего не происходило и все только ждали этой директивы. На протяжении десяти последних лет хемометрики активно сотрудничали с производственниками, на практике применяя эти методы в пищевой, фармацевтической, металлургической промышленности. Однако теперь это стало законом - без соответствующего контроля уже нельзя произвести препарат, предназначенный для продажи на американском рынке. Хемометрику признали!

Парадоксально, что первые ростки новой парадигмы мы находим в скромной, малоизвестной науке — хемометрике, рассматриваемой многими всего лишь как подраздел аналитической химии, призванный решать метрологические и методические задачи в обработке химического эксперимента. Кто знает, может быть, через 100 лет эта парадигма так и будет называться — хемометрической. Проживем — увидим.

Что еще можно почитать о хемометрике

М.А.Шараф, Д.Л.Иллмэн, Б.Р.Ковальски. Хемометрика, Пер. с англ. М.:Мир, 1987.

К.Эсбенсен. Анализ многомерных данных. Сокр. пер. с англ. под ред. О.Родионовой. М.: Изд. ИПХФ РАН, 2005. В.И.Димитров. Простая кинетика. Новосибирск: Наука, 1982.